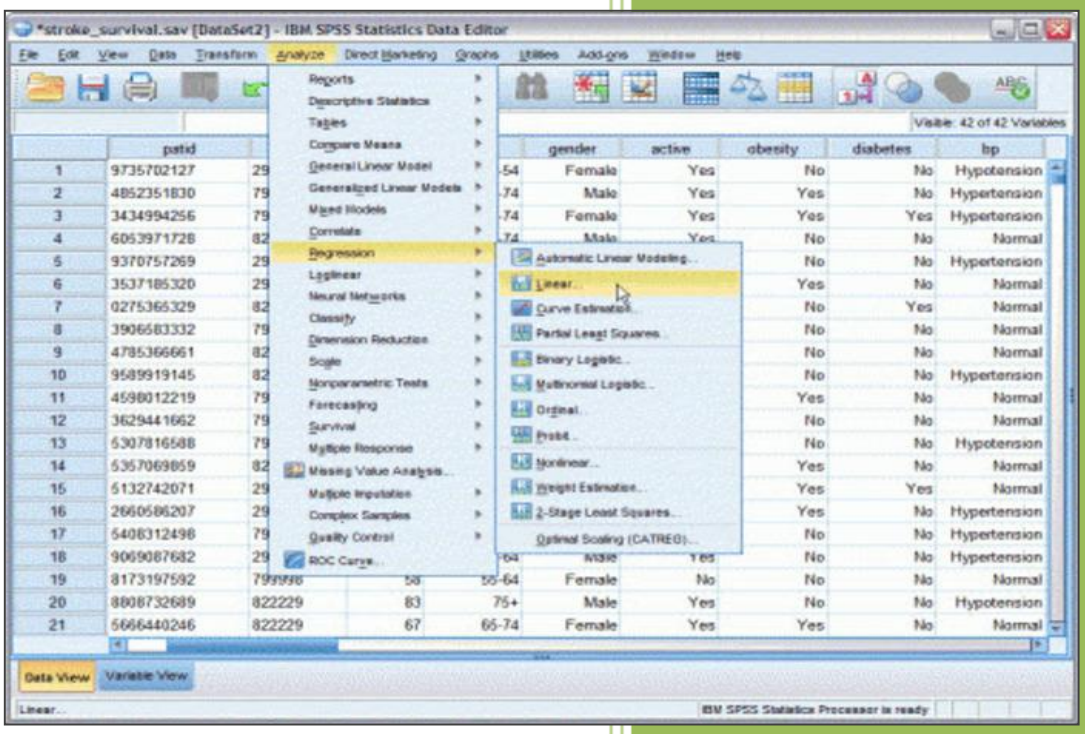
2017

ΣΗΜΕΙΩΣΕΙΣ

Χειρισμού των στατιστικών προγραμμάτων STATA και SPSS



Δημοσθένης Β. ΠΑΝΑΓΙΩΤΑΚΟΣ
Καθηγητής

Σημειώσεις

Χειρισμού του στατιστικού προγράμματος SPSS και του STATA

Δημοσθένης Β. ΠΑΝΑΓΙΩΤΑΚΟΣ, DrMed, FRSPH, FACE
Καθηγητής Βιοστατιστικής & Επιδημιολογίας
Τμήμα Επιστήμης Διαιτολογίας – Διατροφής, Χαροκόπειο Πανεπιστήμιο

Δημήτρης ΠΑΝΑΡΕΤΟΣ, MSc
Στατιστικός -Βιοστατιστικός, Υποψήφιος Διδάκτωρ
Τμήμα Επιστήμης Διαιτολογίας – Διατροφής, Χαροκόπειο Πανεπιστήμιο

« … όταν μπορείς να μετρήσεις αυτό για το οποίο μιλάς και να το εκφράσεις με αριθμούς, τότε γνωρίζεις για αυτό…»
***Lord* Kelvin** (1824 – 1907)

Ο σωστός σχεδιασμός μιας έρευνας αποτελεί πλέον μια αναγκαιότητα παρά μια πολυτέλεια για τους επιστήμονες των επαγγελμάτων υγείας. Η μεθοδολογία της έρευνας, με κύριο εκφραστή της τη Στατιστική, υποστηρίζει την πρακτική εφαρμογή και προσφέρει νέες κατευθύνσεις, διευρύνοντας τους επιστημονικούς ορίζοντες και τη γνώση στο χώρο των επιστημών της υγείας. Η Στατιστική, επιτρέπει αντικειμενικές μετρήσεις από πολύπλοκα επιστημονικά πεδία, προσφέρει ποσοτικές εκτιμήσεις από τα αποτελέσματα των ερευνητικών διαδικασιών και συμβάλει στη λήψη αποφάσεων. Η χρήση των Η/Υ έχει συνεισφέρει στην ταχύτερη και ορθότερη εξαγωγή αποτελεσμάτων και συμπερασμάτων.

Στο εγχειρίδιο αυτό θα βρείτε οδηγίες παράλληλου χειρισμού των στατιστικών λογισμικών STATA & SPSS, και συλλογή λυμένων ασκήσεων, καθώς και συμπλήρωμα θεωρίας, όπου αυτό κρίνεται απαραίτητο.

Αθήνα, Σεπτέμβριος 2017

Δημοσθένης Β. Παναγιωτάκος

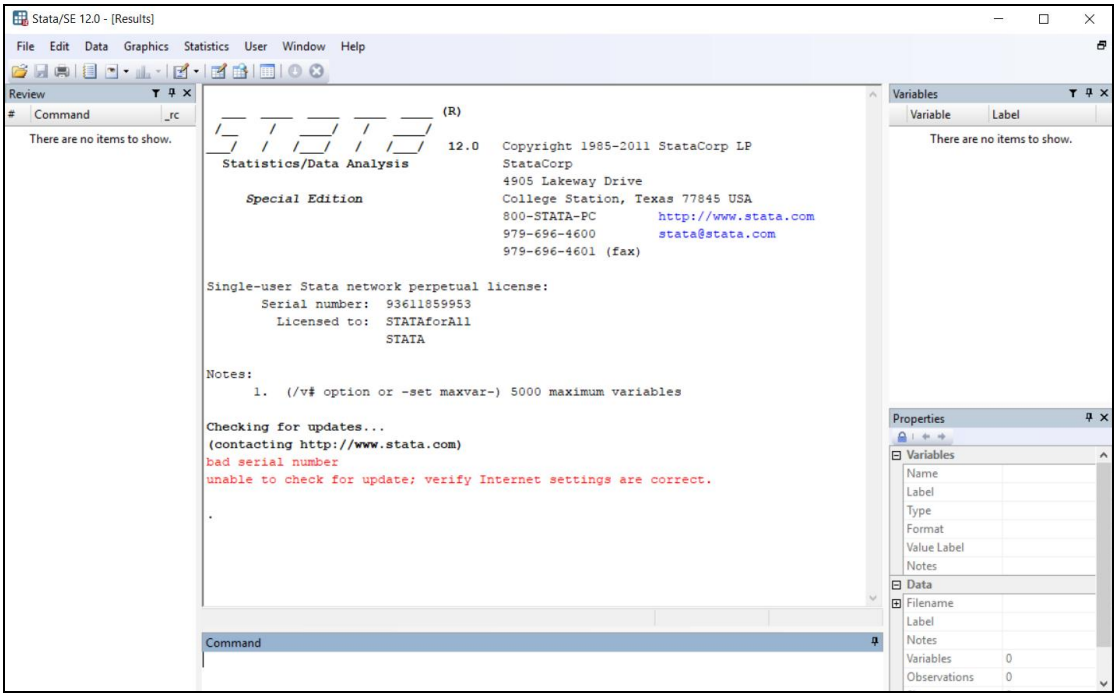
ΠΕΡΙΕΧΟΜΕΝΑ

1. Εισαγωγή στο STATA

Ανοίγοντας το STATA με τον εξής τρόπο:

Έναρξη ⟶ Προγράμματα → STATA

εμφανίζεται το STATA **default screen** (*Εικόνα 1.1*)



Εικόνα 1.1: Το default screen που εμφανίζεται, αμέσως μόλις ανοίξουμε το STATA

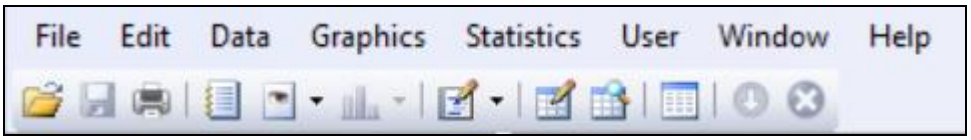
Στο συγκεκριμένο παράθυρο, υπάρχει το μενού επιλογών και η μπάρα εργαλείων (Εικόνα 1.2). Πρέπει να σημειωθεί ότι από το μενού επιλογών, η πιο χρήσιμη επιλογή είναι το «Statistics», γιατί από την συγκεκριμένη επιλογή μπορούν να πραγματοποιηθούν όλες οι στατιστικές αναλύσεις.

Η μπάρα εργαλειών περιέχει εικονίδια (icons) τα οποία επιτρέπουν στον χρήστη να να ανοίξει (Open) και να σώσει αρχεία (Save), να εκτυπώσει (print results) και να χειριστεί όπως θέλει το βασικό του παράθυρο (Bring Graph Window to front).

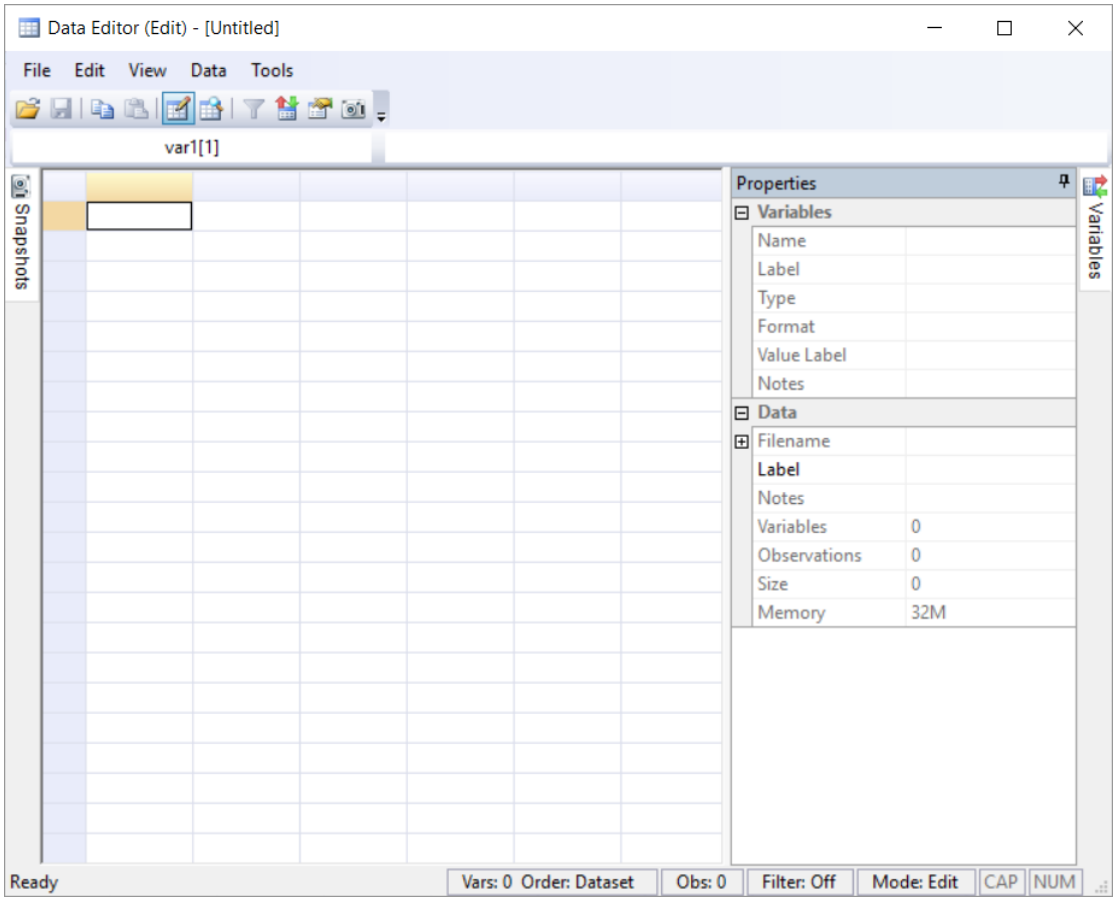
Τα πιο σημαντικά όμως εργαλεία της μπάρας εργαλειών είναι το Data Editor  και το Do – file Editor  . Στο **Do – file** editor μπορεί ο χρήστης να γράψει τις διάφορες εντολές και στη συνέχεια να τις τρέξει με το Execute  .Τέλος μπορεί να τις σώσει σε αρχείο για περαιτέρω χρήση. Τα do -file αρχεία έχουν την κατάληξη "do".

Στο **data editor window** εμφανίζονται τα δεδομένα του χρησιμοποιούμενου αρχείου και το οποίο είναι αρχικά κενό (*Εικόνα 1.3*). Αποτελείται από κελία που σχηματίζονται από γραμμές και στήλες όπου οι γραμμές αντιπροσωπεύουν τις εγγραφές (cases) και οι στήλες τις μεταβλητές (variables) του αρχείου. Στο παράθυρο αυτό μπορούν να δημιουργηθούν νέα αρχεία (να καταχωρηθούν δηλαδή δεδομένα) ή

να ανοιχθούν ήδη υπάρχοντα (βλ. κεφάλαιο 2). Όλα αυτά τα αρχεία έχουν την κατάληξη “dta”.



Εικόνα 1.2: Μενού επιλογών και μπάρα εργαλείων του STATA



Εικόνα 1.3: Data editor window

2. Εισαγωγή δεδομένων και αποθήκευση

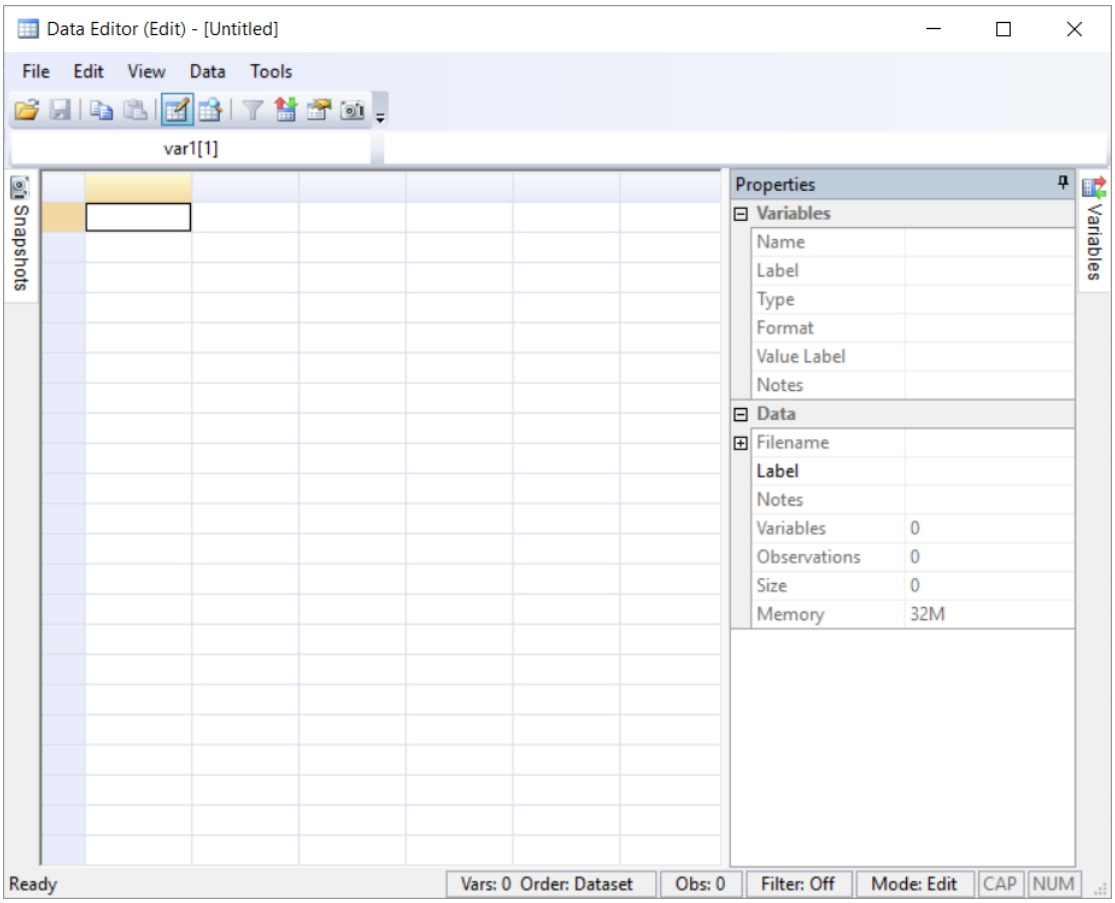
Η εισαγωγή των δεδομένων στο STATA μπορεί να γίνει είτε απευθείας στο ίδιο το στατιστικό πρόγραμμα είτε εισάγοντας τα δεδομένα από κάποιο ήδη υπάρχον αρχείο (π.χ. αρχείο excel).

2.1 Εισαγωγή δεδομένων απευθείας στο STATA

Για το φύλλο εργασίας του STATA ακολουθούμε τα εξής βήματα

Data → Data Editor → Data Editor (Edit)

Κάθε στήλη αντιστοιχεί σε μία μεταβλητή και κάθε γραμμή αντιστοιχεί σε ένα άτομο/περιστατικό.



Εικόνα 2.1: Το φύλο εργασίας του STATA

Στο «***Properties***»: γίνεται η διαχείριση των δεδομένων, ορίζοντας τα χαρακτηριστικά των μεταβλητών της βάσης μας:

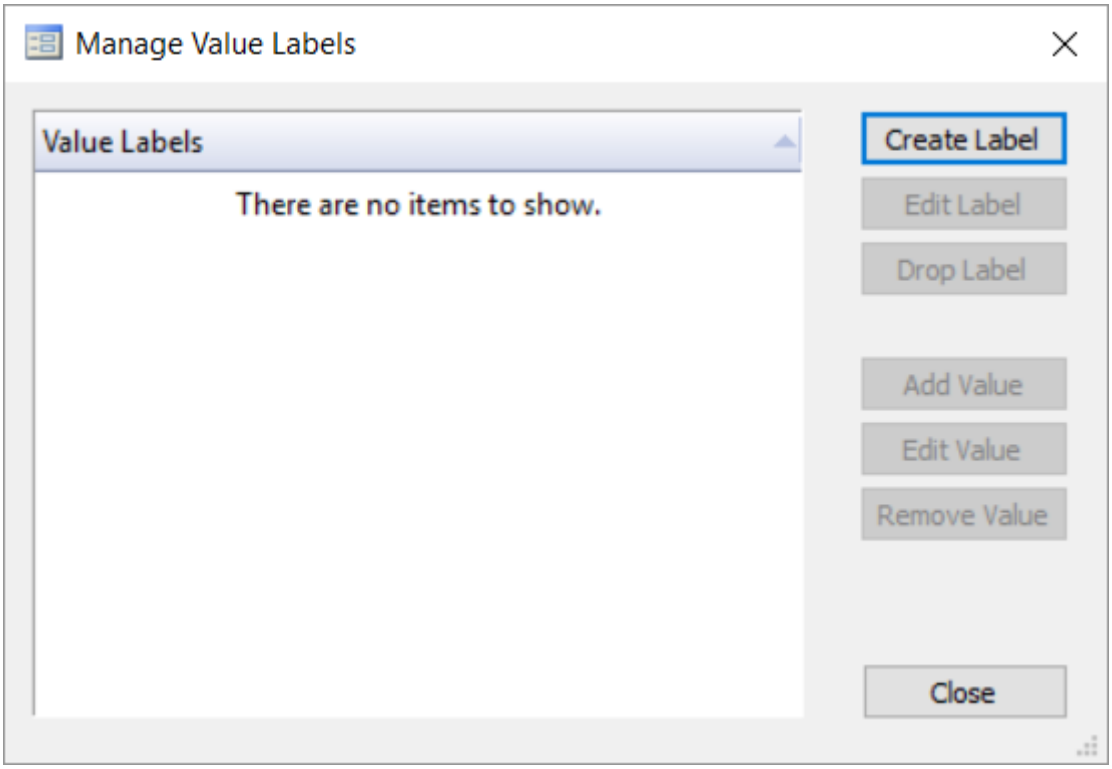
i. **Name**: Το όνομα της μεταβλητής (π.χ. φύλο, ηλικία κ.τ.λ.).
ii. **Label:** Λεζάντα της μεταβλητής (Έχει νόημα στην περίπτωση που το όνομα της μεταβλητής είναι κάποιος κωδικός, έτσι ώστε τοποθετώντας τον κέρσορα επάνω στο όνομα της μεταβλητής να φαίνεται ακριβώς σε τι αφορά η συγκεκριμένη μεταβλητή).

iii. **Type**: Τύπος της μεταβλητής, δηλαδή τύπο δεδομένων που θα εισαχθούν στη συγκεκριμένη μεταβλητή (π.χ. δηλώνουμε ότι η μεταβλητή θα είναι «float» είτε «double» όταν θέλουμε να εισάγουμε πραγματικούς αριθμούς, ενώ δηλώνουμε «int» είτε «long» όταν θέλουμε να εισάγουμε ακέραιους αριθμούς. Η διαφορά βρίσκεται στο πόσα bytes θα διαθέσουμε για να αποθηκευθεί η μεταβλητή αυτή στη μνήμη).

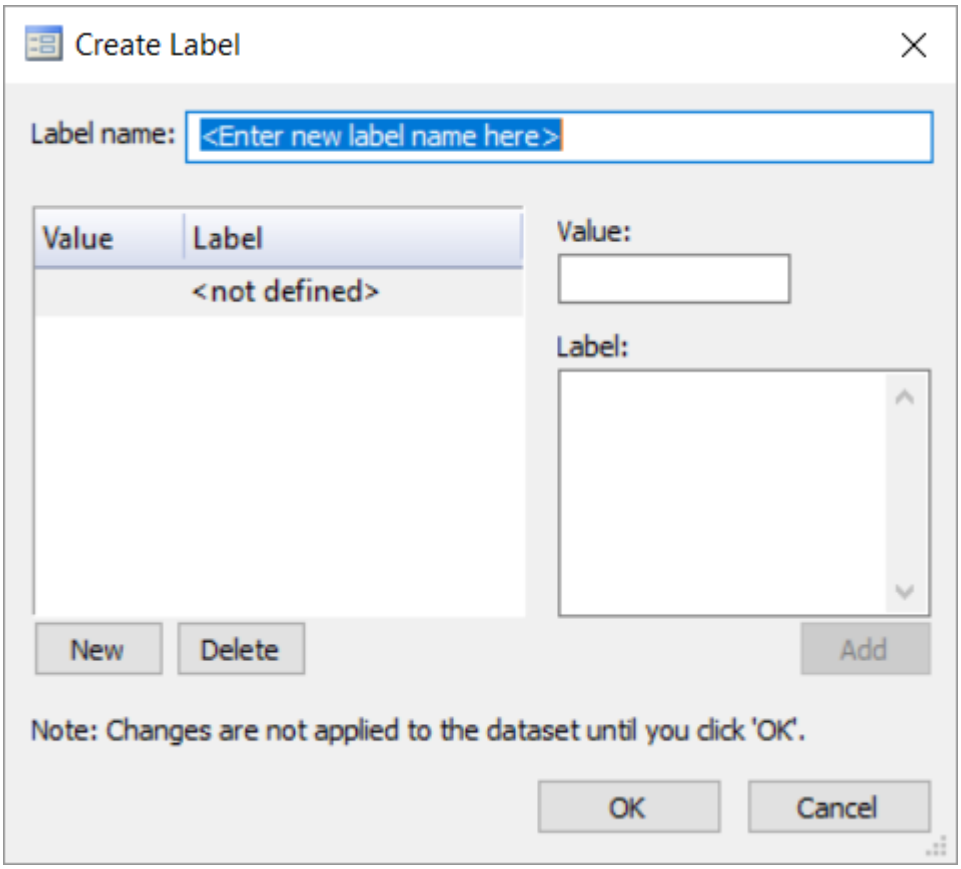
iv. **Format:** Η μορφή της μεταβλητής, δηλαδή δηλώνουμε εάν η μεταβλητή αφορά αριθμητικά δεδομένα ή ημερομηνίες. Εδώ υπάρχει η δυνατότητα να δηλώσουμε την μέγιστη τιμή συμβόλων του δεκαδικού μέρους (decimals) καθώς και τον μέγιστο αριθμό χαρακτήρων που θέλουμε να εισάγουμε σε κάθε μεταβλητή (total digits). Στην περίπτωση που εισάγουμε αλφαριθμητική τιμή τότε αυτομάτως η μορφή της μεταβλητής μετατρέπεται σε string (χαρακτήρας).

v. **Value Label**: Λεζάντα για τις τιμές με τις οποίες έχουν κωδικοποιηθεί οι κατηγορίες μιας κατηγορικής μεταβλητής. Συνεπώς, λοιπόν, η επιλογή Values έχει νόημα μόνο στην περίπτωση που το χαρακτηριστικό που θα καταγράφουμε σε αυτή την μεταβλητή είναι ποιοτικό. Για να ορίσουμε αυτές τις κατηγορίες πατάμε στο δεξιό άκρο του ***Value Label*** κελιού και ανοίγει το πλαίσιο της *Εικόνας 2.2*. Ας υποθέσουμε ότι επιθυμούμε να κωδικοποιήσουμε τις τιμές της μεταβλητής Gender με 1: male & 0: female. Ακολουθούμε λοιπόν τα εξής ***βήματα***:

- Πατάμε ***Create Lable*** και ανοίγει το πλαίσιο της *Εικόνας 2.3*.
- Στο ***Label Name*** πληκτρολογούμε το όνομα της ποιοτικής μεταβλητής που επιθυμούμε να κωδικοποιήσουμε
- Στο ***Value*** πληκτρολογούμε «1»
- Στο ***Label*** πληκτρολογούμε Male και
- Πατάμε «***Add***»
- Συνεχίσουμε ακολουθώντας την ίδια διαδικασία μέχρι να ορίσουμε τα ***Value labels*** για όλους τις κατηγορίες της μεταβλητής και
- Πατάμε «***Ok***»



Εικόνα 2.2: Προσδιορισμός των labels για τις τιμές με τις οποίες έχει κωδικοποιηθεί κάθε κατηγορία μιας κατηγορικής μεταβλητής.



Εικόνα 2.3: Προσδιορισμός των labels για τις τιμές με τις οποίες έχει κωδικοποιηθεί κάθε κατηγορία μιας κατηγορικής μεταβλητής.

ΠΑΡΑΔΕΙΓΜΑ:

Ας υποθέσουμε ότι θέλουμε να φτιάξουμε μία βάση δεδομένων στο SPSS και ας υποθέσουμε ότι τα πρώτα χαρακτηριστικά που καταγράφονται στο ερωτηματολόγιο και τα οποία θα αποτελέσουν τις πρώτες μεταβλητές της βάσης μας είναι:
a) κωδικός ερωτηματολογίου,
b) όνομα συμμετέχοντος,
c) φύλο,
d) ημερομηνία γέννησης.

Τα ***βήματα*** που πρέπει να ακολουθήσουμε είναι:

i. Ανοίγουμε το πρόγραμμα του STATA με τον εξής τρόπο:

Έναρξη ⟶ Προγράμματα → STATA

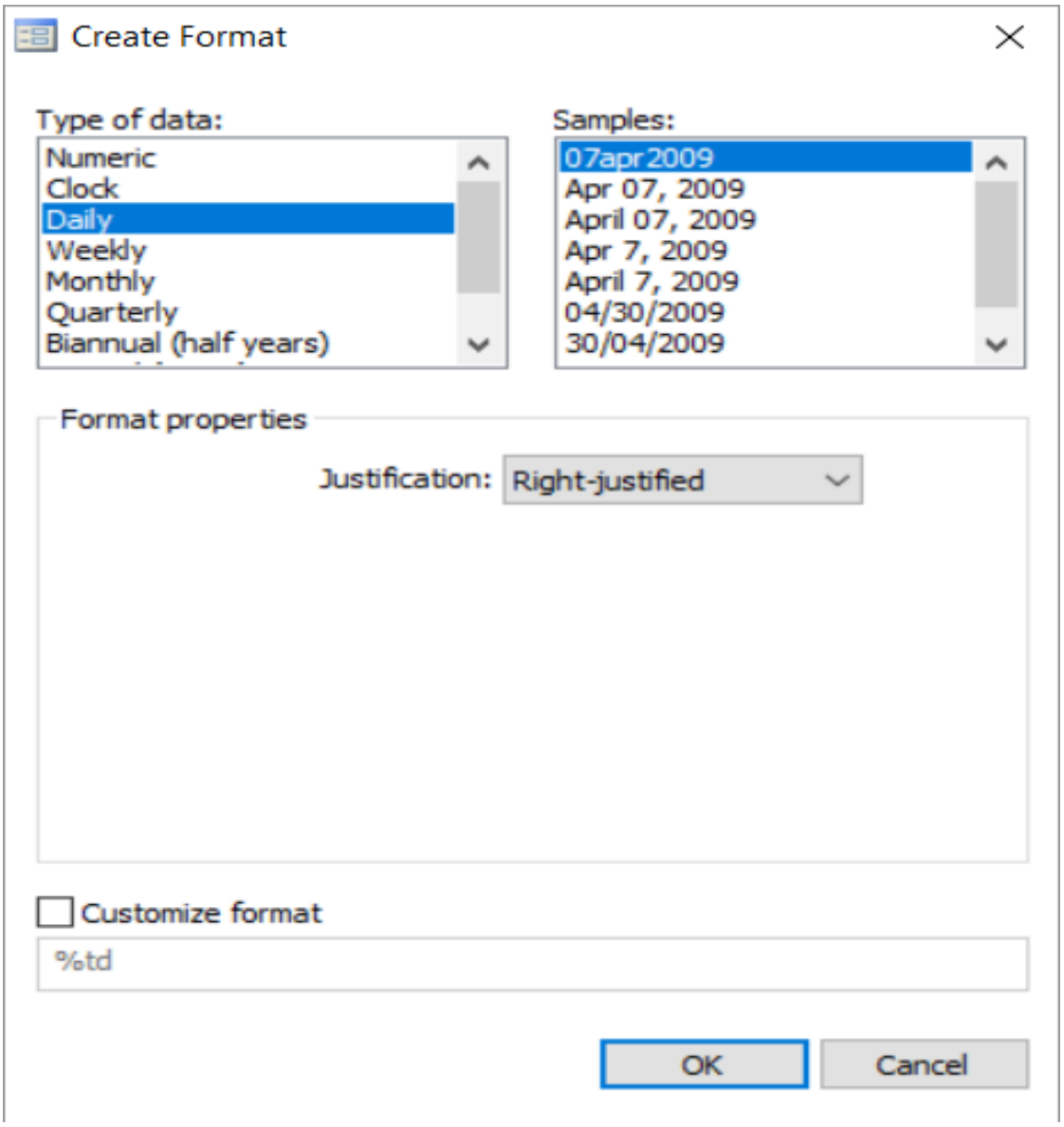
ii. Πατώντας το κουμπί «***Data Editor***» μεταβαίνουμε στη «σελίδα» του STATA που θα ορίσουμε τις μεταβλητές που θα περιέχει το αρχείο μας (*Εικόνα 2.1*).
iii. Δηλώνουμε το όνομα της πρώτης μεταβλητής που είναι ο κωδικός του ερωτηματολογίου ως «id».
iv. Αμέσως, μετά, δηλώνουμε τον τύπο της μεταβλητής, δηλαδή αν τα στοιχεία που θα καταχωρούνται στη συγκεκριμένη μεταβλητή είναι ακέραιοι (int ή long) ή πραγματικοί αριθμοί (float ή double).
v. Ταυτόχρονα δηλώνουμε την μορφή της μεταβλητής, δηλαδή δηλώνουμε εάν η μεταβλητή αφορά αριθμητικά δεδομένα ή ημερομηνίες.
vi. Στη συνέχεια δηλώνουμε το μέγιστο αριθμό χαρακτήρων που θα καταχωρηθεί στη συγκεκριμένη μεταβλητή. Δηλαδή, αν ο μεγαλύτερος κωδικός που έχει δοθεί σε ερωτηματολόγιο είναι ο αριθμός 13325, ο μέγιστος αριθμός ψηφίων είναι 5, άρα στο "Total digits" δηλώνουμε 5.
vii. Στη συνέχεια, δηλώνουμε τον αριθμό των δεκαδικών ψηφίων που ενδέχεται να καταχωρηθούν σε στη συγκεκριμένη μεταβλητή ("Decimals": 0, αφού στον κωδικό δεν έχουν χρησιμοποιηθεί δεκαδικά στοιχεία).
viii. Ορίζουμε ως «Label» για τη συγκεκριμένη μεταβλητή το «Questionnaire number» έτσι ώστε να είναι σαφές ότι σε αυτή τη μεταβλητή καταχωρείται ο αριθμός του ερωτηματολογίου

(***Σημείωση***: το Label δεν είναι αναγκαίο σε μεταβλητές που από το όνομα τους είναι σαφές το περιεχόμενο της μεταβλητής).

ix. Με αντίστοιχο τρόπο, στην ακριβώς επόμενη γραμμή ορίζουμε ότι η δεύτερη μεταβλητή του αρχείου μας είναι το όνομα του συμμετέχοντα ορίζοντας τα εξής στοιχεία: Name: Name, Format: String, Width: 25.
x. Στη συνέχεια ορίζουμε ότι η τρίτη μεταβλητή του αρχείου μας είναι το φύλο του συμμετέχοντα δίνοντας: Name: Gender, Format: Numeric, Width: 1, Decimals: 0, Label: δεν είναι απαραίτητο, αφού από το όνομα της μεταβλητής είναι πλήρως κατανοητό το περιεχόμενο της, Values: 1=Male & 0=Female

(***Σημείωση***: αν και οι πιθανές καταχωρήσεις στη συγκεκριμένη μεταβλητή είναι «άνδρας» ή «γυναίκα», στο πρόγραμμα είναι προτιμότερο να καταχωρούνται οι απαντήσεις με αριθμούς, δεδομένου ότι ο υπολογιστής μπορεί να χρησιμοποιεί σε στατιστικές αναλύσεις μόνο νούμερα. Συνεπώς, για την συγκεκριμένη μεταβλητή είναι απαραίτητο να έχει προηγηθεί ένα είδος κωδικοποίησης, π.χ., άνδρας: 1 & γυναίκα: 0, η οποία κωδικοποίηση μπορεί να είναι εμφανής και στην βάση του STATA ορίζοντας τα κατάλληλα Value labels).

xi. Τέλος, απομένει να ορίσουμε την μεταβλητή στην οποία θα καταχωρούμε την ηλικία των συμμετεχόντων. Ορίζουμε ως «Name» το «Age» και ως «Type» το «Daily». Τότε το πλαίσιο διαλόγου που εμφανίζεται όταν επιθυμούμε να ορίσουμε το «Type» φαίνεται παρακάτω (*Εικόνα 2.3*) όπου θα πρέπει να επιλέξουμε τη μορφή με την οποία επιθυμούμε να καταχωρείται και να εμφανίζεται η ημερομηνία στο αρχείο μας.



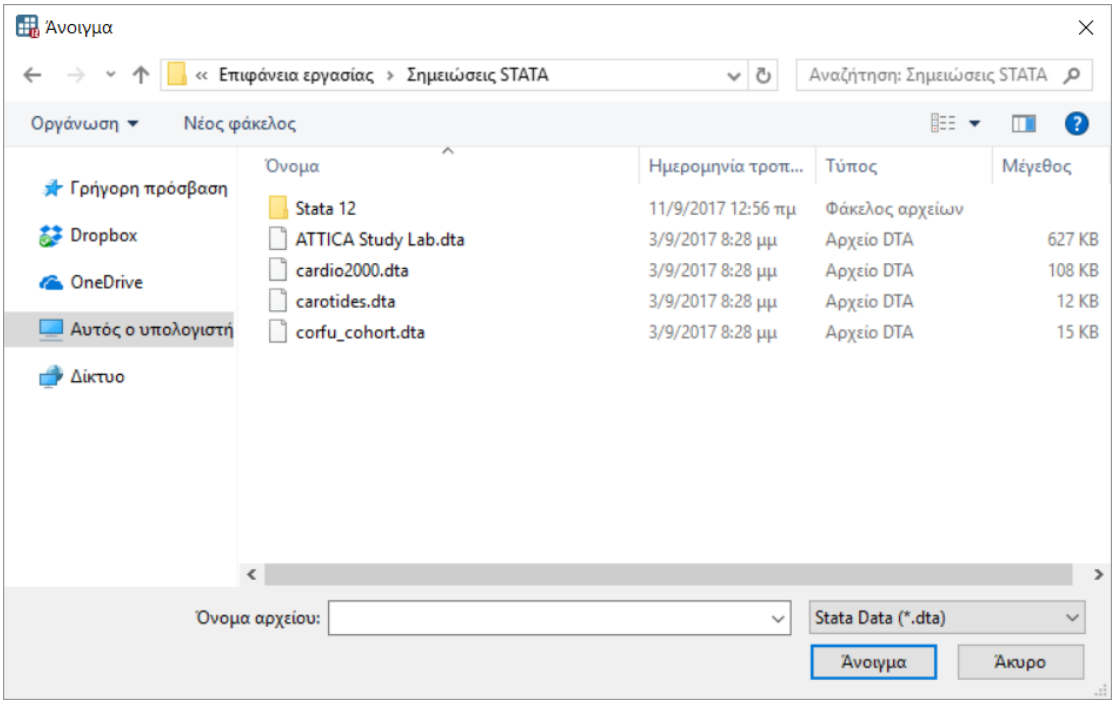
Εικόνα 2.3: Προσδιορισμός της μορφής με την οποία θα καταχωρείται η ημερομηνία

Σημείωση 1: Στον μέγιστο αριθμό χαρακτήρων που θα καταχωρούνται σε μία μεταβλητή (Total digits) πρέπει να συνυπολογίζουμε και την υποδιαστολή στην περίπτωση που δηλώνουμε ότι το χαρακτηριστικό που θα καταχωρείται στην συγκεκριμένη μεταβλητή θα έχει τη μορφή δεκαδικών αριθμών.

2.2 Εισαγωγή δεδομένων από υπάρχον αρχείο στο STATA

Ας υποθέσουμε ότι τα δεδομένα μιας έρευνας είναι καταχωρημένα σε ένα αρχείο excel και επιθυμούμε αυτά τα δεδομένα να τα εισάγουμε σε ένα αρχείο STATA προκειμένου να γίνουν οι απαραίτητες στατιστικές αναλύσεις. Αυτό πραγματοποιείται ακολουθώντας τα εξής ***βήματα***:

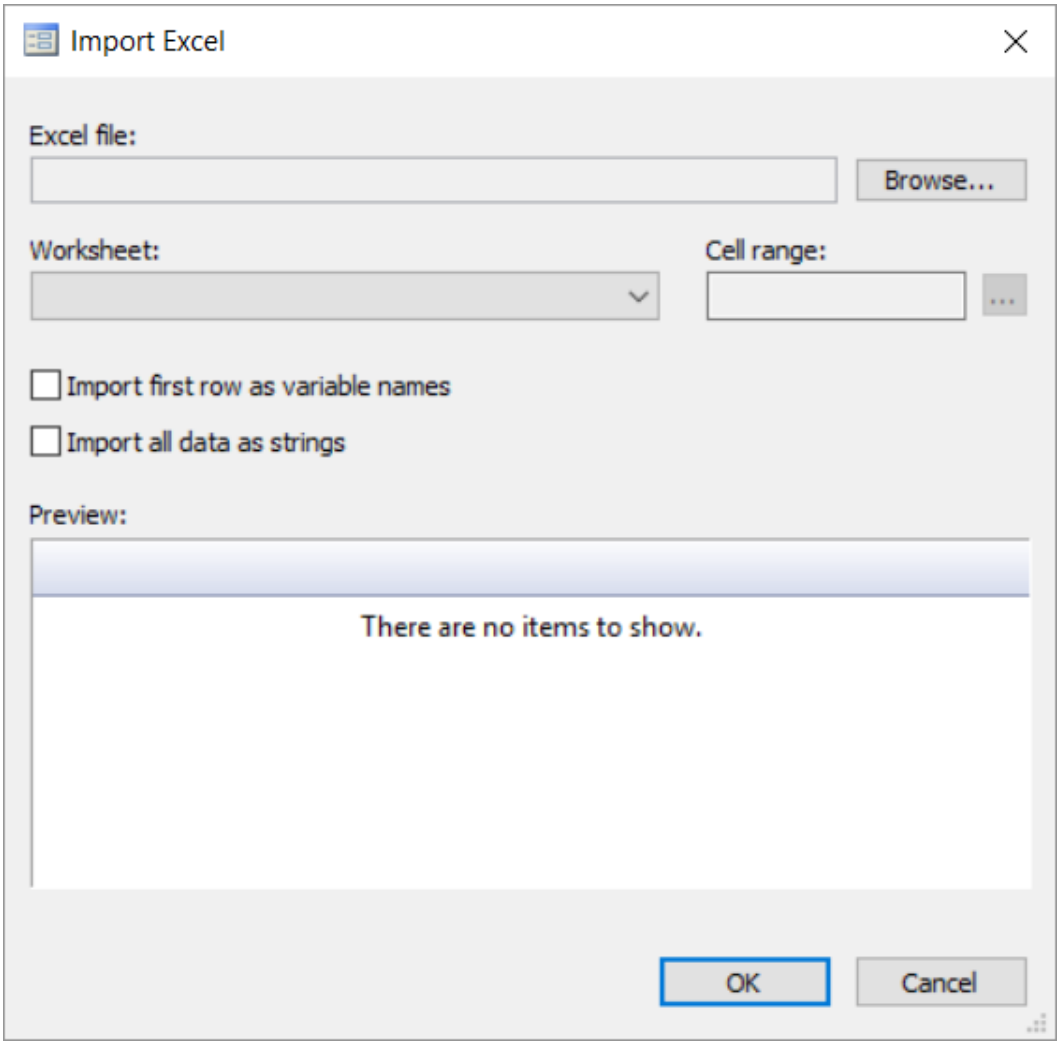
a) Ανοίγουμε ένα κενό αρχείο STATA με τον τρόπο που αναφέρεται παραπάνω.
b) Από το «***File***», επιλέγουμε το «***Open***» και ανοίγει το πλαίσιο διαλόγου της *Εικόνας 2.4.*



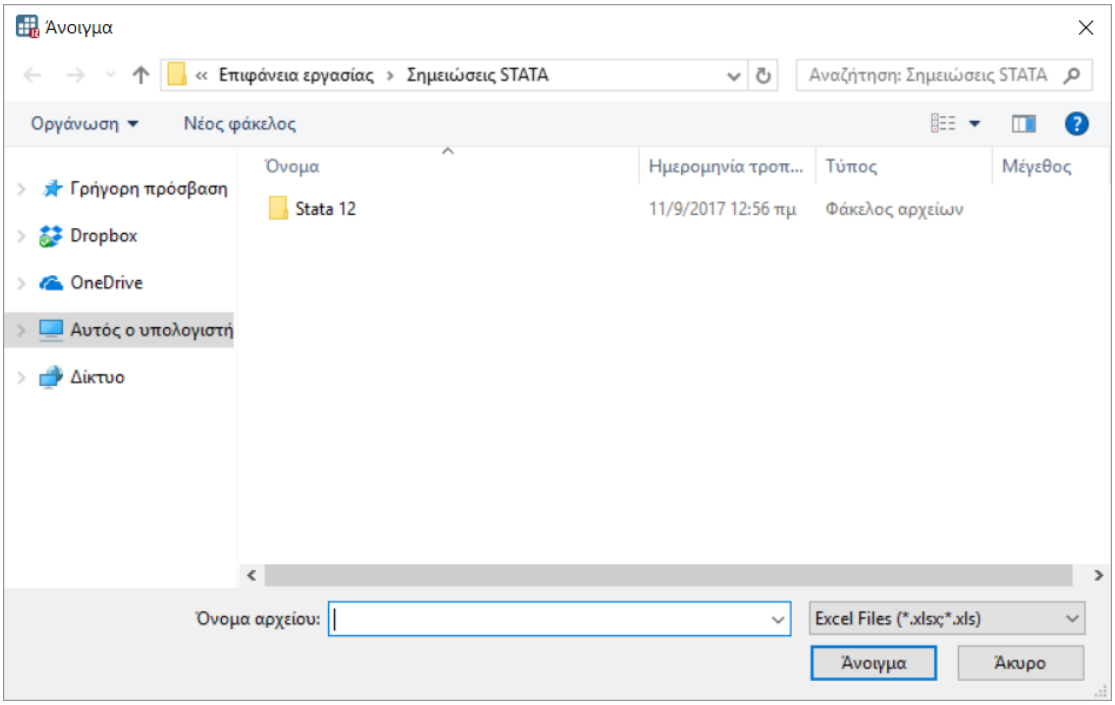
Εικόνα 2.4: Εισαγωγή δεδομένων στο STATA από ένα ήδη υπάρχον αρχείο

c) Στο πλαίσιο της *Εικόνας 2.4* επιλέγουμε τον τύπο του αρχείου που θέλουμε να ανοίξουμε μέσω του STATA
d) Πατάμε «***Άνοιγμα ή Open***» και τα δεδομένα έχουν εισαχθεί στο STATA.

- Στην περίπτωση που θέλουμε να εισάγουμε ένα αρχείο Excel επιλέγουμε File → Import → Excel spreadsheet και εμφανίζεται το επόμενο πλαίσιο όπως φαίνεται στην *Εικόνα 2.5*. Πατώντας Browse εμφανίζεται το πλαίσιο της εικόνας 2.6 και ο χρήστης ακολουθεί τα ίδια βήματα με προηγουμένως.



Εικόνα 2.5: Εισαγωγή δεδομένων στο STATA από ένα ήδη υπάρχον αρχείο Excel

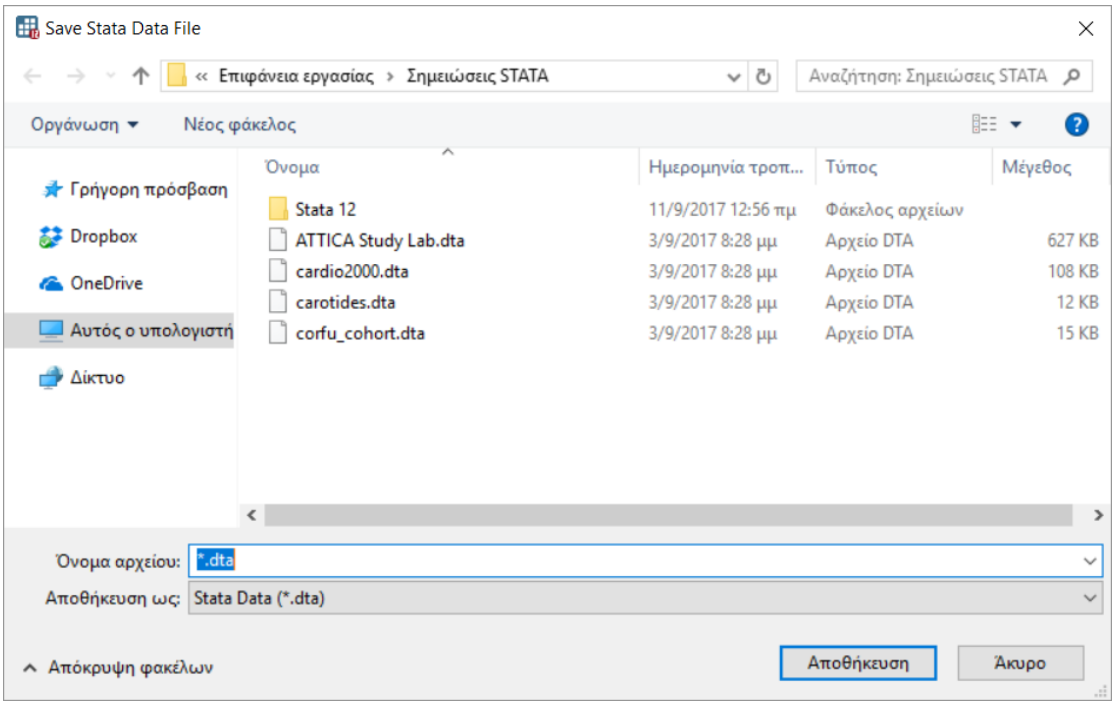


Εικόνα 2.6: Εισαγωγή δεδομένων στο STATA από ένα ήδη υπάρχον αρχείο Excel

2.3 Αποθήκευση αρχείου στο STATA

Για να αποθηκεύσουμε ένα αρχείο STATA ακολουθούμε τα εξής ***βήματα***:

a. Από το μενού του STATA επιλέγουμε "***File***" → "***Save as***" και εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 2.7*.
b. Δηλώνουμε το όνομα που επιθυμούμε να πάρει το αρχείο μας στη θέση «***Όνομα αρχείου***» και
c. Την περιοχή που επιθυμούμε να αποθηκευτεί το συγκεκριμένο αρχείο μέσα στον υπολογιστή, χρησιμοποιώντας το πλαίσιο δεξιά, και
d. Την μορφή που επιθυμούμε να αποθηκευτεί το συγκεκριμένο αρχείο, χρησιμοποιώντας το βελάκι που υπάρχει στο «***Αποθήκευση ως***».



Εικόνα 2.7: Αποθήκευση ενός αρχείου STATA

3. Διαχείριση δεδομένων

3.1 Υπολογισμός νέων μεταβλητών στο STATA

Τα ***βήματα*** που πρέπει να ακολουθήσουμε είναι:

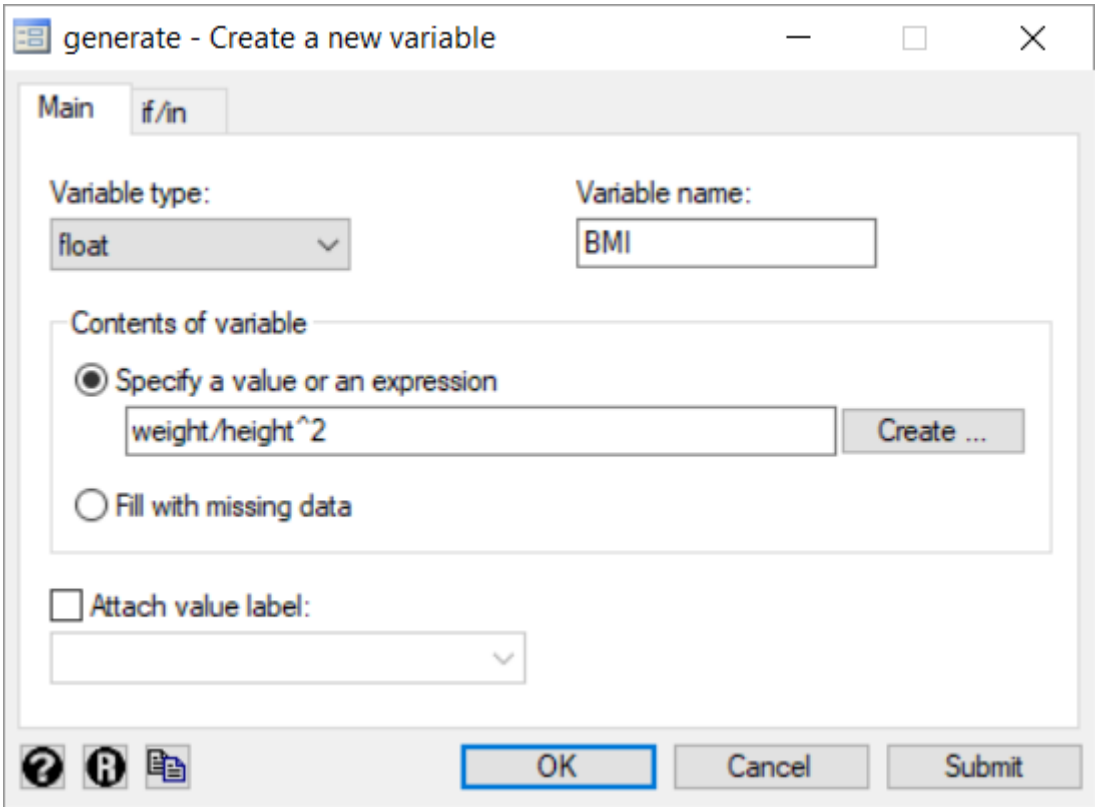
Data ⟶ Create or change data ⟶ Create new variable

Ανοίγει το πλαίσιο διαλόγου (*Εικόνα 3.1*).

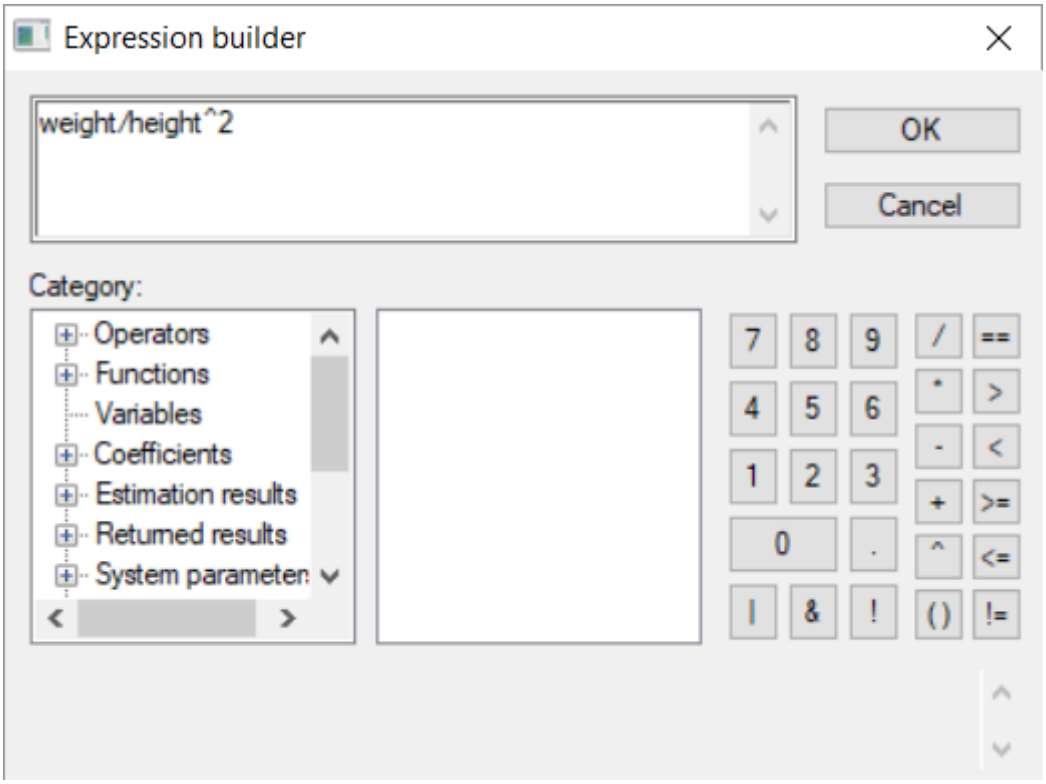
Με την επιλογή "***Create new variable***" στο STATA μπορούμε να υπολογίζουμε τις τιμές μιας νέας μεταβλητής χρησιμοποιώντας τις τιμές μίας ή περισσοτέρων μεταβλητών που υπάρχουν ήδη στο αρχείο μας. Για παράδειγμα, ας υποθέσουμε ότι υπάρχουν καταχωρημένες οι τιμές βάρους (weight) και ύψους (height) στη βάση μας, και επιθυμούμε να υπολογίσουμε τις τιμές του δείκτη μάζας σώματος (BMI).
Τα ***βήματα*** που ακολουθούμε είναι τα εξής:

Στο πλαίσιο διαλόγου της *Εικόνας 3.1*:

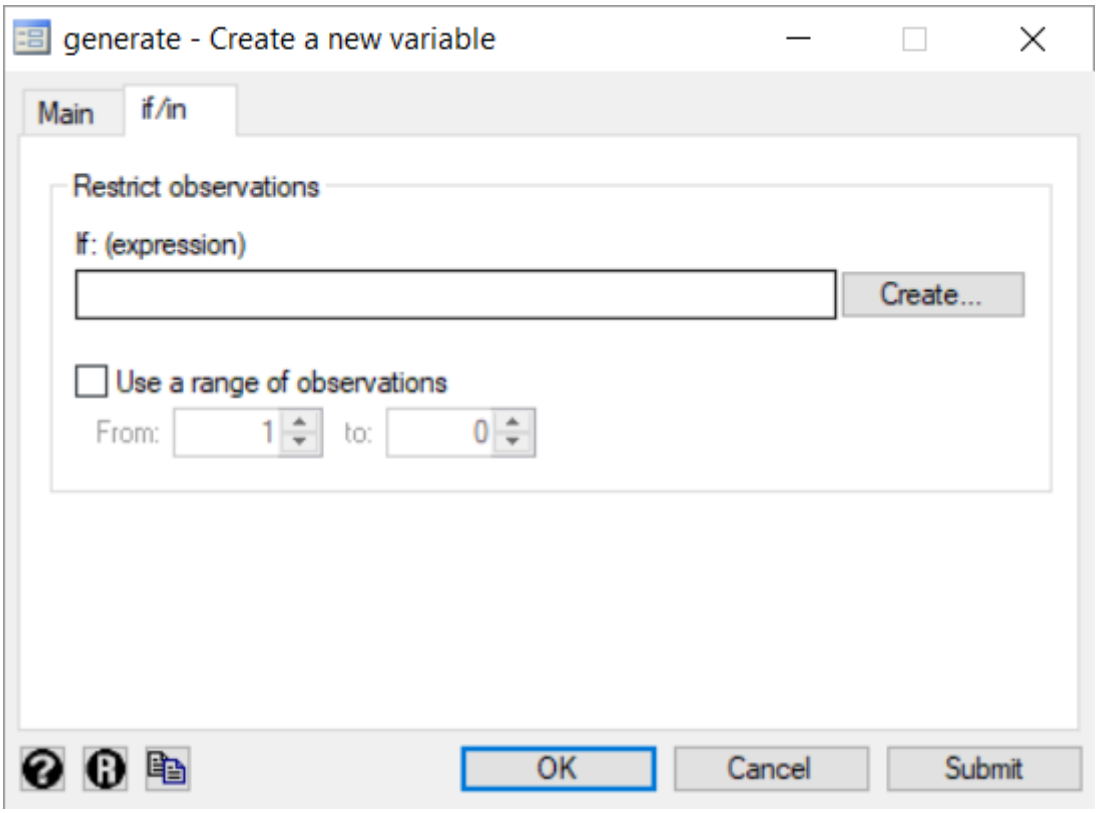
a) στο «***Variable name***» δηλώνουμε το όνομα της νέας μεταβλητής (π.χ. BMI).
b) Στο «***Specify a value or an expression***» δηλώνουμε τον μαθηματικό τύπο ή συνάρτηση βάση της οποίας θα υπολογιστούν οι τιμές που θα αποθηκευτούν στην μεταβλητή «BMI» (π.χ. weight/(height*height) ή weight/height^2)
c) Πατάμε **«Create»** και ανοίγει το πλαίσιο διαλόγου (*Εικόνα 3.2)*



Εικόνα 3.1: Επιλογή "Create new variable".



Εικόνα 3.2: Επιλογή "Expression builder".



Εικόνα 3.3: Επιλογή "Expression builder – if/in".

Σημείωση 1: Στο πλαίσιο «***Category***» της *Εικόνας 3.2* υπάρχουν έτοιμες συναρτήσεις που μπορούν να χρησιμοποιηθούν.
Σημείωση 2: Στην *Εικόνα 3.1* φαίνεται ότι η επιλογή «***Create new variable***» δίνει τη δυνατότητα μέσω της επιλογής «***If/in***» να ζητήσουμε οι τιμές της νέας μεταβλητής να υπολογιστούν μόνο όταν ισχύει κάποια συγκεκριμένη προϋπόθεση (π.χ., μόνο για τους άνδρες, δηλαδή αν gender = 1). (Εικόνα 3.3)

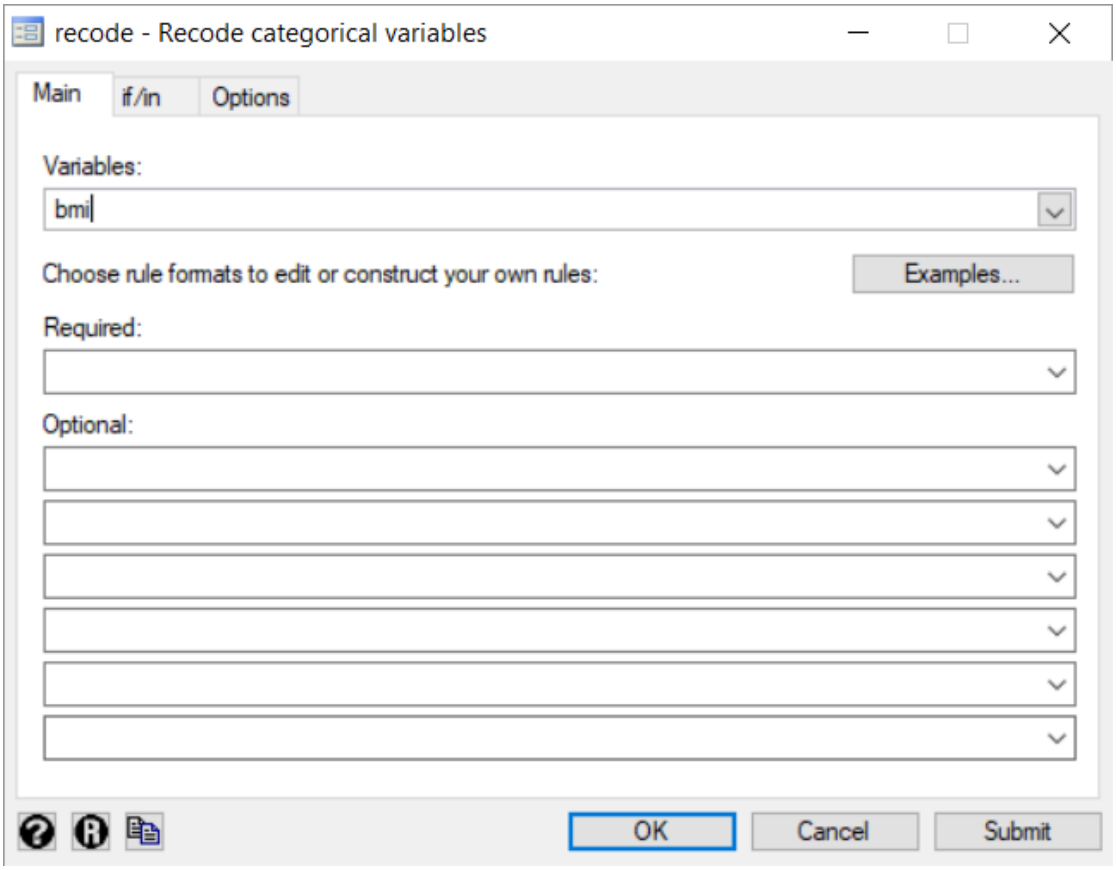
3.2 Επανα-κωδικοποίηση μεταβλητών στο STATA

Η επιλογή "***Recode***" στο STATA δίνει την δυνατότητα να συγχωνεύσουμε τις κατηγορίες μίας ήδη κατηγορικής μεταβλητής ή να κατηγοριοποιήσουμε μία ποσοτική μεταβλητή χρησιμοποιώντας συγκεκριμένα κατώφλια. Ας υποθέσουμε ότι επιθυμούμε να κατηγοριοποιήσουμε τον δείκτη μάζα σώματος (bmi) και να δημιουργήσουμε μία νέα μεταβλητή (π.χ., bmi_group) με 3 κατηγορίες.

Τα ***βήματα*** που πρέπει να ακολουθήσουμε είναι τα εξής:

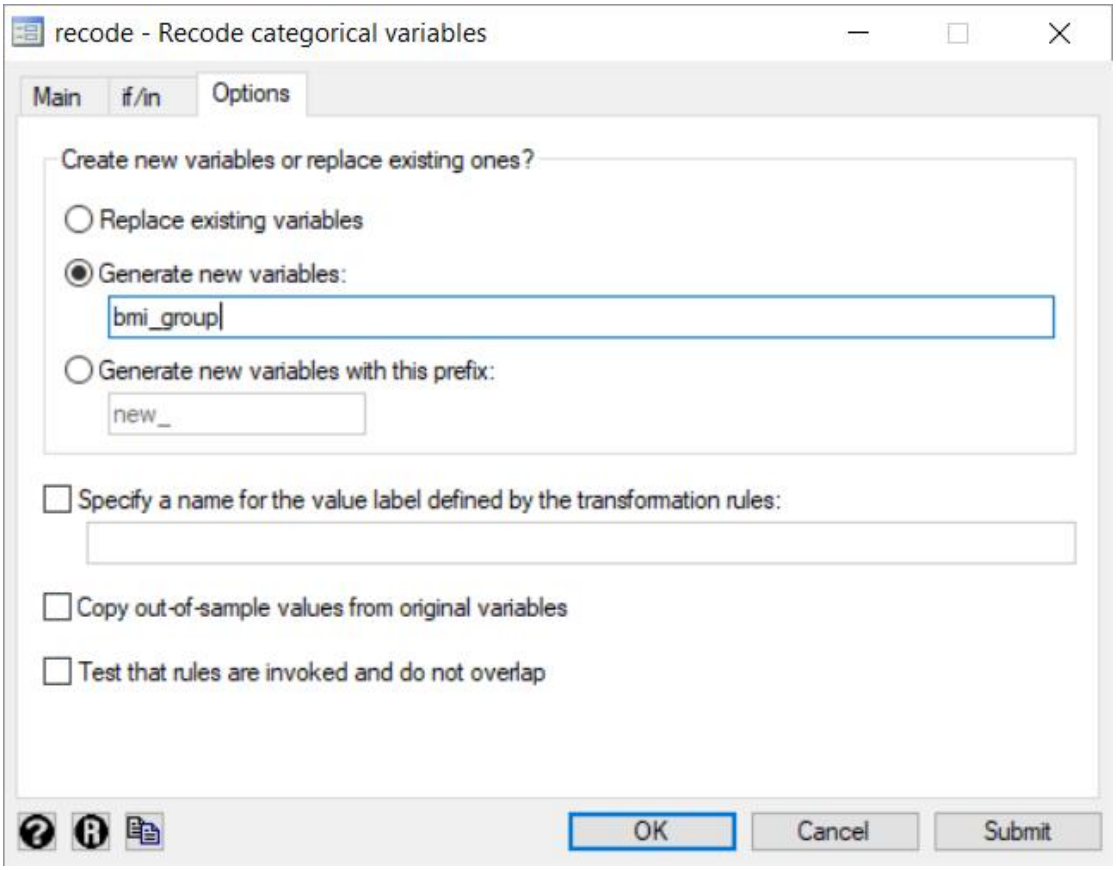
Data ⟶ Create or change data ⟶ Other variable–transformation commands ⟶ Recode categorical variable

Ανοίγει το πλαίσιο που φαίνεται στην *Εικόνα 3.4*.



Εικόνα 3.4: Επιλογή "Recode"

Το STATA μας δίνει τη δυνατότητα αυτές τις αλλαγές να τις πραγματοποιήσουμε είτε αντικαθιστώντας την υπάρχουσα μεταβλητή με την νέα (***Replace existing variables***) είτε δημιουργώντας μία νέα μεταβλητή (***Generate new variables***), όπως φαίνεται στην Εικόνα 3.5.



Εικόνα 3.4: Επιλογή "Recode"

Μερικές από τις δυνατότητες της επιλογής ***Recode*** εμφανίζονται στην *Εικόνα 3.5*.

Understanding recode rules

There can be an arbitrary number of rules, each contained in parentheses.
The most common forms for *rule* are

rule	Example	Meaning
(# = #)	(3 = 1)	3 recoded to 1
(# # = #)	(2 . = 9)	2 and . recoded to 9
(#/# = #)	(1/5 = 4)	1 through 5 recoded to 4
(nonmissing = #)	(nonmiss = 8)	all other nonmissing to 8
(missing = #)	(miss = 9)	all other missings to 9

Εικόνα 3.5: Προσδιορισμός των τιμών της νέας μεταβλητής που αντιστοιχούν σε κάθε μία τιμή της παλιάς μεταβλητής.

Συμβουλή: Είναι προτιμότερο να επιλέγουμε να δημιουργούμε νέες μεταβλητές (Recode into different variables) για να μην χαθεί η αρχική πληροφορία που μελλοντικά μπορεί να φανεί χρήσιμη.

3.3 Συγχώνευση αρχείων στο STATA

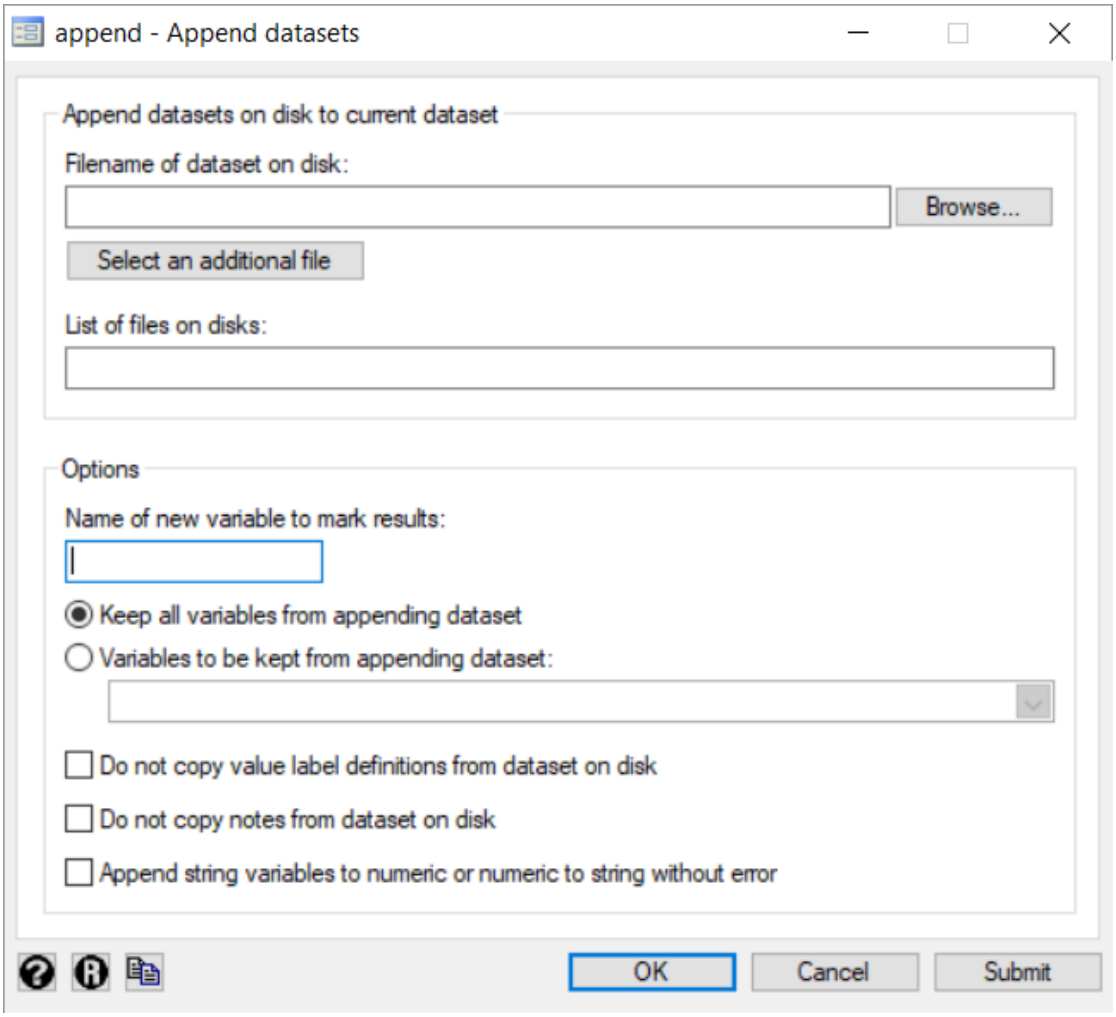
Υπάρχουν 2 διαφορετικές περιπτώσεις που ενδέχεται να επιθυμούμε τη συνένωση 2 αρχείων:

a. 2 αρχεία που να περιέχουν κοινές μεταβλητές, αλλά διαφορετικές εγγραφές/άτομα (π.χ. στο ένα αρχείο να είναι καταχωρημένα τα στοιχεία των γυναικών και στο άλλο τα στοιχεία των ανδρών που συμμετέχουν στη μελέτη). Σε αυτή την περίπτωση πρέπει να κάνουμε είναι «***Append two datasets***».
b. 2 αρχεία που να περιέχουν διαφορετικά στοιχεία για τα ίδια άτομα (π.χ. κάποιες μεταβλητές για τους συμμετέχοντες της μελέτης να έχουν καταχωρηθεί σε ένα αρχείο και κάποιες άλλες σε ένα άλλο αρχείο). Σε αυτή την περίπτωση αυτό που επιθυμούμε να κάνουμε είναι «***Merge two datasets***».

3.3.1. Προσθήκη ατόμων

Στην περίπτωση που επιθυμούμε να ***ενώσουμε 2 αρχεία με διαφορετικές εγγραφές***, τα ***βήματα*** που θα ακολουθήσουμε είναι τα εξής:

i. Ανοίγουμε το ένα από τα 2 αρχεία που επιθυμούμε να συνενώσουμε (π.χ. αρχείο «Males»).
ii. **Data ⟶ Combine datasets → Append two datasets** (*Εικόνα 3.5*)
iii. Στο «***Browse***» επιλέγουμε το αρχείο που περιέχει τα άτομα που θέλουμε να συνενώσουμε
iv. Πατάμε «***Άνοιγμα***» και
v. Έχω την επιλογή να κρατήσω όλες τις μεταβλητές στο νέο αρχείο (**Kept all variables from appending dataset**) είτε να επιλέξω ποιες θέλω να κρατήσω (**Variables to be kept from appending dataset**).
vi. Πατάμε «***Ok***» και το νέο αρχείο έχει δημιουργηθεί.

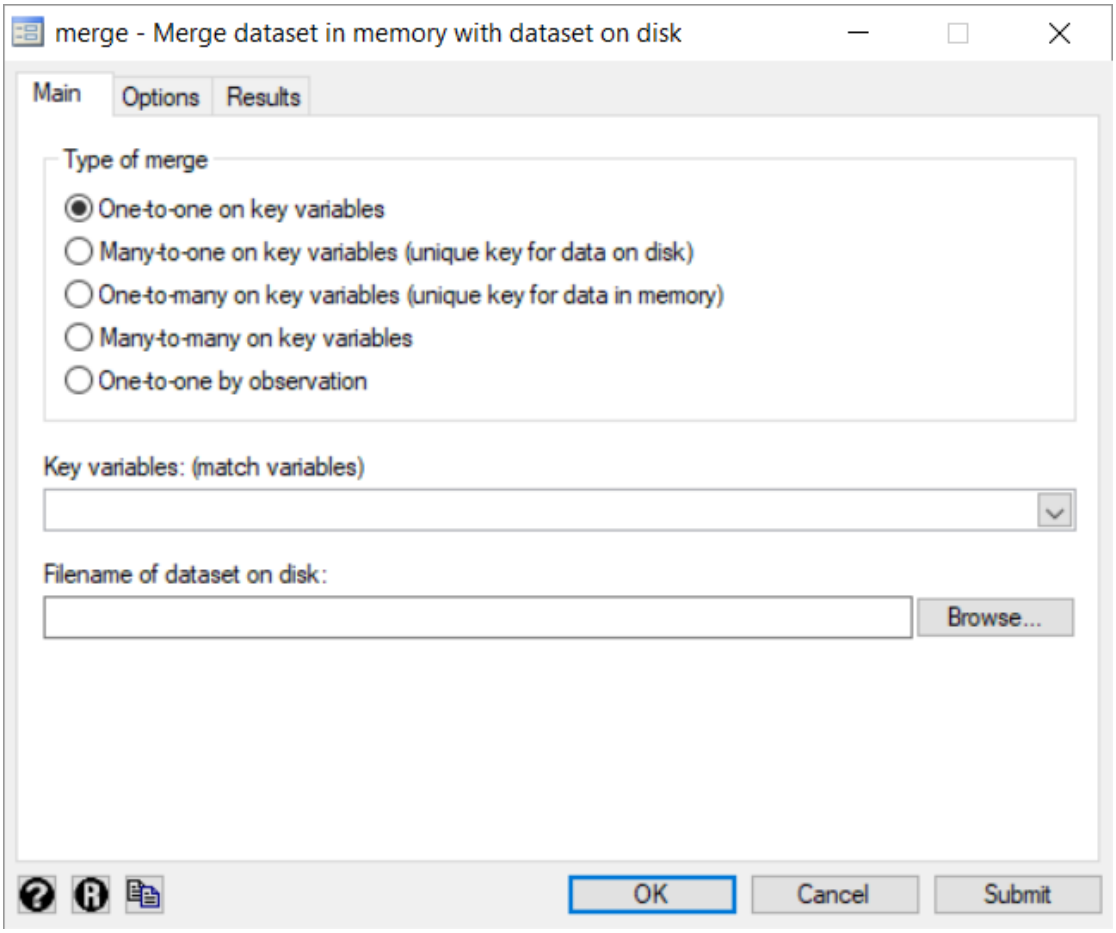


Εικόνα 3.5: Συνένωση αρχείων με ίδιες μεταβλητές αλλά διαφορετικές εγγραφές.

3.3.2 Προσθήκη μεταβλητών

Στην περίπτωση που επιθυμούμε να συνενώσουμε 2 αρχεία με διαφορετικές μεταβλητές αλλά ίδιες εγγραφές/άτομα, τα βήματα που θα ακολουθήσουμε είναι τα εξής:

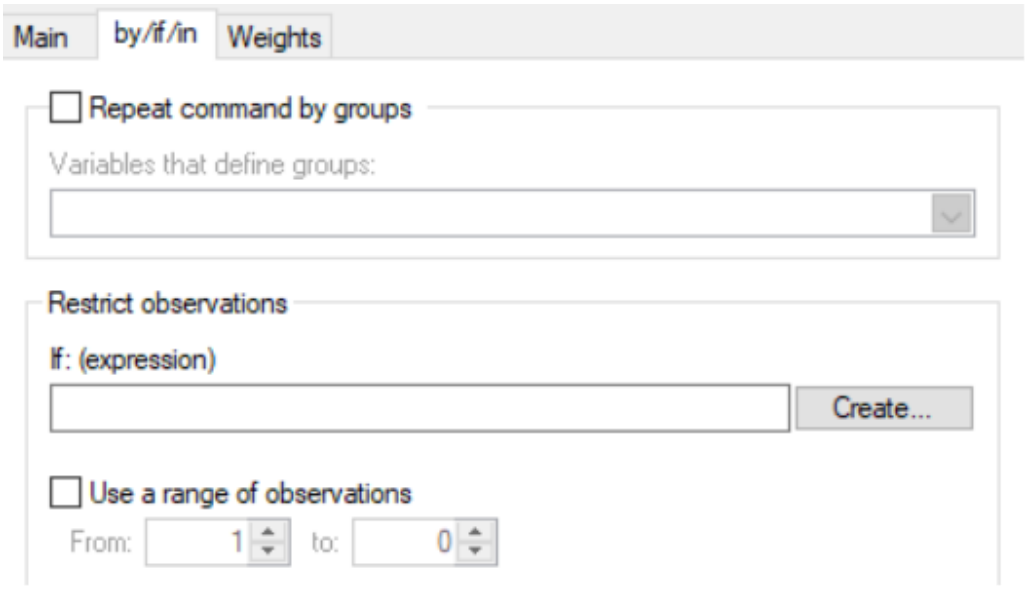
i. Ανοίγουμε το ένα από τα 2 αρχεία που επιθυμούμε να συνενώσουμε.
vii. Στο **Data ⟶ Combine datasets → Append two datasets** (*Εικόνα 3.6*)
viii. Στο «**Browse**» επιλέγουμε το αρχείο που περιέχει τις μεταβλητές που θέλουμε να συνενώσουμε
ii. Πατάμε «***Άνοιγμα***» και
iii. Στο «***Key variables***» βάζω την μεταβλητή βάσει της οποίας θα γίνει η συνένωση και πρέπει να είναι κοινή στα δύο αρχεία.
iv. Πατάμε «***Ok***» και το νέο αρχείο έχει δημιουργηθεί.



Εικόνα 3.6: Συνένωση αρχείων με διαφορετικές μεταβλητές αλλά ίδιες εγγραφές

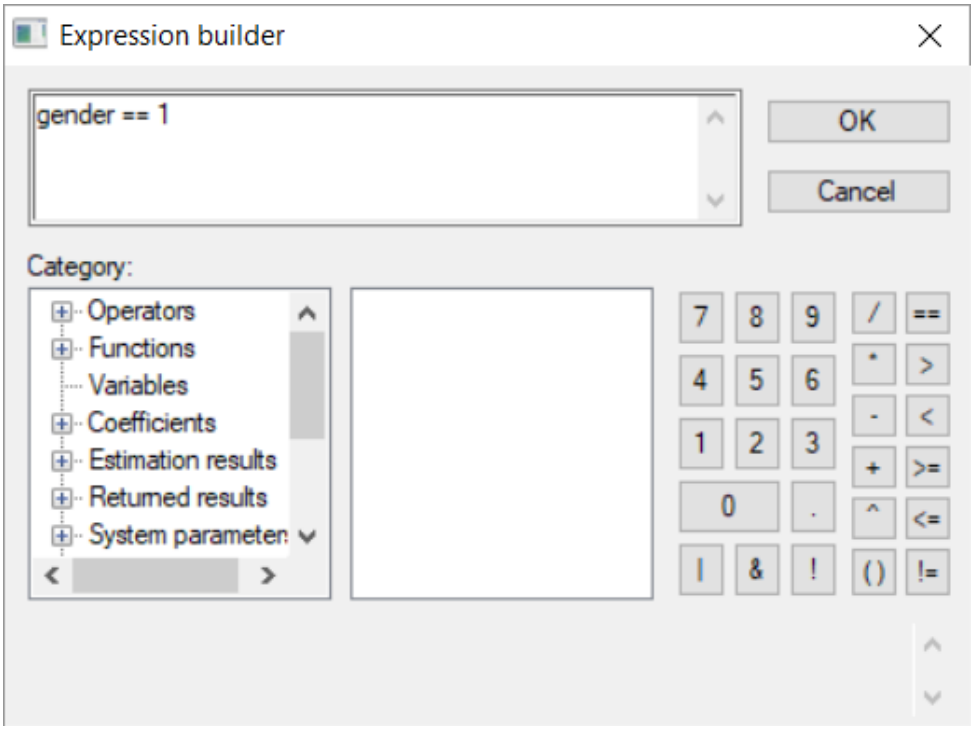
3.4 Επιλογή υπό-ομάδας δείγματος στο STATA

Με την επιλογή «**if**» του μενού by/if/in μπορούμε να επιλέξουμε μία υποομάδα ατόμων χρησιμοποιώντας ως φίλτρο μία ή περισσότερες μεταβλητές του αρχείου και να δημιουργήσουμε μία νέα μικρότερη βάση. Ας υποθέσουμε ότι διαθέτουμε μία μεγάλη βάση, από την οποία επιθυμούμε να δημιουργήσουμε μία μικρότερη βάση που να περιέχει μόνο τους άνδρες. Ανοίγουμε το αρχείο μας (π.χ. final_greecs)
Η επιλογή **by/if/in** δίνεται μέσα σε κάθε πλαίσιο, οποιασδήποτε εντολής χρησιμοποιήσουμε, όπως φαίνεται στην *Εικόνα 3.7*.



Εικόνα 3.7: Επιλογή συγκεκριμένων εγγραφών ενός αρχείου χρησιμοποιώντας ως φίλτρο κάποια μεταβλητή του αρχείου.

Όταν επιλέξουμε ***create*** ανοίγει το πλαίσιο **Expression builder** όπου γράφουμε τον περιορισμό μας. (*Εικόνα 3.8*).



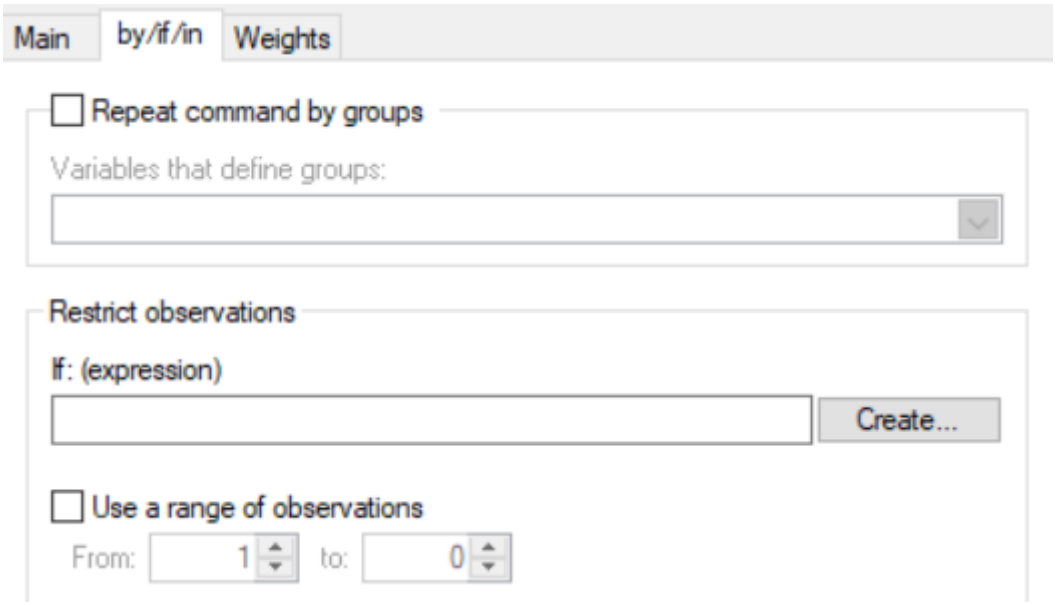
Εικόνα 3.8: Επιλογή Expression builder

3.5. Οργάνωση των αποτελεσμάτων της ανάλυσης ανά ομάδες στο STATA

Η επιλογή «***Repeat command by groups***» από το μενού by/if/in μας δίνει την δυνατότητα να διαιρέσουμε το αρχείο μας σε ομάδες χρησιμοποιώντας μία κατηγορική μεταβλητή (π.χ. φύλο). Χρησιμοποιώντας αυτή την επιλογή, όλες οι αναλύσεις που θα γίνουν στη συνέχεια στο συγκεκριμένο αρχείο θα πραγματοποιούνται και θα παρουσιάζονται σε ομάδες (π.χ. για άνδρες και γυναίκες, ξεχωριστά).

Η επιλογή **by/if/in** δίνεται μέσα σε κάθε πλαίσιο, οποιασδήποτε εντολής χρησιμοποιήσουμε, όπως φαίνεται στην *Εικόνα 3.9*.

Στο πλαίσιο Variables that define groups, παντώντας το βελάκι δεξιά, επιλέγουμε την μεταβλητή ή τις μεταβλητές για τις οποίες θέλουμε να γίνει ο διαμερισμός.



Εικόνα 3.9: Εφαρμογή της επιλογής ***Repeat command by groups***.

Σημείωση: Η επιλογή «Repeat command by groups» οδηγεί όχι μόνο στην εμφάνιση των αποτελεσμάτων των αναλύσεων ανά ομάδα, αλλά και στην σύγκριση των αποτελεσμάτων μεταξύ των ομάδων της μεταβλητής που ορίσαμε.

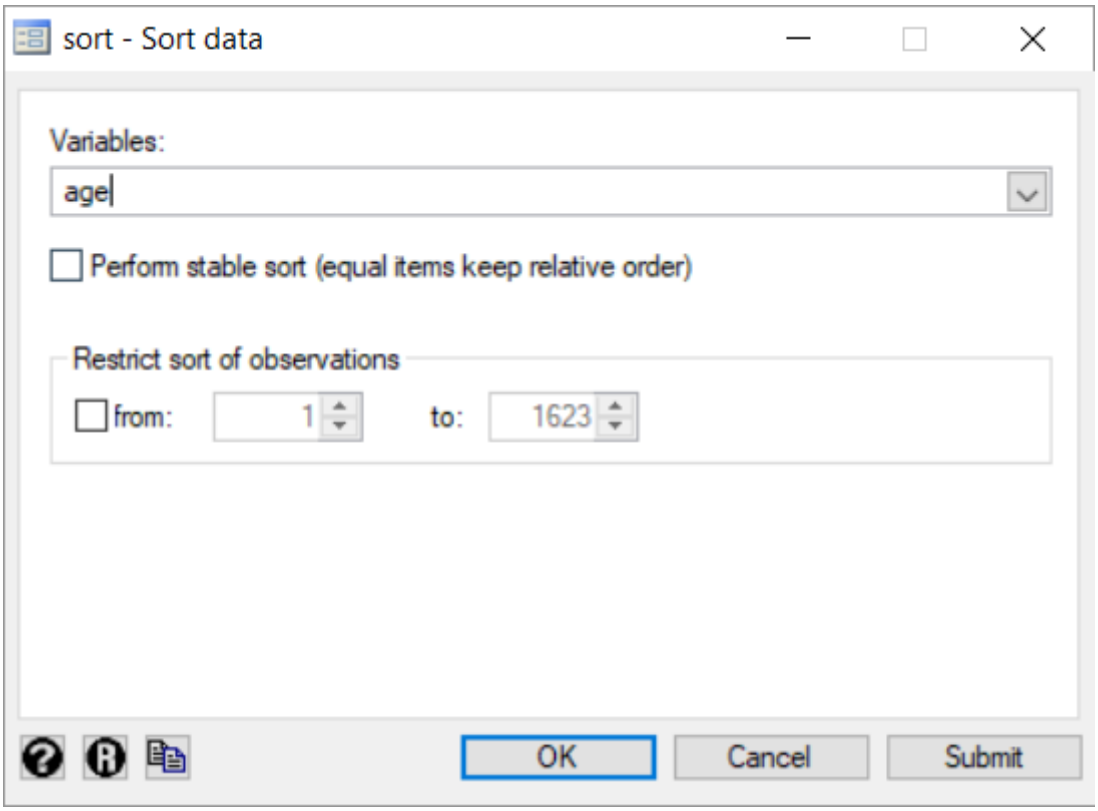
3.6 Ταξινόμηση παρατηρήσεων στο STATA

Η επιλογή «Sort» του STATA μας δίνει τη δυνατότητα να ταξινομήσουμε τις παρατηρήσεις του αρχείου μας βάσει των τιμών μιας συγκεκριμένης μεταβλητής. Ας υποθέσουμε, λοιπόν, ότι επιθυμούμε να ταξινομήσουμε τα άτομα του αρχείου βάσει της ηλικίας τους.

Τα βήματα που ακολουθούμε είναι τα εξής:

Data → Sort → Ascending Sort

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 3.10*
ii. Στο «***Variables***» τοποθετούμε την μεταβλητή βάσει της οποίας επιθυμούμε να γίνει η ταξινόμηση (π.χ. age).
iii. Στο «***Restrict sort of observations***» προεραιτικά επιλέγουμε σε ποιο εύρος παρατηρήσεων επιθυμούμε να γίνει η ταξινόμηση.



Εικόνα 3.10: Ταξινόμηση των παρατηρήσεων του αρχείου

4. Περιγραφική Στατιστική

4.1 Εισαγωγή

Περιγραφική στατιστική ορίζεται ο συνοπτικός τρόπος παρουσίασης των δεδομένων μας και περιλαμβάνει τα: περιγραφικά στατιστικά μέτρα και τα γραφήματα. Η περιγραφική στατιστική διαφέρει ανάλογα με τη φύση των μεταβλητών (π.χ. ποσοτικά χαρακτηριστικά ή ποιοτικά/κατηγορικά).

- Ως περιγραφικά στατιστικά μέτρα για τις ποιοτικές/κατηγορικές μεταβλητές χρησιμοποιούνται η απόλυτη (n) και η σχετική συχνότητα (%), ενώ τα πιο κατάλληλα γραφήματα για την παρουσίαση των ποιοτικών/κατηγορικών χαρακτηριστικών είναι το ραβδόγραμμα και οι πίτες.
- Ως περιγραφικά στατιστικά μέτρα για τις ποσοτικές μεταβλητές χρησιμοποιούνται μέτρα θέσης (π.χ. μέση τιμή, διάμεσος, κορυφή) και μέτρα διασποράς (π.χ. διακύμανση, τυπική απόκλιση, εύρος, ενδοτεταρτημοριακό εύρος), ενώ το ιστόγραμμα και το θηκόγραμμα είναι τα πιο κατάλληλα γραφήματα για την γραφική παουσίαση ποσοτικών δεδομένων.

4.2 Περιγραφικά μέτρα κατηγορικών μεταβλητών

Τα περιγραφικά στατιστικά μέτρα ποιοτικών/κατηγορικών μεταβλητών στο **SPSS** μπορούμε να τα αποκτήσουμε ακολουθώντας τα εξής ***βήματα***:

Analyze ⟶ Descriptive Statistics → Frequencies

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 4.1*.

ii. Επιλέγουμε την ή τις ποιοτικές /κατηγορικές μεταβλητές και τις περνάμε με το βελάκι στο κενό πλαίσιο «**Variable(s)**» (π.χ. φύλο (sex) και τύπος διάγνωσης οξέος στεφανιαίου συνδρόμου (ecg)).

iii. Αν επιθυμούμε ταυτόχρονα να δημιουργηθεί και ένα κατάλληλο γράφημα (π.χ. ραβδόγραμμα ή πίτα), τότε πατάμε το κουμπί επιλογών «***Charts***» και ανοίγει το πλαίσιο διαλόγου της *Εικόνας 4.2*.

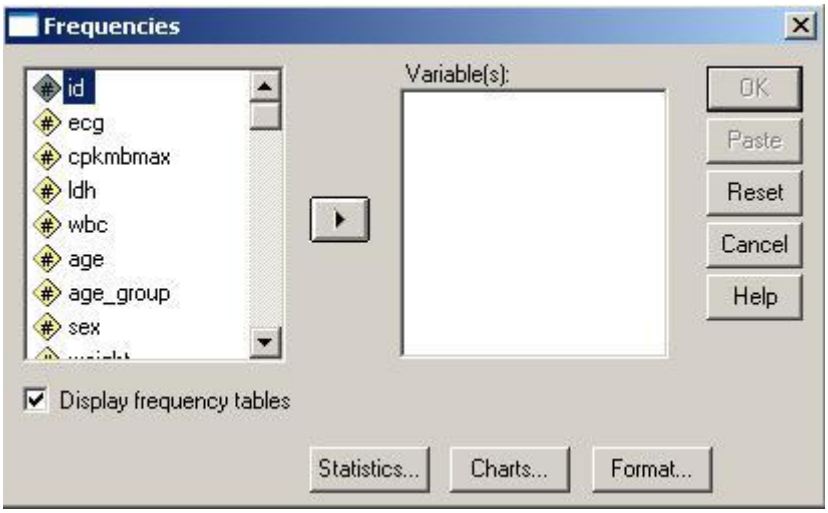
iv. Επιλέγουμε «***Bar charts***» ή «***Pie charts***»

v. Πατάμε "***Ok"***.

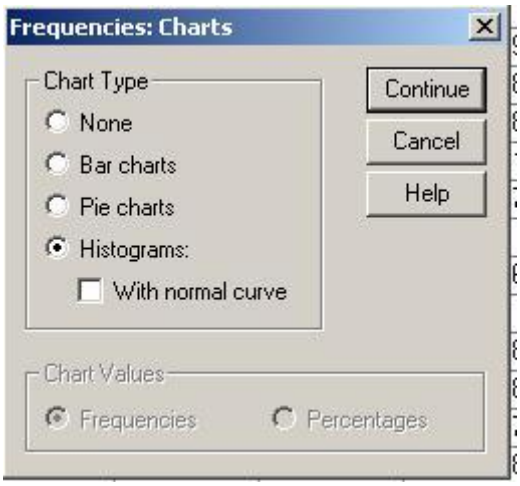
vi. Τα αποτελέσματα παρουσιάζονται στους *Πίνακες 4.1 και 4.2* και στις *Εικόνες 4.3 & 4.4*.

Στους *Πίνακες 4.1 και 4.2* παρουσιάζεται η απόλυτη και η σχετική συχνότητα για κάθε κατηγορία του φύλου και του τύπου διάγνωσης, αντίστοιχα. Στην στήλη «***Frequency***» παρουσιάζεται ο απόλυτος αριθμός των ατόμων του δείγματος που ανήκουν σε κάθε κατηγορία της μεταβλητής, ο αριθμός των ελλειπουσών τιμών (των ατόμων για τα οποία δεν υπάρχει καταχωρημένη τιμή για τη συγκεκριμένη μεταβλητή), το σύνολο των ατόμων συμπεριλαμβανομένων των ελλειπουσών τιμών και το σύνολο των ατόμων εκτός αυτών με ελλείπουσα τιμή για τη συγκεκριμένη μεταβλητή. Στην στήλη «***Percent***» των πινάκων, παρουσιάζονται οι σχετικές συχνότητες (%) για κάθε κατηγορία της μεταβλητής και για τις ελλείπουσες τιμές, ξεχωριστά, ενώ στην στήλη «***Valid percent***» παρουσιάζεται το ποσοστό μόνο μεταξύ των έγκυρων τιμών δηλ. μεταξύ των ατόμων που έχουν καταχωρημένη τιμή στην συγκεκριμένη μεταβλητή. Έτσι, λοιπόν, από τον *Πίνακα 4.1* διαπιστώνουμε ότι το 75,9% των ατόμων της μελέτης είναι άνδρες και από τον *Πίνακα 4.2* διαπιστώνουμε

ότι το 37,5% των ατόμων της μελέτης είναι με έμφραγμα STEMI. Τέλος, στις *Εικόνες 4.3 και 4.4* παρουσιάζονται τα ραβδογράμματα για το φύλο και τον τύπο της διάγνωσης, αντίστοιχα.



Εικόνα 4.1: Απόκτηση περιγραφικών στατιστικών μέτρων για ποιοτικές /κατηγορικές μεταβλητές στο SPSS



Εικόνα 4.2: Επιλογή της εμφάνισης γραφημάτων στο SPSS

sex

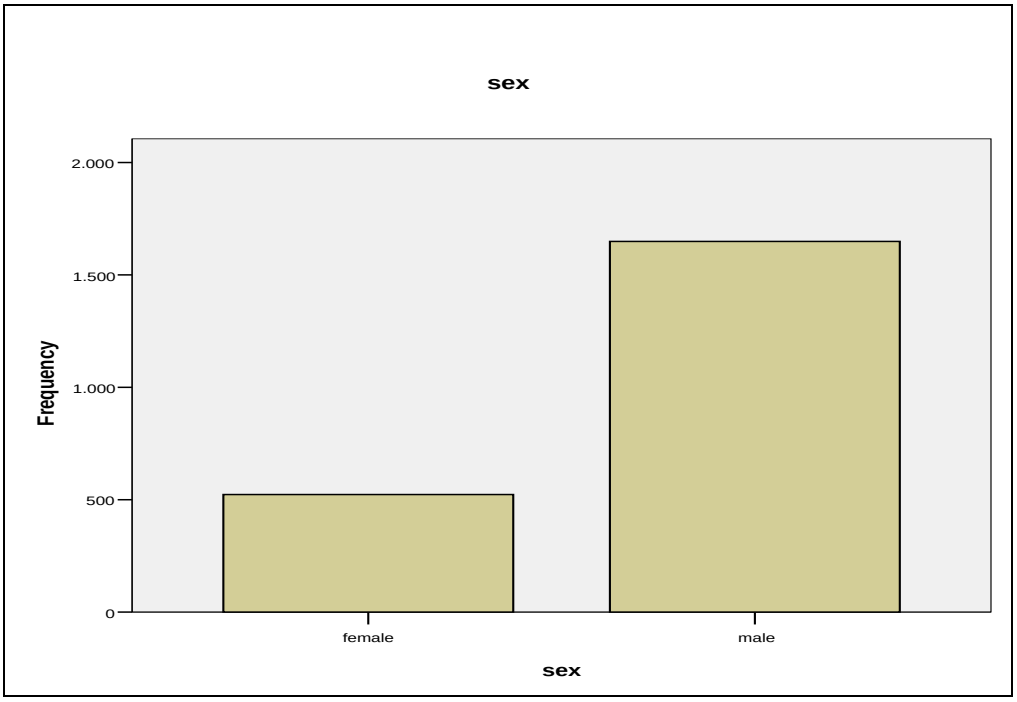
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	523	24,1	24,1	24,1
	male	1649	75,9	75,9	100,0
	Total	2172	100,0	100,0	

Πίνακας 4.1: Περιγραφικά στατιστικά μέτρα για το φύλο

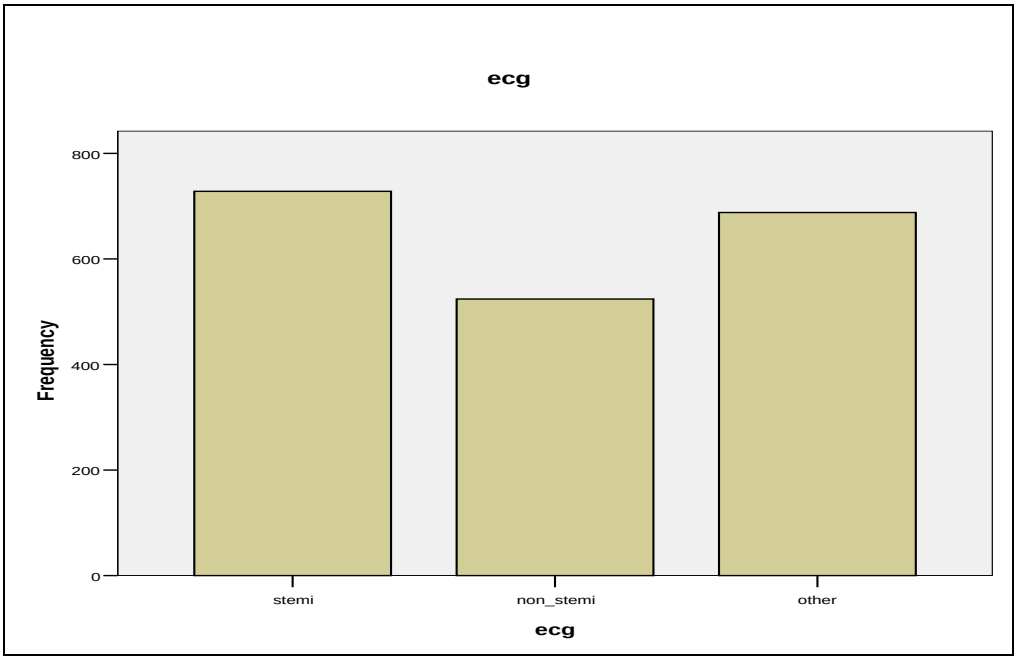
ecg

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	stemi	728	33,5	37,5	37,5
	non_stemi	524	24,1	27,0	64,5
	other	688	31,7	35,5	100,0
	Total	1940	89,3	100,0	
Missing	System	232	10,7		
Total		2172	100,0		

Πίνακας 4.2: Περιγραφικά στατιστικά μέτρα για τον τύπο διάγνωσης



Εικόνα 4.3: Ραβδόγραμμα για το φύλο



Εικόνα 4.4: Ραβδόγραμμα για τον τύπο διάγνωσης

Εκτός από τον παραπάνω τρόπο δημιουργίας των γραφημάτων, γραφήματα μπορούμε να φτιάξουμε χρησιμοποιώντας της επιλογές του «***Graphs***» από το μενού επιλογών του SPSS. Ας υποθέσουμε ότι επιθυμούμε να δημιουργήσουμε μία «πίτα» για την μεταβλητή «φύλο». Τα ***βήματα*** που ακολουθούμε είναι:

Graphs → Pie

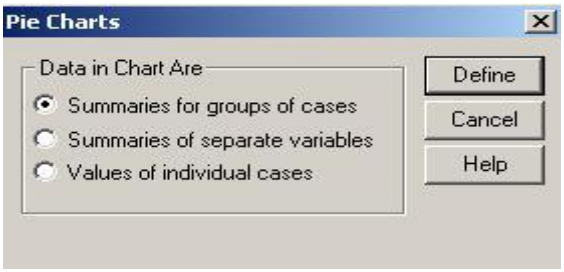
i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 4.5*.

ii. Πατάμε «***Define***» και

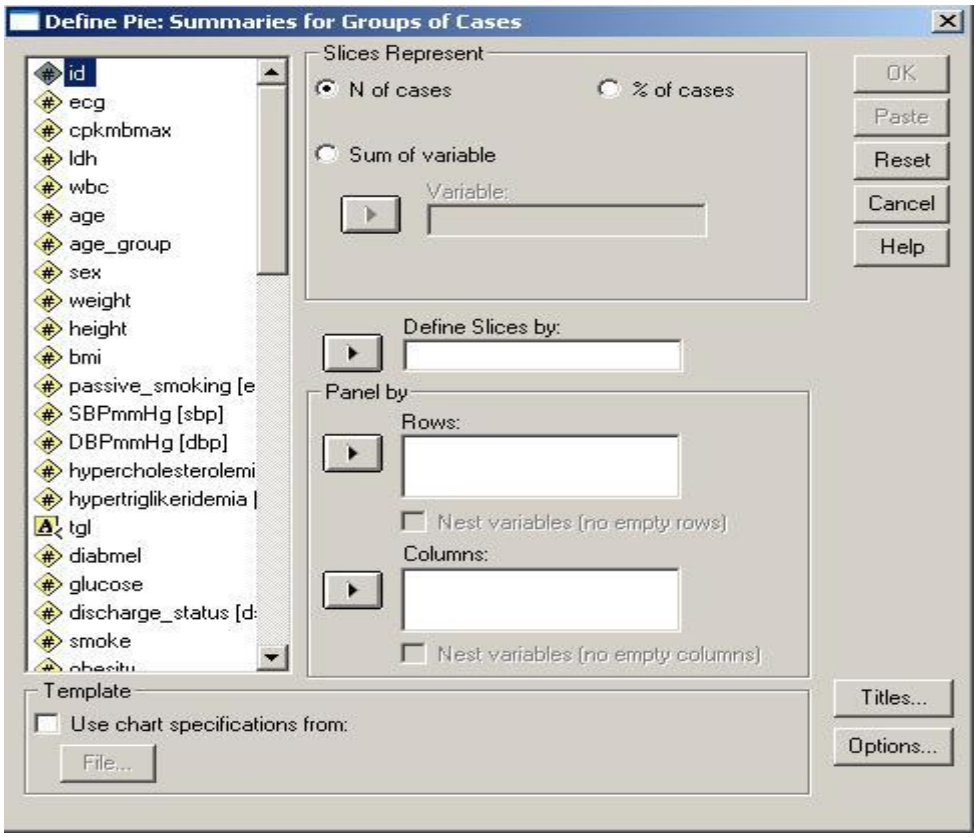
iii. Ανοίγει το πλαίσιο της *Εικόνας 4.6*.

iv. Στο «Define Slices by» τοποθετούμε με το βελάκι την μεταβλητή για την οποία θέλουμε να φτιάξουμε την «πίτα» (π.χ. sex).

v. «***Ok***»



Εικόνα 4.5: Δημιουργία πίτας μέσω του μενού επιλογών «Graphs» του SPSS



Εικόνα 4.6: Προσδιορισμός της μεταβλητής για την οποία επιθυμούμε να φτιάξουμε την πίτα.

Αντίστοιχα, τα περιγραφικά στατιστικά μέτρα ποιοτικών/κατηγορικών μεταβλητών στο **STATA** μπορούμε να τα αποκτήσουμε ακολουθώντας τα εξής ***βήματα***:

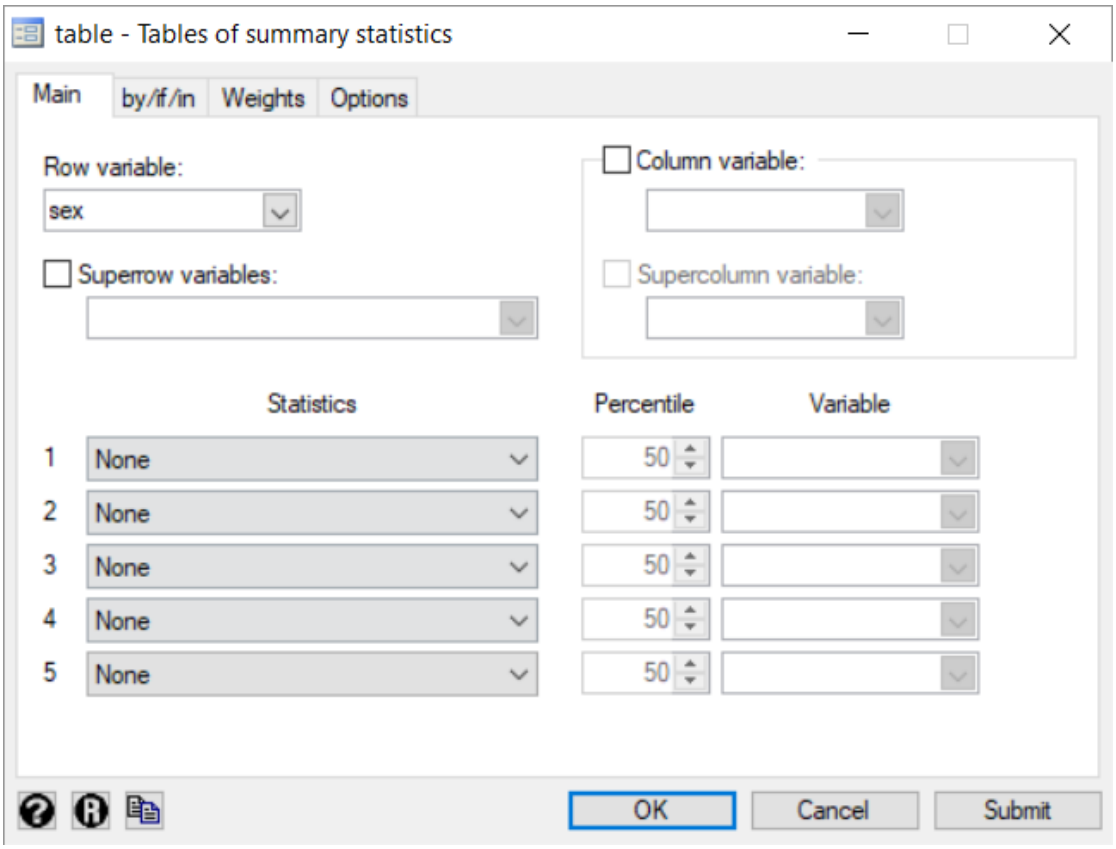
Statistics ⟶ Summaries, tables and tests → Tables → Table of summary statistics (table)

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 4.7*.

ii. Επιλέγουμε την ή τις ποιοτικές /κατηγορικές μεταβλητές και τις περνάμε με το βελάκι δεξιά στο πλαίσιο «**Variable(s)**» (π.χ. φύλο (sex))

iii. Επίσης, έχουμε την δυνατότητα επιλογής υποομάδας δείγματος ή την ανάλυση αποτελέσματών ανά ομάδα από το μενού **by/if/in**.

iv. Πατάμε "***OK***".



Εικόνα 4.7: Απόκτηση περιγραφικών στατιστικών μέτρων για ποιοτικές /κατηγορικές μεταβλητές στο STATA

Στο STATA χρησιμοποιείται η εντολή ***table*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 4.8*:

```
    [R] table — Tables of summary statistics


Syntax

        table rowvar [colvar [supercolvar]] [if] [in] [weight] [, options]


    options                Description
    ---------------------------------------------------------------------------
    Main
      contents(clist)      contents of table cells; select up to five statistics; default is contents(freq)
      by(superrowvarlist)  superrow variables

    Options
      cellwidth(#)         cell width
      csepwidth(#)         column-separation width
      stubwidth(#)         stub width
      scsepwidth(#)        supercolumn-separation width
      center               center-align table cells; default is right-align
      left                 left-align table cells; default is right-align
      cw                   perform casewise deletion
      row                  add row totals
      col                  add column totals
      scol                 add supercolumn totals
      concise              suppress rows with all missing entries
      missing              show missing statistics with period
      replace              replace current data with table statistics
      name(string)         name new variables with prefix string
      format(%fmt)         display format for numbers in cells; default is format(%9.0g)
```

Εικόνα 4.8: Βασική σύνταξη της εντολής table

Αν επιθυμούμε ταυτόχρονα να δημιουργηθεί και ένα κατάλληλο γράφημα (π.χ. ραβδόγραμμα ή πίτα), τότε πατάμε το κουμπί επιλογών «Graphics» και επιλέγουμε το κατάλληλο γράφημα (Bar chart ή Pie chart).

Ας υποθέσουμε ότι επιθυμούμε να δημιουργήσουμε μία «πίτα» για την μεταβλητή «φύλο». Τα ***βήματα*** που ακολουθούμε είναι:

Graphics ⟶ Pie Chart

i. Ανοίγει το πλαίσιο της *Εικόνας 4.9*.

ii. Στο «Category variable» τοποθετούμε με το βελάκι την μεταβλητή για την οποία θέλουμε να φτιάξουμε την «πίτα» (π.χ. sex).

iii. «***OK***»

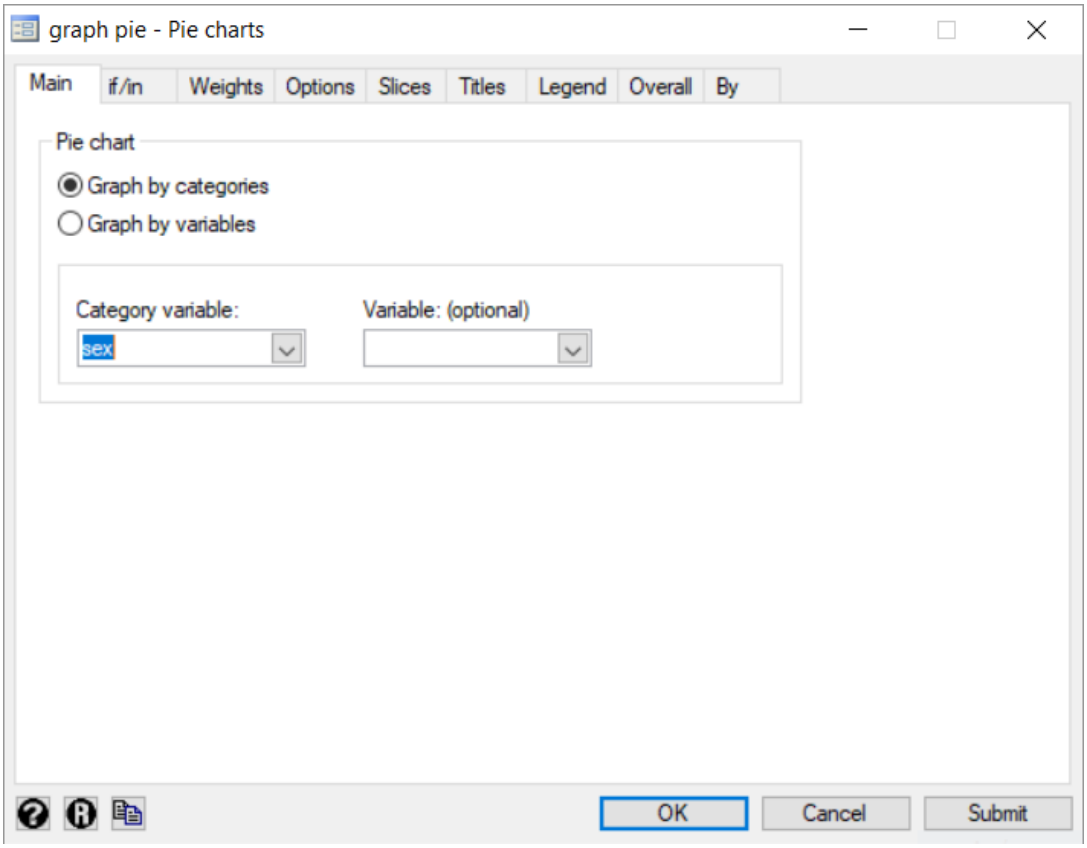
Ας υποθέσουμε επιπλέον ότι επιθυμούμε να δημιουργήσουμε ένα «ραβδόγραμμα» για την μεταβλητή «φύλο». Τα ***βήματα*** που ακολουθούμε είναι:

Graphics ⟶ Bar Chart

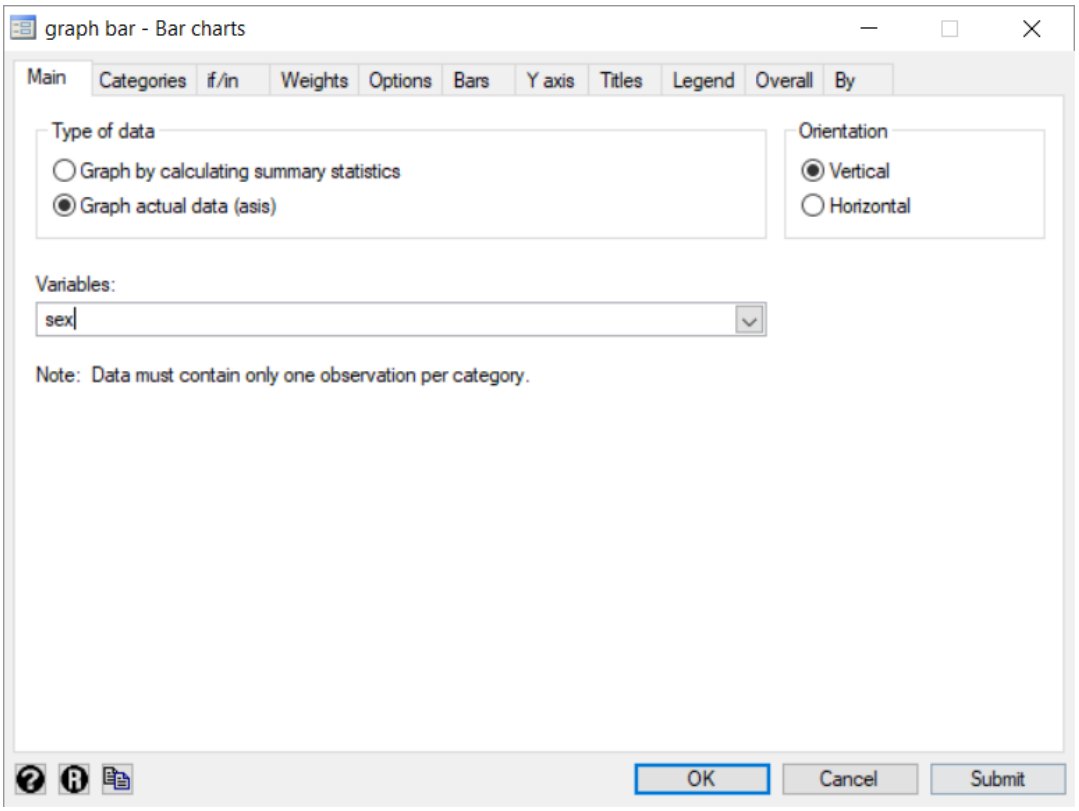
iv. Ανοίγει το πλαίσιο της *Εικόνας 4.10*.

v. Στο «Variables» τοποθετούμε με το βελάκι την μεταβλητή για την οποία θέλουμε να φτιάξουμε το «ραβδόγραμμα» (π.χ. sex).

vi. «***OK***»



Εικόνα 4.9: Δημιουργία πίτας μέσω του μενού επιλογών «Graphics» του STATA



Εικόνα 4.10: Δημιουργία ραβδογράμματος μέσω του μενού επιλογών «Graphics» του STATA

4.3 Περιγραφικά μέτρα ποσοτικών μεταβλητών

Τα περιγραφικά στατιστικά μέτρα ποσοτικών μεταβλητών μπορούμε να τα αποκτήσουμε ακολουθώντας τα εξής ***βήματα***:

Analyze ⟶ Descriptive Statistics → Frequencies

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 4.1*.

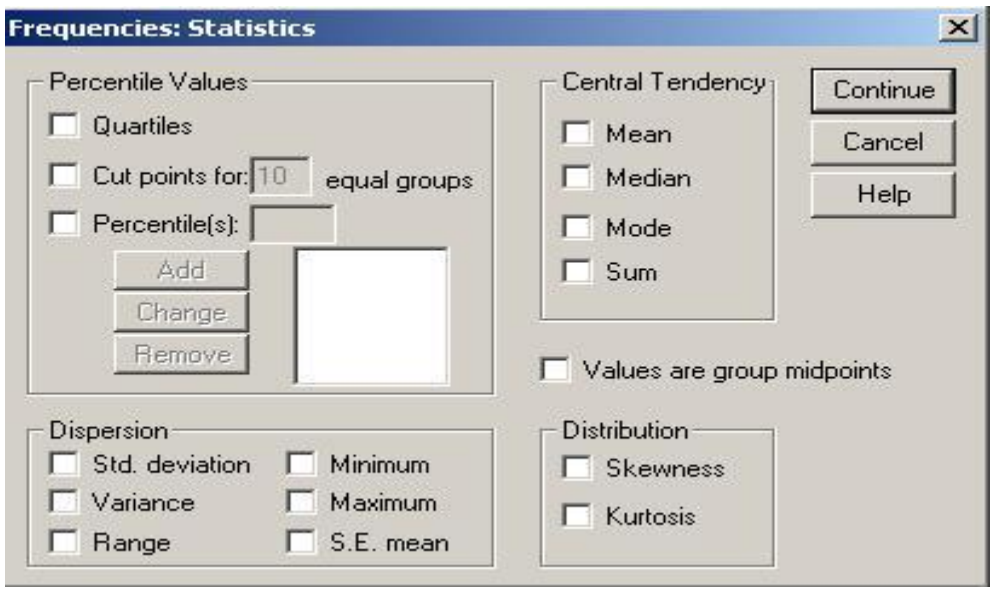
ii. Επιλέγουμε την ή τις ποσοτικές μεταβλητές και τις περνάμε με το βελάκι στο κενό πλαίσιο «**Variable(s)**».

iii. Πατάμε το κουμπί επιλογών ***Statistics*** και εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 4.11*.

iv. Διαλέγουμε τα στατιστικά μέτρα που μας ενδιαφέρουν (π.χ. Mean, median, Std. Deviation, κλπ).

v. Τα αποτελέσματα για την μεταβλητή ηλικία (Age of Subjects) φαίνονται στον *Πίνακα 4.3*.

vi. Από το ίδιο παράθυρο επιλογών (*Εικόνα 4.1*) μπορούμε μέσω του κουμπιού επιλογών ***Charts*** να εμφανίσουμε μια σειρά από Γραφήματα. Το πλαίσιο διαλόγου που εμφανίζεται πατώντας το κουμπί επιλογών ***Charts*** φαίνεται στην *Εικόνα 4.2*.



Εικόνα 4.11: Επιλογή των περιγραφικών στατιστικών για ποσοτικές μεταβλητές.

Statistics

age

N	Valid	2172
	Missing	0
Mean		66,10
Median		68,00
Mode		70
Std. Deviation		13,050
Range		77
Minimum		23
Maximum		100
Percentiles	10	47,00
	20	54,00
	25	57,00
	30	59,00
	40	65,00
	50	68,00
	60	71,00
	70	74,00
	75	76,00
	80	77,40
	90	82,00

Πίνακας 4.3: Περιγραφικά Στατιστικά Μέτρα για συνεχείς μεταβλητές (π.χ. την ηλικία «Age of subjects»).

Από τον *Πίνακα 4.3* διαπιστώνουμε ότι η μέση τιμή της ηλικίας του δείγματος μας είναι 66,10 έτη, η διάμεσος είναι 68 και η ελάχιστη και η μέγιστη ηλικία είναι 23 και 100 έτη, αντίστοιχα.

Αντίστοιχα, τα περιγραφικά στατιστικά μέτρα ποσοτικών μεταβλητών στο **STATA** μπορούμε να τα αποκτήσουμε ακολουθώντας τα εξής ***βήματα***:

Statistics → Summaries, tables and tests → Summary and descriptive statistics → Summary statistics

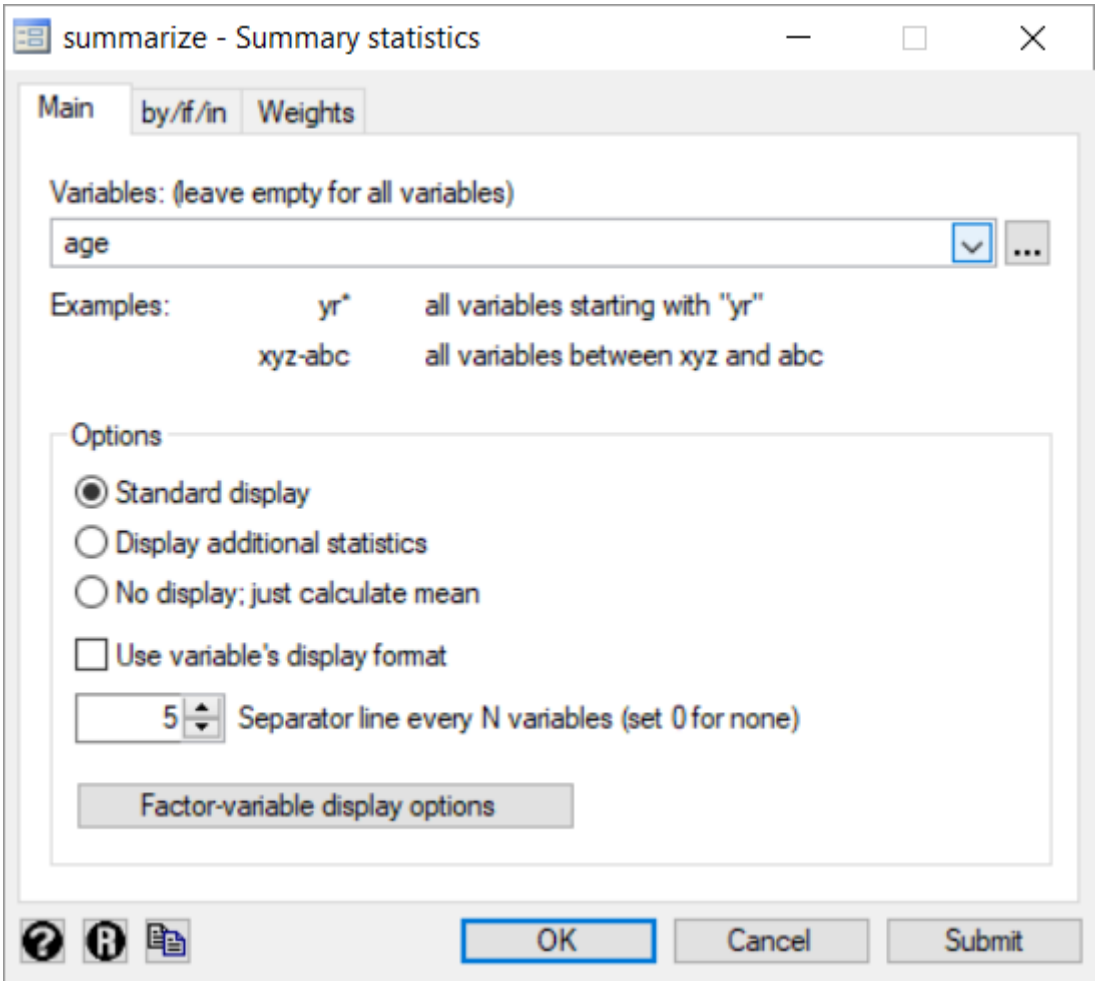
i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 4.12*.

ii. Επιλέγουμε την ή τις ποσοτικές μεταβλητές και τις περνάμε με το βελάκι δεξιά στο πλαίσιο «**Variable(s)**» (π.χ. ηλικία (age))

iii. Επίσης, έχουμε την δυνατότητα επιλογής υποομάδας δείγματος ή την ανάλυση αποτελέσματών ανά ομάδα από το μενού **by/if/in**.

iv. Πατάμε "***Ok"***.

v. Τέλος, από το menu επιλογών μπορούμε μέσω του ***Graphics*** να εμφανίσουμε μια σειρά από γραφήματα.



Εικόνα 4.12: Απόκτηση περιγραφικών στατιστικών μέτρων για ποιοτικές /κατηγορικές μεταβλητές στο STATA

Στο STATA χρησιμοποιείται η εντολή ***summarize*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 4.13*:

```
    [R] summarize — Summary statistics


Syntax

        summarize [varlist] [if] [in] [weight] [, options]

    options            Description
    ---------------------------------------------------------------------------
    Main
      detail           display additional statistics
      meanonly         suppress the display; calculate only the mean;
                         programmer's option
      format           use variable's display format
      separator(#)     draw separator line after every # variables; default is
                         separator(5)
      display_options  control spacing and base and empty cells
```

Εικόνα 4.13: Βασική σύνταξη της εντολής summarize

5. Έλεγχος ύπαρξης συσχέτισης μεταξύ δύο κατηγορικών μεταβλητών.

5.1 Εισαγωγή

Ως γνωστόν, το πιο δημοφιλές στατιστικό κριτήριο για τον έλεγχο ύπαρξης συσχέτισης μεταξύ 2 κατηγορικών/ποιοτικών μεταβλητών με k και s κατηγορίες, αντίστοιχα (π.χ. φύλο και κατάσταση υγείας – καλή, μέτρια, κακή) είναι το $\mathbf{X^2}$ που πρότεινε ο **Pearson**. Συνεπώς, η μηδενική και εναλλακτική υπόθεση είναι:

H_0: ΔΕΝ υπάρχει καμία συσχέτιση ανάμεσα στις 2 μεταβλητές (π.χ. φύλο και κατάσταση υγείας)

H_1: ΥΠΑΡΧΕΙ συσχέτιση ανάμεσα στις 2 μεταβλητές (π.χ. φύλο και κατάσταση υγείας)

Για να υπολογιστεί αυτό το κριτήριο, κρίνεται αναγκαία η δημιουργία ενός «πίνακα συνάφειας». Αυτός ο πίνακας αποτελείται από k*s κελιά, μέσα στα οποία κατανέμονται οι ν παρατηρήσεις του δείγματος. Όπου ν_{ij} είναι η παρατηρούμενη συχνότητα των ατόμων που ανήκουν στην κατηγορία i της μεταβλητής Α (π.χ. φύλο) και στην κατηγορία j της μεταβλητής Β (π.χ. κατάσταση υγείας). Με τη βοήθεια του νόμου ανεξαρτησίας των πιθανοτήτων, υπολογίζεται ο αναμενόμενος αριθμός ατόμων που θα έπρεπε να υπάρχει σε κάθε κελί του «πίνακα συνάφειας» αν οι 2 μεταβλητές ήταν ανεξάρτητες.
Αν οι αναμενόμενες συχνότητες που υπολογίζονται δεδομένου ότι οι δύο μεταβλητές είναι ανεξάρτητες (μηδενική υπόθεση), δεν διαφέρουν στατιστικά σημαντικά από τις παρατηρούμενες συχνότητες, τότε δεν υπάρχει σημαντική απόδειξη για την απόρριψη της μηδενικής υπόθεσης και συνεπώς θεωρείται ότι οι δύο μεταβλητές είναι ανεξάρτητες. Αυτό θα ελεγχθεί, υπολογίζοντας την τιμή του χ^2 κριτηρίου:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{s}\frac{\left(\nu_{ij} - \mathrm{E}_{ij}\right)^2}{\mathrm{E}_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{s}\frac{\left(\nu_{ij} - \nu p_i q_j\right)^2}{\nu p_i q_j}$$

Το παραπάνω κριτήριο ακολουθεί ασυμπτωτικά την χ^2 κατανομή με ks-1 βαθμούς ελευθερίας. Για να είναι ικανοποιητική, όμως, η προσέγγιση του χ^2 κριτηρίου από την χ^2 κατανομή θα πρέπει να ισχύουν οι εξής προϋποθέσεις:

- Οι αναμενόμενες συχνότητες σε όλα τα κελιά να είναι μεγαλύτερες του 5, ή
- Όλες οι αναμενόμενες τιμές να είναι μεγαλύτερες του 1 και το πολύ 20% από αυτές να είναι μικρότερες του 5.

Για να καταλήξουμε στο συμπέρασμα ότι υπάρχει συσχέτιση ανάμεσα στις 2 μεταβλητές (π.χ. φύλο και κατάσταση υγείας), θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value) να είναι < α = 0,05 ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).

5.2 Έλεγχος X^2 με τη χρήση του SPSS

Ας υποθέσουμε ότι στόχος μας είναι να ελέγξουμε αν υπάρχει στατιστικά σημαντική συσχέτιση ανάμεσα στο φύλο (sex) και τον τύπο διάγνωσης του οξέος στεφανιαίου συνδρόμου (ecg: 1=STEMI, 2=NSTEMI & 3=other). Για να υπολογίσουμε το στατιστικό κριτήριο αλλά και την αντίστοιχη πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή την p) ακολουθούμε τα εξής ***βήματα***:

Analyze ⟶ Descriptive Statistics → Cross Tabs

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 5.1*

ii. Τοποθετούμε μια ή περισσότερες κατηγορικές μεταβλητές στα παράθυρα "*Rows*" (π.χ. ecg, δηλ. τύπος διάγνωσης οξέος στεφανιαίου συνδρόμου (ΟΣΣ) από το ΗΚΓ) & "*Columns*" (π.χ. sex, δηλ. φύλο).

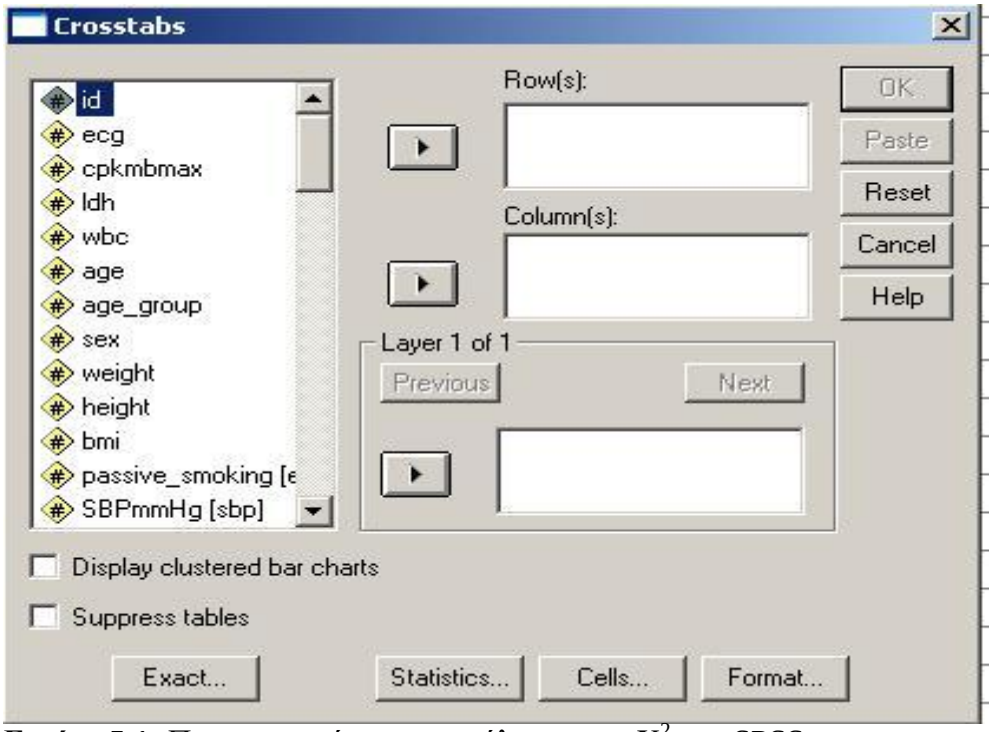
iii. Έπειτα πατάμε το κουμπί επιλογών ***Statistics*** και ανοίγει το πλαίσιο διαλόγου της *Εικόνας 5.2*.

iv. Επιλέγουμε το ***chi-square*** για να υπολογιστεί το στατιστικό κριτήριο X^2 (βλ. *Εικόνα 5.2*). Από το μενού ***Statistics*** μπορούμε επίσης να ζητήσουμε να υπολογιστούν και άλλα στατιστικά μέτρα συσχέτισης κατηγορικών μεταβλητών, όπως το *φ*, το *V* του Cramer, κλπ., καθώς επίσης και στατιστικά κριτήρια για τον έλεγχο της πιθανής συσχέτισης μεταξύ διατάξιμων μεταβλητών, όπως t-Kendal κλπ.

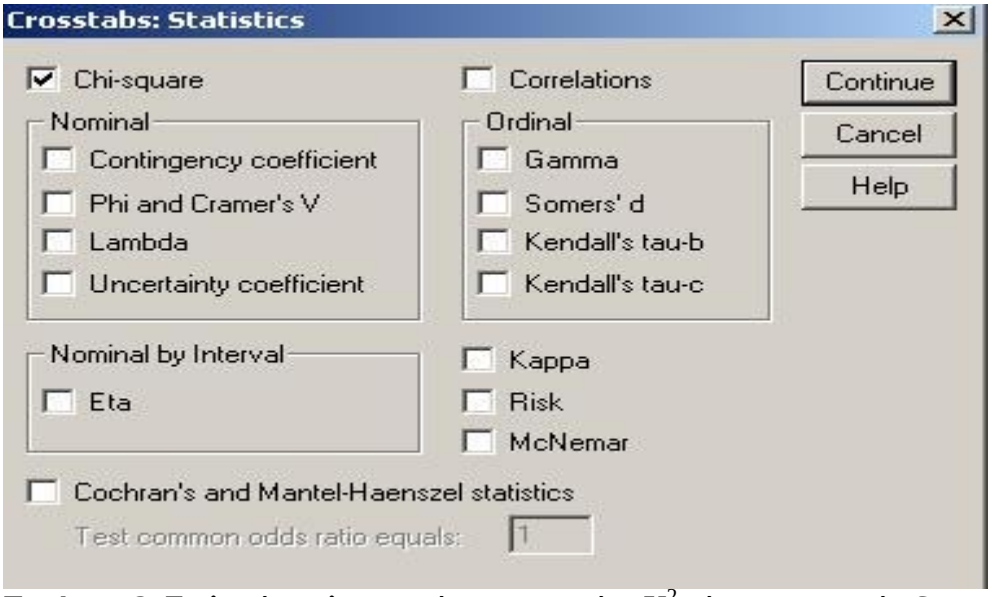
v. Πατάμε «***Continue***».

vi. Έπειτα πατάμε το κουμπί επιλογών «***Cells***» και εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 5.3*. Από το κουμπί επιλογών ***Cells*** μπορούμε να ζητήσουμε να υπολογιστούν ποσοστά επί των 2 μεταβλητών, καθώς επίσης να εμφανιστούν και οι αναμενόμενες συχνότητες που είναι απαραίτητες για τον υπολογισμό του κριτηρίου X^2.

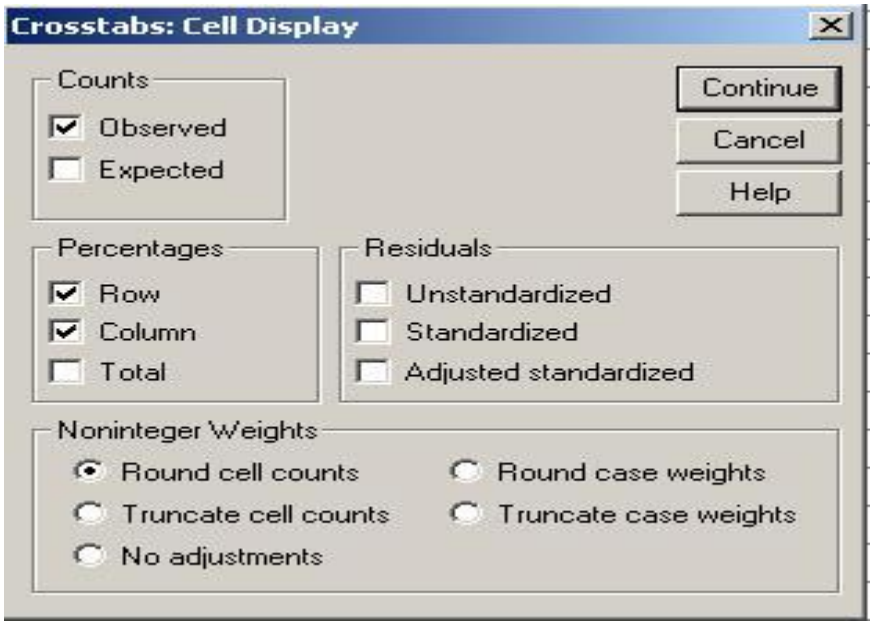
vii. Στους *Πίνακες 5.1 & 5.2* εμφανίζονται τα αποτελέσματα από τον έλεγχο για την ύπαρξη συσχέτισης μεταξύ του φύλου (sex: male/female) και του τύπου διάγνωσης ΟΣΣ (ecg: STEMI/NSTEMI/other). Από τον *Πίνακα 5.2* παρατηρούμε ότι το κριτήριο X^2 βρέθηκε ίσο με 23,230 (2 β.ε.) και η αντίστοιχη τιμή της πιθανότητας p (Asymp. Sig. (2-sided)) = 0,00….0<0,001. Αν δεχτούμε ως επίπεδο στατιστικής σημαντικότητας το $\alpha = 0,05$, τότε επειδή $p < \alpha$, απορρίπτουμε την H_o, γεγονός που σημαίνει ότι υπάρχει συσχέτιση μεταξύ φύλου και τον τύπο διάγνωσης ΟΣΣ στον πληθυσμό αναφοράς της μελέτης. Επίσης, από τον *Πίνακα 5.1* παρατηρούμε ότι το ποσοστό των ατόμων που διαγνώσθηκαν με STEMI είναι υψηλότερο μεταξύ των ανδρών (40,3%) σε σχέση με αυτό που παρατηρείται μεταξύ των γυναικών (28,6%).



Εικόνα 5.1: Πραγματοποίηση της ανάλυσης του X^2 στο SPSS



Εικόνα 5.2. Επιλογή υπολογισμού του κριτηρίου X^2 μέσω του μενού «Statistics».



Εικόνα 5.3: Πλαίσιο επιλογών του κουμπιού «Cells»

ecg * sex Crosstabulation

			sex		
			female	male	Total
ecg	stemi	Count	133	595	728
		% within ecg	18,3%	81,7%	100,0%
		% within sex	28,6%	40,3%	37,5%
	non_stemi	Count	132	392	524
		% within ecg	25,2%	74,8%	100,0%
		% within sex	28,4%	26,6%	27,0%
	other	Count	200	488	688
		% within ecg	29,1%	70,9%	100,0%
		% within sex	43,0%	33,1%	35,5%
Total		Count	465	1475	1940
		% within ecg	24,0%	76,0%	100,0%
		% within sex	100,0%	100,0%	100,0%

Πίνακας 5.1: Αποτελέσματα της ανάλυσης X^2 για τον έλεγχο συσχέτισης ανάμεσα στο φύλο (sex) και τον τύπο διάγνωσης βάση του ΗΚΓ (ecg). Πίνακας συνάφειας

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	23,230[a]	2	,000
Likelihood Ratio	23,616	2	,000
Linear-by-Linear Association	22,732	1	,000
N of Valid Cases	1940		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 125,60.

Πίνακας 5.2. Αποτελέσματα από τον έλεγχο συσχέτισης μεταξύ φύλου (sex) και του τύπου διάγνωσης βάση του ΗΚΓ (ecg). χρησιμοποιώντας το κριτήριο X^2

Αντίστοιχα, για να υπολογίσουμε το στατιστικό κριτήριο αλλά και την αντίστοιχη πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης Ho (δηλαδή την p) στο STATA ακολουθούμε τα εξής βήματα:

Statistics → Summaries, tables and tests → Tables → Two way tables with measures of association

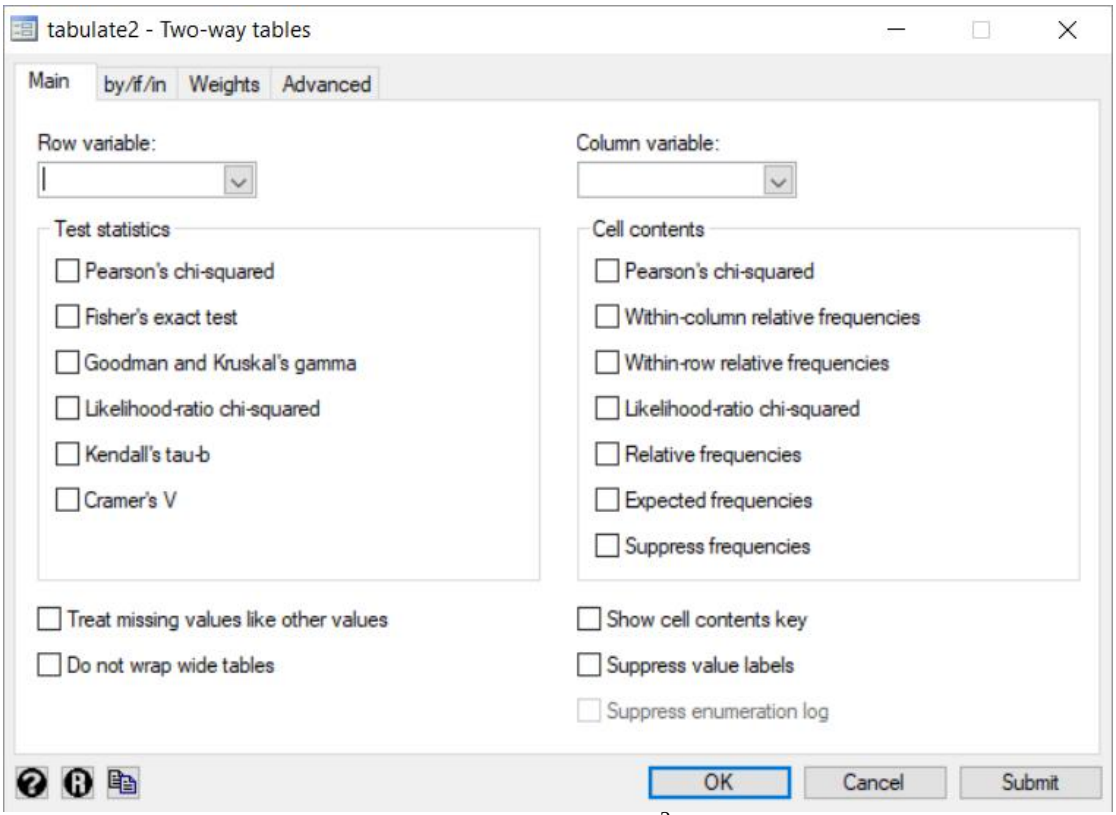
i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 5.4*

ii. Τοποθετούμε την μια κατηγορική μεταβλητή στο πλαίσιο "***Row Variable***" και την άλλη στο πλαίσιο "***Column Variable***".

iii. Από το μενού **Test *Statistics*** επιλέγουμε το **Pearson's *chi-squared*** για να υπολογιστεί το στατιστικό κριτήριο X^2. Από το ίδιο μενού μπορούμε επίσης να ζητήσουμε να υπολογιστούν και άλλα στατιστικά μέτρα συσχέτισης κατηγορικών μεταβλητών, όπως το φ, το V του Cramer, κλπ., καθώς επίσης και στατιστικά κριτήρια για τον έλεγχο της πιθανής συσχέτισης μεταξύ διατάξιμων μεταβλητών, όπως t-Kendal κλπ.

iv. Από το μενού ***Cell contects*** μπορούμε να ζητήσουμε να υπολογιστούν ποσοστά επί των 2 μεταβλητών, καθώς επίσης να εμφανιστούν και οι αναμενόμενες συχνότητες που είναι απαραίτητες για τον υπολογισμό του κριτηρίου X^2 (Expected frequencies).

v. «**OK**»



Εικόνα 5.4: Πραγματοποίηση της ανάλυσης του X^2 στο STATA

Στο STATA χρησιμοποιείται η εντολή ***tabulate*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 5.5*:

```
    [R] tabulate twoway ── Two-way tables of frequencies


Syntax

    Two-way tables

        tabulate varname1 varname2 [if] [in] [weight] [, options]


    Two-way tables for all possible combinations ─ a convenience tool

        tab2 varlist [if] [in] [weight] [, options]


    Immediate form of two-way tabulations

        tabi #11 #12 [...] \ #21 #22 [...] [\ ...] [, options]


    options               Description
    ──────────────────────────────────────────────────────────────────────────
    Main
      chi2                report Pearson's chi-squared
      exact[(#)]          report Fisher's exact test
      gamma               report Goodman and Kruskal's gamma
      lrchi2              report likelihood-ratio chi-squared
      taub                report Kendall's tau-b
      V                   report Cramır's V
      cchi2               report Pearson's chi-squared in each cell
      column              report relative frequency within its column of each cell
      row                 report relative frequency within its row of each cell
      clrchi2             report likelihood-ratio chi-squared in each cell
      cell                report the relative frequency of each cell
      expected            report expected frequency in each cell
      nofreq              do not display frequencies
      missing             treat missing values like other values
      wrap                do not wrap wide tables
```

Εικόνα 5.5: Βασική σύνταξη της εντολής tabulate

6. Έλεγχος ύπαρξης συσχέτισης ανάμεσα σε μία ποσοτική και μία κατηγορική μεταβλητή.

6.1 Εισαγωγή

Τα κριτήριο που χρησιμοποιούνται για να πραγματοποιηθεί ο έλεγχος ύπαρξης συσχέτισης ανάμεσα σε μία ποσοτική και μία ποιοτική μεταβλητή, ποικίλουν ανάλογα με τον αριθμό των κατηγοριών της ποιοτικής μεταβλητής (2 ή περισσότερες από 2) και τις ιδιότητες της ποσοτικής μεταβλητής (κανονική κατανομή ή όχι και ίσες διασπορές ή όχι σε όλες τις κατηγορίες της ποιοτικής μεταβλητής). Έτσι τα πιθανά κριτήρια που μπορούν να εφαρμοστούν είναι 4:

i. ***Student's t-test***
ii. ***One way ANOVA***
iii. ***Mann – Whitney test***
iv. ***Kruskal-Wallis test***

Τα κριτήρια (i) & (ii) είναι τα **παραμετρικά κριτήρια**, ενώ τα κριτήρια (iii) & (iv) είναι τα **μη παραμετρικά**.

Το Student's t-test & Mann-Whitney είναι τα κριτήρια που εφαρμόζονται όταν η ποιοτική μεταβλητή έχει 2 κατηγορίες. Η επιλογή του κατάλληλου κριτηρίου μεταξύ των 2 απαιτεί τον έλεγχο της παρακάτω προϋπόθεσης:

- **Κανονικότητα:** Η ποσοτική μεταβλητή ακολουθεί την κανονική κατανομή και στις 2 κατηγορίες της ποιοτικής μεταβλητής ή όχι;

Στην περίπτωση που ακολουθεί την κανονική κατανομή, εφαρμόζεται το Student's t-test κριτήριο. Διαφορετικά, εφαρμόζεται το αντίστοιχο του Student's t-test κριτήριο, που είναι το Mann-Whitney.

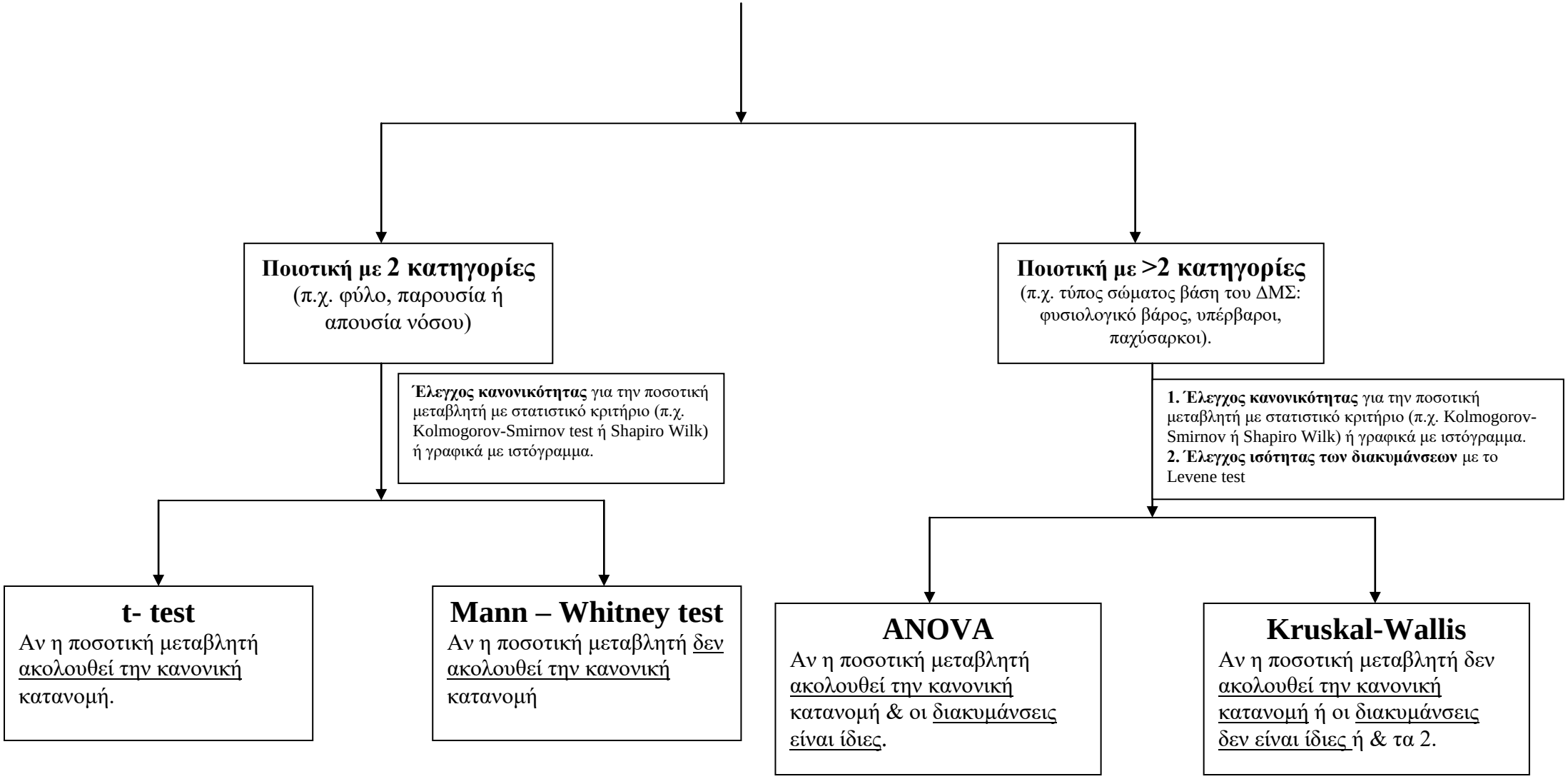
Παρομοίως, τα κριτήρια one way ANOVA & Kruskal-Wallis εφαρμόζονται όταν η ποιοτική μεταβλητή έχει περισσότερες από 2 κατηγορίες. Σε αυτήν την περίπτωση, η επιλογή του κατάλληλου κριτηρίου απαιτεί τον έλεγχο 2 προϋποθέσεων:

- **Κανονικότητα:** Η ποσοτική μεταβλητή ακολουθεί την κανονική κατανομή και στις 2 κατηγορίες της ποιοτικής μεταβλητής ή όχι;
- **Ομοσκεδαστικότητα:** Οι διακυμάνσεις της ποσοτικής μεταβλητής είναι ίσες μεταξύ των κατηγοριών της ποιοτικής ή όχι;

Στην περίπτωση, που η κατανομή της ποσοτικής μεταβλητής είναι η κανονική, και οι διασπορές της είναι ίσες μεταξύ των κατηγοριών της ποιοτικής, εφαρμόζουμε την ανάλυση διακύμανσης κατά ένα παράγοντα (ANOVA). Διαφορετικά, αν έστω και μία από τις 2 προϋποθέσεις δεν ισχύουν, εφαρμόσουμε το κριτήριο Kruskal-Wallis.

Στο σχεδιάγραμμα *6.1* παρουσιάζονται σχηματικά τα παραπάνω:

Σχεδιάγραμμα 6.1: Έλεγχος ύπαρξης συσχέτισης ανάμεσα σε μία ποσοτική και μία ποιοτική μεταβλητή



6.2. Έλεγχος ύπαρξης συσχέτισης ανάμεσα σε μία ποσοτική μεταβλητή και μία ποιοτική με 2 κατηγορίες.

6.2.1 Student's t-test με το SPSS

Ας υποθέσουμε ότι επιθυμούμε να ελέγξουμε αν υπάρχει συσχέτιση ανάμεσα στον δείκτη μάζας σώματος και το φύλο σε ενήλικες Έλληνες και ας υποθέσουμε πώς έχει πραγματοποιηθεί ο έλεγχος κανονικότητας (βλ. *κεφάλαιο 8*) του ΔΜΣ και έχει βρεθεί ότι η κατανομή του είναι κανονική τόσο στους άνδρες όσο και στις γυναίκες. Σύμφωνα με όσα αναφέρονται στην *παράγραφο 6.1*, το κατάλληλο στατιστικό κριτήριο για την πραγματοποίηση αυτού του ελέγχου είναι το Student's t-test.

Με αυτό το κριτήριο, ο έλεγχος που πραγματοποιείται είναι αν η μέση τιμή του ΔΜΣ είναι ίση μεταξύ των ανδρών και των γυναικών. Πιο συγκεκριμένα, η μηδενική και εναλλακτική υπόθεση είναι:

H_0: Η μέση τιμή του ΔΜΣ των ανδρών είναι ίση με αυτή των γυναικών και συνεπώς ΔΕΝ ΥΠΑΡΧΕΙ συσχέτιση ανάμεσα στον ΔΜΣ και το φύλο.

H_1: Η μέση τιμή του ΔΜΣ των ανδρών διαφέρει από αυτή των γυναικών και συνεπώς ΥΠΑΡΧΕΙ συσχέτιση ανάμεσα στον ΔΜΣ και το φύλο.

Για να καταλήξουμε στο συμπέρασμα ότι υπάρχει συσχέτιση ανάμεσα στον ΔΜΣ και το φύλο, θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value) να είναι $< \alpha = 0,05$ ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).

Για να υπολογίσουμε τόσο την τιμή του στατιστικού κριτηρίου όσο και την αντίστοιχη τιμή p-value με το SPSS ακολουθούμε τα ***βήματα***:

Analyze ⟶ Compare Means → Independent-samples T test

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 6.1*

ii. Τοποθετούμε την ή τις ποσοτικές μεταβλητές στο ***Test Variable(s)***,

iii. Την κατηγορική μεταβλητή που διαιρεί το δείγμα σε 2 ομάδες στο ***Grouping Variable***.

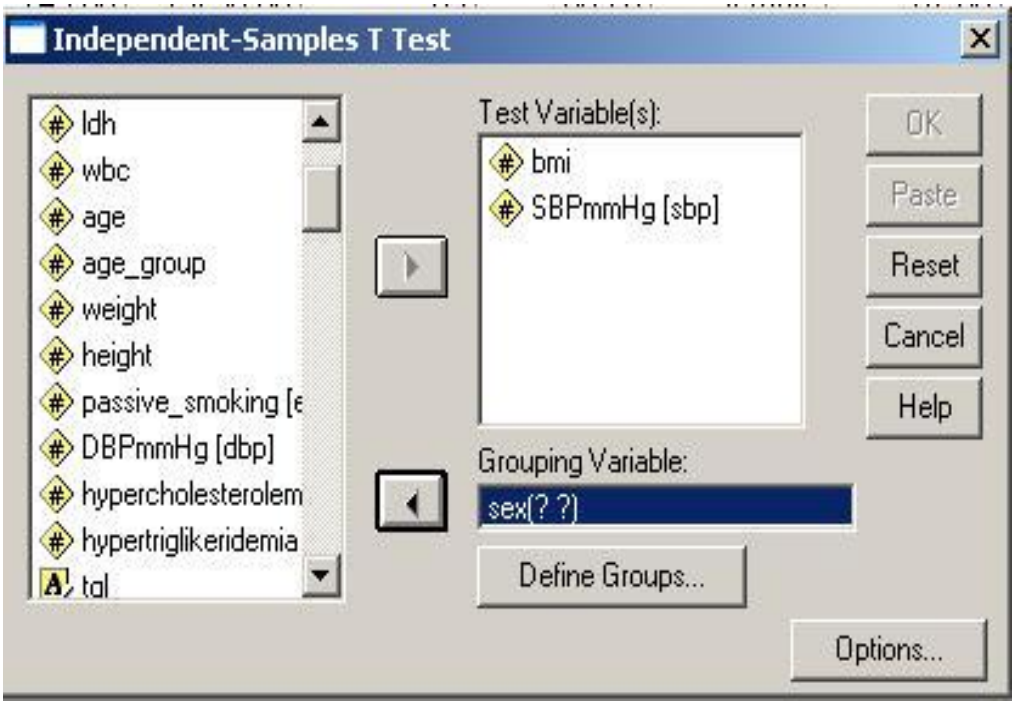
iv. Επίσης, είμαστε υποχρεωμένοι να δηλώσουμε με ποιες αριθμητικές τιμές έχουμε ορίσει τις 2 ομάδες της κατηγορικής μεταβλητής πιέζοντας το κουμπί "***Define groups…***", οπότε ανοίγει το πλαίσιο διαλόγου της *Εικόνας 6.2*.

v. Πληκτρολογούμε π.χ. Group 1: 1 = άνδρες και Group 2:0 = γυναίκες

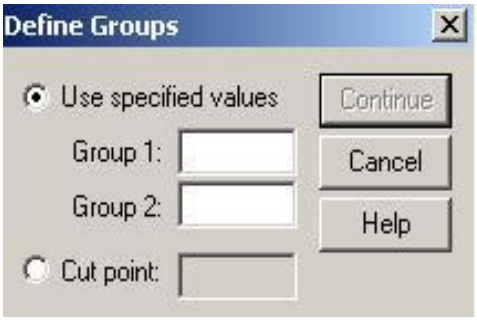
vi. Στη συνέχεια πατάμε το κουμπί "***Continue***".

vii. «***Ok***»

viii. Στους *Πίνακες 6.1 & 6.2* παρουσιάζονται τα αποτελέσματα από την πραγματοποίηση του ελέγχου t-test ανάμεσα στον δείκτη μάζας σώματος (ΔΜΣ) και το φύλο και ανάμεσα στην συστολική αρτηριακή πίεση (ΣΑΠ) και το φύλο.



Εικόνα 6.1: Εφαρμογή Student's t-test στο SPSS



Εικόνα 6.2: Προσδιορισμός των αριθμών με τους οποίους έχουν κωδικοποιηθεί οι 2 ομάδες της ποιοτικής μεταβλητής.

Η τιμή p-value για τον έλεγχο σύγκρισης των μέσων τιμών

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
bmi	Equal variances assumed	20,101	,000	-1,526	2019	,127	-,30876	,20238	-,70565	,08813
	Equal variances not assumed			-1,373	690,985	,170	-,30876	,22493	-,75040	,13287
SBPmmHg	Equal variances assumed	,980	,322	-3,587	1718	,000	-5,288	1,474	-8,179	-2,397
	Equal variances not assumed			-3,500	653,849	,000	-5,288	1,511	-8,254	-2,321

Η τιμή του p-value για τον έλεγχο ισότητας

Πίνακας 6.1. Αποτελέσματα από τον έλεγχο t-test του Student για τη διαφορά των αριθμητικών μέσων της μεταβλητής BMI (δείκτης μάζας σώματος) και της μεταβλητής SBP (συστολική αρτηριακή πίεση) μεταξύ ανδρών και γυναικών.

Group Statistics

	sex	N	Mean	Std. Deviation	Std. Error Mean
bmi	male	1540	27,4628	3,66187	,09331
	female	481	27,7715	4,48867	,20467
SBPmmHg	male	1312	136,29	25,707	,710
	female	408	141,57	26,940	1,334

Πίνακας 6.2: Περιγραφικά στοιχεία για την μεταβλητή BMI (δείκτης μάζας σώματος) και την μεταβλητή SBP (συστολική αρτηριακή πίεση) για άνδρες και γυναίκες, ξεχωριστά: Αποτελέσματα από την πραγματοποίηση t-test.

Πιο συγκεκριμένα από τον *Πίνακα 6.1* παρατηρούμε:

- ***Sig. του Levene test*** για την ισότητα των διασπορών του ΔΜΣ και της ΣΑΠ μεταξύ ανδρών και γυναικών ($2^{η}$ στήλη). Αν sig.<0,05 , τότε απορρίπτεται ότι ισχύει η ισότητα των διασπορών.
- ***Sig(2-tailed). του t-test*** ($5^{η}$ στήλη) όπου αν sig.<0,05 τότε απορρίπτουμε ότι ισχύει η ισότητα των μέσων τιμών της ποσοτική μεταβλητής μεταξύ ανδρών και γυναικών. Όπως, φαίνεται από τον *Πίνακα 6.1*, ο έλεγχος Student's t-test υπολογίζει 2 διαφορετικές πιθανότητες εσφαλμένης απόρριψης της μηδενικής υπόθεσης (p-value), λαμβάνοντας υπόψη αν ισχύει η ισότητα των διασπορών (πρώτη γραμμή) ή όχι (δεύτερη γραμμή). Συνεπώς, πιο από τα 2 p-value, θα χρησιμοποιήσουμε για να εξάγουμε το συμπέρασμά μας εξαρτάται από το αποτέλεσμα του Levene test.
- Συγκεκριμένα, αν το sig. από το Levene test < α=0,05, τότε εμπιστευόμαστε το $2^{ο}$ p-value της $5^{ης}$ στήλης. Αντίθετα, αν το sig. από το Levene test είναι > α=0,05, τότε εμπιστευόμαστε το $1^{ο}$ p-value της $5^{ης}$ στήλης.
- Συνεπώς, αναφορικά με τον ΔΜΣ του παραδείγματος του *Πίνακα 6.1*, διαπιστώνουμε ότι δεν ισχύει η ισότητα των διακυμάνσεων (sig. από το Levene test = 0,00…1<α=0,05), άρα η σωστή πιθανότητα εσφαλμένης απόρριψης της υπόθεσης ότι δεν υπάρχει συσχέτιση ανάμεσα στον ΔΜΣ και το φύλο είναι το δεύτερο p-value (sig(2-tailed)=0,170) > α=0,05. Άρα, δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση και άρα το συμπέρασμά μας είναι ότι η μέση τιμή του ΔΜΣ δεν διαφέρει ανάμεσα στους άνδρες και τις γυναίκες και συνεπώς δεν υπάρχει στατιστικά σημαντική συσχέτιση ανάμεσα στο φύλο και τον ΔΜΣ. Αναφορικά με την ΣΑΠ, διαπιστώνουμε ότι ισχύει η ισότητα των διασπορών (sig.=0,322), άρα το κατάλληλο p-value για να εξάγουμε το συμπέρασμά μας αναφορικά με το αρχικό μας ερώτημα (υπάρχει συσχέτιση ανάμεσα στην ΣΑΠ και το φύλο;) είναι το $1^{ο}$. Από το sig.(2-tailed) = 0,000…1 <0,05 διαπιστώνουμε ότι η μέση τιμή της ΣΑΠ διαφέρει μεταξύ ανδρών και γυναικών.

Στον *Πίνακα 6.2* παρατηρούμε την μέση τιμή (mean) και την τυπική απόκλιση (st.deviation) του ΔΜΣ και της ΣΑΠ σε άνδρες και γυναίκες, ξεχωριστά και μπορούμε να βγάλουμε το συμπέρασμα ότι η ΣΑΠ είναι σημαντικά υψηλότερη στις γυναίκες (mean=141,57) σε σχέση με τους άνδρες (mean=136,29).

Σημείωση: Από το ίδιο μενού επιλογών (**Analyze ⟶ Compare Means → One-Sample T test**) μπορεί να εκτελεστεί και το στατιστικό κριτήριο t-test για τη διαφορά του αριθμητικού μέσου ενός δείγματος με μια δεδομένη τιμή, τοποθετώντας στο «***Test Variable(s)***» την ή τις μεταβλητές και στο «***Test Values***» την τιμή με την οποία επιθυμούμε να συγκριθεί η μέση τιμή των παραπάνω μεταβλητών του δείγματός μας.

6.2.2 Student's t-test με το STATA

Για να υπολογίσουμε τόσο την τιμή του στατιστικού κριτηρίου όσο και την αντίστοιχη τιμή p-value με το STATA ακολουθούμε τα εξής ***βήματα***:

Statistics → Summaries, tables and tests → Classical tests of hypotheses → Two -group mean - comparison test

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 6.3*

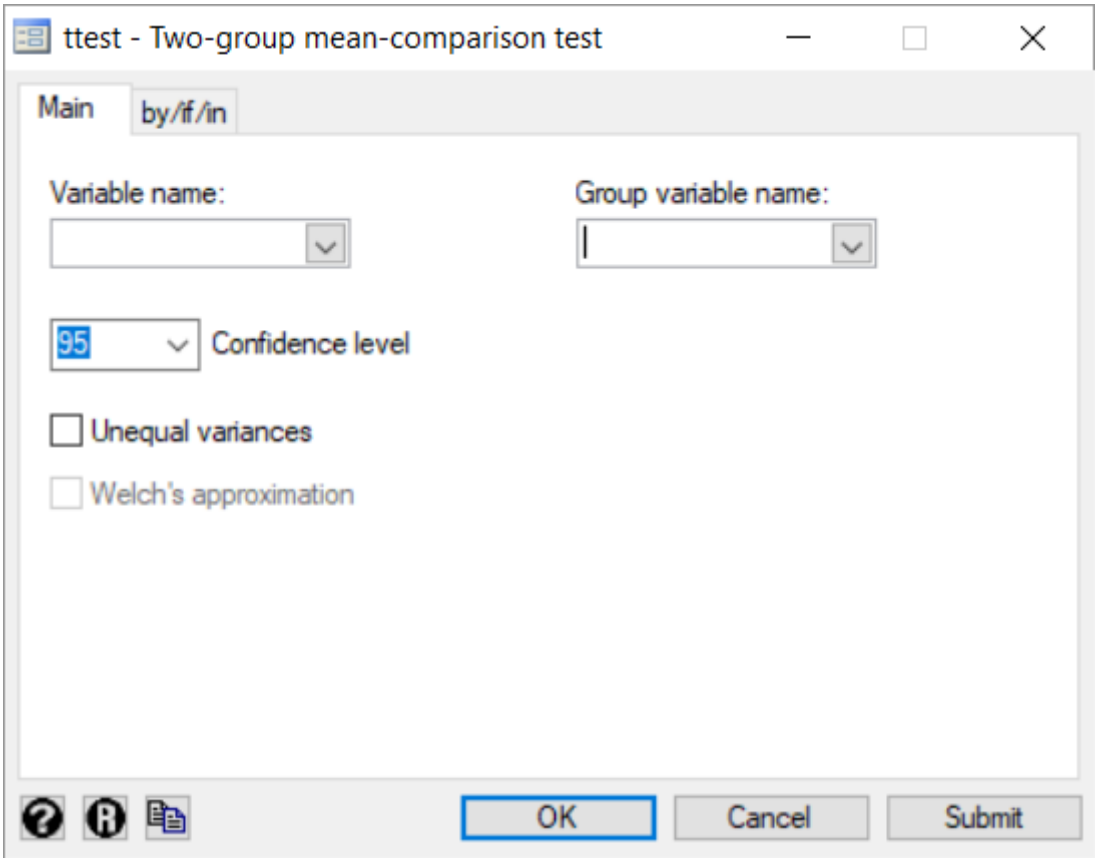
ii. Τοποθετούμε την ποσοτική μεταβλητή στο ***Variable name***,

iii. Την κατηγορική μεταβλητή που διαιρεί το δείγμα σε 2 ομάδες στο ***Group variable name***.

iv. Στο πλαίσιο **Confidence level** έχουμε την δυνατότητα να ζητήσουμε διάστημα εμπιστοσύνης, με διαφορετικό επίπεδο στατιστικής σημαντικότητας από 5% που είναι συνήθως.

v. Τέλος, έχουμε την δυνατότητα να δηλώσουμε εάν θεωρούμε ίσες ή άνισες τις δύο πληθυσμιακές διακυμάνσεις. Στην περίπτωση που επιλέξουμε ίσες διακυμάνσεις (***Equal variances*** – εξ ορισμού επιλογή) θα πραγματοποιηθεί ο έλεγχος του Student, ενώ στην περίπτωση που επιλέξουμε άνισες διακυμάνσεις (***Unequal variances***) θα πραγματοποιηθεί ο έλεγχος του Welch (βλ. *Εικόνα 6.3*).

vi. «***OK***»



Εικόνα 6.3: Εφαρμογή t-test στο STATA

Ο έλεγχος της ισότητας των δύο πληθυσμιακών διακυμάνσεων, προκειμένου να αποφασιστεί το είδος του ελεγχου που θα πραγματοποιηθεί γίνεται ακολουθώντας τα εξής ***βήματα***:

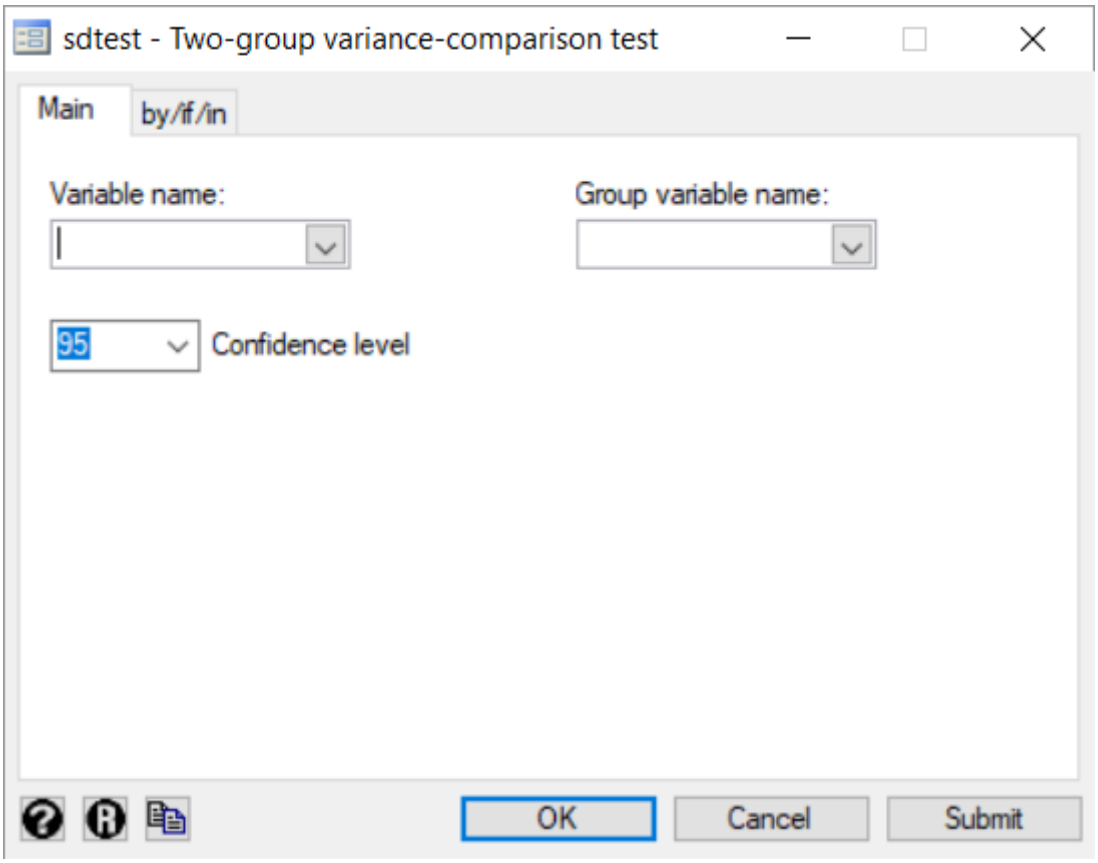
Statistics → Summaries, tables and tests → Classical tests of hypotheses → Two –group variance - comparison test

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 6.4*

ii. Τοποθετούμε την ποσοτική μεταβλητή στο ***Variable name***, και

iii. Την κατηγορική μεταβλητή που διαιρεί το δείγμα σε 2 ομάδες στο ***Group variable name***.

iv. «***OK***»



Εικόνα 6.4: Έλεγχος ισότητας πληθυσμιακών διακυμάνσεων στο STATA

Στο STATA χρησιμοποιείται η εντολή ***ttest*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 6.5*:

```
    [R] ttest ── Mean-comparison tests


Syntax

    One-sample mean-comparison test

        ttest varname == # [if] [in] [, level(#)]


    Two-sample mean-comparison test (unpaired)

        ttest varname1 == varname2 [if] [in], unpaired [unequal welch level(#)]


    Two-sample mean-comparison test (paired)

        ttest varname1 == varname2 [if] [in] [, level(#)]


    Two-group mean-comparison test

        ttest varname [if] [in] , by(groupvar) [options1]


    Immediate form of one-sample mean-comparison test

        ttesti #obs #mean #sd #val [, level(#)]


    Immediate form of two-sample mean-comparison test

        ttesti #obs1 #mean1 #sd1 #obs2 #mean2 #sd2 [, options2]

  options1          Description
  ---------------------------------------------------------------------
  Main
  * by(groupvar)    variable defining the groups
    unequal         unpaired data have unequal variances
    welch           use Welch's approximation
    level(#)        set confidence level; default is level(95)
  ---------------------------------------------------------------------
  * by(groupvar) is required.

  options2          Description
  ---------------------------------------------------------------------
  Main
    unequal         unpaired data have unequal variances
    welch           use Welch's approximation
    level(#)        set confidence level; default is level(95)
  ---------------------------------------------------------------------
```

Εικόνα 6.5: Βασική σύνταξη της εντολής ttest

6.2.3 Mann-Whitney test με το SPSS

Ας υποθέσουμε ότι στόχος μας είναι να ελέγξουμε αν τα επίπεδα του ενζύμου μυοκαρδιακής νέκρωσης CPKMBmax διαφέρει μεταξύ ανδρών και γυναικών με Οξύ Στεφανιαίο Σύνδρομο (ΟΣΣ). Ας υποθέσουμε, επίσης, ότι πραγματοποιήσαμε τον έλεγχο κανονικότητας (βλ. *κεφάλαιο 8*) και διαπιστώσαμε ότι το ένζυμο CPKMBmax δεν ακολουθεί την κανονική κατανομή ούτε στους άνδρες ούτε στις γυναίκες. Σύμφωνα, με όσα αναφέρονται στην *παράγραφο 6.1*, ο κατάλληλος στατιστικός έλεγχος είναι το Mann-Whitney test.

Με αυτό το κριτήριο, ο έλεγχος που πραγματοποιείται είναι αν η κατανομή του ενζύμου CPKMBmax είναι ίδια στους άνδρες και τις γυναίκες. Πιο συγκεκριμένα, η μηδενική και εναλλακτική υπόθεση είναι:

H_0: Η κατανομή του ενζύμου CPKMBmax των ανδρών είναι ίση με αυτή των γυναικών και συνεπώς ΔΕΝ ΥΠΑΡΧΕΙ συσχέτιση ανάμεσα στο συγκεκριμένο ένζυμο και το φύλο.

H_1: Η κατανομή του ενζύμου CPKMBmax των ανδρών διαφέρει από αυτή των γυναικών και συνεπώς ΥΠΑΡΧΕΙ συσχέτιση ανάμεσα στο συγκεκριμένο ένζυμο και το φύλο.

Για να καταλήξουμε στο συμπέρασμα ότι υπάρχει συσχέτιση ανάμεσα στο ένζυμο CPKMBmax και το φύλο, θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value) να είναι $< \alpha = 0,05$ ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).

Ο έλεγχος Mann-Whitney test μπορεί να πραγματοποιηθεί ακολουθώντας τα εξής ***βήματα***:

Analyze ⟶ Nonparametric Tests ⟶ 2 Independent Samples

i. Ανοίγει το παράθυρο διαλόγου της *Εικόνας 6.6*

ii. Στο «***Test Variable List***» τοποθετούμε την συνεχή μεταβλητή (π.χ. CPKMBmax)

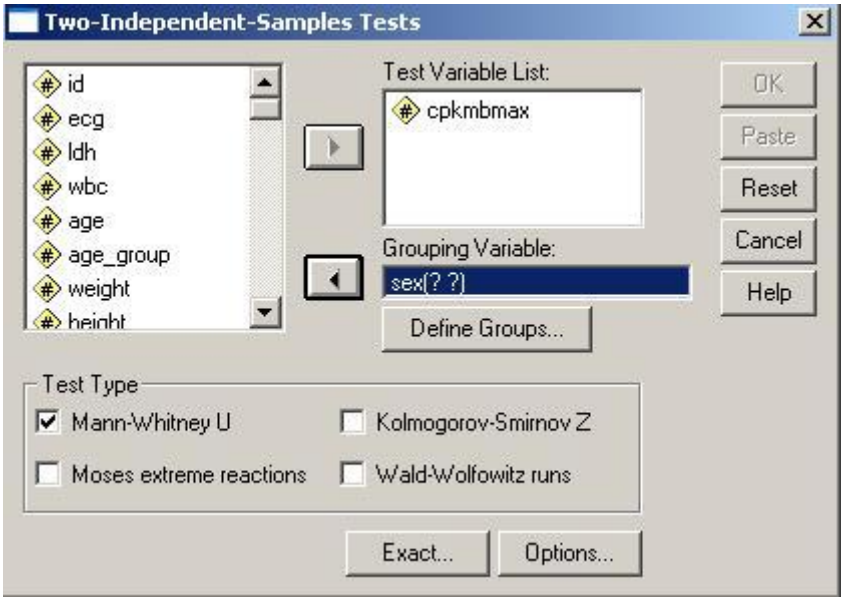
iii. Στο «***Grouping Variable***» τοποθετούμε την κατηγορική μεταβλητή (π.χ. sex)

iv. Πιέζουμε το κουμπί «***Define***» έτσι ώστε να ορίσουμε τους αριθμούς με τους οποίους έχουν κωδικοποιηθεί οι 2 κατηγορίες της κατηγορικής μεταβλητής (π.χ. 1: male & 0: female) (*Εικόνα 6.7*)

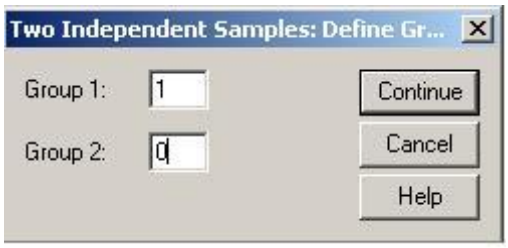
v. «***Continue***» &

vi. «***Ok***»

vii. Στον *Πίνακα 6.3* παρουσιάζονται τα αποτελέσματα. Από τον Πίνακα «***test Statistics***» του output διαπιστώνουμε ότι τα επίπεδα CPKMB διαφέρουν στατιστικά σημαντικά μεταξύ ανδρών και γυναικών με ΟΣΣ, αφού η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης (ότι δηλ. δεν υπάρχει συσχέτιση ανάμεσα στην CPKMB και το φύλο) είναι Asymp. Sig.<0,001<a=0,05.



Εικόνα 6.6: Πραγματοποίηση του μη παραμετρικού ελέγχου Mann-Whitney στο SPSS



Εικόνα 6.7: Προσδιορισμός των αριθμών με τις οποίες έχουν κωδικοποιηθεί οι 2 ομάδες της κατηγορικής/ ποιοτικής μεταβλητής (1: male, 0: female)

Test Statistics[a]

	cpkmbmax
Mann-Whitney U	286623,50
Wilcoxon W	387648,50
Z	-3,532
Asymp. Sig. (2-tailed)	,000

a. Grouping Variable: sex

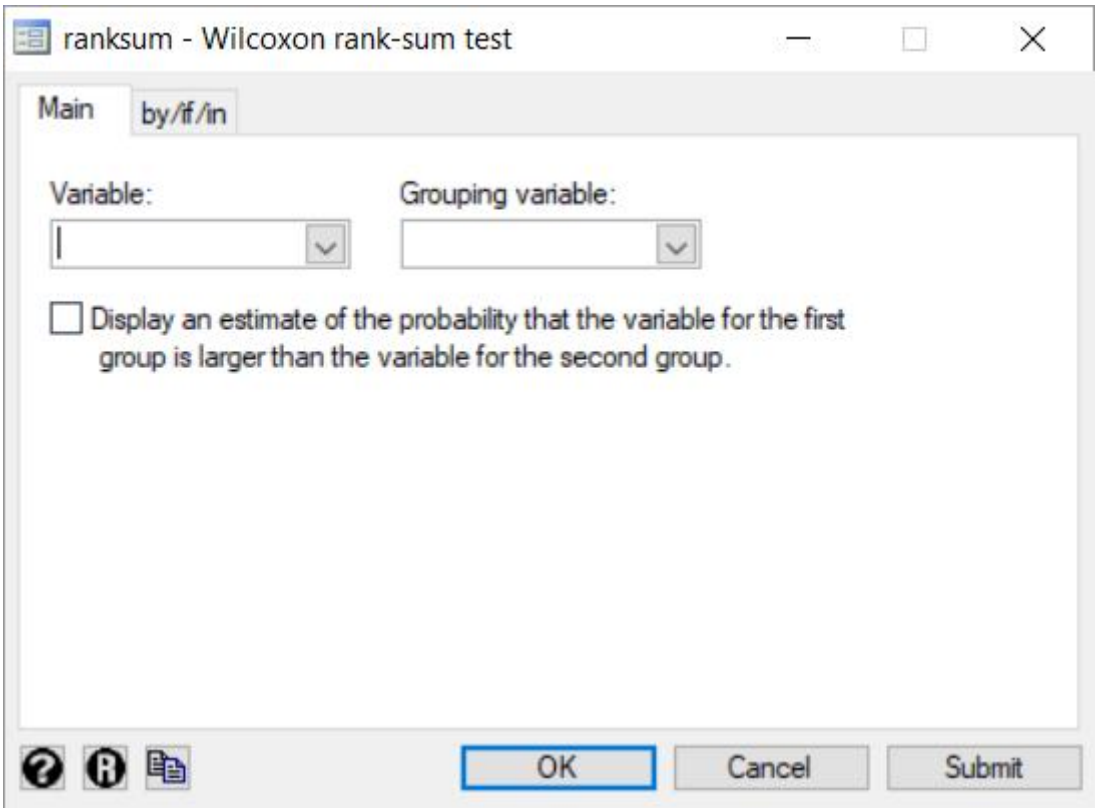
Πίνακας 6.3: Αποτελέσματα από τον έλεγχο Mann-Whitney για τον έλεγχο ύπαρξης συσχέτισης ανάμεσα στο ένζυμο μυοκαρδιακής νέκρωσης (CPKMB) και το φύλο.

6.2.4 Mann-Whitney test με το STATA

Ο έλεγχος Wilcoxon rank sum test (γνωστό και ως Mann Whitney test) στο STATA μπορεί να πραγματοποιηθεί ακολουθώντας τα εξής ***βήματα***:

Statistics → Summaries, tables and tests → Nonparametric analysis → Tests of hypotheses → Wilcoxon rank – sum test

i. Ανοίγει το παράθυρο διαλόγου της *Εικόνας 6.8*
ii. Στο «***Variable***» τοποθετούμε την συνεχή μεταβλητή
iii. Στο «***Grouping Variable***» τοποθετούμε την κατηγορική μεταβλητή
iv. «***OK***»



Εικόνα 6.8: Πραγματοποίηση του μη παραμετρικού ελέγχου Mann-Whitney στο STATA

Στο STATA χρησιμοποιείται η εντολή ***ranksum*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 6.9*:

Syntax

Wilcoxon rank-sum test

`ranksum` *varname* [*if*] [*in*] , `by(`*groupvar*`)` [`porder`]

Nonparametric equality-of-medians test

`median` *varname* [*if*] [*in*] [*weight*] , `by(`*groupvar*`)` [*median_options*]

ranksum_options	Description
Main	
* `by(`*groupvar*`)`	grouping variable
`porder`	probability that variable for first group is larger than variable for second group

median_options	Description
Main	
* `by(`*groupvar*`)`	grouping variable
`exact`	perform Fisher's exact test
`medianties(below)`	assign values equal to the median to below group
`medianties(above)`	assign values equal to the median to above group
`medianties(drop)`	drop values equal to the median from the analysis
`medianties(split)`	split values equal to the median equally between the two groups

*`by(`*groupvar*`)` is required.
`by` is allowed with `ranksum` and `median`; see [D] **by**.
`fweights` are allowed with `median`; see [U] **11.1.6 weight**.

Εικόνα 6.9: Βασική σύνταξη της εντολής ranksum

6.3 Έλεγχος ύπαρξης συσχέτισης ανάμεσα σε μία ποσοτική και μία κατηγορική μεταβλητή με περισσότερες από δύο κατηγορίες

6.3.1 Ανάλυση διακύμανσης κατά έναν παράγοντα (One way ANOVA)

Ας υποθέσουμε ότι επιθυμούμε να ελέγξουμε την ύπαρξη συσχέτισης ανάμεσα σε μια κατηγορική μεταβλητή με περισσότερες από 2 κατηγορίες (π.χ. τύπος διάγνωσης οξέος στεφανιαίου συνδρόμου: STEMI/NSTEMI/other) και μιας ποσοτικής μεταβλητής που ακολουθεί την κανονική κατανομή σε κάθε κλάση της κατηγορικής (π.χ. ηλικία). Ας υποθέσουμε, επίσης, ότι η ποσοτική μας μεταβλητή ακολουθεί την κανονική κατανομή σε όλες τις κατηγορίες της ποιοτικής μας μεταβλητής (βλ. *Κεφάλαιο 7*) και ότι οι διασπορές της ποσοτικής μας μεταβλητής είναι ίσες σε όλες τις κατηγορίες της ποιοτικής. Ο κατάλληλος στατιστικός έλεγχος είναι η **Ανάλυση Διακύμανσης Κατά Ένα Παράγοντα (One-Way ANOVA).** Αυτό που ελέγχουμε με τον συγκεκριμένο έλεγχο είναι ότι αν οι μέσες τιμές της ποσοτικής μας μεταβλητής είναι ίσες σε όλες τις ομάδες της ποιοτικής /κατηγορικής μεταβλητής. Πιο συγκεκριμένα, η μηδενική και εναλλακτική υπόθεση είναι:

H_0: Η μέση τιμή της ηλικίας των ατόμων με STEMI είναι ίση με αυτή των ατόμων με NSTEMI και ίση με αυτή των ατόμων με other διάγνωση και συνεπώς ΔΕΝ ΥΠΑΡΧΕΙ συσχέτιση ανάμεσα στην ηλικία και τον τύπο διάγνωσης οξέος στεφανιαίου συνδρόμου.

H_1: Οι μέσες τιμές της ηλικίας διαφέρουν τουλάχιστον ανάμεσα σε 2 από τους 3 τύπους διάγνωσης του οξέος στεφανιαίου συνδρόμου και συνεπώς ΥΠΑΡΧΕΙ συσχέτιση ανάμεσα στην ηλικία και τον τύπο διάγνωσης οξέος στεφανιαίου συνδρόμου.

Για να καταλήξουμε στο συμπέρασμα ότι υπάρχει συσχέτιση ανάμεσα στην ηλικία και τον τύπο διάγνωσης οξέος στεφανιαίου συνδρόμου, θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value) να είναι $< \alpha = 0,05$ ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).

Για να υπολογίσουμε την τιμή του στατιστικού κριτηρίου (F) που δίνεται από τον παραπάνω έλεγχο, αλλά και την αντίστοιχη πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (p) ακολουθούμε τα ***βήματα***:

Analyze ⟶ Compare Means → One-Way ANOVA

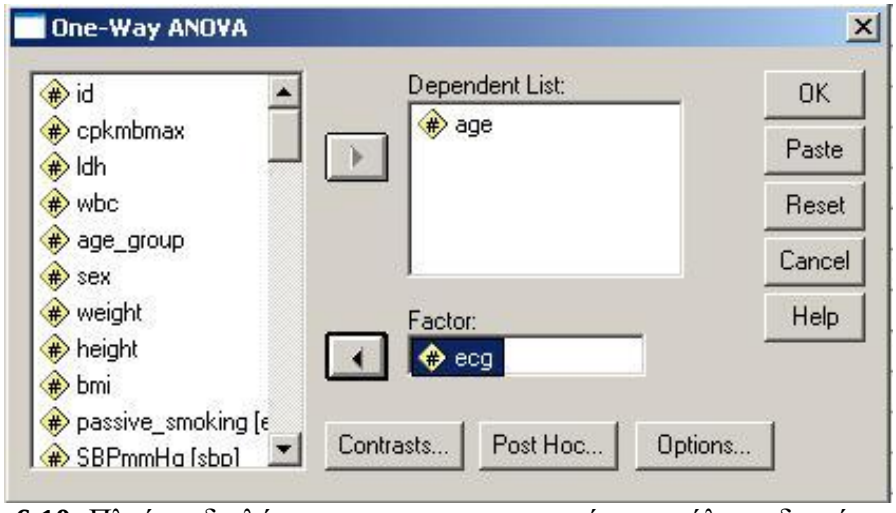
i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 6.10*.
ii. Στο ***Dependent List*** τοποθετούμε την ποσοτική μεταβλητή (μπορούμε να εισάγουμε και περισσότερες από μία ποσοτικές μεταβλητές)
iii. Στο ***Factor*** τοποθετούμε την κατηγορική μεταβλητή που διαχωρίζει το δείγμα σε περισσότερες από 2 ομάδες.
iv. Αν επιθυμούμε να εμφανιστούν στα αποτελέσματα του παραπάνω ελέγχου και τα περιγραφικά χαρακτηριστικά της ποσοτικής μεταβλητής σε κάθε κλάση της κατηγορικής μεταβλητής, πατάμε το κουμπί επιλογών ***Options*** που υπάρχει στο πλαίσιο διαλόγου της *Εικόνας 6.10*, και ανοίγει ένα νέο πλαίσιο διαλόγου (*Εικόνα 6.11*)
v. Επιλέγουμε το ***Descriptive*** και ***Continue***.

Σημείωση: Από το πλαίσιο διαλόγου της *Εικόνας 6.11* μπορούμε να επιλέξουμε να πραγματοποιηθεί και ο έλεγχος ισότητας των διακυμάνσεων που είναι μία από τις προϋποθέσεις που πρέπει να ισχύουν για την ορθή εφαρμογή της ανάλυσης διακύμανσης κατά έναν παράγοντα, επιλέγοντας την επιλογή «***Homogeneity of variance test***». Με αυτό τον τρόπο διαπιστώνουμε αν ορθώς εφαρμόσαμε τη συγκεκριμένη ανάλυση οπότε και χρησιμοποιούμε τα αποτελέσματα που προέκυψαν για την διεξαγωγή συμπερασμάτων, διαφορετικά επαναλαμβάνουμε την ανάλυση χρησιμοποιώντας τον αντίστοιχο με την ανάλυση διακύμανσης κατά έναν παράγοντα έλεγχο (Kruskal-Wallis).

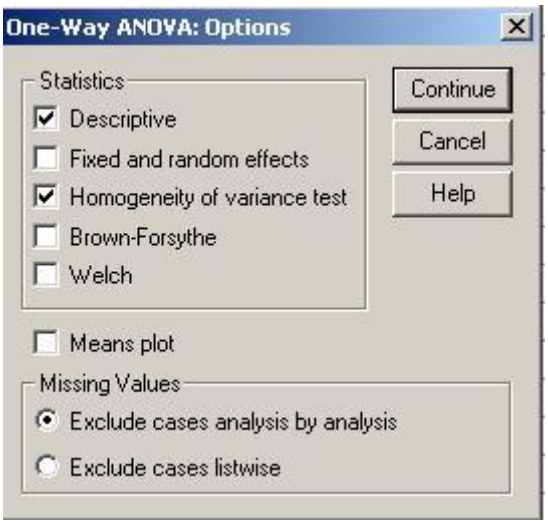
vi. Σε αυτή την ανάλυση, όμως, υπεισέρχεται το πρόβλημα των πολλαπλών συγκρίσεων, το οποίο ξεπερνάτε χρησιμοποιώντας κατάλληλη διόρθωση. Αυτό μπορεί να πραγματοποιηθεί, πατώντας το κουμπί επιλογών ***Post Hoc*** του αρχικού πλαισίου διαλόγου (*Εικόνα 6.10*), οπότε και ανοίγει ένα νέο πλαίσιο διαλόγου (*Εικόνα 6.11*).

vii. Μπορούμε να διαλέξουμε ανάμεσα σε μία σειρά μεθόδων κατάλληλων για την πραγματοποίηση διορθώσεων για πολλαπλούς ελέγχους. Η πιο συνήθης μέθοδος είναι η διόρθωση κατά ***Bonferroni***.

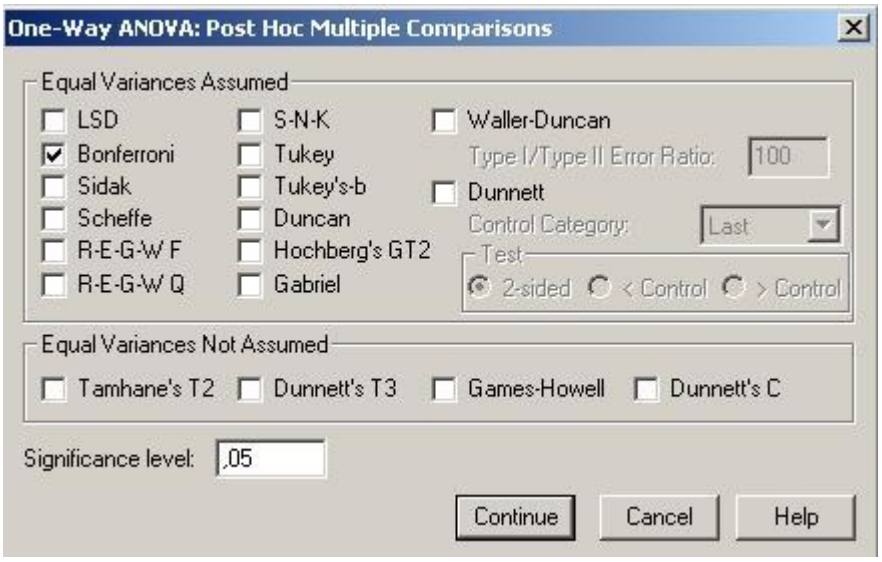
viii. Στους *Πίνακες 6.4, 6.5, 6.6 & 6.7* παρουσιάζονται τα αποτελέσματα της παραπάνω ανάλυσης για τον έλεγχο ύπαρξης συσχέτισης ανάμεσα στην ηλικία και τον τύπο διάγνωσης οξέος στεφανιαίου συνδρόμου.



Εικόνα 6.10: Πλαίσιο διαλόγου για την πραγματοποίηση ανάλυση διακύμανσης κατά έναν παράγοντα (one way ANOVA) στο SPSS



Εικόνα 6.11: Το πλαίσιο διαλόγου που εμφανίζεται πατώντας το κουμπί επιλογών «Options».



Εικόνα 6.12: Επιλογές για την πραγματοποίηση των πολλαπλών ελέγχων (κουμπί επιλογών «Post Hoc»).

ANOVA

age

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4299,278	2	2149,639	12,942	,000
Within Groups	321722,8	1937	166,093		
Total	326022,1	1939			

Πίνακας 6.4: Αποτελέσματα από τον έλεγχο One-Way ANOVA για τη διαφορά των αριθμητικών μέσων της μεταβλητής Age μεταξύ αυτών που πάσχουν από τους τρεις

διαφορετικούς τύπους του Οξέος Στεφανιαίου Συνδρόμου (ECG findings (STEMI/non-STEMI/other)).

Από τον *Πίνακα 6.4* παρατηρούμε ότι η τιμή του κριτηρίου F που υπολογίζεται από τον έλεγχο One-Way ANOVA είναι 12,942, και η αντίστοιχη πιθανότητα (p) εσφαλμένης απόρριψης της H_o είναι sig. = 0,000…1, σαφώς μικρότερη του επιπέδου σημαντικότητας α = 0,05. Συνεπώς, απορρίπτουμε την μηδενική υπόθεση περί ισότητας των αριθμητικών μέσων της ηλικίας στις 3 ομάδες. Επίσης, από τον *Πίνακα 6.5* όπου παρουσιάζονται τα περιγραφικά μέτρα της ηλικίας για κάθε τύπο διάγνωσης οξέος στεφανιαίου συνδρόμου, ξεχωριστά, διαπιστώνουμε πως αυτοί που έπασχαν από STEMI ήταν σχεδόν κατά 3 χρόνια νεότεροι συγκριτικά με αυτούς που ανήκαν σε κάποια από τις άλλες 2 ομάδες (Non-STEMI ή other).

Descriptives

age

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
stemi	728	64,16	13,413	,497	63,18	65,13	23	100
non_stemi	524	67,45	12,861	,562	66,35	68,56	27	96
other	688	67,03	12,328	,470	66,10	67,95	26	91
Total	1940	66,06	12,967	,294	65,49	66,64	23	100

Πίνακας 6.5: Περιγραφικά χαρακτηριστικά που προκύπτουν κατά την ανάλυση διακύμανσης κατά έναν παράγοντα.

Οι *Πίνακες 6.4 & 6.5* μας δίνουν τις πληροφορίες που χρειάζονται για να εξάγουμε το συμπέρασμα πως η μέση ηλικία των ατόμων διαφέρει μεταξύ των 3 τύπων διάγνωσης του οξέος στεφανιαίου συνδρόμου, όμως δεν μας διευκρινίζει μεταξύ ποιων τύπων υπάρχει η διαφορά (ποιοι είναι οι μέσοι όροι που διαφέρουν). Αυτή η πληροφορία παρουσιάζεται στον *Πίνακα 6.6*, από όπου προκύπτει ότι η ηλικία διαφέρει ανάμεσα σε αυτούς με STEMI & Non-STEMI και σε αυτούς με STEMI & other, ενώ δεν υπάρχει διαφορά στην ηλικία αυτών με Non-STEMI & αυτών με other διάγνωση.

Multiple Comparisons

Dependent Variable: age

Bonferroni

(I) ecg	(J) ecg	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
stemi	non_stemi	-3,297*	,738	,000	-5,07	-1,53
	other	-2,871*	,685	,000	-4,51	-1,23
non_stemi	stemi	3,297*	,738	,000	1,53	5,07
	other	,426	,747	1,000	-1,36	2,22
other	stemi	2,871*	,685	,000	1,23	4,51
	non_stemi	-,426	,747	1,000	-2,22	1,36

*. The mean difference is significant at the .05 level.

Πίνακας 6.6. Αποτελέσματα πολλαπλών ελέγχων χρησιμοποιώντας τη διόρθωση Bonferroni, για τη σύγκριση των μέσων τιμών της ηλικίας στους διάφορους τύπους Οξέος Στεφανιαίου Συνδρόμου.

Τέλος, από την *Πίνακα 6.7* διαπιστώνουμε ότι η προϋπόθεση της ομοσκεδαστικότητας απορρίπτεται, αφού από τον πίνακα ***Test of Homogeneity of variance*** το sig. = 0,018 < 0,05. Συνεπώς, λοιπόν, ακόμα και αν ίσχυε η προϋπόθεση της κανονικής κατανομής της ηλικίας σε κάθε μία κλάση της κατηγορικής μεταβλητής η ANOVA δεν είναι η σωστή ανάλυση στην συγκεκριμένη περίπτωση (βλ. *παράγραφο 6.3.2* για την ιδανική ανάλυση).

Test of Homogeneity of Variances

age

Levene Statistic	df1	df2	Sig.
4,012	2	1937	,018

Πίνακας 6.7: Αποτελέσματα για τον έλεγχο της ισότητας των διασπορών της ηλικίας στους 3 τύπους διάγνωσης του οξέος στεφανιαίου συνδρόμου.

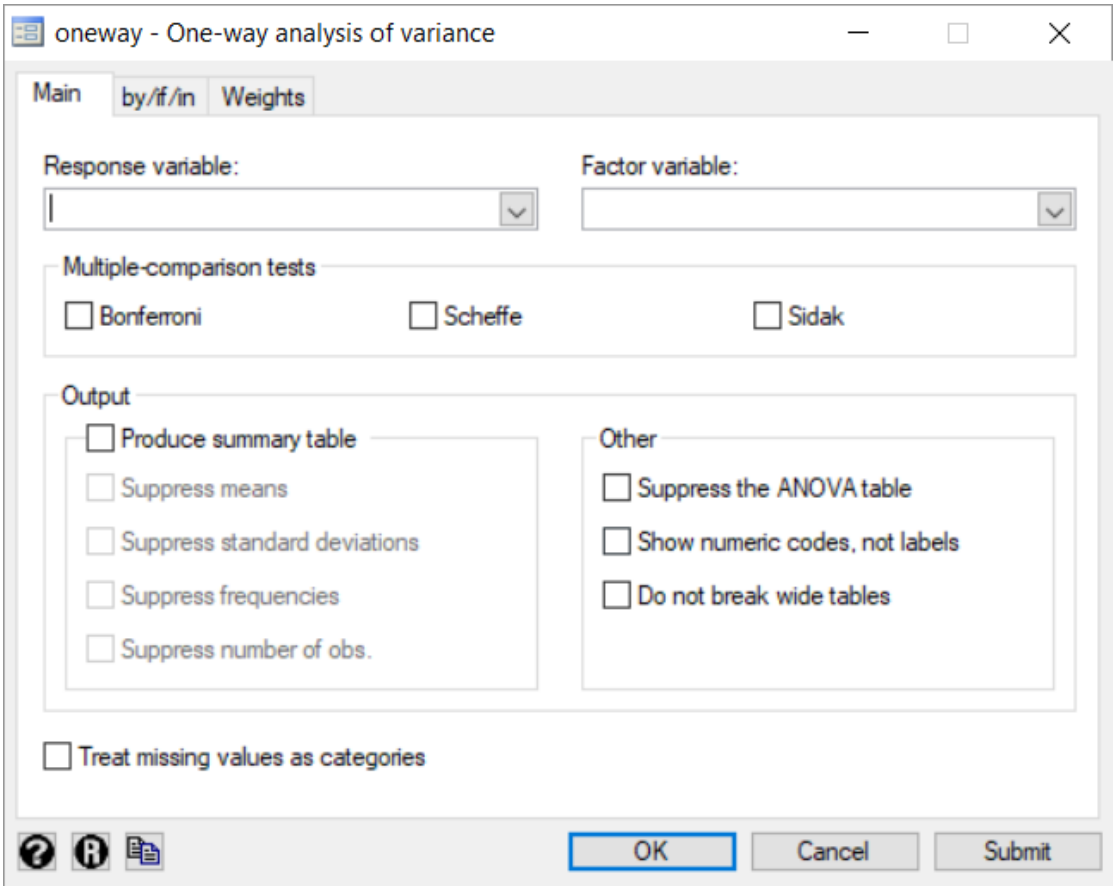
6.3.2 Ανάλυση διακύμανσης κατά έναν παράγοντα (One way ANOVA) στο STATA

Για να υπολογίσουμε την τιμή του στατιστικού κριτηρίου αλλά και την αντίστοιχη πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o ακολουθούμε τα ***βήματα***:

Statistics → Linear models and related → ANOVA/MANOVA → One-Way ANOVA

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 6.13*.
ii. Στο ***Response variable*** τοποθετούμε την ποσοτική μεταβλητή
iii. Στο ***Factor variable*** τοποθετούμε την κατηγορική μεταβλητή που διαχωρίζει το δείγμα σε περισσότερες από 2 ομάδες.
iv. Αν επιθυμούμε να εμφανιστούν στα αποτελέσματα του παραπάνω ελέγχου και τα περιγραφικά χαρακτηριστικά της ποσοτικής μεταβλητής σε κάθε κλάση της κατηγορικής μεταβλητής, επιλέγουμε ***Produce summary table*** που υπάρχει στο πλαίσιο Output.
v. Σε αυτή την ανάλυση, όμως, υπεισέρχεται το πρόβλημα των πολλαπλών συγκρίσεων, το οποίο όμως ξεπερνάτε χρησιμοποιώντας κατάλληλη διόρθωση. Αυτό μπορεί να πραγματοποιηθεί, επιλέγοντας κάποιον διαθέσιμο έλεγχο από το πλαίσιο ***Multiple – comparison test***.
vi. Μπορούμε να διαλέξουμε ανάμεσα στις 3 διαθέσιμες μεθόδους για την πραγματοποίηση διορθώσεων για πολλαπλούς ελέγχους. Η πιο συνήθης μέθοδος είναι η διόρθωση κατά ***Bonferroni***.

Σημείωση: Ταυτόχρονα με το One way Anova test πραγματοποιείται και ο έλεγχος ισότητας των διακυμάνσεων που είναι μία από τις προϋποθέσεις που πρέπει να ισχύουν για την ορθή εφαρμογή της ανάλυσης διακύμανσης κατά έναν παράγοντα. Με αυτό τον τρόπο διαπιστώνουμε αν ορθώς εφαρμόσαμε τη συγκεκριμένη ανάλυση οπότε και χρησιμοποιούμε τα αποτελέσματα που προέκυψαν για την διεξαγωγή συμπερασμάτων, διαφορετικά επαναλαμβάνουμε την ανάλυση χρησιμοποιώντας τον αντίστοιχο με την ανάλυση διακύμανσης κατά έναν παράγοντα έλεγχο (Kruskal-Wallis).



Εικόνα 6.13: Πλαίσιο διαλόγου για την πραγματοποίηση ανάλυση διακύμανσης κατά έναν παράγοντα στο STATA

Στο STATA χρησιμοποιείται η εντολή ***oneway*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 6.14*:

Syntax

`oneway` *response_var factor_var* [*if*] [*in*] [*weight*] [, *options*]

options	Description
Main	
`bonferroni`	Bonferroni multiple-comparison test
`scheffe`	Scheffé multiple-comparison test
`sidak`	Šidák multiple-comparison test
`tabulate`	produce summary table
[`no`]`means`	include or suppress means; default is `means`
[`no`]`standard`	include or suppress standard deviations; default is `standard`
[`no`]`freq`	include or suppress frequencies; default is `freq`
[`no`]`obs`	include or suppress number of obs; default is `obs` if data are weighted
`noanova`	suppress the ANOVA table
`nolabel`	show numeric codes, not labels
`wrap`	do not break wide tables
`missing`	treat missing values as categories

`by` is allowed; see [D] **by**.

`aweights` and `fweights` are allowed; see [U] **11.1.6 weight**.

Εικόνα 6.14: Βασική σύνταξη της εντολής oneway

6.3.3 Kruskal–Wallis test στο SPSS

Ας υποθέσουμε ότι στόχος μας είναι να ελέγξουμε αν τα επίπεδα του ενζύμου μυοκαρδιακής νέκρωσης CPKMB διαφέρει μεταξύ των ασθενών με Οξύ Στεφανιαίο Σύνδρομο που έχουν διαγνωσθεί με «STEMI», «non-STEMI» ή «Other» διάγνωση (π.χ μεταβλητή ecg). Ας υποθέσουμε, επίσης, ότι το ένζυμο CPKMB δεν ακολουθεί την κανονική κατανομή σε καμία από τις τρεις πιθανές διαγνώσεις (βλ. *Κεφάλαιο 8*) ή δεν ισχύει η ισότητα των διακυμάνσεων της ποσοτικής μεταβλητής σε όλες τις κατηγορίες της κατηγορικής μεταβλητής (ομοσκεδαστικότητα). Σε αυτή την περίπτωση, ο κατάλληλος στατιστικός έλεγχος είναι το Kruskal-Wallis test που είναι ο αντίστοιχος μη παραμετρικός έλεγχος του ελέγχου one-way ANOVA. Πιο συγκεκριμένα, η μηδενική και εναλλακτική υπόθεση είναι:

H_0: Η κατανομή του ενζύμου CPKMBmax είναι ίδια και στους 3 τύπος διάγνωσης του οξέος στεφανιαίου συνδρόμου και συνεπώς ΔΕΝ ΥΠΑΡΧΕΙ συσχέτιση ανάμεσα στο συγκεκριμένο ένζυμο και τον τύπο διάγνωσης.

H_1: Η κατανομή του ενζύμου CPKMBmax διαφέρει μεταξύ τουλάχιστον 2 εκ των 3 τύπων διάγνωσης οξέος στεφανιαίου συνδρόμου και συνεπώς ΥΠΑΡΧΕΙ συσχέτιση ανάμεσα στο συγκεκριμένο ένζυμο και τον τύπο διάγνωσης.

Για να καταλήξουμε στο συμπέρασμα ότι υπάρχει συσχέτιση ανάμεσα στο ένζυμο CPKMBmax και τον τύπο διάγνωσης, θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value) να είναι $< \alpha = 0,05$ ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).

Ο έλεγχος Kruskal-Wallis, μπορεί να πραγματοποιηθεί ακολουθώντας τα εξής ***βήματα***:

Analyze ⟶ Nonparametric Tests ⟶ K Independent Samples

i. Ανοίγει το παράθυρο διαλόγου της *Εικόνας 6.15*

ii. Στο «***Test Variable List***» τοποθετούμε την συνεχή μεταβλητή (π.χ. CPKMBmax)

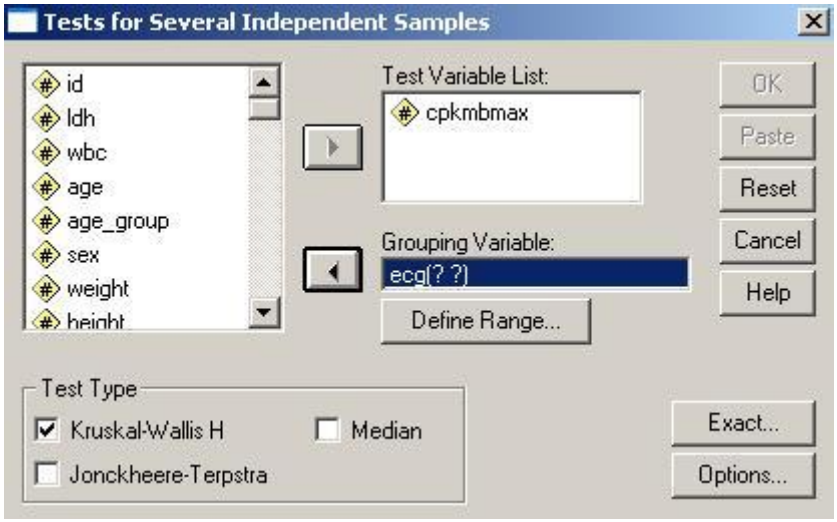
iii. Στο «***Grouping Variable***» τοποθετούμε την κατηγορική μεταβλητή (π.χ. ecg)

iv. Πιέζουμε το κουμπί «***Define***» έτσι ώστε να ορίσουμε τον μικρότερο και μεγαλύτερο αριθμό με τους οποίους έχουν κωδικοποιηθεί οι κατηγορίες της κατηγορικής μεταβλητής (π.χ. 1 για την διάγνωση «STEMI» & 3 για την διάγνωση «Other») και ανοίγει το πλαίσιο διαλόγου της *Εικόνας 6.16*.

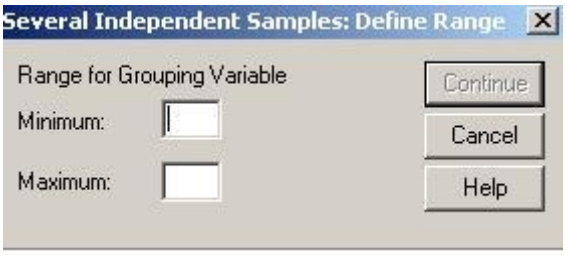
v. Πληκτρολογούμε για παράδειγμα «1» στο «Minimum» & 3 στο «Maximum».

vi. «***Continue***» & «***Ok***»

vii. Στον *Πίνακα 6.8* παρουσιάζονται τα αποτελέσματα. Από τον Πίνακα «***test Statistics***» του output διαπιστώνουμε ότι τα επίπεδα CPKMB διαφέρουν στατιστικά σημαντικά μεταξύ των τριών διαγνώσεων ΟΣΣ, αφού η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης (ότι δηλ. δεν υπάρχει συσχέτιση ανάμεσα στα επίπεδα της CPKMB και της διάγνωσης ΟΣΣ) είναι Asymp.sig.$<0,001<\alpha=0,05$.



Εικόνα 6.15: Πραγματοποίηση του ελέγχου Kruskal-Wallis στο SPSS



Εικόνα 6.16: Προσδιορισμός της μικρότερης και της μεγαλύτερης τιμής με τις οποίες έχουν κωδικοποιηθεί οι κατηγορίες της κατηγορικής /ποιοτικής μεταβλητής.

Test Statistics[a,b]

	cpkmbmax
Chi-Square	452,425
df	2
Asymp. Sig.	,000

a. Kruskal Wallis Test

b. Grouping Variable: ecg

Πίνακας 6.8: Αποτελέσματα από τον έλεγχο Kruskal-Wallis για τον έλεγχο ύπαρξης συσχέτισης ανάμεσα στο ένζυμο μυοκαρδιακής νέκρωσης (CPKMB) και τον τύπο διάγνωσης ΟΣΣ (π.χ. STEMI, non-STEMI ή other).

6.3.4 Kruskal–Wallis test στο STATA

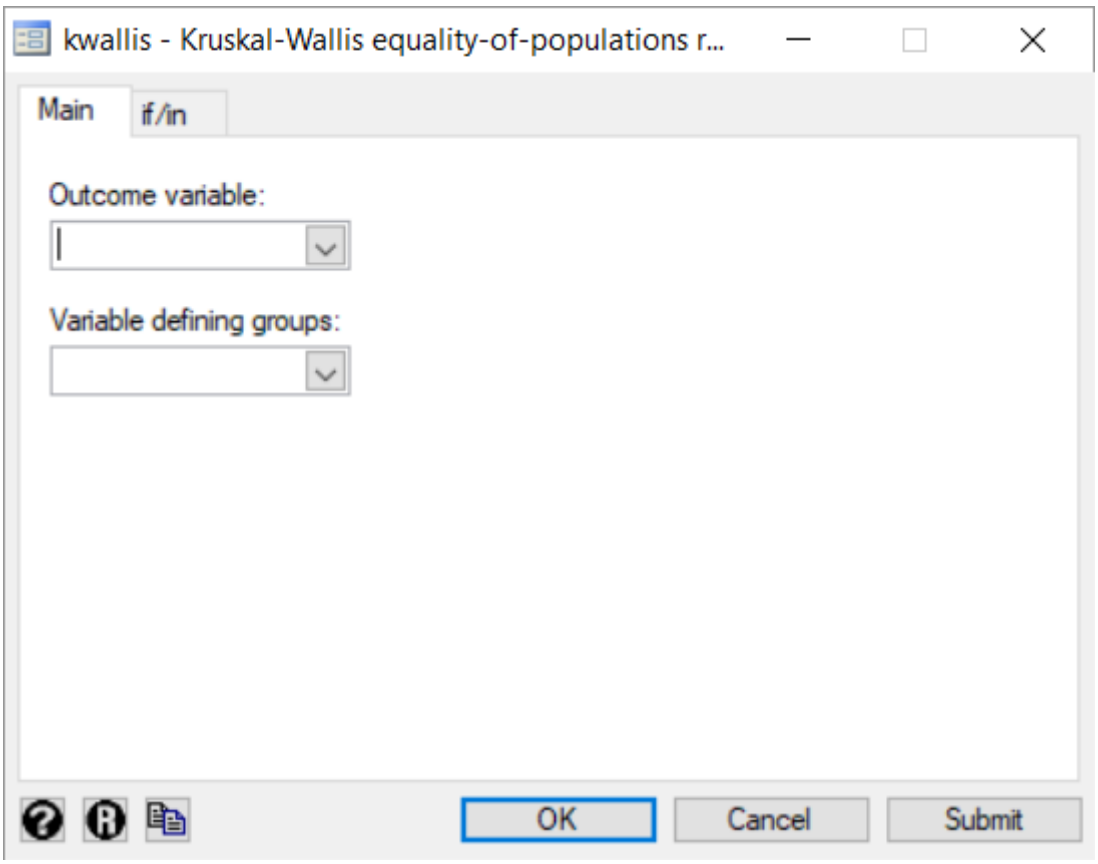
Ο έλεγχος Kruskal-Wallis, μπορεί να πραγματοποιηθεί ακολουθώντας τα εξής ***βήματα***:

Statistics → Summaries, tables and tests → Nonparametric analysis → Tests of hypotheses → Kruskal Wallis rank test

i. Ανοίγει το παράθυρο διαλόγου της *Εικόνας 6.17*

ii. Στο «***Outcome Variable***» τοποθετούμε την συνεχή μεταβλητή

iii. Στο «***Variable defining groups***» τοποθετούμε την κατηγορική μεταβλητή



Εικόνα 6.17: Πραγματοποίηση του ελέγχου Kruskal-Wallis στο STATA

Στο STATA χρησιμοποιείται η εντολή ***kwallis*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 6.18*:

Syntax

`kwallis` *varname* [*if*] [*in*] , `by(`*groupvar*`)`

Menu

Statistics > Nonparametric analysis > Tests of hypotheses > Kruskal-Wallis rank test

Description

`kwallis` tests the hypothesis that several samples are from the same population. In the syntax diagram above, *varname* refers to the variable recording the outcome, and *groupvar* refers to the variable denoting the population. `by()` is required.

Option

`by(`*groupvar*`)` is required. It specifies a variable that identifies the groups.

Εικόνα 6.18: Βασική σύνταξη της εντολής kwallis

7. Έλεγχος ύπαρξης γραμμικής συσχέτισης ανάμεσα σε δύο ποσοτικές μεταβλητές.

7.1 Εισαγωγή

Ο έλεγχος ύπαρξης γραμμικής συσχέτισης πραγματοποιείται χρησιμοποιώντας τους συντελεστές γραμμικής συσχέτισης. Αυτοί οι συντελεστές εφαρμόζονται στις περιπτώσεις που διαθέτουμε 2 ποσοτικές μεταβλητές (διακριτές ή συνεχείς) και στόχος μας είναι να ελέγξουμε αν αυτές οι 2 μεταβλητές συσχετίζονται γραμμικά.

Υπάρχουν **2 συντελεστές** γραμμικής συσχέτισης:

- Συντελεστής συσχέτισης του **Pearson,** &
- Συντελεστής συσχέτισης του **Spearman**

Η επιλογή του κατάλληλου συντελεστή εξαρτάται από το αν:

- οι ποσοτικές μεταβλητές είναι συνεχείς οι διακριτές, και από το αν
- οι συνεχείς μεταβλητές ακολουθούν την κανονική κατανομή ή όχι.

Στον *Πίνακα 7.1* παρουσιάζονται συνοπτικά οι συνθήκες κάτω από τις οποίες εφαρμόζεται κάθε ένας από αυτούς τους 2 συντελεστές γραμμικής συσχέτισης.

		Ποσοτικές συνεχείς μεταβλητές		**Ποσοτικές διακριτές μεταβλητές**
		Κανονική κατανομή	***Μη κανονική κατανομή***	
Ποσοτικές συνεχείς μεταβλητές	***Κανονική κατανομή***	Pearson	Spearman	Spearman
	Μη κανονική κατανομή	Spearman	Spearman	Spearman
Ποσοτικές διακριτές μεταβλητές		Spearman	Spearman	Spearman

Πίνακας 7.1: Συνθήκες κάτω από τις οποίες εφαρμόζεται κάθε ένας εκ των 2 συντελεστές γραμμικής συσχέτισης.

Το **εύρος τιμών** και των 2 συντελεστών γραμμικής συσχέτισης είναι από **-1 έως +1**. Τιμές των συντελεστών προς το +1 ή το -1 υποδηλώνουν ισχυρή γραμμική συσχέτιση (θετική και αρνητική, αντίστοιχα), ενώ τιμές κοντά στο μηδέν υποδηλώνουν μη ύπαρξη γραμμικής συσχέτισης.

Πιο συγκεκριμένα, τιμές των συντελεστών συσχέτισης:

- $r > |0,8|$ εκφράζουν **ισχυρή** συσχέτιση,
- $|0,6| < r < |0,8|$ εκφράζουν **μέτρια** συσχέτιση
- $|0,3| < r < |0,6|$ εκφράζουν **ελαφριά** συσχέτιση

- $r < |0,3|$ εκφράζουν **μηδενική** συσχέτιση (ασυσχέτιστες μεταβλητές)

ΠΡΟΣΟΧΗ!!!

1. Οι συντελεστές γραμμικής συσχέτισης εξετάζουν μόνο κατά πόσο 2 ποσοτικές μεταβλητές συσχετίζονται **γραμμικά** μεταξύ τους. Συνεπώς, ένας συντελεστής γραμμικής συσχέτισης κοντά στο μηδέν δεν υποδηλώνει ότι αυτές οι 2 μεταβλητές δεν συσχετίζονται μεταξύ τους, αλλά ότι δεν συσχετίζονται γραμμικά. Ενδεχομένως, αυτές οι 2 μεταβλητές να συσχετίζονται παραβολικά (για παράδειγμα).
2. Οι συντελεστές γραμμικής συσχέτισης **δεν** μπορούν να προσδιορίσουν αν η σχέση που συνδέει τις 2 μεταβλητές είναι **σχέση αιτίας αιτιατού** (αιτιολογική σχέση).

7.2 Γραφική διερεύνηση της ύπαρξης γραμμικής συσχέτισης στο SPSS: Στικτόγραμμα

Ένας από τους πιο εύκολους τρόπους να ελέγξουμε αν 2 μεταβλητές συσχετίζονται γραμμικά είναι το στικτόγραμμα (scatterplot). Ας υποθέσουμε ότι επιθυμούμε να ελέγξουμε αν υπάρχει γραμμική συσχέτιση ανάμεσα στο βάρος (weight) και την ηλικία (Age) ασθενών που εισήχθησαν στο νοσοκομείο με οξύ στεφανιαίο σύνδρομο. Για να δημιουργήσουμε το στικτόγραμμα στο SPSS ακολουθούμε τα εξής ***βήματα***:

Graphs → Scatter/Dot…

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 7.1*

ii. Πατάμε «***Simple Scatter***» και «***Define***» και εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 7.2*.

iii. Επιλέγουμε τη μία εκ των 2 μεταβλητών και την τοποθετούμε στον άξονα X (***X axis***) και την άλλη στον άξονα Y (***Y axis***) (π.χ., ηλικία και βάρος αντίστοιχα).

iv. Πατάμε «***Ok***» και εμφανίζεται το γράφημα της *Εικόνας 7.3*

v. Τοποθετούμε τον κέρσορα πάνω στο γράφημα της *Εικόνας 7.3* και πατώντας διπλό κλικ εμφανίζεται το παράθυρο του SPSS «***Chart Editor***» (*Εικόνα 7.4*)

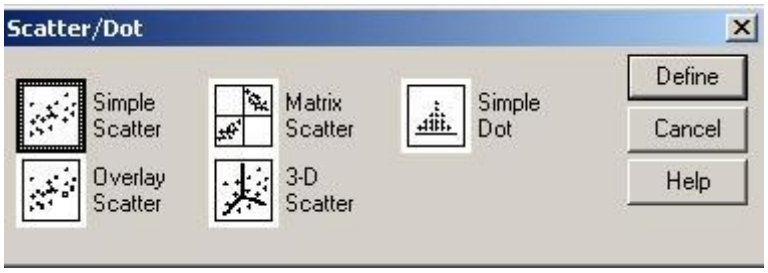
vi. Από την μπάρα εργαλείων του ***Chart Editor*** επιλέγουμε «***Elements»***, «***Fit line at total»*** και

vii. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 7.5* και πατάμε «***Linear***»

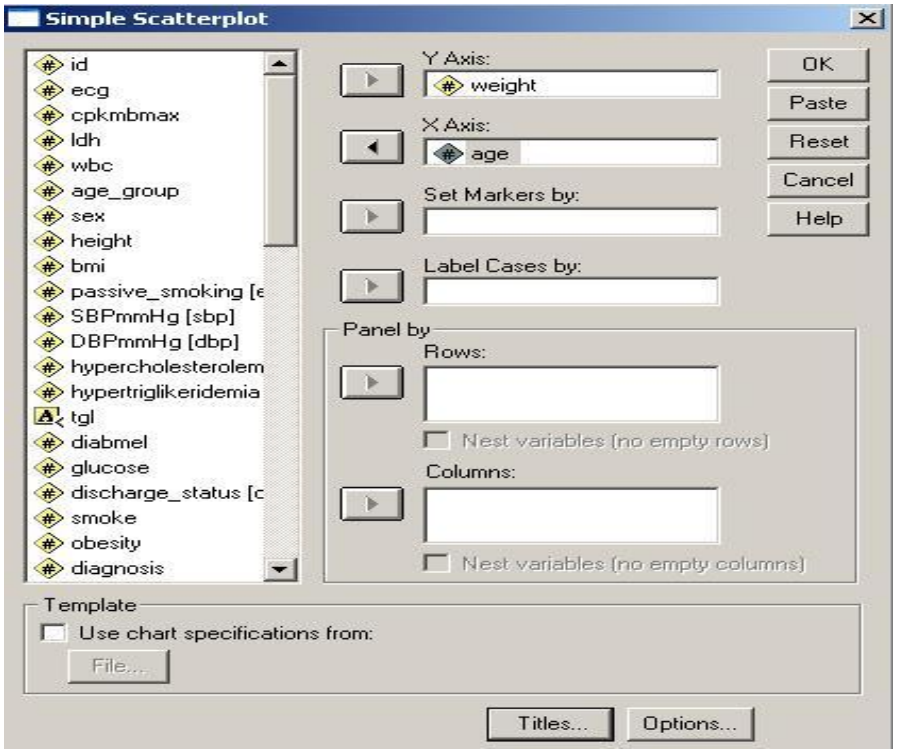
viii. Πατάμε «***Apply***» και

ix. Εμφανίζεται το γράφημα της *Εικόνας 7.6*.

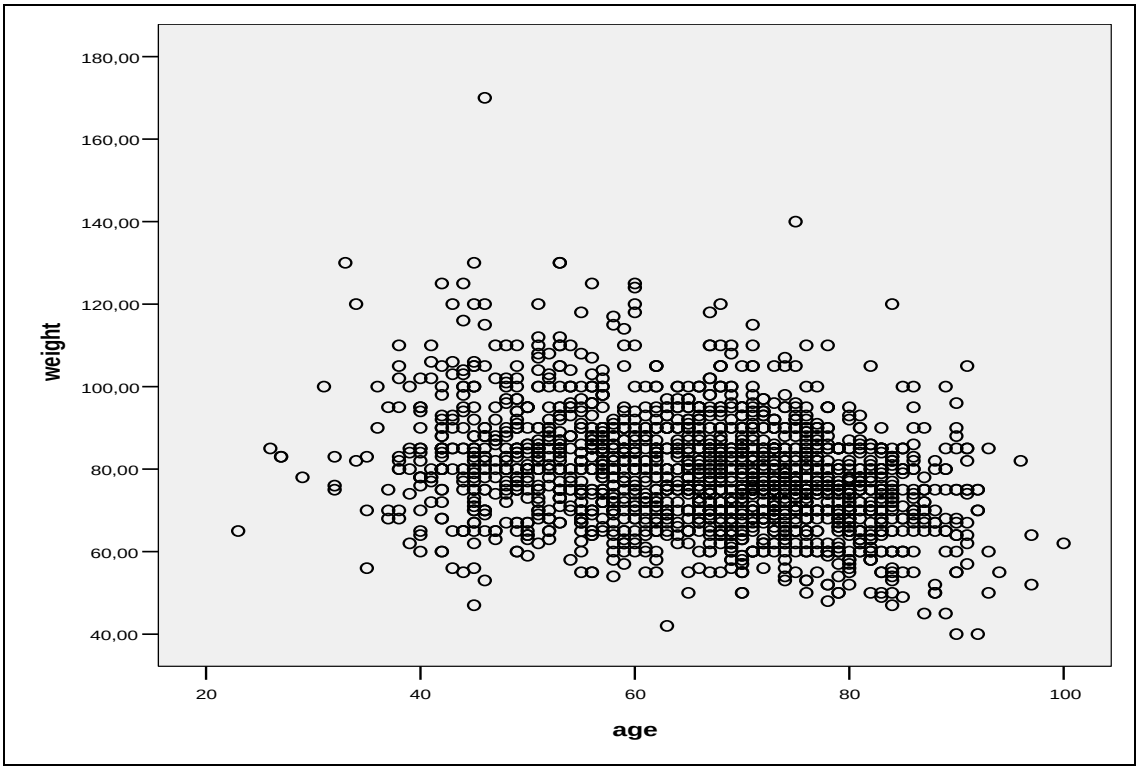
x. Από την *Εικόνα 7.6* διαπιστώνουμε ότι η κλίση της γραμμής που προσαρμόσαμε και εκφράζει την ένταση της γραμμικής σχέσης ανάμεσα στο βάρος και την ηλικία είναι μικρή, υποδηλώνοντας ότι η γραμμική συσχέτιση είναι ελαφριά και μάλιστα αρνητική (όσο αυξάνεται η ηλικία μειώνεται το βάρος).



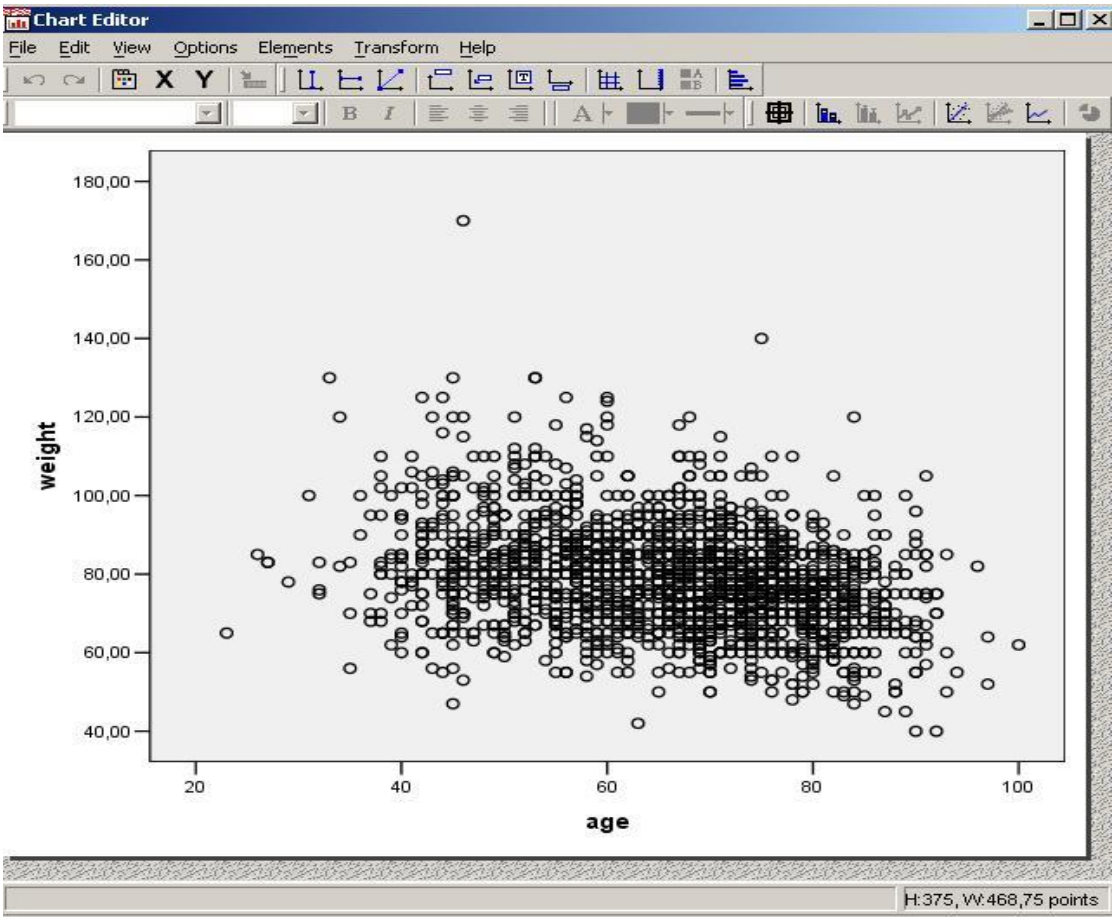
Εικόνα 7.1: Δημιουργία στικτογράμματος στο SPSS



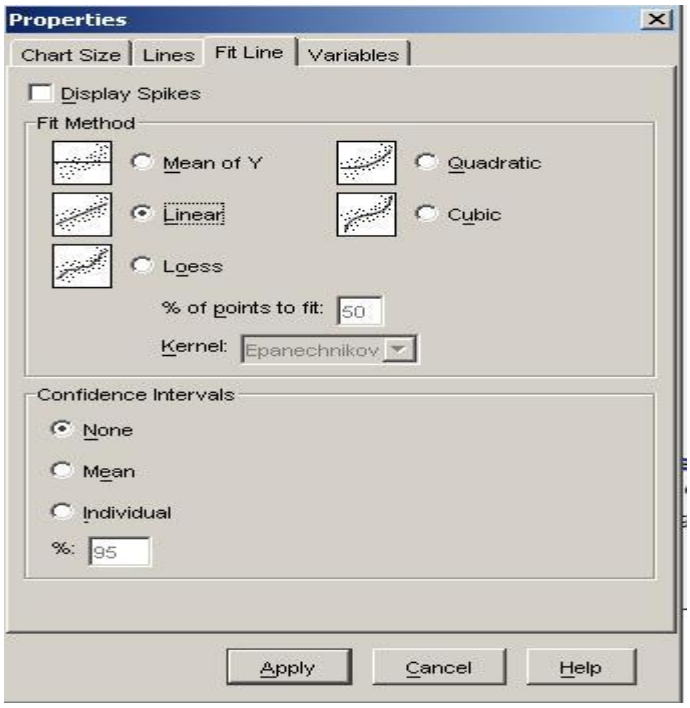
Εικόνα 7.2: Προσδιορισμός των μεταβλητών μεταξύ των οποίων επιθυμούμε να δημιουργηθεί το στικτόγραμμα (π.χ. ηλικία και βάρος).



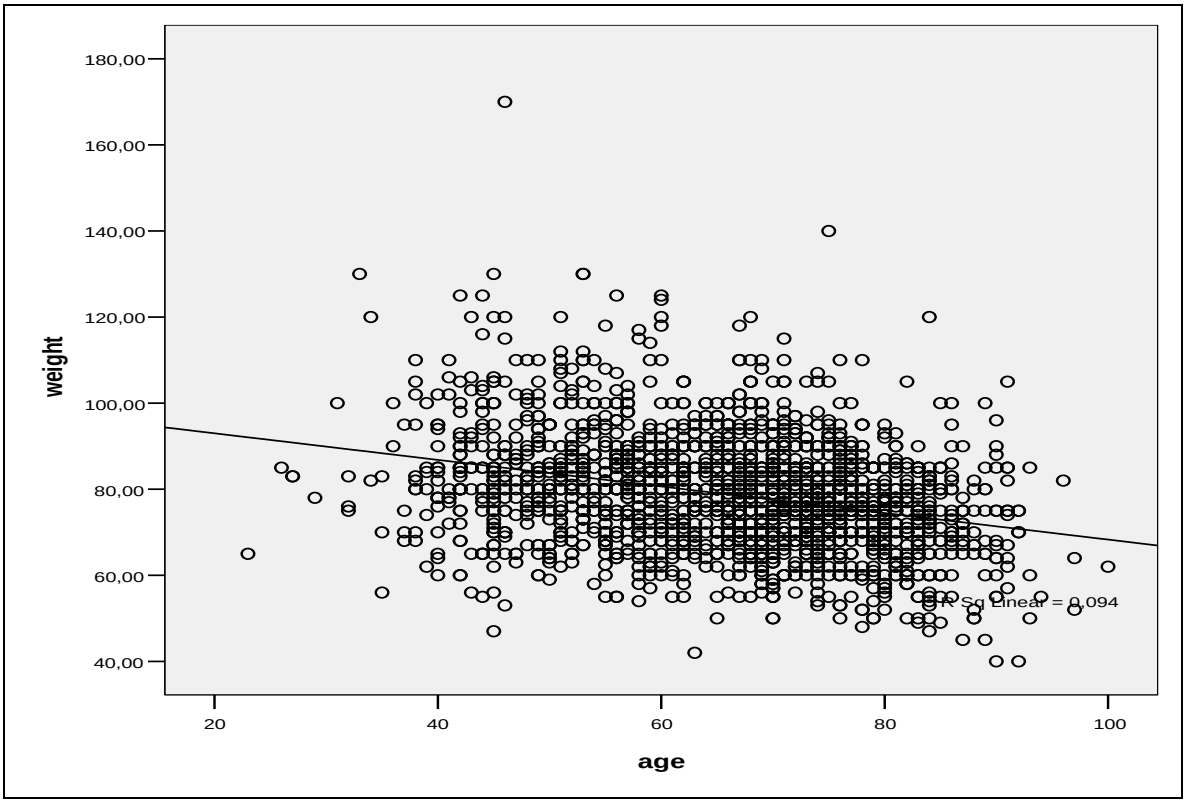
Εικόνα 7.3: Στικτόγραμμα ανάμεσα στο βάρος και την ηλικία.



Εικόνα 7.4: Chart editor παράθυρο που εμφανίζεται κάνοντας διπλό κλικ πάνω στο γράφημα της Εικόνας 7.3.



Εικόνα 7.5: Προσαρμογή της γραμμής που εκφράζει τα δεδομένα



Εικόνα 7.6: Στικτόγραμμα ανάμεσα στην ηλικία και το βάρος έχοντας εφαρμόσει και τη γραμμή που εκφράζει τα δεδομένα μας.

7.3 Γραφική διερεύνηση της ύπαρξης γραμμικής συσχέτισης στο STATA: Στικτόγραμμα

Για να δημιουργήσουμε το στικτόγραμμα στο STATA ακολουθούμε τα εξής ***βήματα***:

Graphics ⟶ Twoway graph (scatter, line, etc)

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 7.7*

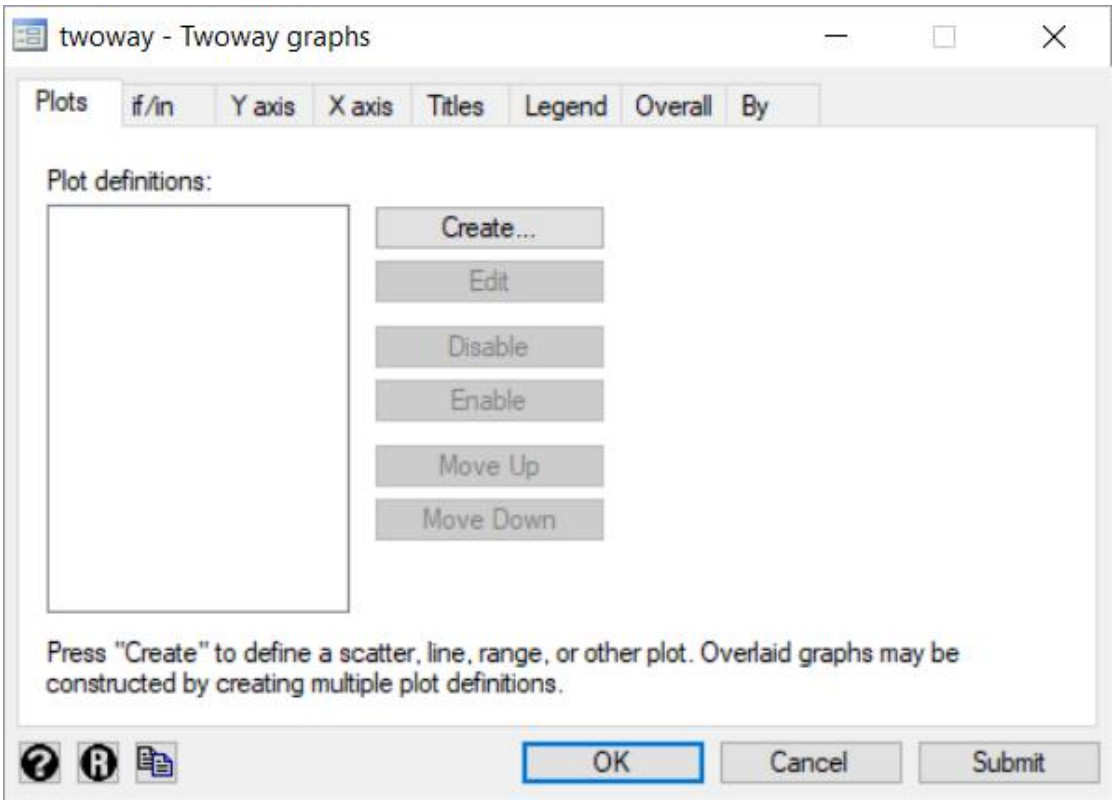
ii. Όπως φαίνεται έχουμε την δυνατότητα να ορίσουμε τίτλο στον οριζόντιο άξονα (***X axis***), στον κάθετο άξονα (***Y axis***) ή στο γράφιμα γενικά (***Titles***).

iii. Πατάμε «***Create***» και εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 7.8*.

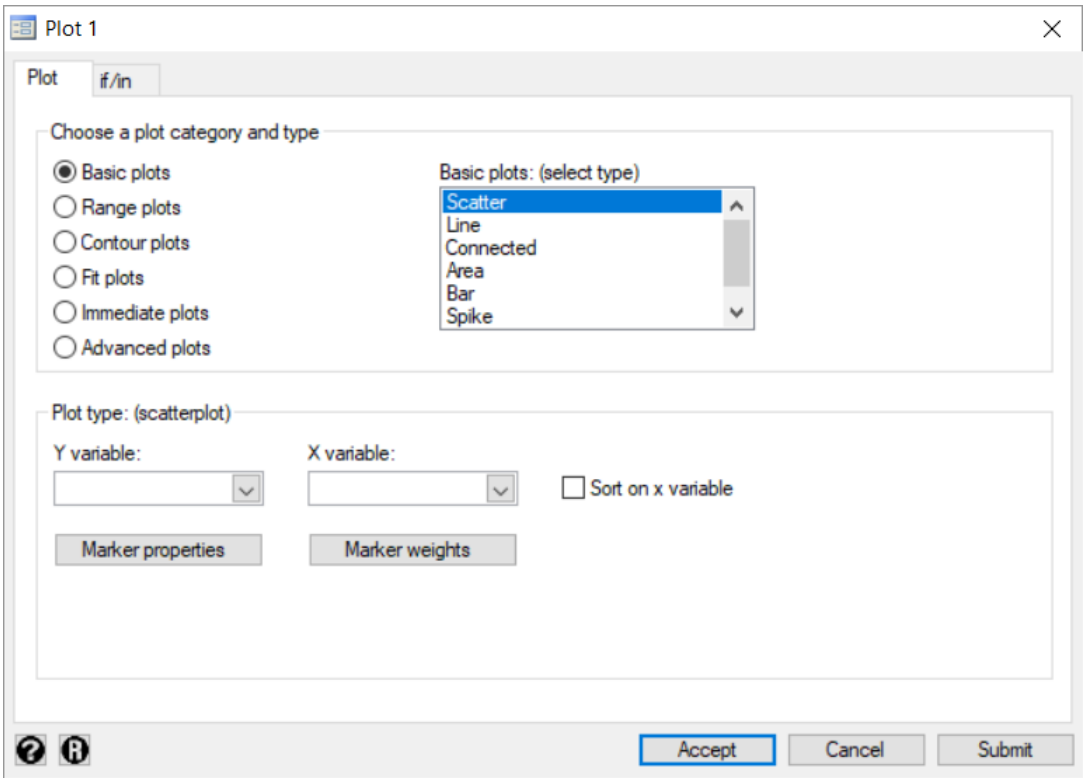
iv. Επιλέγουμε τη μία εκ των 2 μεταβλητών και την τοποθετούμε στο ***Y variable*** και την άλλη στο ***X variable***.

v. Πατάμε «***Accept***» και στην συνέχεια «***OK***»

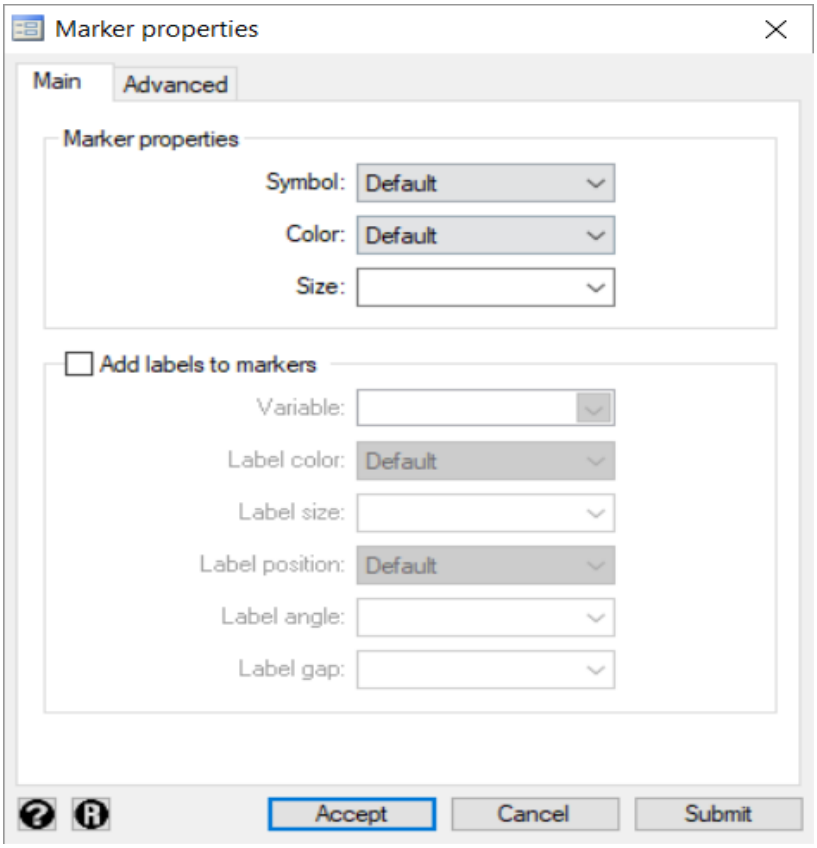
vi. Στο «***market properties***» έχουμε την δυνατότητα να αλλάξουμε την εμφάνιση του γραφήματος, όπως φαίνεται στην *Εικόνα 7.9*.



Εικόνα 7.7: Δημιουργία στικτογράμματος στο STATA



Εικόνα 7.8: Προσδιορισμός των μεταβλητών μεταξύ των οποίων επιθυμούμε να δημιουγηθεί το στικτόγραμμα.



Εικόνα 7.9: Η επιλογή Marker properties

Στο STATA χρησιμοποιείται η εντολή ***scatter*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 7.10*:

Syntax

[twoway] scatter *varlist* [*if*] [*in*] [*weight*] [, *options*]

where *varlist* is

y_1 [y_2[...]] x

options	Description
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
connect_options	change look of lines or connecting method
composite_style_option	overall style of the plot
jitter_options	jitter marker positions using random noise
axis_choice_options	associate plot with alternative axis
twoway_options	titles, legends, axes, added lines and text, by, regions, name, aspect ratio, etc.

Each is defined below.

marker_options	Description
msymbol(*symbolstylelist*)	shape of marker
mcolor(*colorstylelist*)	color of marker, inside and out
msize(*markersizestylelist*)	size of marker
mfcolor(*colorstylelist*)	inside or "fill" color
mlcolor(*colorstylelist*)	color of outline
mlwidth(*linewidthstylelist*)	thickness of outline
mlstyle(*linestylelist*)	overall style of outline
mstyle(*markerstylelist*)	overall style of marker

See [G-3] ***marker_options***.

Εικόνα 7.10: Βασική σύνταξη της εντολής scatter

7.3 Συντελεστής συσχέτισης του Pearson ή Spearman στο SPSS

Ας υποθέσουμε ότι επιθυμούμε να ελέγξουμε αν υπάρχει γραμμική συσχέτιση ανάμεσα σε 2 ποσοτικές συνεχείς μεταβλητές (π.χ. ηλικία και βάρος). Επίσης, ας υποθέσουμε ότι έχουμε πραγματοποιήσει έλεγχο κανονικής κατανομής (βλ. *κεφάλαιο 8*) και ότι και οι 2 μεταβλητές ακολουθούν την κανονική κατανομή. Το κατάλληλο κριτήριο για τον έλεγχο αυτής της συσχέτισης είναι ο ***συντελεστής συσχέτισης του Pearson "Pearson Correlation"***. Με αυτήν την ανάλυση υπολογίζουμε μία τιμή του συντελεστή συσχέτισης Pearson, όπως αυτή προκύπτει χρησιμοποιώντας τα δεδομένα από το δείγμα μας και στη συνέχεια ελέγχουμε αν αυτή η τιμή διαφέρει σημαντικά από το μηδέν. Πιο συγκεκριμένα, η μηδενική και εναλλακτική υπόθεση είναι:

H_0: Ο συντελεστής συσχέτισης του Pearson είναι ίσος με το μηδέν και άρα ΔΕΝ ΥΠΑΡΧΕΙ γραμμική συσχέτιση ανάμεσα στο βάρος και την ηλικία.

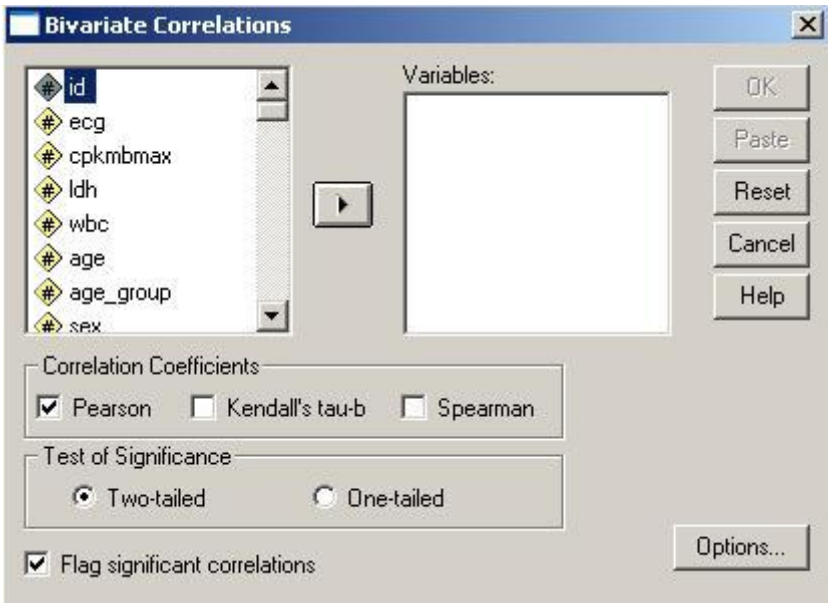
H_1: Ο συντελεστής συσχέτισης του Pearson διαφέρει από το μηδέν και άρα ΥΠΑΡΧΕΙ γραμμική συσχέτιση ανάμεσα στο βάρος και την ηλικία.

Για να καταλήξουμε στο συμπέρασμα ότι υπάρχει συσχέτιση ανάμεσα στο βάρος και την ηλικία, θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value) να είναι $< \alpha = 0,05$ ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).

Η ανάλυση συσχέτισης στο SPSS πραγματοποιείται ακολουθώντας τα ***βήματα***:

Analyse → Correlate → Bivariate

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 7.11*.

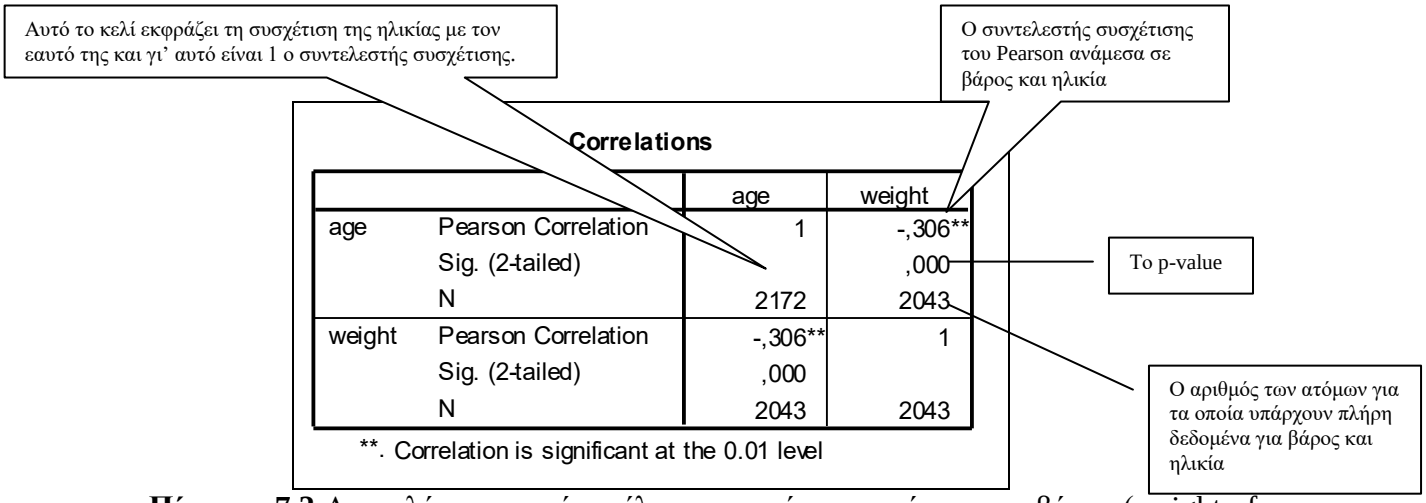


Εικόνα 7.11: Υπολογισμός των συντελεστών συσχέτισης.

ii. Στο «***Variables***» τοποθετούμε τις ποσοτικές μεταβλητές μεταξύ των οποίων επιθυμούμε να ελέγξουμε την ύπαρξη συσχέτισης (π.χ., ηλικία και φύλο).

iii. Στο «***Correlation Coefficients***» είναι προεπιλεγμένο το Pearson, οπότε,

iv. Πατώντας «***Ok***» εμφανίζεται ο *Πίνακας 7.2*, στον οποίο παρουσιάζονται τα αποτελέσματα. Συγκεκριμένα, στα αποτελέσματα φαίνεται η εκτίμηση του πραγματικού συντελεστή συσχέτισης (r) μεταξύ αυτών των μεταβλητών (Pearson Correlation = - 0,306) και η αντίστοιχη πιθανότητα (p) εσφαλμένης απόρριψης της H_o (sig.(2-tailed) = 0,000…1), η οποία, σαφώς, είναι μικρότερη του επιπέδου σημαντικότητας α=0,05. Συνεπώς, απορρίπτουμε τη μηδενική υπόθεση, δηλαδή συμπεραίνουμε πως υπάρχει ***γραμμική συσχέτιση*** ανάμεσα στο βάρος και την ηλικία και συγκεκριμένα υπάρχει αρνητική συσχέτιση, (δηλαδή αύξηση του ενός συνεπάγεται μείωση του άλλου) στον πληθυσμό αναφοράς της μελέτης. Επίσης, παρατηρούμε ότι πλήρη δεδομένα και για τις 2 μεταβλητές υπάρχουν για 2043 άτομα (N).



Πίνακας 7.2 Αποτελέσματα από το έλεγχο συσχέτισης ανάμεσα σε βάρος (weight of subjects) και ηλικία (age) χρησιμοποιώντας το συντελεστή συσχέτισης του Pearson.

Παρατήρηση: Στον πίνακα που προκύπτει από την ανάλυση συσχέτισης (π.χ. *Πίνακας 7.2*) τα διαγώνια κελιά εκφράζουν τη συσχέτιση όλων των μεταβλητών με τον εαυτό τους. Αυτά τα κελιά δεν μας παρέχουν καμία χρήσιμη πληροφορία.

7.4 Συντελεστής συσχέτισης του Pearson ή Spearman στο STATA

Ο συντελεστής συσχέτισης του Pearson στο STATA βρίσκεται ακολουθώντας τα ***βήματα***:

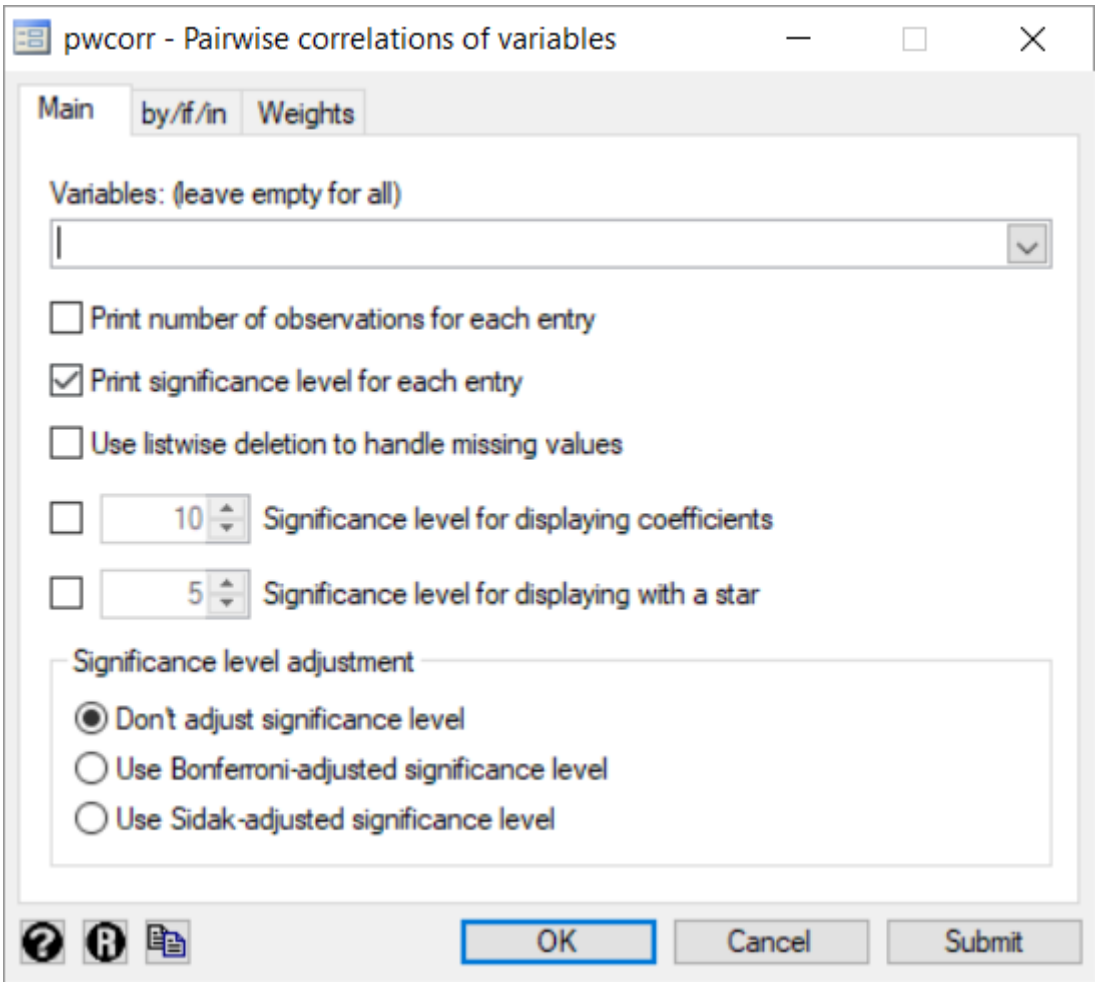
Analyse Statistics → Summaries, tables, and tests → Summary and descriptive statistics → Pairwise correlations

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 7.12*

ii. Στο «***Variables***» τοποθετούμε τις ποσοτικές μεταβλητές μεταξύ των οποίων επιθυμούμε να ελέγξουμε την ύπαρξη συσχέτισης.

iii. Επιλέγουμε «***Print significance level for each entry***» με σκοπό να μας εμφανίσει την πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value).

iv. «***OK***»



Εικόνα 7.12: Υπολογισμός του συντελεστή συσχέτισης του Pearson

Στο STATA χρησιμοποιείται η εντολή ***pwcorr*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 7.13*:

Syntax

Display correlation matrix or covariance matrix

`correlate` [*varlist*] [*if*] [*in*] [*weight*] [, *correlate_options*]

Display all pairwise correlation coefficients

`pwcorr` [*varlist*] [*if*] [*in*] [*weight*] [, *pwcorr_options*]

correlate_options	Description
Options	
`means`	display means, standard deviations, minimums, and maximums with matrix
`noformat`	ignore display format associated with variables
`covariance`	display covariances
`wrap`	allow wide matrices to wrap

pwcorr_options	Description
Main	
`obs`	print number of observations for each entry
`sig`	print significance level for each entry
`listwise`	use listwise deletion to handle missing values
`casewise`	synonym for `listwise`
`print(#)`	significance level for displaying coefficients
`star(#)`	significance level for displaying with a star
`bonferroni`	use Bonferroni-adjusted significance level
`sidak`	use Šidák-adjusted significance level

varlist may contain time-series operators; see [U] **11.4.4 Time-series varlists**.
`by` is allowed with `correlate` and `pwcorr`; see [D] **by**.
`aweights` and `fweights` are allowed; see [U] **11.1.6 weight**.

Εικόνα 7.13: Βασική σύνταξη της εντολής pwcorr

Ο συντελεστής συσχέτισης του Spearman στο STATA βρίσκεται ακολουθώντας τα ***βήματα***:

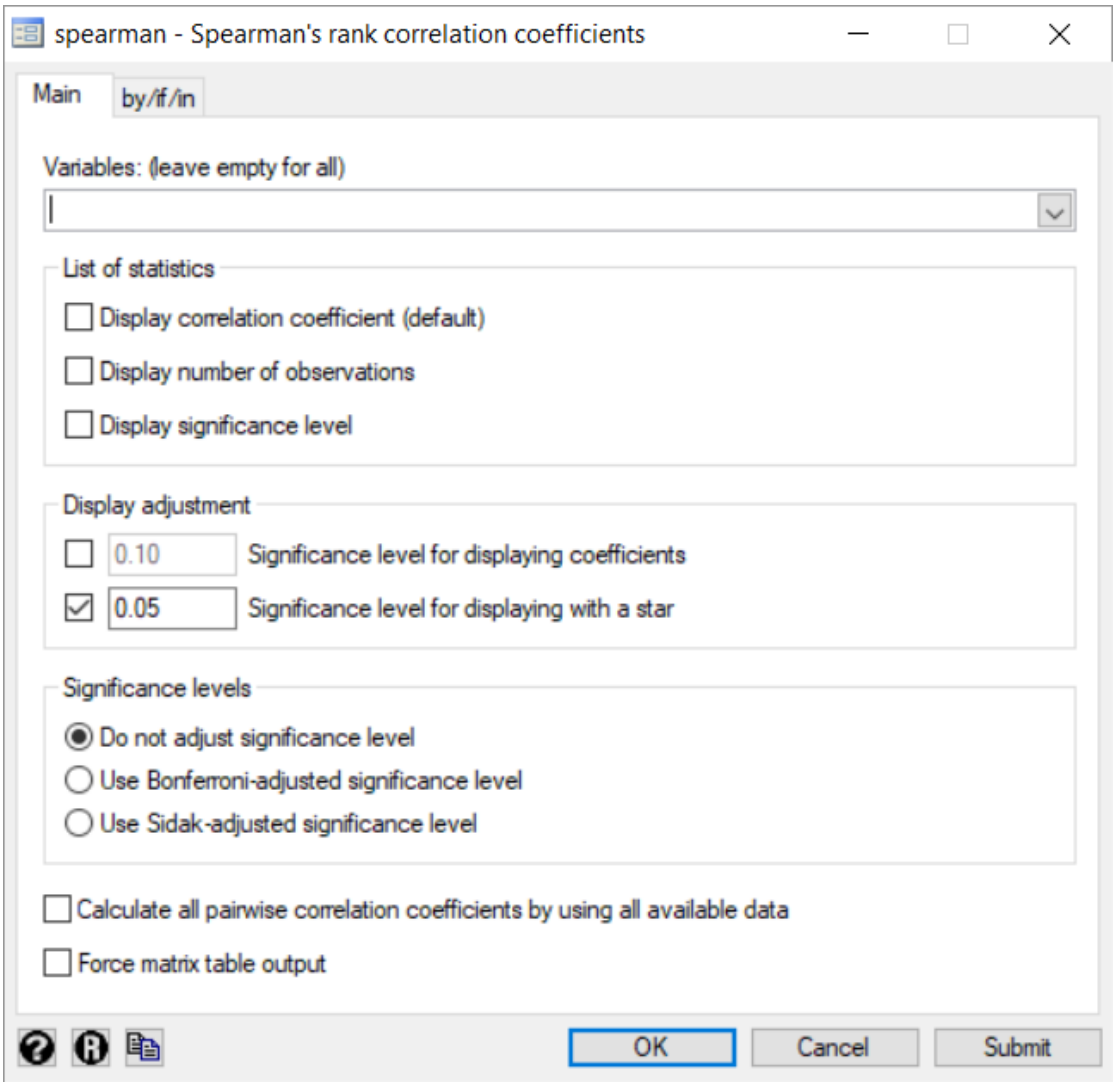
Statistics → Nonparametric analysis → Tests of hypotheses → Spearman's rank correlation

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 7.14*

ii. Στο «***Variables***» τοποθετούμε τις ποσοτικές μεταβλητές μεταξύ των οποίων επιθυμούμε να ελέγξουμε την ύπαρξη συσχέτισης.

iii. Επιλέγουμε «***Significance level for displaying with a star***» με σκοπό να μας εμφανίσει την πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value).

iv. «***OK***»



Εικόνα 7.14: Υπολογισμός του συντελεστή συσχέτισης του Spearman

Στο STATA χρησιμοποιείται η εντολή ***spearman*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 7.15*:

Syntax

Spearman's rank correlation coefficients

`spearman` [*varlist*] [*if*] [*in*] [`,` *spearman_options*]

Kendall's rank correlation coefficients

`ktau` [*varlist*] [*if*] [*in*] [`,` *ktau_options*]

spearman_options	Description
Main	
`stats(`*spearman_list*`)`	list of statistics; select up to three statistics; default is `stats(rho)`
`print(#)`	significance level for displaying coefficients
`star(#)`	significance level for displaying with a star
`bonferroni`	use Bonferroni-adjusted significance level
`sidak`	use Šidák-adjusted significance level
`pw`	calculate all pairwise correlation coefficients by using all available data
`matrix`	display output in matrix form

ktau_options	Description
Main	
`stats(`*ktau_list*`)`	list of statistics; select up to six statistics; default is `stats(taua)`
`print(#)`	significance level for displaying coefficients
`star(#)`	significance level for displaying with a star
`bonferroni`	use Bonferroni-adjusted significance level
`sidak`	use Šidák-adjusted significance level
`pw`	calculate all pairwise correlation coefficients by using all available data
`matrix`	display output in matrix form

`by` is allowed with `spearman` and `ktau`; see [D] **by**.

where the elements of *spearman_list* may be

`rho`	correlation coefficient
`obs`	number of observations
`p`	significance level

Εικόνα 7.15: Βασική σύνταξη της εντολής spearman

8. Έλεγχος κανονικότητας και έλεγχος ομοσκεδαστικότητας

8.1 Έλεγχος κανονικότητας στο SPSS

Από τα παραπάνω φαίνεται πως για την πραγματοποίηση πολλών στατιστικών ελέγχων, απαραίτητη προϋπόθεση είναι η συνεχής μεταβλητή να ακολουθεί την κανονική κατανομή. Ο έλεγχος κανονικής κατανομής μπορεί να πραγματοποιηθεί με 2 τρόπους:

a. **Γραφικός έλεγχος*: Ιστόγραμμα*** όπου θα πρέπει να έχει τη μορφή της κανονικής κατανομής ή το ***Normal Q-Q plot*** όπου τα σημεία θα πρέπει να βρίσκονται πάνω σε μία διαγώνια ευθεία γραμμή.
b. **Στατιστικός έλεγχος**: ***Kolmogorov – Smirnov*** και ***Shapiro – Wilk.*** Η μηδενική και εναλλακτική υπόθεση αυτών των ελέγχων είναι:

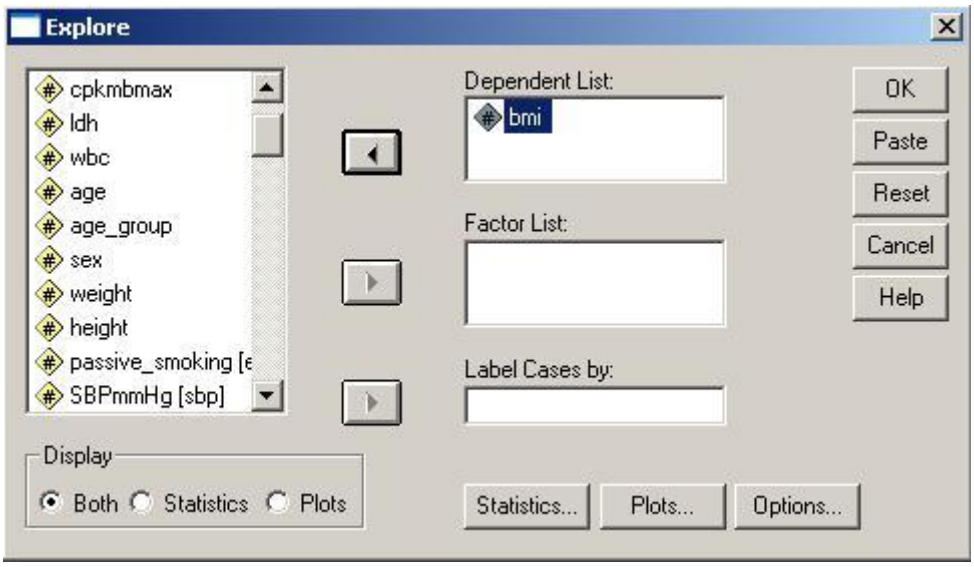
   H_0: Η κατανομή της μεταβλητής μας ΕΙΝΑΙ κανονική.

   H_1: Η κατανομή της μεταβλητής μας ΔΕΝ ΕΙΝΑΙ κανονική.

Το SPSS έχει τη δυνατότητα να πραγματοποιούμε τον έλεγχο κανονικότητας εφαρμόζοντας ταυτόχρονα γραφικό και στατιστικό έλεγχο ακολουθώντας τα εξής ***βήματα***:

Analyze ⟶ Descriptive Statistics → Explore

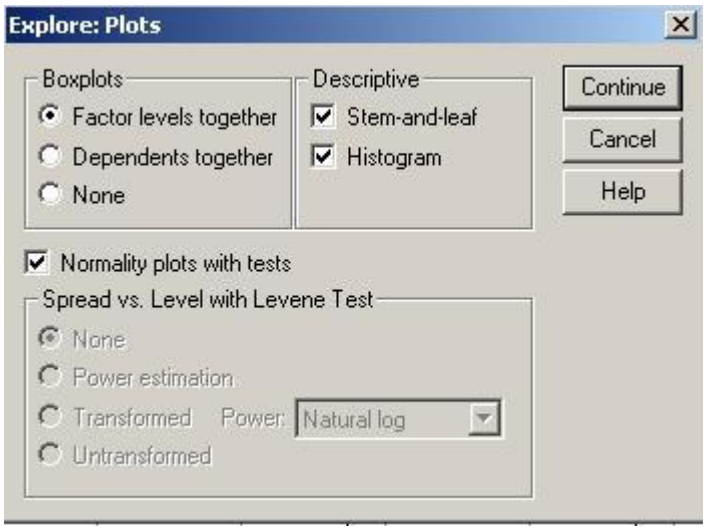
i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 8.1*,



Εικόνα 8.1: Πραγματοποίηση ελέγχου κανονικής κατανομής μιας ποσοτικής μεταβλητής (π.χ. bmi).

ii. Τοποθετούμε στο "***Dependent List***" όλες τις συνεχείς μεταβλητές, των οποίων την κανονικότητα επιθυμούμε να ελέγξουμε.
iii. Πατώντας το κουμπί επιλογών "***Plots***" ανοίγει ένα άλλο παράθυρο διαλόγου (*Εικόνα 8.2*), και
iv. Επιλέγουμε "***Normality Plots with tests***" & "***Histogram***" .

v. "***Continue***" & "***Ok***"



Εικόνα 8.2: Επιλογή πραγματοποίησης των κατάλληλων στατιστικών ελέγχων και γραφημάτων για τον έλεγχο της κανονικής κατανομής.

vi. Στον *Πίνακα 8.1* παρουσιάζονται τα αποτελέσματα του στατιστικού ελέγχου για τον έλεγχο κανονικής κατανομής, όπου και διαπιστώνουμε ότι και τα 2 στατιστικά κριτήρια απορρίπτουν την μηδενική υπόθεση, δηλ. την προϋπόθεση της κανονικής κατανομής, αφού Sig.<0,001 και για τα 2 κριτήρια.

Tests of Normality

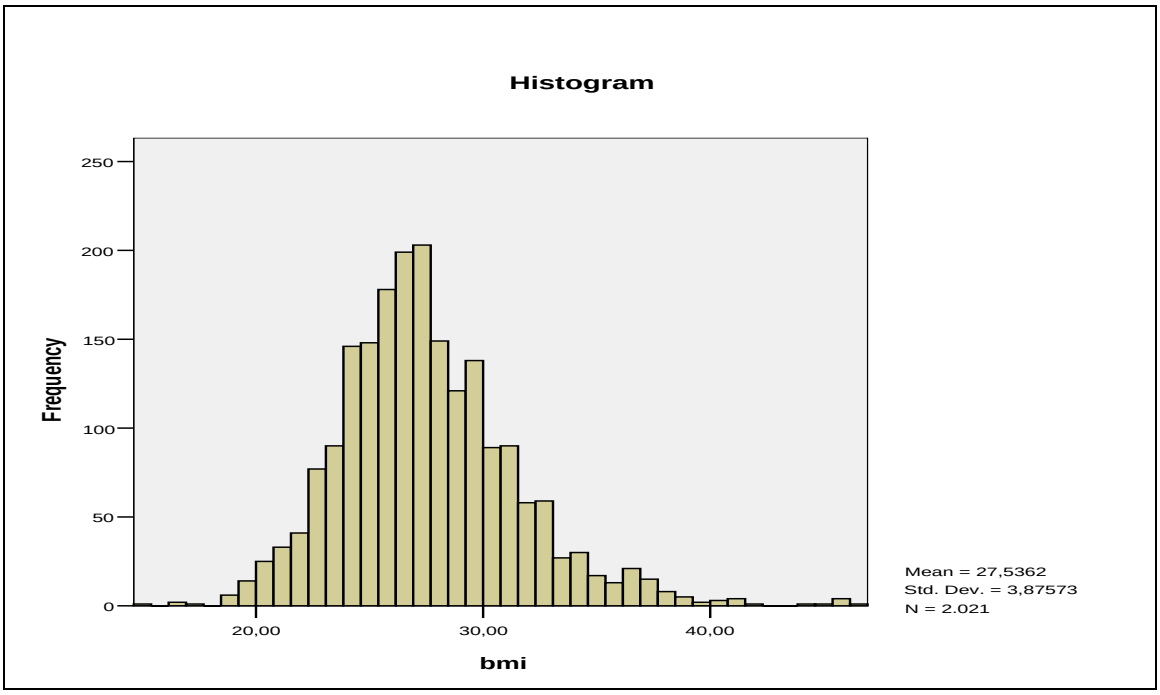
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
bmi	,075	2021	,000	,966	2021	,000

a. Lilliefors Significance Correction

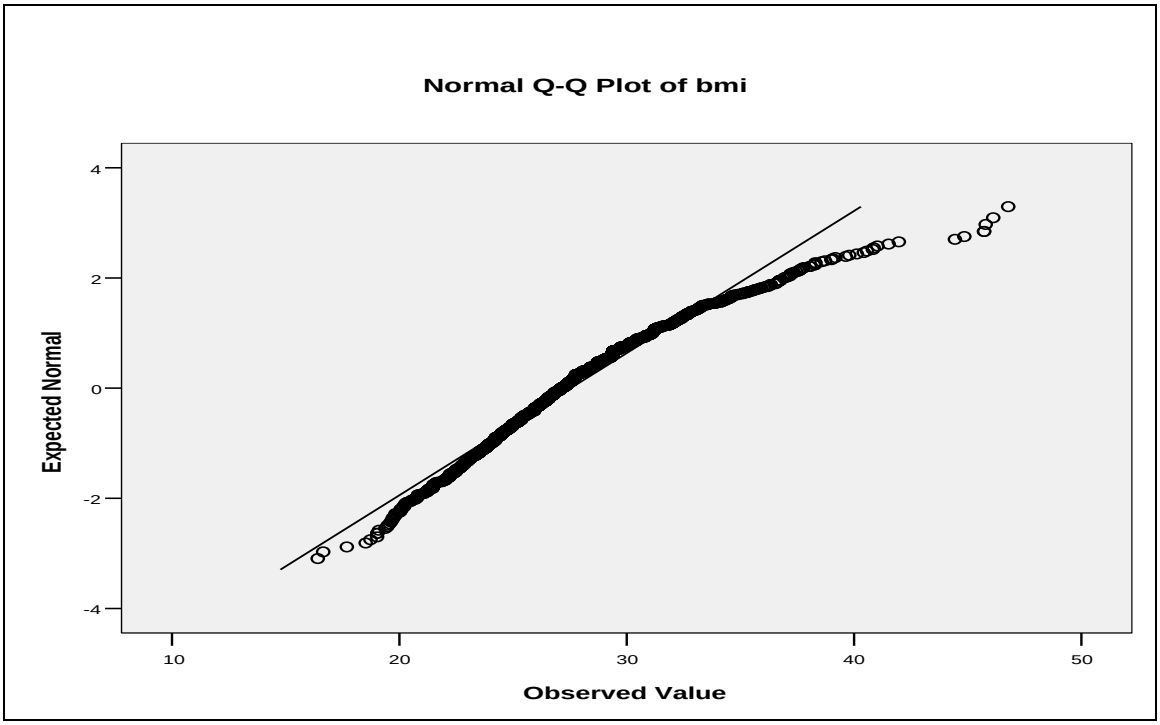
Πίνακας 8.1: Αποτελέσματα στατιστικού ελέγχου για τον έλεγχο κανονικής κατανομής της μεταβλητής ΔΜΣ (bmi).

ΠΡΟΣΟΧΗ:
Σε αυτό το σημείο πρέπει να αναφερθεί ότι τα στατιστικά κριτήρια για τον έλεγχο της κανονικότητας έχουν χαμηλή ισχύ, και αυτό μας οδηγεί στο συμπέρασμα ότι είναι προτιμότερο να εμπιστευόμαστε τα αποτελέσματα του γραφικού ελέγχου.

vii. Οι *Εικόνες 8.1* & *8.2* παρουσιάζουν το Ιστόγραμμα και το Normal Q-Q plot για τον ΔΜΣ. Από τα γραφήματα διαπιστώνουμε ότι η κατανομή του ΔΜΣ δεν απέχει πολύ από την κανονική κατανομή. Συνεπώς, λοιπόν, διαπιστώνουμε ότι υπάρχει διαφωνία των 2 διαφορετικών μεθόδων (γραφικός και στατιστικός) όσον αφορά στον έλεγχο κανονικότητας της κατανομής.



Εικόνα 8.1: Ιστόγραμμα για τον έλεγχο της κανονικής κατανομής της μεταβλητής ΔΜΣ.



Εικόνα 8.2: Normal Q-Q plot για τον έλεγχο της κανονικής κατανομής της μεταβλητής ΔΜΣ.

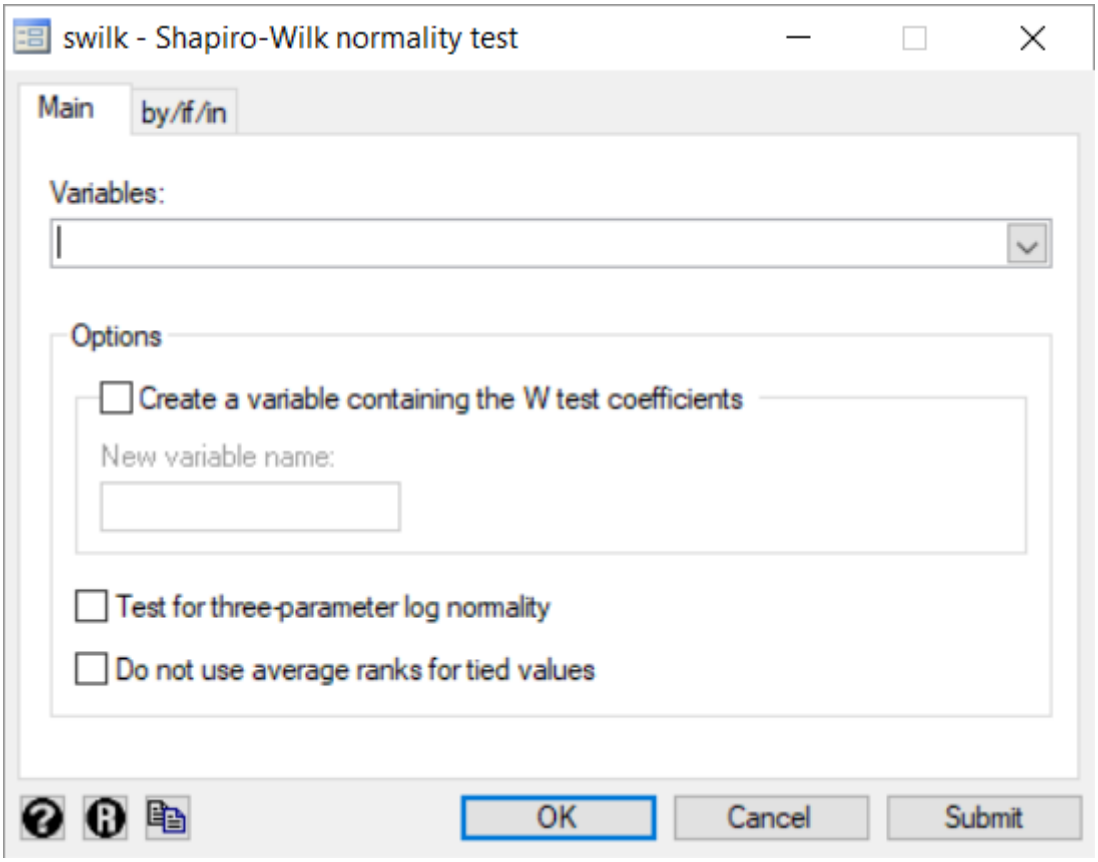
8.2 Έλεγχος κανονικότητας στο STATA

Ο στατιστικός έλεγχος της κανονικότητας γίνεται ακολουθώντας τα εξής ***βήματα***:

Statistics → Summaries, tables, and tests → Distributional plots and tests → Shapiro-Wilk normality test

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 8.3*

ii. Τοποθετούμε στο "***Variables***" όλες τις συνεχείς μεταβλητές, των οποίων την κανονικότητα επιθυμούμε να ελέγξουμε.



Εικόνα 8.3: Πραγματοποίηση ελέγχου κανονικής κατανομής μιας ποσοτικής μεταβλητής

Στο STATA χρησιμοποιείται η εντολή ***swilk*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 8.4*:

Syntax

Shapiro–Wilk normality test

`swilk` *varlist* [*if*] [*in*] [, *swilk_options*]

Shapiro–Francia normality test

`sfrancia` *varlist* [*if*] [*in*] [, *sfrancia_options*]

swilk_options	Description
Main	
`generate(`*newvar*`)`	create *newvar* containing W test coefficients
`lnnormal`	test for three-parameter lognormality
`noties`	do not use average ranks for tied values

sfrancia_options	Description
Main	
`boxcox`	use the Box–Cox transformation for W'; the default is to use the log transformation
`noties`	do not use average ranks for tied values

`by` is allowed with `swilk` and `sfrancia`; see [D] **by**.

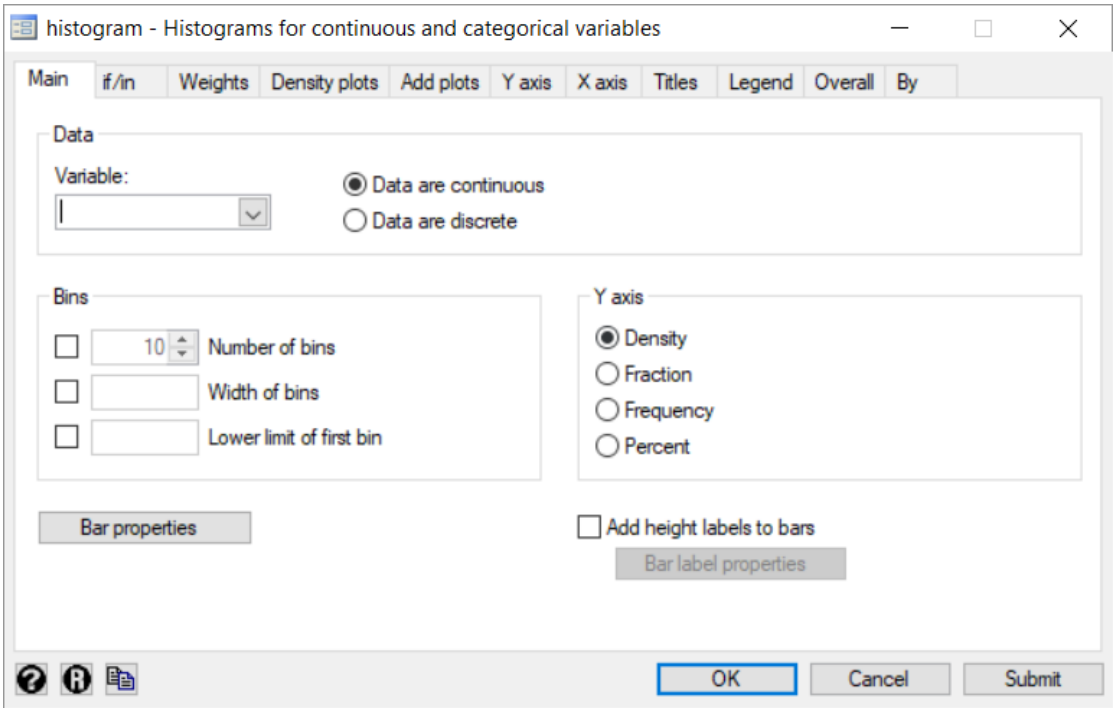
Εικόνα 8.4: Βασική σύνταξη της εντολής swilk.

Ο γραφικός έλεγχος της κανονικότητας γίνεται ακολουθώντας τα εξής ***βήματα***:

Graphics →Histogram

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 8.5*,

ii. Τοποθετούμε στο "***Variable***" την συνεχή μεταβλητή, για την οποία επιθυμούμε να κατασκευάσουμε το Ιστόγραμμα της.



Εικόνα 8.5: Πραγματοποίηση ελέγχου κανονικής κατανομής μιας ποσοτικής μεταβλητής μέσω Ιστογράμματος

Στο STATA χρησιμοποιείται η εντολή ***histogram*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 8.6*:

Syntax

`histogram` *varname* [*if*] [*in*] [*weight*] [, [*continuous_opts* | *discrete_opts*] *options*]

continuous_opts	Description
Main	
`bin(#)`	set number of bins to #
`width(#)`	set width of bins to #
`start(#)`	set lower limit of first bin to #

discrete_opts	Description
Main	
`discrete`	specify that data are discrete
`width(#)`	set width of bins to #
`start(#)`	set theoretical minimum value to #

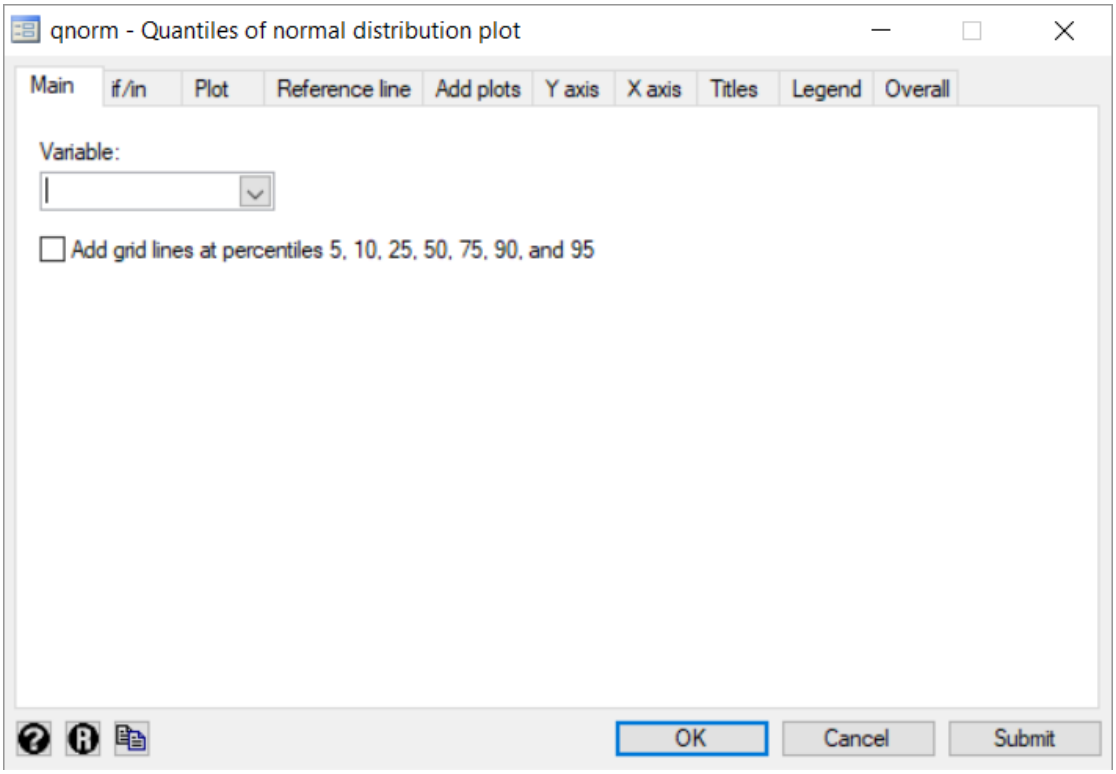
Εικόνα 8.6: Βασική σύνταξη της εντολής histogram

Εναλλακτικά μπορούμε να κατασκευάσουμε το ***Normal Q-Q plot*** ακολουθώντας τα εξής ***βήματα***:

Statistics → Summaries, tables, and tests → Distributional plots and tests → Normal quantile plot

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 8.7*,

ii. Τοποθετούμε στο "***Variable***" την συνεχή μεταβλητή, για την οποία επιθυμούμε να κατασκευάσουμε το Normal quantile plot.



Εικόνα 8.7: Normal Q-Q plot για τον έλεγχο της κανονικής κατανομής της

Στο STATA χρησιμοποιείται η εντολή ***qqplot*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 8.8*:

Quantiles of varname$_1$ *against quantiles of* varname$_2$

`qqplot` *varname*$_1$ *varname*$_2$ [*if*] [*in*] [, *options*$_1$]

Εικόνα 8.8: Βασική σύνταξη της εντολής qqplot

8.2 Έλεγχος ομοσκεδαστικότητας

Ο έλεγχος της ομοσκεδαστικότητας μιας ποσοτικής μεταβλητής έχει ενδιαφέρον μόνο στην περίπτωση που επιθυμούμε να κάνουμε έναν έλεγχο συσχέτισης ανάμεσα σε μία ποσοτική (π.χ. BMI) και μία ποιοτική μεταβλητή (π.χ. sex). Η μηδενική και εναλλακτική υπόθεση αυτού του ελέγχου είναι:

H_0: Η διακύμανση του BMI είναι ίδια και στα 2 φύλα.

H_1: Η διακύμανση του BMI διαφέρει μεταξύ των 2 φύλων.

Ο παραπάνω στατιστικός έλεγχος πραγματοποιείται ακολουθώντας τα εξής ***βήματα***:

Analyze ⟶ Descriptive Statistics → Explore

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 8.1*,

ii. Τοποθετούμε στο "***Dependent List***" όλες τις συνεχείς μεταβλητές, των οποίων την κανονικότητα επιθυμούμε να ελέγξουμε (π.χ. BMI).

iii. Τοποθετούμε στο «***Factor list***» την ποιοτική μεταβλητή (π.χ. sex)

iv. Πατώντας το κουμπί επιλογών "***Plots***" ανοίγει ένα άλλο παράθυρο διαλόγου (*Εικόνα 8.2*), και

v. Όταν τσεκάρουμε τις επιλογές "***Normality Plots with tests***" & "***Histogram***" προκειμένου να πραγματοποιηθεί ο έλεγχος για κανονική κατανομή της ποσοτικής μεταβλητής σε κάθε μία από τις κατηγορίες της ποιοτικής μεταβλητής αυτόματα ενεργοποιείται και η επιλογή «***Spread vs. Level with Levene test***».

vi. Τσεκάρουμε «***Untransformed***»

vii. "***Continue***" & "***Ok***"

viii. Στον *Πίνακα 8.2* παρουσιάζονται τα αποτελέσματα του στατιστικού ελέγχου για τον έλεγχο ισότητας των διακυμάνσεων του BMI μεταξύ ανδρών και γυναικών. Από το sig. =0,000…..1<0,05 διαπιστώνουμε ότι απορρίπτεται η μηδενική υπόθεση, δηλ. απορρίπτεται η προϋπόθεση της ομοσκεδαστικότητας.

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
bmi	Based on Mean	20,101	1	2019	,000
	Based on Median	18,175	1	2019	,000
	Based on Median and with adjusted df	18,175	1	1937,702	,000
	Based on trimmed mean	19,163	1	2019	,000

Πίνακας 8.2: Αποτελέσματα του ελέγχου για ομοσκεδαστικότητα της μεταβλητής BMI μεταξύ ανδρών και γυναικών.

Τον έλεγχο ομοσκεδαστικότητας στο STATA, τον μελετήσαμε στην Ενότητα 6.2.2 (*Εικόνα 6.3*).

9. Γραμμική Παλινδρόμηση

9.1 Εισαγωγή

9.1.1 Απλή γραμμική παλινδρόμηση.

Η γραμμική παλινδρόμηση είναι μία στατιστική τεχνική που μας δίνει την δυνατότητα να διαπιστώσουμε τον τρόπο με τον οποίο μια μεταβλητή που ονομάζεται ***ανεξάρτητη μεταβλητή (X)*** επηρεάζει τις τιμές μιας άλλης μεταβλητής που ονομάζεται ***εξαρτημένη μεταβλητή (Y),*** και αυτή η μορφή γραμμικής παλινδρόμησης ονομάζεται ***«απλή γραμμική παλινδρόμηση»***. Η γραμμική παλινδρόμηση, λοιπόν, μοιάζει αρκετά με την απλή συσχέτιση που αναφέρεται στο *Κεφάλαιο 7*. Όμως, ενώ η απλή συσχέτιση μας πληροφορεί μόνο για το αν υπάρχει γραμμική συσχέτιση ανάμεσα στις 2 μεταβλητές (ένταση και διεύθυνση της σχέσης), η γραμμική παλινδρόμηση απαντά και στο ερώτημα «Πόσο πολύ θα μεταβληθεί η Y όταν θα αλλάξει η X». Με άλλα λόγια, με την γραμμική παλινδρόμηση μπορούμε να εκτιμήσουμε πόσο πολύ θα αλλάξει η Y για συγκεκριμένη μεταβολή της X. Συνεπώς, λοιπόν, η γραμμική παλινδρόμηση μας δίνει την δυνατότητα να προβλέψουμε τις τιμές της εξαρτημένης μεταβλητής όταν η ανεξάρτητη μεταβλητή παίρνει συγκεκριμένες τιμές**.** Για να επιτευχθεί αυτό, το μόνο που απαιτείται είναι να εκφραστεί αυτή η σχέση μεταξύ των X και Y με μία κατάλληλη ***μαθηματική συνάρτηση***.

Σημείωση: Μπορούν να χρησιμοποιηθούν και περισσότερες από μία ανεξάρτητες μεταβλητές για να προβλεφθούν οι τιμές της εξαρτημένης μεταβλητής και σε αυτή την περίπτωση η γραμμική παλινδρόμηση ονομάζεται «***πολλαπλή γραμμική παλινδρόμηση»***.

9.1.1.1 Προσαρμογή της απλής γραμμικής παλινδρόμησης

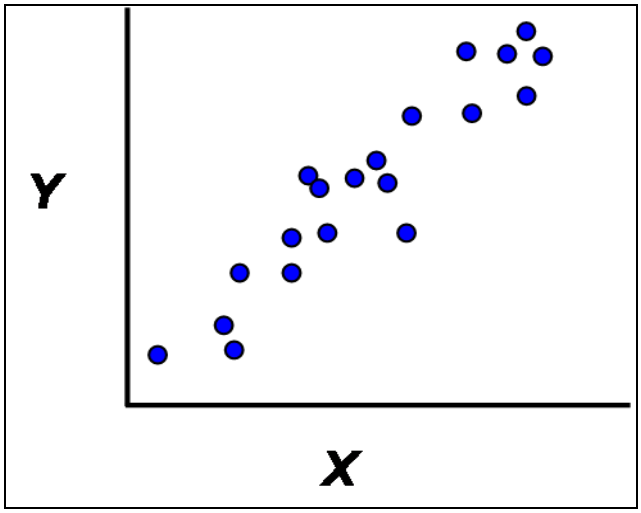
Αυτή η κατάλληλη ***μαθηματική συνάρτηση*** είναι της μορφής:

$$\hat{Y} = \beta_0 + \beta_1 X_1 \qquad (1)$$

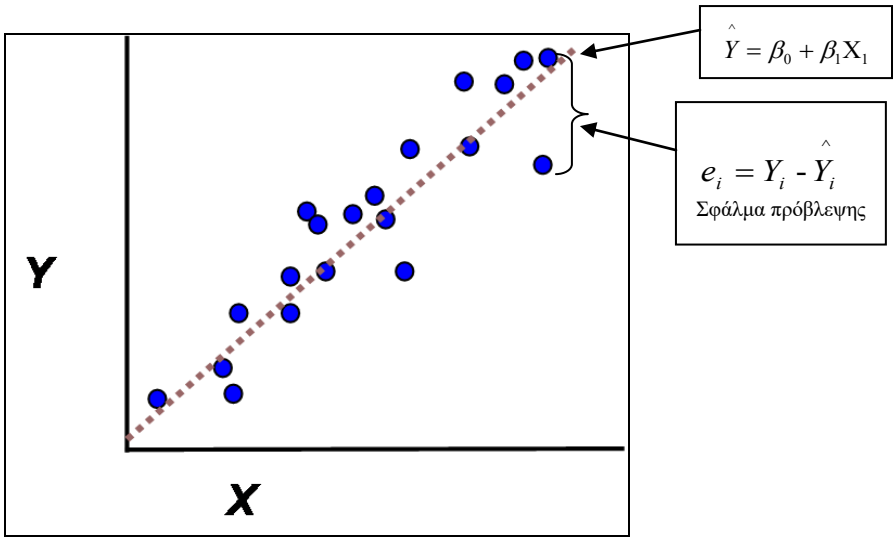
και ***γεωμετρικά***, μεταφράζεται ως μία «***ευθεία γραμμή***» που θα διέρχεται μέσα από τα πραγματικά δεδομένα των μεταβλητών X και Y. Οι συντελεστές β_0, β_1 είναι άγνωστοι και πρέπει να εκτιμηθούν προκειμένου για κάθε τιμή της X να είναι εφικτή ο υπολογισμός της Y. Είναι γεγονός, ότι οι μαθηματικές συναρτήσεις που μπορούν να υπολογιστούν είναι άπειρες (άπειροι οι συνδυασμοί των β_0, β_1 που μπορούν να χρησιμοποιηθούν), όμως, μόνο μία είναι η ***βέλτιστη,*** δηλαδή αυτή που περιγράφει με τον καλύτερο δυνατό τρόπο την πραγματική σχέση ανάμεσα στην X και την Y. Συνεπώς, λοιπόν, οι β_0, β_1 θα πρέπει να εκτιμηθούν με σκοπό η μαθηματική συνάρτηση που θα προκύψει να είναι η βέλτιστη.

Ας υποθέσουμε, λοιπόν, ότι στην *Εικόνα 9.1* παρουσιάζονται τα πραγματικά δεδομένα των μεταβλητών X και Y σε ένα δείγμα. Και ας υποθέσουμε ότι στην *Εικόνα 9.2* παρουσιάζεται η «ευθεία γραμμή» που προέκυψε από την προσαρμογή μίας τυχαία γραμμικής μαθηματικής συνάρτησης. Για να είναι η γραμμική συνάρτηση της *Εικόνας 9.2* η βέλτιστη, θα πρέπει η απόσταση της «ευθείας γραμμής» από όλα τα

σημεία να είναι η ελάχιστη δυνατή. Λαμβάνοντας υπόψη ότι η απόσταση κάθε σημείου της *Εικόνας 9.2* από την ευθεία γραμμή ονομάζεται «***σφάλμα πρόβλεψης***», αντιλαμβανόμαστε ότι η «ευθεία γραμμή» που περιγράφει καλύτερα τα δεδομένα είναι αυτή που ελαχιστοποιεί τα σφάλματα πρόβλεψης. Συνεπώς, οι συντελεστές β_0, β_1 της γραμμικής συνάρτησης (1) θα πρέπει να εκτιμηθούν με γνώμονα να ελαχιστοποιούνται τα «σφάλματα πρόβλεψης». Η μέθοδος που χρησιμοποιείται για την εκτίμηση των συντελεστών β_0, β_1 ονομάζεται «***μέθοδος των ελαχίστων τετραγώνων***».



Εικόνα 9.1: Στικτόγραμμα των μεταβλητών X και Y.



Εικόνα 9.2: Προσαρμογή της ευθείας γραμμής και το σφάλμα πρόβλεψης από την προσαρμογή αυτής.

9.1.1.2 Ερμηνεία συντελεστών της απλής γραμμικής παλινδρόμησης και ο έλεγχος υποθέσεων για τους συντελεστές.

Οι συντελεστές β_0, β_1 ερμηνεύονται ως εξής:

β_0: η αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y, **όταν η τιμή** της μεταβλητής X είναι μηδέν.

β_1: η μεταβολή στην εξαρτημένη μεταβλητή Y, **για κάθε μονάδα αύξηση** της ανεξάρτητης μεταβλητής X1.

Επιπλέον, ενδιαφέρον παρουσιάζει ο στατιστικός έλεγχος για τις πραγματικές τιμές των β_0, β_1 στον πληθυσμό. Πιο συγκεκριμένα, μπορούμε να ελέγξουμε αν οι τιμές των συντελεστών β_0, β_1 διαφέρουν από το μηδέν στον πραγματικό πληθυσμό. Ιδιαίτερο ενδιαφέρον, παρουσιάζει ο έλεγχος υποθέσεων για το β_1, από όπου θα προκύψει και το συμπέρασμα αν η ανεξάρτητη μεταβλητή X συνεισφέρει σημαντικά στην πρόβλεψη των τιμών της μεταβλητής Y. Συνεπώς, λοιπόν, οι μηδενικές και εναλλακτικές υποθέσεις για του συντελεστές β_0, β_1 είναι:

H_0: $\beta_0 = 0$ και H_1: $\beta_0 \neq 0$

H_0: $\beta_1 = 0$ και H_1: $\beta_1 \neq 0$

Για να καταλήξουμε στο συμπέρασμα ότι διαφέρουν τα β_0, β_1 από το μηδέν, θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value) να είναι < α = 0,05 ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).

Αν δεν απορρίψουμε την H_o από τον έλεγχο υποθέσεων για το β_1 θα συμβαίνει ένα από τα 2:
α) η μεταβλητή X θα είναι ελάχιστα ή καθόλου σημαντική για την πρόβλεψη της Y
β) η πραγματική σχέση ανάμεσα στην X και την Y δεν είναι γραμμική.

9.1.2 Πολλαπλή γραμμική παλινδρόμηση

Όπως έχει ήδη αναφερθεί στην *σημείωση* της *παραγράφου 9.1.1,* η πολλαπλή παλινδρόμηση είναι η επέκταση της απλής παλινδρόμησης στην περίπτωση που έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές. Η εισαγωγή περισσότερων ερμηνευτικών μεταβλητών έχει ως σκοπό την ερμηνεία όλο και μεγαλύτερου τμήματος της συνολικής μεταβλητότητας των τιμών της εξαρτημένης μεταβλητής Y, μειώνοντας κατ' αυτό τον τρόπο τις τιμές των σφαλμάτων e_i, άρα και την διακύμανσή τους $\hat{\sigma}^2$.

Με την πολλαπλή γραμμική παλινδρόμηση καθίστανται εφικτά τα εξής:

- Εκτίμηση της επίδρασης της μεταβολής κάποιας ανεξάρτητης (ερμηνευτικής) μεταβλητής στην εξαρτημένη μεταβλητή Y, ελέγχοντας για την ενδεχόμενη επίδραση άλλων μεταβλητών, δηλαδή εκτίμηση της απ' ευθείας επίδρασης μιας μεταβλητής στην τιμή της Y.

- Πιο ακριβής πρόβλεψη της τιμής της εξαρτημένης μεταβλητής για κάποια μελλοντική παρατήρηση.

Στην απλή γραμμική παλινδρόμηση προσαρμόζουμε στο δείγμα των παρατηρήσεων (X_i, Y_i), $i = 1,2,...,n$ την καλύτερη ευθεία. Αν έχουμε δύο ανεξάρτητες μεταβλητές, τότε οι παρατηρήσεις μας θα είναι διατεταγμένες τριάδες $(Y_i, X_{i1}, X_{i2}), i = 1,2,...,n$ και θα αντιπροσωπεύουν σημεία στο χώρο των τριών διαστάσεων. Στα σημεία αυτά θα προσαρμόσουμε το **«επίπεδο ελάχιστων τετραγώνων».** Γενικά, αν έχουμε k ανεξάρτητες μεταβλητές, τότε οι παρατηρήσεις $(Y_i, X_{i1}, X_{i2},..., X_{ik}), i = 1,2,...,n$, θα είναι σημεία του χώρου των k+1 διαστάσεων και στα σημεία αυτά θα προσαρμόσουμε το **«πολυεπίπεδο ελάχιστων τετραγώνων».**

9.1.2.1 Εκτίμηση των παραμέτρων του υποδείγματος της πολλαπλής γραμμικής παλινδρόμησης.

Στην πολλαπλή γραμμική παλινδρόμηση, όπως και στην απλή, θεωρούμε ότι η εξαρτημένη μεταβλητή Y μπορεί να εκφραστεί ως μία γραμμική συνάρτηση των k ανεξάρτητων μεταβλητών $X_1, X_2,..., X_k$.

Το μοντέλο της γραμμικής παλινδρόμησης επιδιώκει τον εντοπισμό της κατάλληλης γραμμικής σχέσης μεταξύ της εξαρτημένης μεταβλητής Y και των k ανεξάρτητων μεταβλητών $X_1, X_2,..., X_k$. Ας υποθέσουμε ότι η γραμμική σχέση που συνδέει τις παραπάνω μεταβλητές, εκφράζεται από την παρακάτω συνάρτηση:

$$\hat{Y}_i \equiv \mu_{Y \setminus X_1, X_2, ..., X_k,} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} \qquad (2)$$

όπου

- Y_i : η τιμή της εξαρτημένης μεταβλητής (για την οποία ενδιαφερόμαστε να διατυπώσουμε προβλέψεις) στην i παρατήρηση
- X_{ij} : η τιμή της j ανεξάρτητης μεταβλητής για την i παρατήρηση, i = 1, 2, . . . , n & j = 1, 2, k
- β_0: σταθερά, που εκφράζει την μέση τιμή της εξαρτημένης μεταβλητής όταν όλες οι ερμηνευτικές μεταβλητές έχουν τιμή μηδέν.
- $\beta_1, \beta_2, \ldots, \beta_k$: οι μερικοί συντελεστές παλινδρόμησης για τις μεταβλητές $X_1, \ldots, X_k$, αντίστοιχα, κάθε ένας από τους οποίους εκφράζει την κατά μέσο όρο μεταβολή της τιμής της εξαρτημένης μεταβλητής όταν η αντίστοιχη μεταβλητή μεταβάλλεται κατά μία μονάδα και οι υπόλοιπες παραμένουν σταθερές.

9.1.2.2 Έλεγχος υποθέσεων στην πολλαπλή γραμμική παλινδρόμηση

Σε ένα μοντέλο πολλαπλής παλινδρόμησης μπορεί να ανακύψουν τα εξής ερωτήματα:

- Κατά πόσο ολόκληρο το μοντέλο συνεισφέρει στατιστικά σημαντικά στην πρόβλεψη της εξαρτημένης μεταβλητής Y ή αλλιώς στην ερμηνεία της μεταβλητότητάς της.
- Κατά πόσο μία συγκεκριμένη τυχαία μεταβλητή παρέχει στατιστικά σημαντική πληροφορία για την πρόβλεψη της Y, δεδομένης της παρουσίας άλλων ερμηνευτικών μεταβλητών στο μοντέλο.
- Κατά πόσο μία ομάδα τυχαίων μεταβλητών παρέχει στατιστικά σημαντική πληροφορία για την πρόβλεψη της Y, δεδομένης της παρουσίας άλλων

ερμηνευτικών μεταβλητών στο μοντέλο.
Η απάντηση σε όλα τα παραπάνω ερωτήματα, δίνεται πραγματοποιώντας τους κατάλληλους στατιστικούς ελέγχους υποθέσεων.

Α. Έλεγχος για ολόκληρο το μοντέλο

Η διατύπωση των κατάλληλων υποθέσεων για την πραγματοποίηση αυτού του ελέγχου είναι η εξής:

H_0: Όλες οι μεταβλητές που συμμετέχουν στο μοντέλο δεν ερμηνεύουν στατιστικά σημαντικό μέρος της μεταβλητότητας των δεδομένων, δηλαδή δεν συμβάλλουν στην πρόβλεψη της εξαρτημένης μεταβλητής, ή αλλιώς

$$\beta_1 = \beta_2 = ... = \beta_k = 0$$

H_1: Έστω και μία μεταβλητή ερμηνεύει στατιστικά σημαντικό μέρος της μεταβλητότητας των δεδομένων, δηλαδή έστω και ένας από τους συντελεστές $\beta_1, \beta_2, ..., \beta_k$ είναι διάφορος του μηδενός.

Β. Έλεγχος για την προσθήκη μιας μόνο μεταβλητής

Έστω ότι σε ένα αρχικό μοντέλο που περιέχει k ερμηνευτικές μεταβλητές προσθέτουμε άλλη μία (X_{k+1}). Αυτό που μας ενδιαφέρει είναι να ελέγξουμε αν αυτή η μεταβλητή προσφέρει στατιστικά σημαντική πληροφορία στην πρόβλεψη της εξαρτημένης μεταβλητής Y, δεδομένης της παρουσίας των άλλων ερμηνευτικών μεταβλητών.

Άρα, ο κατάλληλος έλεγχος υποθέσεων διατυπώνεται ως εξής:

H_0: Η προσθήκη της μεταβλητής X_{k+1} στο μοντέλο, δεδομένης της παρουσίας των υπολοίπων k μεταβλητών, δε βελτιώνει στατιστικά σημαντικά την πρόβλεψη της εξαρτημένης μεταβλητής, δηλαδή δεν αυξάνει στατιστικά σημαντικά την ερμηνευτική ικανότητα του μοντέλου παλινδρόμησης, ή αλλιώς : $\beta_{k+1} = 0$

H_1: Η προσθήκη της μεταβλητής X_{k+1} στο μοντέλο βελτιώνει στατιστικά σημαντικά την πρόβλεψη της εξαρτημένης μεταβλητής, ή αλλιώς : $\beta_{k+1} \neq 0$

Γ. Έλεγχος για την προσθήκη μιας ομάδας μεταβλητών

Έστω ότι σε ένα μοντέλο παλινδρόμησης που περιέχει k ερμηνευτικές μεταβλητές, προσθέτουμε άλλες m μεταβλητές και θέλουμε να ελέγξουμε αν αυτές οι επιπλέον m μεταβλητές βελτιώνουν την ερμηνευτική - προβλεπτική ικανότητα του μοντέλου.

Ο κατάλληλος έλεγχος υποθέσεων διατυπώνεται ως εξής:

H_0: Και οι m μεταβλητές, συνολικά, δεν βελτιώνουν στατιστικά σημαντικά την ερμηνευτική - προβλεπτική ικανότητα του μοντέλου, ή αλλιώς

$$\beta_{k+1} = \beta_{k+2} = ... = \beta_m = 0$$

H_1: Τουλάχιστον μία από τις m μεταβλητές βελτιώνει στατιστικά σημαντικά την προβλεπτική ικανότητα του μοντέλου, ή αλλιώς ένας από τους παραπάνω συντελεστές ($\beta_{k+1}, \beta, ..., \beta_m$) είναι διάφορος του μηδενός.

9.1.3 Προϋποθέσεις ορθής εφαρμογής της γραμμικής παλινδρόμησης

Προκειμένου να εφαρμοστεί ορθά η γραμμική παλινδρόμηση (απλή η πολλαπλή) θα πρέπει να ισχύουν οι εξής ***προϋποθέσεις***:

i. Γραμμικότητα: Η συσχέτιση ανάμεσα σε κάθε μία από τις ανεξάρτητες μεταβλητές και την εξαρτημένη μεταβλητή θα πρέπει να είναι γραμμική.
ii. Κανονικότητα: Η κατανομή των σφαλμάτων να είναι κανονική για κάθε τιμή των ανεξάρτητων μεταβλητών.
iii. Ομοσκεδαστικότητα: Η τυπική απόκλιση των σφαλμάτων να είναι ίση για όλες τις τιμές της κάθε ανεξάρτητης μεταβλητής.
iv. Ανεξαρτησία: Οι παρατηρήσεις θα πρέπει να είναι ανεξάρτητες, δηλ. να προέρχονται από διαφορετικά άτομα.
v. Πολυσυγγραμικότητα: Οι ανεξάρτητες μεταβλητές δεν πρέπει να συσχετίζονται ισχυρά μεταξύ τους. Η συσχέτιση μεταξύ των μεταβλητών που χρησιμοποιούνται σε πολλαπλή γραμμική παλινδρόμηση ως ανεξάρτητες θα πρέπει να είναι η μικρότερη δυνατή, δεδομένου ότι ισχυρή συσχέτιση μεταξύ των ανεξάρτητων μεταβλητών δημιουργεί το πρόβλημα της πολυσυγγραμμικότητας το οποίο με τη σειρά του έχει σαν αποτέλεσμα τον υπολογισμό εκτιμητών με αυξημένα τυπικά σφάλματα. Συνεπώς, λοιπόν, είναι απαραίτητο να ελέγχουμε το βαθμό συσχέτισης των ανεξάρτητων μεταβλητών που χρησιμοποιούνται σε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης. Τα κατάλληλα στατιστικά γι' αυτό τον έλεγχο είναι το ***Tolerance*** και το ***VIF*.**
    - ***Tolerance***: όσο μεγαλύτερο είναι τόσο μικρότερη είναι η συσχέτισή του με όλες τις υπόλοιπες ανεξάρτητες μεταβλητές του μοντέλου. Το εύρος τιμών του ***Tolerance*** διακυμαίνεται μεταξύ 0 και 1. Τιμές πολύ κοντά στην 1 υποδηλώνουν την έλλειψη συσχέτισης ανάμεσα στις ανεξάρτητες μεταβλητές
    - ***VIF***: όσο μεγαλύτερο είναι τόσο μεγαλύτερη είναι η συσχέτιση του παράγοντα με τις υπόλοιπες ανεξάρτητες μεταβλητές του μοντέλου. Το εύρος τιμών του ***VIF*** είναι από 1 έως +άπειρο. Τιμές πάνω από 2 ή 3 υποδηλώνουν ισχυρή συσχέτιση.

Οι προϋποθέσεις (i) έως (iv), μπορούν να ελεγχθούν με τους εξής τρόπους:

i. Παριστάνοντας γραφικά τα κατάλοιπα έναντι των εκτιμούμενων τιμών που έχουν προκύψει βάση του μοντέλου, για να ελέγξουμε αν υπάρχει κάποια καμπύλη στο γράφημα και για να δούμε αν τα κατάλοιπα βρίσκονται γύρο από το μηδέν και έχουν ίση διακύμανση κατά μήκος όλων των εκτιμούμενων τιμών.
ii. Φτιάχνοντας ιστόγραμμα ή γράφημα κανονικής πιθανότητας των καταλοίπων. Το ιστόγραμμα, αν ισχύει η προϋπόθεση της κανονικότητας θα πρέπει να είναι συμμετρικό, ενώ στο γράφημα κανονικής πιθανότητας, τα κατάλοιπα θα πρέπει να βρίσκονται πάνω σε μία ευθεία διαγώνια γραμμή.

9.1.4 Ερμηνευτικότητα του μοντέλου

Τόσο στην απλή όσο και στην πολλαπλή γραμμική παλινδρόμηση, ο συντελεστής προσδιορισμού (R^2) εκφράζει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητή Y που ερμηνεύεται από το μοντέλο ή αλλιώς από την μία ανεξάρτητη μεταβλητή (στην απλή γραμμική παλινδρόμηση) ή όλες τις ανεξάρτητες μεταβλητές (στην πολλαπλή γραμμική παλινδρόμηση). Το εύρος τιμών του R^2 είναι:

$$0 \le R^2 \le 1$$

και όσο η τιμή του R^2 πλησιάζει στη μονάδα, τόσο μεγαλύτερο είναι το ποσοστό της μεταβλητότητας της Y που ερμηνεύεται και συνεπώς τόσο πιο ακριβείς θα είναι οι προβλέψεις για τις τιμές της Y χρησιμοποιώντας τη συγκεκριμένη συνάρτηση.

9.1.5 Επιλογή βέλτιστου μοντέλου

Σε μερικές περιπτώσεις, δεν υπάρχει ανεξάρτητη μεταβλητή στόχος. Πιο συγκεκριμένα, δεν υπάρχει μία συγκεκριμένη ανεξάρτητη μεταβλητή, όπου στόχος μας είναι να μελετήσουμε την επίδραση της πάνω στην εξαρτημένη μας μεταβλητή μετά από έλεγχο για πιθανούς συγχυτικούς παράγοντες (πολλαπλή γραμμική παλινδρόμηση). Σε αυτές τις περιπτώσεις, στόχος μας είναι να διερευνήσουμε ποιες από τις πιθανές ανεξάρτητες μεταβλητές, ασκούν στατιστικά σημαντική επίδραση στην εξαρτημένη μας μεταβλητή. Ουσιαστικά, λοιπόν, στόχος μας είναι να επιλέξουμε το βέλτιστο μοντέλο μεταξύ όλων των πιθανών που μπορούν να προκύψουν αν χρησιμοποιήσουμε μία σειρά από συγκεκριμένες ανεξάρτητες μεταβλητές. Στη διαδικασία επιλογής μιας εξίσωσης συνήθως εμπλέκονται δύο αντιτιθέμενα κριτήρια:

1. Για την κατασκευή μιας εξίσωσης χρήσιμης για σκοπούς πρόβλεψης θα πρέπει το μοντέλο μας να περιλαμβάνει όσο το δυνατόν περισσότερες ανεξάρτητες (ερμηνευτικές) μεταβλητές έτσι ώστε οι προσαρμοσμένες τιμές (εκτιμήσεις) να είναι αξιόπιστες.
2. Επειδή η συγκέντρωση πληροφοριών για ένα μεγάλο αριθμό ανεξάρτητων μεταβλητών και η επακόλουθη επεξεργασία τους κοστίζουν, θα θέλαμε η εξίσωση να περιλαμβάνει όσο το δυνατό λιγότερες μεταβλητές.

Ο συμβιβασμός μεταξύ των δύο αυτών ακραίων περιπτώσεων είναι αυτό που συνήθως ονομάζεται επιλογή της καλύτερης εξίσωσης παλινδρόμησης. Για να πετύχουμε την καλύτερη εξίσωση δεν υπάρχει μία και μοναδική διαδικασία. Προκαλεί σύγχυση το γεγονός ότι όλες αυτές οι διαδικασίες όταν εφαρμόζονται στο ίδιο πρόβλημα δεν οδηγούν αναγκαστικά στην ίδια λύση, μολονότι για πολλά προβλήματα επιτυγχάνουν την ίδια απάντηση.

Οι μέθοδοι που χρησιμοποιούνται είναι οι εξής:

1. Ανάδρομη διαδικασία (backward procedure)
2. Πρόδρομη διαδικασία (forward procedure)
3. Βηματική διαδικασία (stepwise procedure)

9.1.5.1 Η μέθοδος της διαδοχικής απαλοιφής (ανάδρομη διαδικασία ή Backward-elimination procedure)

Τα βήματα για την επιλογή της καλύτερης παλινδρόμησης με αυτή τη μέθοδο είναι τα εξής:

1. Προ-επιλέγεται ένα μέγιστο παρατηρούμενο επίπεδο σημαντικότητας (p-value) για την απομάκρυνση των μεταβλητών
2. Προσαρμόζεται το πλήρες μοντέλο, δηλαδή το μοντέλο που περιλαμβάνει όλες τις δυνατές ερμηνευτικές μεταβλητές
3. Υπολογίζονται για όλες τις μεταβλητές τα partial F test (τύπου III), δηλαδή το F test για κάθε μία μεταβλητή θεωρώντας ότι αυτή εισάγεται τελευταία ή αλλιώς το F test για κάθε μεταβλητή δεδομένης της παρουσίας όλων των υπόλοιπων μεταβλητών στο μοντέλο

4. Αν το μεγαλύτερο p – value (που αντιστοιχεί στη λιγότερο στατιστικά σημαντική μεταβλητή) είναι μεγαλύτερο από το p-value που ορίστηκε αρχικά, τότε η συγκεκριμένη μεταβλητή απομακρύνεται
5. Τα βήματα 2, 3, 4 επαναλαμβάνονται μέχρις ότου καμία από τις μεταβλητές του μοντέλου να μην έχει p – value μεγαλύτερο από αυτό που εμείς ορίσαμε, αρχικά, ως κριτήριο απόρριψης των μεταβλητών

9.1.5.2 Η μέθοδος της διαδοχικής εισαγωγής πρόσθετων ερμηνευτικών μεταβλητών (Forward-selection procedure)

Τα βήματα για την επιλογή της καλύτερης εξίσωσης παλινδρόμησης με αυτή τη μέθοδο είναι τα εξής:

1. Προ-επιλέγουμε ένα μέγιστο παρατηρούμενο επίπεδο σημαντικότητας (p-value) για την εισαγωγή των μεταβλητών
2. Πραγματοποιούμε απλή γραμμική παλινδρόμηση για κάθε μία από τις ανεξάρτητες μεταβλητές που έχουμε στη διαθεσή μας
3. Υπολογίζονται οι τιμές του στατιστικού F (F-test) ή του στατιστικού T (t-test) για κάθε μία από τις απλές παλινδρομήσεις (όπως αναφέρεται και παραπάνω και τα δύο στατιστικά οδηγούν στο ίδιο συμπέρασμα όσον αφορά στη στατιστική σημαντικότητα των ερμηνευτικών μεταβλητών)
4. Αν το μικρότερο p-value από το F-test ή το t-test (αντιστοιχεί στην πιο στατιστικά σημαντική μεταβλητή), είναι μεγαλύτερο από το αρχικό p-value που ορίστηκε ως όριο για την εισαγωγή των μεταβλητών, δεν εισάγεται καμία μεταβλητή. Αντίθετα, αν αυτό είναι μικρότερο από το p-value της εισαγωγής, τότε εισάγεται στο μοντέλο αυτή η μεταβλητή
5. Για κάθε μία από τις υπόλοιπες μεταβλητές, υπολογίζουμε το partial F test (τύπου III), δηλαδή το F test για κάθε μία από αυτές δεδομένης της παρουσίας της πρώτης μεταβλητής στο μοντέλο
6. Η μεταβλητή με το μικρότερο p-value, σύμφωνα με τα F-test που υπολογίστηκαν παραπάνω, εισάγεται στο μοντέλο με την προϋπόθεση ότι αυτό το p-value είναι μικρότερο από αυτό που ορίστηκε αρχικά
7. Τα βήματα 5 και 6 επαναλαμβάνονται μέχρις ότου καμία επιπλέον μεταβλητή να μην μπορεί να εισαχθεί στο μοντέλο.

9.1.5.3 Η μέθοδος επιλογής μεταβλητών κατά βήματα (Stepwise regression procedure)

Στη διαδικασία της διαδοχικής επιλογής των μεταβλητών για την επιλογή της καλύτερης εξίσωσης παλινδρόμησης, ελλοχεύει ο κίνδυνος να πάψει να είναι στατιστικά σημαντική κάποια μεταβλητή που έχει ήδη εισαχθεί στο μοντέλο, μετά την εισαγωγή κάποιας άλλης, εξαιτίας της σχέσης που πιθανότητα συνδέει αυτές τις δύο ανεξάρτητες μεταβλητές. Γι' αυτό το λόγο υπάρχει ένας άλλος τρόπος επιλογής του βέλτιστου μοντέλου, που είναι τροποποίηση της προηγούμενης διαδικασίας (forward selection procedure). Χαρακτηριστικό αυτής της μεθόδου είναι η επανεξέταση των ήδη υπαρχόντων μεταβλητών στο μοντέλο μετά την εισαγωγή μιας νέας μεταβλητής.

Τα βήματα αυτής της διαδικασίας είναι τα εξής:

1. Προ-επιλέγονται ένα p-value για την απομάκρυνση των μεταβλητών που ήδη υπάρχουν στο μοντέλο και ένα για την εισαγωγή νέων μεταβλητών. Αυτά τα δύο p-value δεν πρέπει να είναι ίσα, αλλά το p-value της απομάκρυνσης να

είναι μεγαλύτερο από αυτό της εισαγωγής, εξασφαλίζοντας κατ' αυτό τον τρόπο ότι κάποια μεταβλητή που θα απομακρυνθεί από το μοντέλο δεν θα ξαναμπεί.

2. Εισάγεται η πρώτη μεταβλητή στο μοντέλο ακολουθώντας τα στάδια 2, 3 και 4 της διαδικασίας Β.
3. Για κάθε μία από τις υπόλοιπες μεταβλητές που βρίσκονται εκτός μοντέλου, υπολογίζουμε το partial F test (τύπου ΙΙΙ), δηλαδή το F test για κάθε μία από αυτές δεδομένης της παρουσίας της πρώτης μεταβλητής στο μοντέλο. Αν το μικρότερο p-value, σύμφωνα με τα F-test που υπολογίστηκαν παραπάνω είναι μικρότερο από αυτό που ορίστηκε αρχικά για την εισαγωγή των μεταβλητών, εισάγεται στο μοντέλο
4. Υπολογίζεται το partial F test (τύπου ΙΙΙ) της πρώτης μεταβλητής που είχε εισαχθεί στο μοντέλο, δηλαδή το F test δεδομένης της παρουσίας της δεύτερης μεταβλητής του μοντέλου. Αν το p-value είναι μεγαλύτερο από αυτό που ορίστηκε αρχικά για την απομάκρυνση των μεταβλητών, τότε η μεταβλητή απομακρύνεται από το μοντέλο
5. Επαναλαμβάνονται τα βήματα 3 και 4 μέχρις ότου καμία μεταβλητή να μην μπορεί να προστεθεί ή να απομακρυνθεί από το μοντέλο. Πρέπει να σημειωθεί ότι όταν ελέγχεται το ενδεχόμενο απομάκρυνσης κάποιας μεταβλητής από αυτές που ήδη υπάρχουν στο μοντέλο, συγκρίνουμε το μεγαλύτερο p-value (που αντιστοιχεί στη λιγότερο στατιστικά σημαντική μεταβλητή απ' αυτές που ήδη υπάρχουν στο μοντέλο) με αυτό που έχει οριστεί ως όριο για την απομάκρυνση των μεταβλητών.

9.1.6 Έλεγχος συνεργιών μέσω της πολλαπλής παλινδρόμησης

Συνεργία μεταξύ δύο ή περισσοτέρων παραγόντων ή χαρακτηριστικών υπάρχει όταν η σχέση του ενός παράγοντα με κάποιο νόσημα ή χαρακτηριστικό που διερευνάται επηρεάζεται από την παρουσία του άλλου παράγοντα, έτσι ώστε η συμπαρουσία των δύο ή περισσοτέρων παραγόντων να αυξήσει την ερμηνευτική τους ικανότητα στο υπόδειγμα. Για να ελεγχθεί αν για παράδειγμα δύο παράγοντες συνεργούν τότε εκτιμάται το υπόδειγμα:

- $Y = \beta_o + \beta_1 * X_1 + \beta_2 * X_2 + \gamma * X_1*X_2$
  - Αν ο όρος X_1*X_2 είναι στατιστικά σημαντικός, τότε λέμε ότι οι παράγοντες ή τα χαρακτηριστικά X_1, X_2 συνεργούν και το παραπάνω υπόδειγμα είναι το κατάλληλο.
  - Αν ο όρος X_1*X_2 δεν είναι στατιστικά σημαντικός, τότε λέμε ότι οι παράγοντες ή τα χαρακτηριστικά X_1, X_2 δεν συνεργούν και αφαιρείται ο όρος από το υπόδειγμα.

9.1.7 Έλεγχος συγχυτικών επιδράσεων μέσω της πολλαπλής παλινδρόμησης

«Σύγχυση» είναι το συστηματικό σφάλμα που προκύπτει από ένα συγχυτικό παράγοντα. «Συγχυτικός» παράγοντας είναι αυτός που συνήθως, συσχετίζεται με τον υπό μελέτη παράγοντα ή χαρακτηριστικό και ταυτόχρονα επηρεάζει τη σχέση του παράγοντα ή χαρακτηριστικού με την εξαρτημένη μεταβλητή. Για παράδειγμα στη σχέση της κατανάλωσης καφέ (ανεξάρτητος παράγοντας ή ερμηνευτική μεταβλητή) με ένα δείκτη φλεγμονής (εξαρτημένη μεταβλητή) συγχυτικό ρόλο μπορεί να παίζουν

οι καπνιστικές συνήθειες. Μέσω των στατιστικών μοντέλων ο έλεγχος για συγχυτικές επιδράσεις μπορεί να γίνει ως εξής:

- Αρχικά εκτιμάται το γραμμικό υπόδειγμα με τον κύριο παράγοντα (Χ)
  - $Y = \beta_o + \beta_1 * X$
- Στη συνέχεια εισάγεται και ο πιθανός συγχυτικός παράγοντας (Σ)
  - $Y = \beta_o + \beta'_1 * X + \beta_2 * \Sigma$
- Είναι αναμενόμενο ο συντελεστής β_1 να αλλάξει σε β'1. Αν όμως ο συντελεστής β'_1 διαφέρει σημαντικά από τον β_1 (ένας έλεγχος είναι το Student's t-test) τότε ο παράγοντας Σ που εισήχθη στο υπόδειγμα έχει συγχυτικό ρόλο στη σχέση του παράγοντα ή χαρακτηριστικού Χ με τον παράγοντα ή χαρακτηριστικό Υ.

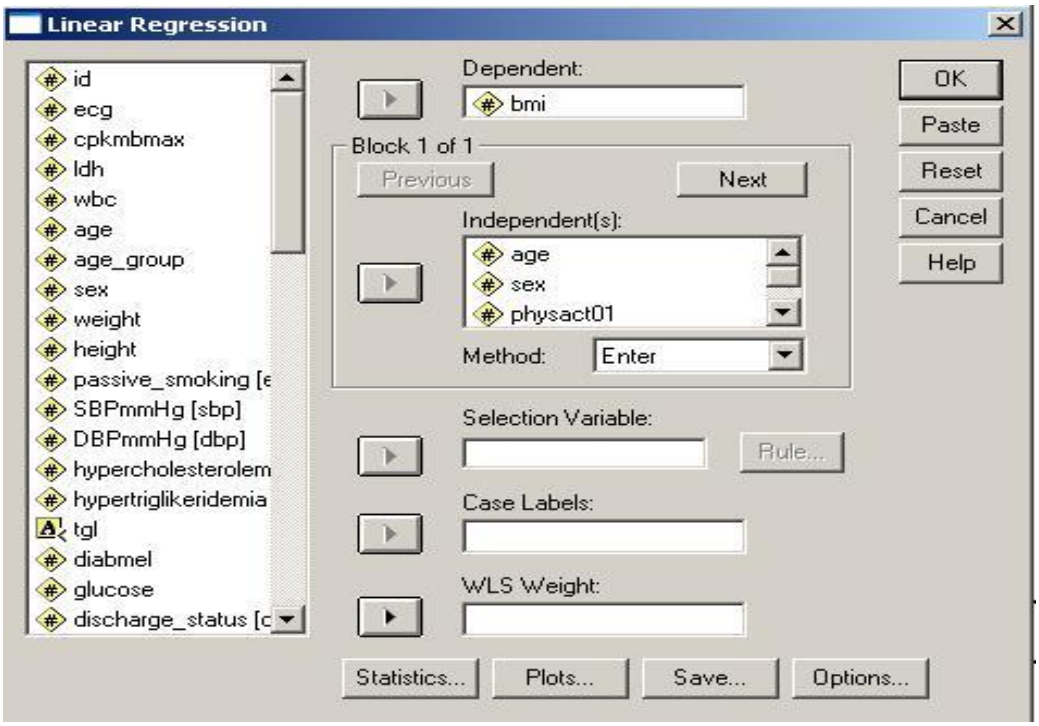
Ανεξάρτητα αν ο συγχυτικός παράγοντας Σ έχει ή όχι στατιστικά σημαντική ερμηνευτική ικανότητα στη μεταβλητότητα του παράγοντα Υ, πρέπει να παραμείνει στο υπόδειγμα. Συνήθως στις ιατρο-βιολογικές μελέτες η ηλικία, το φύλο, καθώς και άλλα χαρακτηριστικά των ατόμων μπορεί να έχουν συγχυτικό ρόλο στη διερεύνηση κάποιων σχέσεων. Τα χαρακτηριστικά αυτά πρέπει να υπάρχουν στο προτεινόμενο υπόδειγμα.

9.2 Πολλαπλή γραμμική παλινδρόμηση με το SPSS.

Ας υποθέσουμε ότι στόχος μας είναι να ελέγξουμε κατά πόσο το φύλο, η ηλικία και η σωματική δραστηριότητα (males, age, physact01, αντίστοιχα) επηρεάζουν τις τιμές του δείκτη μάζας σώματος (ΔΜΣ) και συνεπώς να ελέγξουμε αν αυτοί οι 3 παράγοντες μπορούν να συμβάλουν στην πρόβλεψη των τιμών του ΔΜΣ. Η κατάλληλη στατιστική τεχνική για να διερευνήσουμε αυτό το ερώτημα είναι η ***πολλαπλή γραμμική παλινδρόμηση,*** η οποία μπορεί να πραγματοποιηθεί ακολουθώντας τα παρακάτω ***βήματα***:

Analyse → Regression → Linear

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 9.3*



Εικόνα 9.3: Πραγματοποίηση της πολλαπλής γραμμικής παλινδρόμησης

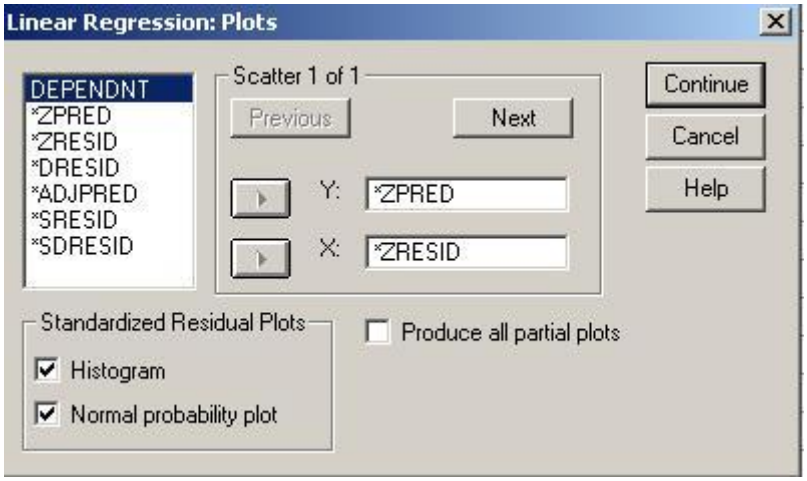
ii. Ορίζουμε ως ***Dependent*** τη συνεχή μεταβλητή /έκβαση (π.χ. Δείκτης Μάζας Σώματος), τις τιμές της οποίας επιθυμούμε να εκτιμήσουμε χρησιμοποιώντας κάποιο άλλο χαρακτηριστικό,

iii. Ως ***Independent(s)*** ορίζουμε τις ανεξάρτητες μεταβλητές /παράγοντες (π.χ. ηλικία, φύλο, σωματική άσκηση)

iv. Πατάμε το κουμπί επιλογών «***Plots***» όπου εμφανίζεται ένα νέο πλαίσιο διαλόγου (*Εικόνα 9.4*) προκειμένου να δώσουμε την εντολή για την πραγματοποίηση του ελέγχου των προϋποθέσεων για την ορθή εφαρμογή της γραμμικής παλινδρόμησης.

v. Επιλέγουμε τα «***ZPRED***» & «***ZRESID***» και τα τοποθετούμε στους ***άξονες x και y***, αντίστοιχα.

vi. Τσεκάρουμε το "***Histogram"*** & το "***Normal Probability Plots"***



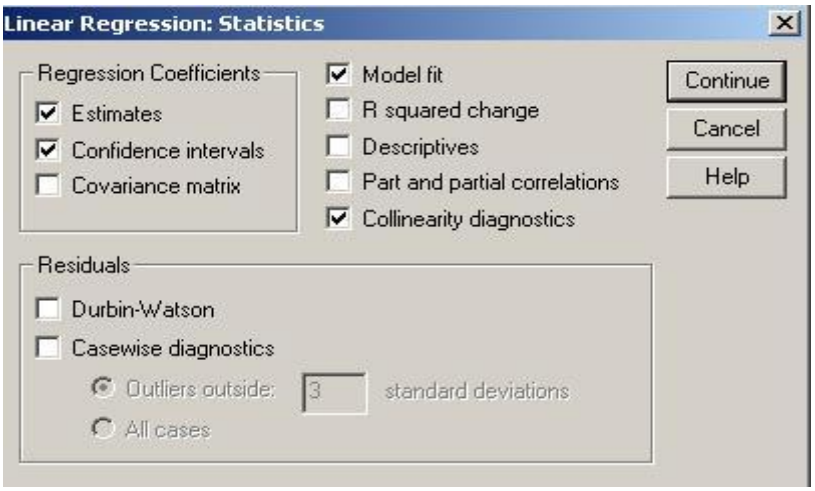
Εικόνα 9.4: Πραγματοποίηση του ελέγχου των απαραίτητων προϋποθέσεων για την ορθή εφαρμογή της γραμμικής παλινδρόμησης.

vii. Πατάμε το κουμπί επιλογών "***Statistics***", προκειμένου να δώσουμε την εντολή για τον υπολογισμό των κατάλληλων στατιστικών για τον έλεγχο ύπαρξης πολυσυγγραμικότητας και ανοίγει ένα νέο πλαίσιο διαλόγου (*Εικόνα 9.5*).

viii. Επιλέγουμε το "***Collinearity diagnostics***" (επίσης, επιλέγουμε και το «Confidence Interval» προκειμένου στο output του SPSS να εμφανιστούν και τα διαστήματα εμπιστοσύνης για τους β-συντελεστές (***unstandardized coefficient B***))

ix. ***"Continue"***

x. ***"Ok"***



Εικόνα 9.5: Υπολογισμός των στατιστικών για τον έλεγχο ύπαρξης πολυσυγγραμικότητας

xi. Οι *Πίνακες 9.1, 9.2 & 9.3* παρουσιάζουν τα αποτελέσματα της ανάλυσης, και συγκεκριμένα τις εκτιμήσεις των ***unstandardized & standardized coefficients,*** τα διαγνωστικά χαρακτηριστικά του συνολικού μοντέλου και τα στατιστικά ***Tolerance*** & ***VIF*** για τον έλεγχο ύπαρξης πολυσυγγραμικότητας.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	31,936	,532		60,008	,000		
	age	-,056	,007	-,184	-7,962	,000	,950	1,053
	sex	-,602	,211	-,066	-2,856	,004	,944	1,060
	physact01	-,630	,182	-,079	-3,468	,001	,975	1,026

a. Dependent Variable: bmi

Πίνακας 9.1: Αποτελέσματα της πολλαπλής γραμμικής παλινδρόμησης; Συντελεστές της γραμμικής παλινδρόμησης

Πιο συγκεκριμένα, από τον *Πίνακα 9.1* παρατηρούμε:

- και οι τρεις ανεξάρτητες μεταβλητές συσχετίζονται στατιστικά σημαντικά με τον ΔΜΣ. Πιο συγκεκριμένα βλέπουμε ότι η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης για την επίδραση της ηλικίας στις τιμές του ΔΜΣ (δηλαδή, ότι δεν υπάρχει συσχέτιση ανάμεσα στην ηλικία και το ΔΜΣ) είναι sig.=0,000…1, δηλαδή πολύ μικρότερη από το επίπεδο σημαντικότητας α=0,05. Συνεπώς, η ηλικία επηρεάζει τις τιμές του ΔΜΣ. Παρομοίως, βλέπουμε ότι η πιθανότητες εσφαλμένης απόρριψης των μηδενικών υποθέσεων για την επίδραση του φύλο και της σωματικής άσκησης στις τιμές του ΔΜΣ (δηλαδή, ότι δεν υπάρχει συσχέτιση ανάμεσα στο φύλο και το ΔΜΣ και ανάμεσα στη σωματική άσκηση και τον ΔΜΣ) είναι sig.=0,004, και sig.=0,001, αντίστοιχα.
- και οι τρεις ανεξάρτητες μεταβλητές συσχετίζονται αντίστροφα με τις τιμές του ΔΜΣ. Αυτό προκύπτει από τους συντελεστές β ***"unstandardized coefficient B"*** της παλινδρόμησης (B= - 0,056 για την ηλικία, B= - 0,602 για το φύλο και B=-0,630 για την σωματική άσκηση). Πιο συγκεκριμένα, αύξηση της ηλικίας κατά ένα χρόνο οδηγεί σε μείωση του ΔΜΣ κατά 0,056 kg/m^2 δεδομένου ότι οι υπόλοιπες 2 ανεξάρτητες μεταβλητές παραμένουν σταθερές. Αναφορικά με το φύλο, διαπιστώνουμε ότι οι άνδρες αναμένεται να έχουν 0,602 kg/m^2 χαμηλότερο ΔΜΣ σε σχέση με τις γυναίκες δεδομένου ότι η ηλικία και η σωματική άσκηση είναι ίδια.
- ποιος από τους 3 ανεξάρτητους παράγοντες επηρεάζει περισσότερο τις τιμές του ΔΜΣ. Αυτό προκύπτει από τους ***standardized coefficient Beta***. Πιο συγκεκριμένα, ο παράγοντας με το μεγαλύτερο Beta κατά απόλυτη τιμή είναι αυτός που επηρεάζει περισσότερο τις τιμές του ΔΜΣ. Συνεπώς, στο παράδειγμά μας, ο πιο σημαντικός ανεξάρτητος παράγοντας για την πρόβλεψη του ΔΜΣ είναι η ηλικία (Beta= -0,184), η σωματική άσκηση (Beta= -0,079) και το φύλο (Beta= -0,066). Επιπλέον, το Beta μιας ανεξάρτητης μεταβλητής εκφράζει την μεταβολή της εξαρτημένης μεταβλητής όταν η ανεξάρτητη μεταβλητή αυξηθεί κατά μία τυπική απόκλιση και οι τιμές των υπόλοιπων ανεξάρτητων μεταβλητών παραμείνουν σταθερές.
- Αν υπάρχει πρόβλημα πολυσυγγραμικότητας ή όχι μεταξύ των ανεξάρτητων μεταβλητών του μοντέλου. Αυτό προκύπτει από τα στατιστικά μέτρα ***Tolerance*** & ***VIF***. Πιο συγκεκριμένα, διαπιστώνουμε ότι οι ανεξάρτητες μεταβλητές είναι και ασυσχέτιστες δεδομένου ότι και το Tolerance & το VIF όλων των ανεξάρτητων μεταβλητών είναι κοντά στη μονάδα.

Επίσης, από τον πίνακα ***«Model Summary»*** (*Πίνακας 9.2*) διαπιστώνουμε ότι και οι τρεις ανεξάρτητες μεταβλητές από κοινού ερμηνεύουν το 3,6% της μεταβλητότητας του ΔΜΣ (Adjusted ***R Square*** = 0,036).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,194[a]	,038	,036	3,82328

a. Predictors: (Constant), physact01, age, sex

Πίνακας 9.2: Ερμηνευτικότητα του μοντέλου

Τέλος, από τον *Πίνακα 9.3* διαπιστώνουμε ότι το μοντέλο που περιλαμβάνει αυτές τις 3 ανεξάρτητες μεταβλητές είναι στατιστικά σημαντικά καλύτερο όσον αφορά στην πρόβλεψη των τιμών της ΣΑΠ σε σχέση με το μοντέλο που δεν περιλαμβάνει καμία ανεξάρτητη μεταβλητή, αφού η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης (ότι δηλ. το μοντέλο μας δεν είναι καλύτερο από το μοντέλο χωρίς ανεξάρτητες μεταβλητές) είναι sig.=0,00…1< α = 0,05 και συνεπώς απορρίπτουμε την μηδενική υπόθεση.

ANOVA[b]

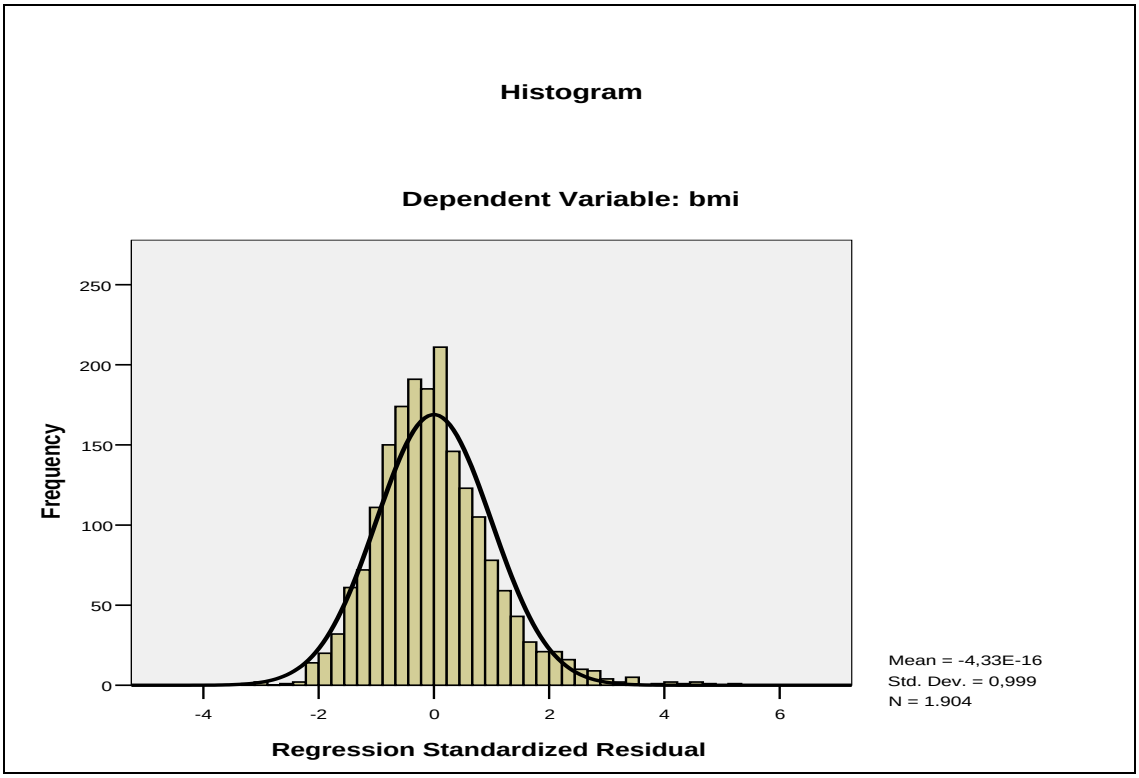
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1084,481	3	361,494	24,730	,000[a]
	Residual	27773,227	1900	14,617		
	Total	28857,708	1903			

a. Predictors: (Constant), physact01, age, sex

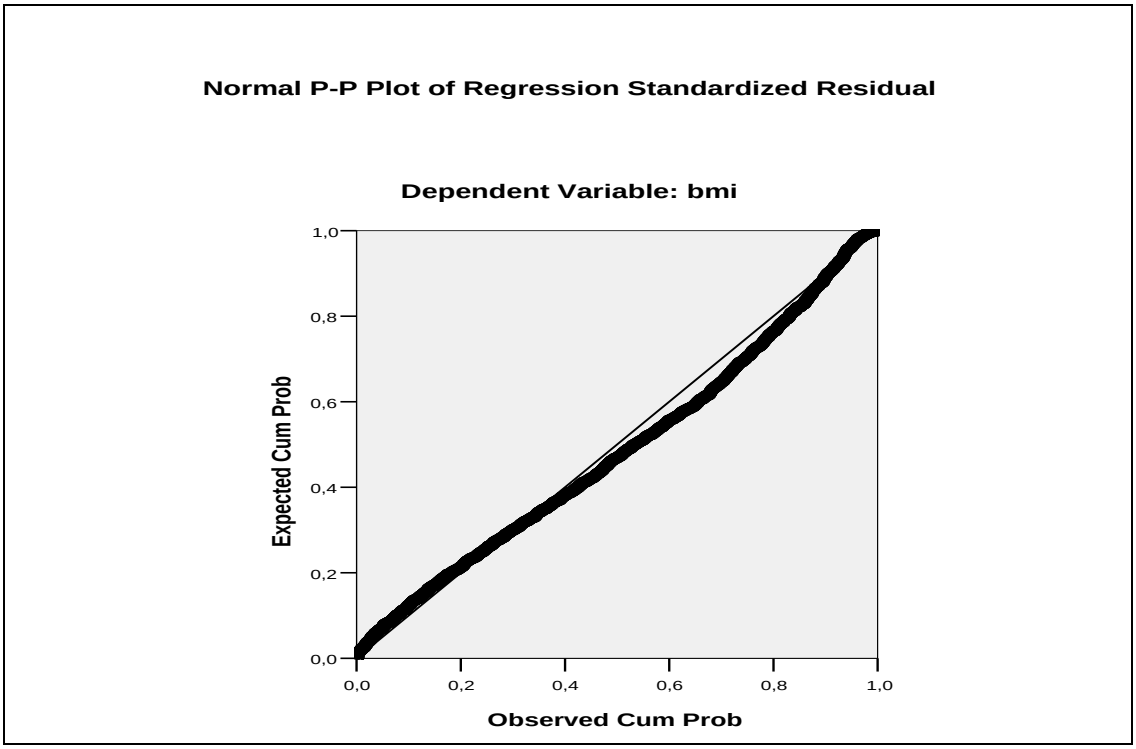
b. Dependent Variable: bmi

Πίνακας 9.3: Έλεγχος για το αν το μοντέλο διαφέρει στατιστικά σημαντικά από ένα μοντέλο που να μην περιέχει καμία ανεξάρτητη μεταβλητή.

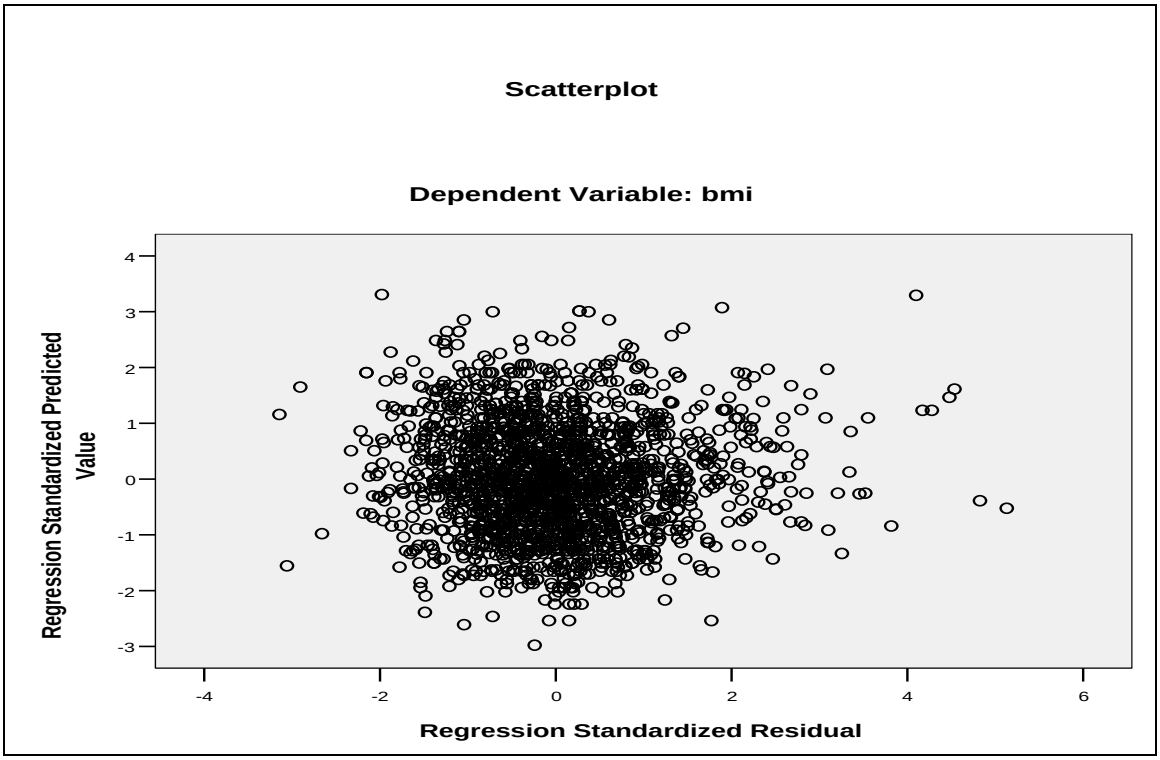
xii. Στις *Εικόνες 9.6, 9.7, & 9.8* παρουσιάζονται τα γραφήματα όσον αφορά στον έλεγχο των προϋποθέσεων για την πολλαπλή γραμμική παλινδρόμηση που αναφέρεται στην *παράγραφο 9.1*. Από αυτές τις *Εικόνες* διαπιστώνουμε ότι και οι τρεις προϋποθέσεις ισχύουν (*βλ. παράγραφο 9.1*).



Εικόνα 9.6: Ιστόγραμμα για τον έλεγχο της κανονικότητας των σφαλμάτων του μοντέλου πολλαπλής γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή τον ΔΜΣ και με ανεξάρτητες μεταβλητές την ηλικία, το φύλο και τη σωματική δραστηριότητα.



Εικόνα 9.7: Γράφημα για τον έλεγχο της κανονικότητας των σφαλμάτων του μοντέλου πολλαπλής γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή τον ΔΜΣ και με ανεξάρτητες μεταβλητές την ηλικία, το φύλο και τη σωματική δραστηριότητα.



Εικόνα 9.8: Γράφημα για τον έλεγχο της ομοσκεδαστικότητας των σφαλμάτων και της γραμμικότητας του μοντέλου πολλαπλής γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή τον ΔΜΣ και με ανεξάρτητες μεταβλητές την ηλικία, το φύλο και τη σωματική δραστηριότητα.

9.3 Πολλαπλή γραμμική παλινδρόμηση με το STATA.

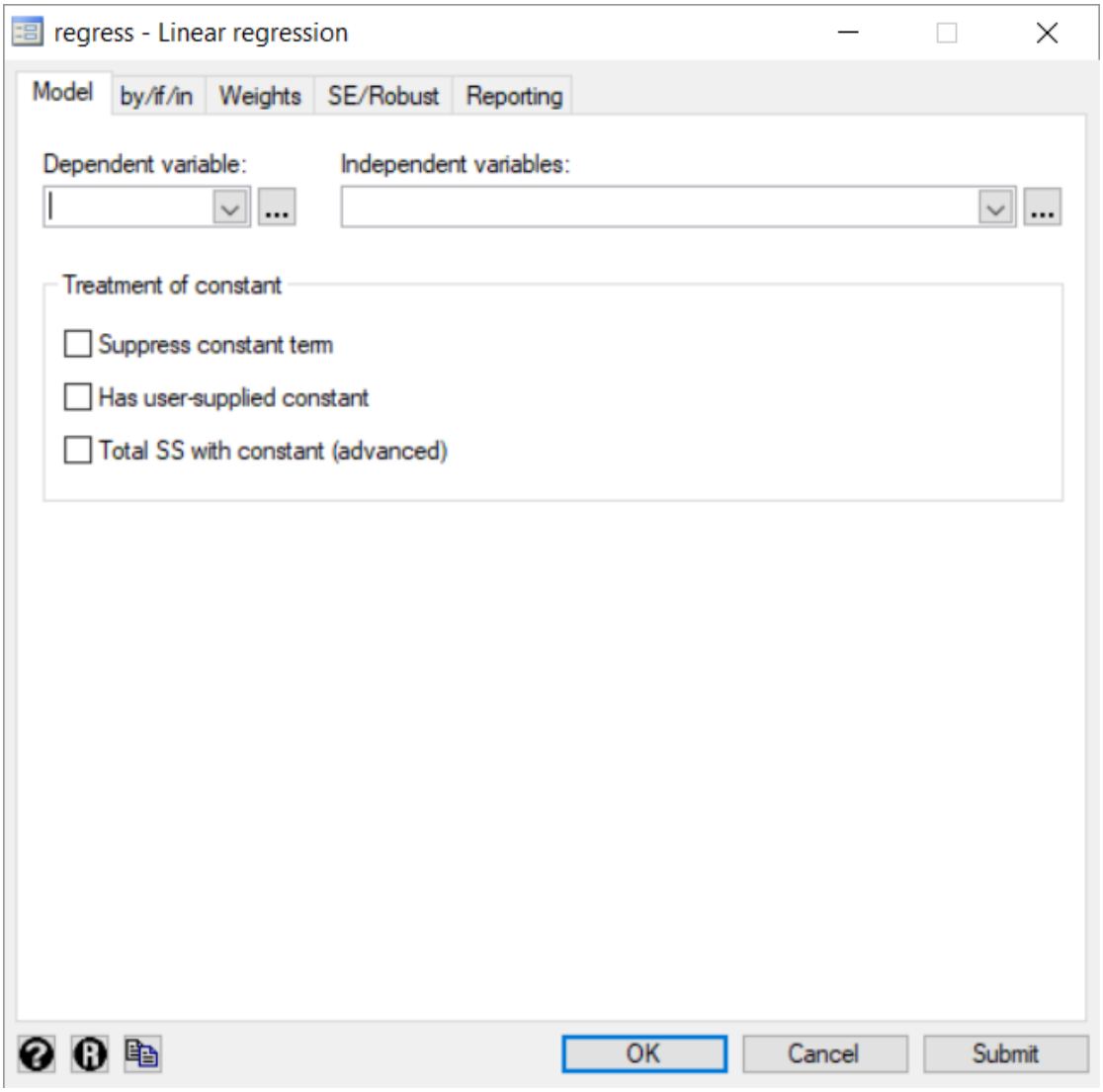
Η πολλαπλή γραμμική παλινδρόμηση με το STATA πραγματοποιείτε ακολουθώντας τα παρακάτω ***βήματα***:

Statistics → Linear models and related → Linear regression

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 9.9*

ii. Ορίζουμε ως ***Dependent variable*** τη συνεχή μεταβλητή /έκβαση, τις τιμές της οποίας επιθυμούμε να εκτιμήσουμε χρησιμοποιώντας κάποιο άλλο χαρακτηριστικό,

iii. Ως ***Independent variables*** ορίζουμε τις ανεξάρτητες μεταβλητές /παράγοντες

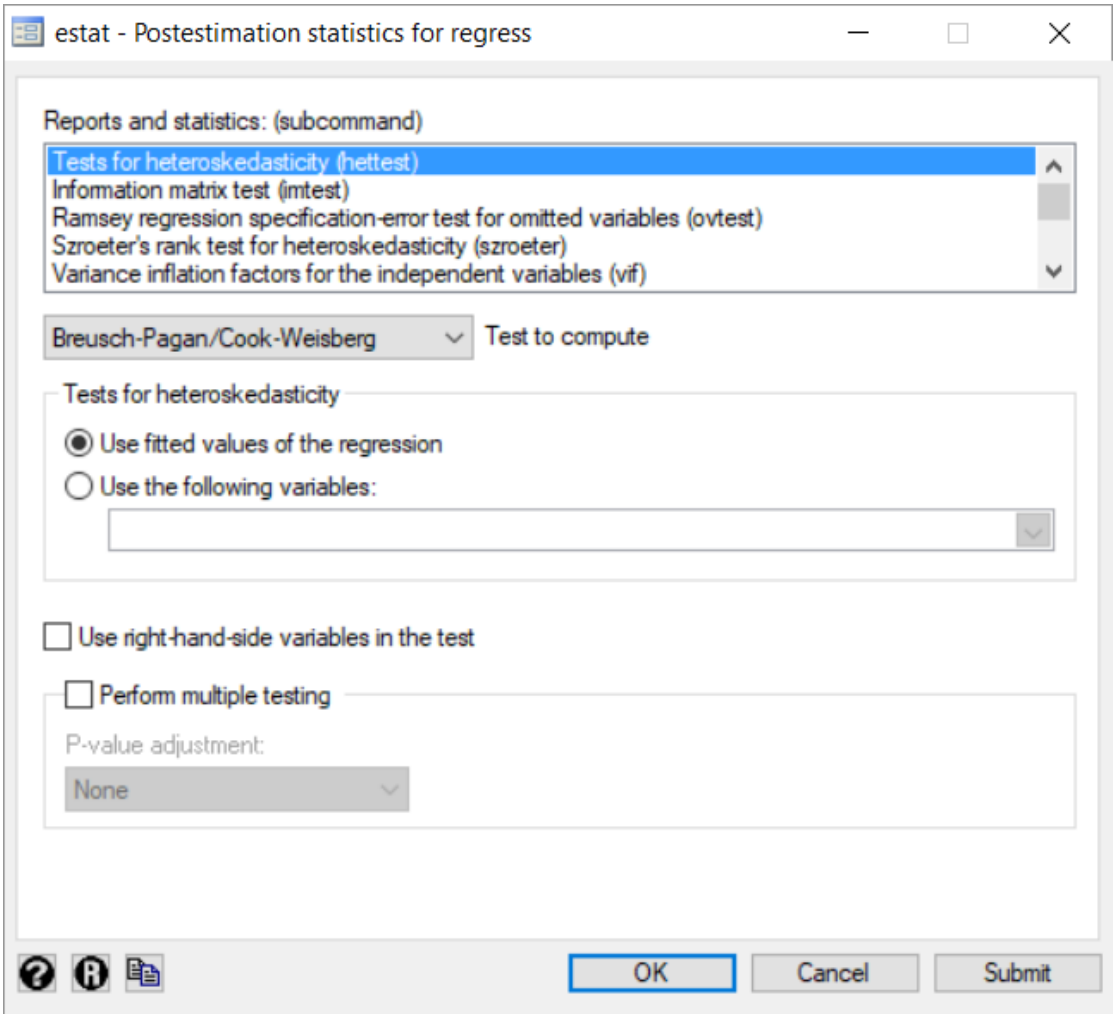


Εικόνα 9.9: Πραγματοποίηση της πολλαπλής γραμμικής παλινδρόμησης

Προκειμένου να δώσουμε την εντολή για την πραγματοποίηση των ελέγχων των προϋποθέσεων για την ορθή εφαρμογή της γραμμικής παλινδρόμησης ακολουθούμε τα εξής ***βήματα***:

Statistics → Linear models and related → Regression diagnostics → Specification tests, etc.

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 9.10*



Εικόνα 9.10: Πραγματοποίηση των απαραίτητων ελέγχων για την ορθή εφαρμογή της γραμμικής παλινδρόμησης.

ii. Από τις επιλογές που μας δίνονται στο πλαίσιο «***Reports and statistics***» τσεκάρουμε τους απαραίτητους ελέγχους. Για παράδειγμα τσεκάρουμε «***Test for heteroscedasticity***» για την εφαρμογή του ελέγχου ετεροσκεδαστικότητας, είτε variance «***inflation factors for the independent variables***» για τον έλεγχο ύπαρξης πολυσυγγραμικότητας.

Στο STATA χρησιμοποιείται η εντολή ***regress*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 9.11*:

Syntax

```
regress depvar [indepvars] [if] [in] [weight] [, options]
```

options	Description
Model	
noconstant	suppress constant term
hascons	has user-supplied constant
tsscons	compute total sum of squares with constant; seldom used
SE/Robust	
vce(*vcetype*)	*vcetype* may be ols, robust, cluster *clustvar*, bootstrap, jackknife, hc2, or hc3
Reporting	
level(#)	set confidence level; default is level(95)
beta	report standardized beta coefficients
eform(*string*)	report exponentiated coefficients and label as *string*
depname(*varname*)	substitute dependent variable name; programmer's option
display_options	control column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
noheader	suppress output header
notable	suppress coefficient table
plus	make table extendable
mse1	force mean squared error to 1
coeflegend	display legend instead of statistics

Εικόνα 9.11: Βασική σύνταξη της εντολής regress

9.4 Επιλογή βέλτιστου μοντέλου με το SPSS

Για να πραγματοποιήσουμε την επιλογή του βέλτιστου μοντέλου, στο παράθυρο διαλόγου που φαίνεται στην *Εικόνα 9.3*, απλά πρέπει να επιλέξουμε την μέθοδο επιλογής μας χρησιμοποιώντας το κουμπί επιλογών «***Methods***».

Ας υποθέσουμε ότι στόχος είναι να επιλέξουμε το βέλτιστο μοντέλο για την πρόβλεψη των τιμών της συστολικής αρτηριακής πίεσης (ΣΑΠ) την στιγμή της εισαγωγής ασθενών με οξύ στεφανιαίο σύνδρομο (ΟΣΣ) χρησιμοποιώντας ως ανεξάρτητες μεταβλητές το φύλο (sex), την ηλικία (age), τον ΔΜΣ (bmi), το ιστορικό σακχαρώδους διαβήτη (diabmel), την έκθεση σε παθητικό κάπνισμα (passive_smoking), το χρόνο που μεσολάβησε μεταξύ της έναρξης των συμπτωμάτων και την άφιξη στο νοσοκομείο (timecat2) και τη σωματική άσκηση (physact01). Επίσης, ας χρησιμοποιήσουμε την «προς τα πίσω απαλοιφή» για την απόκτηση του βέλτιστου μοντέλου. Για να το πραγματοποιήσουμε αυτό ακολουθούμε τα εξής ***βήματα***:

i. Πραγματοποιούμε τα βήματα της πολλαπλής γραμμικής παλινδρόμησης όπως αναφέρονται παραπάνω και επιπλέον.
ii. Στο πλαίσιο διαλόγου της *Εικόνας 9.3* επιλέγουμε την μέθοδο επιλογής μας χρησιμοποιώντας το κουμπί επιλογών «***Methods***» (π.χ. stepwise).
iii. Επίσης, στο πλαίσιο διαλόγου της *Εικόνας 9.5* επιλέγουμε την επιλογή «***R squared change***».
iv. Οι *Πίνακες 9.4, 9.5, 9.6 & 9.7* παρουσιάζουν τα αποτελέσματα.

Από όλους τους Πίνακες διαπιστώνουμε ότι χρειάστηκαν 4 βήματα μέχρι να φτάσουμε στο βέλτιστο μοντέλο. Από τον *Πίνακα 9.4* διαπιστώνουμε ότι το βέλτιστο μοντέλο (model 4) περιλαμβάνει το φύλο, τον ΔΜΣ και το ιστορικό σακχαρώδους διαβήτη, ενώ από τον *Πίνακα 9.5* παρατηρούμε τις μεταβλητές που αφαιρέθηκαν σε κάθε βήμα. Πιο συγκεκριμένα διαπιστώνουμε ότι αρχικά αφαιρέθηκε η σωματική άσκηση, στη συνέχεια ο χρόνος που μεσολάβησε μεταξύ της έναρξης των συμπτωμάτων και της άφιξης στο νοσοκομείο και τέλος η ηλικία. Από τον *Πίνακα 9.6* παρατηρούμε το ποσοστό της μεταβλητότητας της ΣΑΠ που ερμηνεύεται από το κάθε μοντέλο που προκύπτει σε κάθε βήμα της βηματικής διαδικασίας. Επίσης, από αυτό τον Πίνακα παρατηρούμε τη μεταβολή στο ποσοστό της μεταβλητότητας μεταξύ των διαδοχικών μοντέλων που προέρχονται από τα διαδοχικά βήματα της βηματικής διαδικασίας. Τέλος, από τον *Πίνακα 9.7* παρατηρούμε αν το κάθε ένα από τα μοντέλα που προκύπτουν από τη βηματική διαδικασία είναι στατιστικά σημαντικά καλύτερο όσον αφορά στην πρόβλεψη των τιμών της ΣΑΠ σε σχέση με το μοντέλο που δεν περιλαμβάνει καμία ανεξάρτητη μεταβλητή.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	116,383	8,019		14,514	,000	100,651	132,115		
	age	,056	,062	,027	,903	,367	-,066	,177	,894	1,119
	sex	-3,113	1,831	-,050	-1,700	,089	-6,706	,480	,944	1,059
	bmi	,674	,203	,097	3,312	,001	,275	1,072	,951	1,052
	diabmel	2,875	1,644	,051	1,749	,081	-,351	6,101	,972	1,029
	physact01	-1,155	1,557	-,022	-,742	,458	-4,211	1,900	,968	1,033
	timecat2	1,347	1,548	,025	,870	,385	-1,691	4,384	,982	1,019
2	(Constant)	115,484	7,925		14,572	,000	99,936	131,032		
	age	,059	,062	,029	,949	,343	-,063	,180	,897	1,115
	sex	-3,248	1,822	-,052	-1,783	,075	-6,822	,326	,954	1,048
	bmi	,684	,203	,099	3,374	,001	,286	1,082	,955	1,047
	diabmel	2,940	1,642	,052	1,791	,074	-,281	6,160	,975	1,026
	timecat2	1,431	1,544	,027	,927	,354	-1,597	4,460	,987	1,013
3	(Constant)	115,833	7,916		14,634	,000	100,303	131,363		
	age	,063	,062	,031	1,030	,303	-,057	,184	,903	1,107
	sex	-3,323	1,820	-,053	-1,826	,068	-6,894	,247	,956	1,046
	bmi	,694	,202	,100	3,429	,001	,297	1,092	,958	1,044
	diabmel	2,882	1,640	,051	1,757	,079	-,336	6,100	,976	1,025
4	(Constant)	121,404	5,779		21,008	,000	110,066	132,743		
	sex	-3,704	1,782	-,060	-2,079	,038	-7,200	-,208	,997	1,003
	bmi	,652	,198	,094	3,289	,001	,263	1,042	,998	1,002
	diabmel	3,135	1,622	,055	1,933	,053	-,047	6,317	,998	1,002

a. Dependent Variable: SBPmmHg

Πίνακας 9.4: Αποτελέσματα της πολλαπλής γραμμικής παλινδρόμησης; Συντελεστές της γραμμικής παλινδρόμησης για όλα τα μοντέλα που μεσολάβησαν μέχρι να τερματιστεί η διαδικασία.

Οι μεταβλητές που συμπεριλαμβάνονται στο βέλτιστο μοντέλο

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
2	physact01	-,022[a]	-,742	,458	-,021	,968	1,033	,894
3	physact01	-,023[b]	-,808	,419	-,023	,973	1,027	,899
	timecat2	,027[b]	,927	,354	,027	,987	1,013	,897
4	physact01	-,025[c]	-,875	,382	-,025	,978	1,023	,978
	timecat2	,029[c]	1,010	,313	,029	,994	1,006	,993
	age	,031[c]	1,030	,303	,030	,903	1,107	,903

a. Predictors in the Model: (Constant), timecat2, diabmel, bmi, sex, age

b. Predictors in the Model: (Constant), diabmel, bmi, sex, age

c. Predictors in the Model: (Constant), diabmel, bmi, sex

d. Dependent Variable: SBPmmHg

Πίνακας 9.5: Οι μεταβλητές που αφαιρέθηκαν σε κάθε βήμα της βηματικής διαδικασίας.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,135[a]	,018	,013	25,987	,018	3,722	6	1197	,001
2	,134[b]	,018	,014	25,982	,000	,550	1	1197	,458
3	,131[c]	,017	,014	25,981	-,001	,860	1	1198	,354
4	,128[d]	,016	,014	25,982	-,001	1,061	1	1199	,303

a. Predictors: (Constant), timecat2, diabmel, bmi, sex, physact01, age

b. Predictors: (Constant), timecat2, diabmel, bmi, sex, age

c. Predictors: (Constant), diabmel, bmi, sex, age

d. Predictors: (Constant), diabmel, bmi, sex

Η μεταβολή στο R square μεταξύ των μοντέλων που προκύπτουν από τα διαδοχικά βήματα

Πίνακας 9.6: Ερμηνευτικότητα όλων των μοντέλων μέχρι να τερματιστεί η διαδικασία.

ANOVA[e]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15081,233	6	2513,539	3,722	,001[a]
	Residual	808386,3	1197	675,344		
	Total	823467,6	1203			
2	Regression	14709,630	5	2941,926	4,358	,001[b]
	Residual	808757,9	1198	675,090		
	Total	823467,6	1203			
3	Regression	14129,249	4	3532,312	5,233	,000[c]
	Residual	809338,3	1199	675,011		
	Total	823467,6	1203			
4	Regression	13413,203	3	4471,068	6,623	,000[d]
	Residual	810054,3	1200	675,045		
	Total	823467,6	1203			

a. Predictors: (Constant), timecat2, diabmel, bmi, sex, physact01, age

b. Predictors: (Constant), timecat2, diabmel, bmi, sex, age

c. Predictors: (Constant), diabmel, bmi, sex, age

d. Predictors: (Constant), diabmel, bmi, sex

e. Dependent Variable: SBPmmHg

Πίνακας 9.7: Έλεγχος για το αν όλα τα μοντέλα που μεσολάβησαν μέχρι να τερματιστεί η διαδικασία διαφέρουν στατιστικά σημαντικά από ένα μοντέλο που να μην περιέχει καμία ανεξάρτητη μεταβλητή.

9.5 Επιλογή βέλτιστου μοντέλου με το STATA

Για την επιλογή βέλτιστου μοντέλου ακολουθούμε τα εξής βήματα:

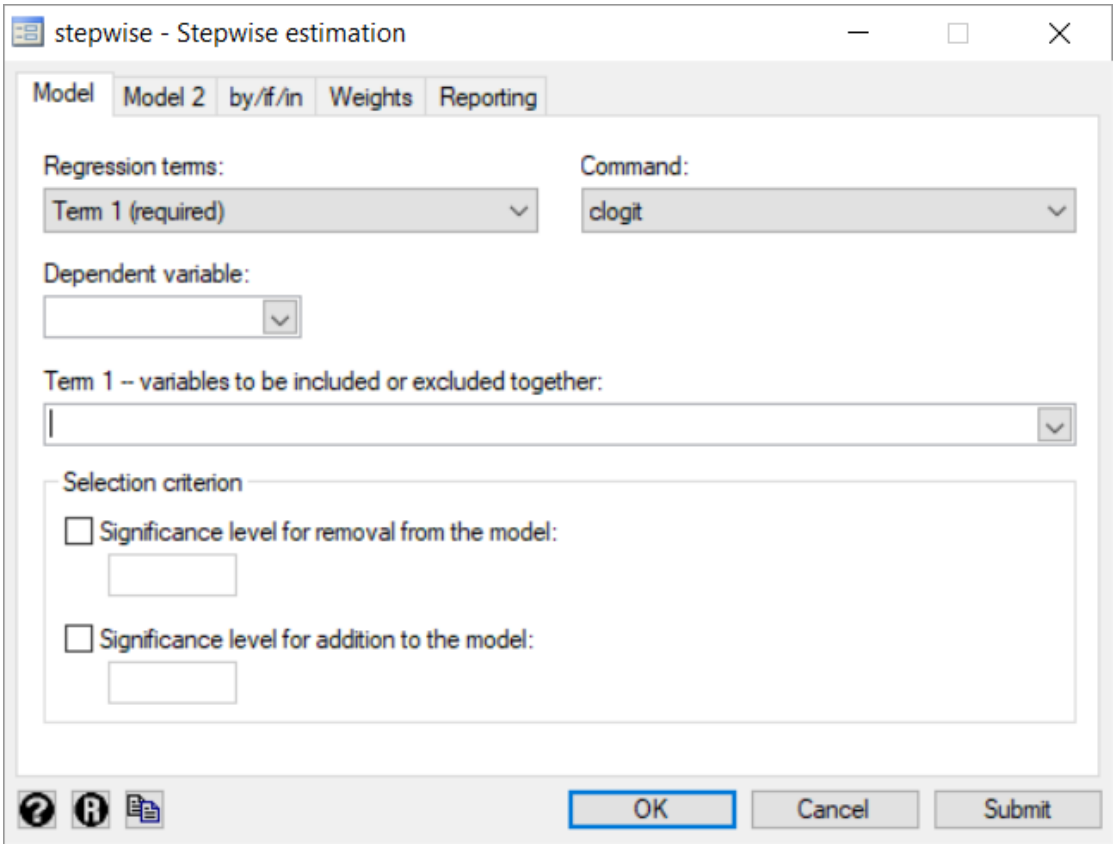
Statistics → Other → Stepwise estimation

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 9.12*

ii. Μέσα στο πλαίσιο «***Dependent variable***» τοποθετούμε την εξαρτημένη μας μεταβλητή.

iii. Στο πλαίσιο «***Term 1***» τοποθετούμε τις ανεξάρτητες μεταβλητές τις οποιες επιθυμούμε να εισάγουμε ή να εξάγουμε από το μοντέλο με κάποια μέθοδο.

iv. Στα πλαίσια «***Significance level for removal from the model***» και «***Significance level for addition to the model***» τοοθετούμε το επίπεδο στατιστικής σημαντικότητας το οποίο θα χρησιμοποιηθεί ως κριτήριο εξαγωγής ή εισαγωγής αντίστοιχα των μεταβλητών στο μοντέλο.



Εικόνα 9.12: Επιλογή βέλτιστου μοντέλου στο STATA

Στο STATA χρησιμοποιείται η εντολή ***stepwise*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 9.13*:

Syntax

`stepwise` [, *options*] : *command*

options	Description
Model	
*`pr(#)`	significance level for removal from the model
*`pe(#)`	significance level for addition to the model
Model2	
`forward`	perform forward-stepwise selection
`hierarchical`	perform hierarchical selection
`lockterm1`	keep the first term
`lr`	perform likelihood-ratio test instead of Wald test
Reporting	
display_options	control column formats and line width

* At least one of `pr(#)` or `pe(#)` must be specified.

`by` and `xi` are allowed; see [U] **11.1.10 Prefix commands**.

Weights are allowed if *command* allows them; see [U] **11.1.6 weight**.

All postestimation commands behave as they would after *command* without the `stepwise` prefix; see the postestimation manual entry for *command*.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

Εικόνα 9.13: Βασική συνταξη της εντολής stepwise

10. Λογαριθμιστική παλινδρόμηση

10.1. Εισαγωγή

Πολλές βιοϊατρικές έρευνες μελετούν παράγοντες που επηρεάζουν την εμφάνιση ή μη μιας συγκεκριμένης κατάστασης (π.χ ενός νοσήματος). Στην περίπτωση αυτή, η εξαρτημένη μεταβλητή είναι ποιοτική με δύο πιθανά αποτελέσματα (π.χ. 0 αν δεν υπάρχει η κατάσταση (το νόσημα) που μελετάται και με 1 αν υπάρχει) και επομένως δεν είναι δυνατή η εφαρμογή της πολλαπλής γραμμικής παλινδρόμησης. Συνήθως, εφαρμόζεται η ***λογαριθμιστική παλινδρόμηση (logistic regression model)***, που στηρίζεται επίσης στην εφαρμογή ενός γραμμικού υποδείγματος στα δεδομένα. Το μοντέλο της λογαριθμιστικής παλινδρόμησης ανήκει στα γενικευμένα γραμμικά μοντέλα, όπως και η γραμμικη παλινδρόμηση. Επομένως, θα είναι της μορφής:

$$g(\mu_i) = \eta_i = \sum_{j=1}^{k} x_{ij}\beta_j \quad (1)$$

Όπου

μ_i = E(Y\X), η μέση τιμή της εξαρτημένης μεταβλητής, δεδομένων των τιμών των ανεξάρτητων μεταβλητών. Η μέση τιμή της εξαρτημένης μεταβλητής στην περίπτωση της λογαριθμιστικής παλινδρόμησης είναι η αναλογία (ή πιθανότητα π) των ατόμων που έχουν την κατάσταση (το νόσημα). Η αναλογία αυτή παίρνει τιμές μεταξύ 0 και 1.

$\eta_i = \sum_{j=1}^{k} x_{ij}\beta_j$, ο γραμμικός συνδυασμός (linear predictor) των ανεξάρτητων μεταβλητών, του οποίου το εύρος τιμών κυμαίνεται μεταξύ $-\infty$ και $+\infty$, και

g(.) : η συνδετική συνάρτηση (link function)

Όπως φαίνεται από τα παραπάνω, το εύρος τιμών των 2 μελών της συνάρτησης (1) δεν συμπίπτουν. Γι' αυτό κρίνεται αναγκαία η χρησιμοποίηση κάποιου κατάλληλου μετασχηματισμού του π(x). Αυτός ο μετασχηματισμός είναι:

$$g(\pi(x)) = \log\{\pi(x)/(1-\pi(x))\},$$

που ονομάζεται logit και είναι ο λογάριθμος του λόγου των συμπληρωματικών πιθανοτήτων (log(odds)). Συνεπώς, η αλγεβρική μορφή του μοντέλου θα είναι:

$$\ln\left\{\frac{\pi_i(x)}{1-\pi_i(x)}\right\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}$$

όπου $X_{i1}, X_{i2} \ldots, X_{ip}$: οι τιμές των ανεξάρτητων μεταβλητών (j = 1, 2, . . . , p) της i – οστής παρατήρησης

$\beta_1, \beta_2, \ldots, \beta_p$: οι συντελεστές παλινδρόμησης.

$\pi_i(x) = \mathbf{P}(Y_i = 1 \setminus x)$, η πιθανότητα να συμβεί το γεγονός δεδομένου συγκεκριμένων τιμών των ανεξάρτητων μεταβλητών

$1 - \pi_i(x) = \mathbf{P}(Y_i = 0 \setminus x)$, η πιθανότητα να μη συμβεί το γεγονός

δεδομένου συγκεκριμένων τιμών των ανεξάρτητων μεταβλητών. Φαίνεται, λοιπόν, ότι μετά από αυτό το μετασχηματισμό, ο λογάριθμος του σχετικού λόγου των συμπληρωματικών πιθανοτήτων είναι γραμμικός ως προς τις ανεξάρτητες μεταβλητές.

10.1.1 Εκτίμηση των β-συντελεστών παλινδρόμησης και ερμηνεία

Η εκτίμηση των συντελεστών παλινδρόμησης πραγματοποιείται χρησιμοποιώντας τη *μέθοδο της μέγιστης πιθανοφάνειας (maximum likelihood estimation),* η οποία δίνει τέτοιες τιμές για τις άγνωστες παραμέτρους που να μεγιστοποιούν την πιθανότητα εμφάνισης των δεδομένων που υπάρχουν ως παρατηρήσεις.

Ένας συντελεστής λογαριθμιστικής παλινδρόμησης, π.χ. ο $\beta_ι$, δείχνει πόσο θα μεταβληθεί η εξαρτημένη μεταβλητή αν η ανεξάρτητη X_1 μεταβληθεί κατά μια μονάδα, ενώ οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν αμετάβλητες. Έτσι, π.χ., για $X_1 = 0$ η εξαρτημένη θα είναι 1ogit (p(0)) και για $X_2= 1$ θα είναι logit(p(1)).

$$\text{Logit } (\pi(0)) - \text{logit}(\pi(1)) =$$
$$\log(\pi(1)/1-\pi(1)) - \log(\pi(0)/1-\pi(0)) =$$
$$\log\{(\pi(1)*(1-\pi(0)) / (\pi(0)*(1-\pi(1))\}$$

Η τελευταία ποσότητα είναι ίση με το λογάριθμο του σχετικού λόγου (ln(odds ratio)) και στην επιδημιολογία είναι ένα από τα κυριότερα μέτρα που συσχετίζει μια ανεξάρτητη μεταβλητή («έκθεση» σε παράγοντα) με ένα αποτέλεσμα (π.χ. νόσημα ή κατάσταση υγείας). Είναι γεγονός ότι η ερμηνεία των συντελεστών της λογαριθμιστικής παλινδρόμησης δεν είναι εύκολη. Όμως, ένα σημαντικό πλεονέκτημα της εφαρμογής της λογαριθμιστικής παλινδρόμησης στη βιοιατρική έρευνα είναι ακριβώς ότι οι συντελεστές $\beta_ι$ που υπολογίζονται, εκφράζονται εύκολα με τη μορφή του σχετικού λόγου (***Σχετικός λόγος*** (ΣΛ) = $\exp(\beta_1)$).

Χρησιμοποιώντας τον παραπάνω μετασχηματισμό, λοιπόν, διαπιστώνουμε ότι από τους συντελεστές β_1 μπορούμε να υπολογίσουμε τον σχετικό λόγο εμφάνισης της νόσου (ή μιας οποιαδήποτε κατάστασης που μελετάμε ως εξαρτημένη μεταβλητή) που συνδέεται με κάθε ανεξάρτητη μεταβλητή. Ο ΣΛ (π.χ. $\exp(\beta_1)$) εκφράζει πόσο μεγαλύτερη ή μικρότερη είναι η πιθανότητα εμφάνισης της κατάστασης που μελετάμε (π.χ. νόσος) όταν η ανεξάρτητη μεταβλητή X_1 αυξηθεί κατά μία μονάδα και οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν αμετάβλητες. ΣΛ μεγαλύτερος της μονάδας συνεπάγεται ότι η αύξηση της μεταβλητής X_1 οδηγεί σε αύξηση της πιθανότητας εμφάνισης της κατάστασης, ενώ ΣΛ μικρότερος της μονάδας συνεπάγεται ότι αύξηση της μεταβλητής X_1 οδηγεί σε μείωση της πιθανότητας εμφάνισης της κατάστασης

Η λογαριθμιστική παλινδρόμηση είναι η σύγχρονη μέθοδος επιλογής για την ανάλυση των μελετών ασθενών-μαρτύρων (case-control), καθώς και κλινικών μελετών με δυνητικό αποτέλεσμα το θάνατο ή την επιβίωση, την εμφάνιση ή όχι μιας κατάστασης υγείας (π.χ. παχυσαρκία), την εισαγωγή ή όχι σε μονάδα εντατικής θεραπείας κλπ. Στη λογαριθμιστική παλινδρόμηση είναι δυνατό να συμπεριληφθούν ως ανεξάρτητες μεταβλητές, ποσοτικές, ποιοτικές και διατάξιμες μεταβλητές.

10.1.2 Έλεγχος υποθέσεων στο μοντέλο πολλαπλής λογαριθμιστικής παλινδρόμησης

Όπως και στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης έτσι και εδώ τα

ερωτήματα που μπορούν να ανακύψουν είναι τα εξής:

- Κατά πόσο ολόκληρο το μοντέλο συνεισφέρει στατιστικά σημαντικά στην πρόβλεψη της εξαρτημένης μεταβλητής Y ή αλλιώς στην ερμηνεία της μεταβλητότητάς της.
- Κατά πόσο μία συγκεκριμένη τυχαία μεταβλητή παρέχει στατιστικά σημαντική πληροφορία για την πρόβλεψη της Y, δεδομένης της παρουσίας άλλων ερμηνευτικών μεταβλητών στο μοντέλο.
- Κατά πόσο μία ομάδα τυχαίων μεταβλητών παρέχει στατιστικά σημαντική πληροφορία για την πρόβλεψη της Y, δεδομένης της παρουσίας άλλων ερμηνευτικών μεταβλητών στο μοντέλο.

Η απάντηση σε όλα τα παραπάνω ερωτήματα, δίνεται πραγματοποιώντας τους κατάλληλους στατιστικούς ελέγχους υποθέσεων. Οι μηδενικές και εναλλακτικές υποθέσεις για κάθε ένα από τα παραπάνω ερωτήματα διατυπώνονται αναλυτικά στο *Κεφάλαιο 9* (*παράγραφος 9.1.2.2*). Παρακάτω παρουσιάζονται συνοπτικά οι κατάλληλοι στατιστικοί έλεγχοι για την διερεύνηση αυτών των υποθέσεων.

Κριτήριο Z ή Wald test

Το κριτήριο Z χρησιμοποιείται για να ελεγχθεί κατά πόσο καθένας από τους συντελεστές εξάρτησης ενός πολλαπλού λογαριθμιστικού μοντέλου είναι στατιστικά σημαντικός, δηλαδή κατά πόσο η συγκεκριμένη ανεξάρτητη μεταβλητή έχει σαν αποτέλεσμα οι προβλεπόμενες τιμές να προσεγγίζουν περισσότερο τις παρατηρηθείσες.

Άρα, η μηδενική και εναλλακτική υπόθεση είναι:

$$H_0: \hat{\beta} = \beta_0 \text{ vs } H_1: \hat{\beta} \neq \beta_0 .$$

Το κριτήριο Z βασίζεται στο γεγονός ότι οι συντελεστές εξάρτησης ως εκτιμητές μέγιστης πιθανοφάνειας ακολουθούν ασυμπτωτικά την κανονική κατανομή, όπως προαναφέρθηκε. Ορίζεται ως εξής:

$$Z = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})}$$

και κάτω από την μηδενική υπόθεση ακολουθεί την τυποποιημένη κανονική κατανομή.

Στην περίπτωση, που θέλουμε να ελέγξουμε ταυτόχρονα τη στατιστική σημαντικότητα πολλών συντελεστών εξάρτησης (όπως στις κατηγορικές μεταβλητές με περισσότερες από δύο κατηγορίες για τον έλεγχο overall της μεταβλητής), το κριτήριο Z υπολογίζεται με τη βοήθεια πινάκων, από την εξής σχέση:

$$W = \left(\hat{\beta} - \beta_0\right)' I(\beta)^{-1} \left(\hat{\beta} - \beta_0\right) \sim \chi^2_n$$

Οι τιμές αυτών των κριτηρίων συγκρίνονται με τις αντίστοιχες κρίσιμες τιμές και έτσι υπολογίζεται η πιθανότητα εσφαλμένης απόρριψης της H_0 ή αλλιώς το p-value. Αν $p<0,05$ που είναι το μέγιστο επιτρεπτό όριο εσφαλμένης απόρριψης της H_0 τότε απορρίπτουμε την H_0 και συμπεραίνουμε ότι ότι η συγκεκριμένη μεταβλητή ή ομάδα μεταβλητών προσφέρει στατιστικά σημαντική πληροφορία στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής.

Έλεγχος λόγου πιθανοφάνειας (LR)

Ο έλεγχος λόγου πιθανοφάνειας είναι ένας άλλος τρόπος ελέγχου των παραπάνω υποθέσεων. Αυτός βασίζεται στη σύγκριση των πιθανοφανειών των δύο μοντέλων, εκ των οποίων το ένα συμπεριλαμβάνει τις ανεξάρτητες μεταβλητές που θέλουμε να ελέγξουμε αν προσφέρουν σημαντική πληροφορία στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής, ενώ το άλλο όχι και ορίζεται ως εξής:

LR = -2 [ln(πιθανοφάνειας χωρίς τη μεταβλητή) – ln(πιθανοφάνειας με τη μεταβλητή)]

Αυτό το κριτήριο (ο λόγος των 2 πθανοφανειών) ακολουθεί την χ^2 κατανομή με βαθμούς ελευθερίας όση είναι και η διαφορά των ανεξάρτητων μεταβλητών των συγκρινόμενων μοντέλων.

Για να εφαρμοστεί ο έλεγχος λόγου πιθανοφάνειας, θα πρέπει να ισχύουν οι εξής προϋποθέσεις:

1. Το ένα μοντέλο να είναι φωλιασμένο (nested) στο άλλο
2. Τα δύο υπό σύγκριση μοντέλα θα πρέπει να προσαρμόζονται στον ίδιο ακριβώς αριθμό παρατηρήσεων.

Η τιμή του παραπάνω κριτηρίου συγκρίνεται με τη αντίστοιχη κρίσιμη τιμή και έτσι υπολογίζεται η πιθανότητα εσφαλμένης απόρριψης της H_0 ή αλλιώς το p-value. Αν $p<0,05$ που είναι το μέγιστο επιτρεπτό όριο εσφαλμένης απόρριψης της H_0 τότε απορρίπτουμε την H_0 και συμπεραίνουμε ότι ότι η συγκεκριμένη μεταβλητή ή ομάδα μεταβλητών προσφέρει στατιστικά σημαντική πληροφορία στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής.

10.1.3 Έλεγχος καλής προσαρμογής του μοντέλου

Ο έλεγχος καλής προσαρμογής του μοντέλου πραγματοποιείται συγκρίνοντας την κατηγορία της εξαρτημένης μεταβλητής στην οποία κάθε άτομο κατατάσσεται βάσει του μοντέλου με την πραγματική κατηγορία στην οποία το συγκεκριμένο άτομο ανήκει. Έχουν προταθεί διάφορα κριτήρια για τον έλεγχο καλής προσαρμογής του υποδείγματος, όπως για παράδειγμα του Pearson και η deviance. Καλή προσαρμογή του υποδείγματος στα πραγματικά δεδομένα έχουμε για μικρές τιμές των κριτηρίων, δηλ. $p>\alpha$ (0,05). Άλλος ένας πολύ συνηθισμένος τρόπος ελέγχου καλής προσαρμογής είναι ο υπολογισμός του κριτηρίου των Hosmer – Lemeshow. Με το κριτήριο αυτό σχηματίζονται κ διατεταγμένα υποσύνολα των ατόμων ανάλογα με την εκτιμούμενη πιθανότητα π. Στη συνέχεια συγκρίνονται ο αριθμός των ατόμων που πραγματικά ανήκει στο κάθε υποσύνολο με τον εκτιμούμενο από το υπόδειγμα αριθμό των ατόμων που θα ανήκε στο κάθε υποσύνολο. Το κριτήριο αυτό ακολουθεί την X^2 κατανομή. Όσο πιο μικρές τιμές έχει το κριτήριο τόσο πιο κοντά βρισκόμαστε στην μηδενική υπόθεση που λέει ότι το εκτιμηθέν υπόδειγμα δεν διαφέρει από το πραγματικό. Όπως είναι φανερό όσο πιο πολλές είναι οι ερμηνευτικές μεταβλητές τόσο πιο πολύπλοκος είναι ο σχηματισμός των υποσυνόλων, γεγονός που κάνει την εφαρμογή του κριτηρίου δύσκολη.

Η καλή προσαρμογή μπορεί να ελεγχθεί και γραφικά με την απεικόνιση της ΔD (μεταβολής της deviance για κάθε συνδυασμό των παραμέτρων) έναντι των εκτιμηθέντων από το υπόδειγμα πιθανοτήτων π.

10.2 Διατάξιμη λογαριθμιστική παλινδρόμηση

Όταν η έκβαση ή αλλιώς η εξαρτημένη μεταβλητή (outcome ή dependent variable) αντί για δίτιμη μεταβλητή είναι διατάξιμη κατηγορική (π.χ βαρύτητα της νόσου: ελαφριά, βαριά, πολύ βαριά) η κατάλληλη στατιστική ανάλυση είναι η ***διατάξιμη λογαριθμιστική παλινδρόμηση (ordinal logistic regression)***. Υπάρχουν τρία διαφορετικά είδη μοντέλων διατάξιμης λογαριθμιστικής παλινδρόμησης ανάλογα με τις κατηγορίες της έκβασης που θέλουμε να συγκρίνουμε.

- Αυτό που συγκρίνει κάθε κατηγορία με την αμέσως επόμενη και ονομάζεται *λογαριθμιστικό μοντέλο παρακείμενης κατηγορίας (adjacent-category logistic model),*
- Αυτό που συγκρίνει κάθε κατηγορία με όλες τις κατηγορίες που προηγούνται αυτής και ονομάζεται *λογαριθμιστικό μοντέλο συνέχειας-αναλογίας (continuation-ratio logistic model)*
- Αυτό που συγκρίνει την πιθανότητα να ισχύει κάποια κατηγορία του outcome μικρότερη ή ίση με την j, με την πιθανότητα να είναι μεγαλύτερη του j, και αυτό ονομάζεται *μοντέλο αναλογικών λόγων συμπληρωματικών πιθανοτήτων (proportional odds model).*

10.2.1 Μοντέλα αναλογικών λόγων συμπληρωματικών πιθανοτήτων

Τα πιο συχνά μοντέλα διατάξιμης λογαριθμιστικής παλινδρόμησης είναι τα μοντέλα των αναλογικών λόγων συμπληρωματικών πιθανοτήτων. Αυτά βασίζονται στην πιθανότητα μία παρατήρηση να ανήκει σε κατηγορία της εξαρτημένης μεταβλητής μικρότερη από την j , δηλαδή στην $\gamma_j = P(\mathrm{Y} \le j)$) και όχι στην πιθανότητα μία παρατήρηση να ανήκει σε κάποια συγκεκριμένη κατηγορία της εξαρτημένης μεταβλητής ($\pi_j = P(U = j)$) όπως ισχύει στην κλασσική λογαριθμιστική παλινδρόμηση που αναφέρεται παραπάνω.

Η αλγεβρική μορφή ενός τέτοιου μοντέλου είναι :

$$\ln\left\{\frac{\gamma_j(x)}{1-\gamma_j(x)}\right\} = \kappa_j - (\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi})$$

όπου j = 1, 2, . . . , I – 1 και I = ο αριθμός των κατηγοριών της εξαρτημένης μεταβλητής

κ_j **:** η σταθερά που παίρνει διαφορετική τιμή για κάθε κατηγορία

$X_1, X_2 \ldots, X_p$: οι ανεξάρτητες μεταβλητές

$\beta_1, \beta_2, \ldots, \beta_p$: οι συντελεστές παλινδρόμησης,

$\gamma_j(x) = P(\mathrm{Y} \le j \setminus x)$: πιθανότητα μία παρατήρηση να ανήκει σε κατηγορία της έκβασης μικρότερη από την j, δεδομένης της τιμής των ανεξάρτητων μεταβλητών

$1-\gamma_j(x) = P(\mathrm{Y} > j \setminus x)$: πιθανότητα μία παρατήρηση να ανήκει σε κατηγορία της έκβασης μεγαλύτερη από την j, δεδομένης της τιμής των ανεξάρτητων μεταβλητών

Ένα μοντέλο αυτής της μορφής ονομάζεται μοντέλο των αναλογικών λόγων συμπληρωματικών πιθανοτήτων (proportional – odds model), γιατί ο λόγος των

λόγων συμπληρωματικών πιθανοτήτων του Y≤ j είναι ανεξάρτητος από την επιλογή της κατηγορίας, δηλαδή ο λόγος των λόγων συμπληρωματικών πιθανοτήτων παραμένει σταθερός για όλες τις τιμές του j. Αυτό αποδεικνύεται παρακάτω.
π.χ Έστω ότι ζητείται να υπολογιστεί ο σχετικός λόγος συμπληρωματικών πιθανοτήτων για κάποια μεταβολή στην τιμή της χ_1 μεταβλητής, διατηρώντας σταθερές τις τιμές όλων των υπόλοιπων μεταβλητών. Τα δύο μοντέλα θα έχουν την εξής μορφή.

$$\ln\left\{\frac{\gamma_j(x_{11})}{1-\gamma_j(x_{11})}\right\} = \kappa_j - (\beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

και

$$\ln\left\{\frac{\gamma_j(x_{12})}{1-\gamma_j(x_{12})}\right\} = \kappa_j - (\beta_1 X_{12} + \beta_2 X_2 + ... + \beta_p X_p)$$

Ο σχετικός λόγος συμπληρωματικών πιθανοτήτων θα είναι:

$$\frac{\gamma_j(x_{12})/1-\gamma_j(x_{12})}{\gamma_j(x_{11})/1-\gamma_j(x_{11})} = \exp\{-\beta_1(X_{12}-X_{11})\}$$

Παρατηρούμε λοιπόν, ότι ο σχετικός λόγος συμπληρωματικών πιθανοτήτων είναι ανεξάρτητο από την επιλογή της κατηγορίας j.
Επίσης, πρέπει να αναφερθεί ότι το αρνητικό πρόσημο στην τελευταία εξίσωση συνεπάγεται ότι η μεταβλητή χ συνδέεται αρνητικά με τις μικρότερες τιμές της διατάξιμης εξαρτημένης μεταβλητής, αν ο συντελεστής συσχέτισης (β) είναι θετικός. Δηλαδή, αύξηση της χ συνεπάγεται μεγαλύτερη πιθανότητα η διατάξιμη μεταβλητή να πάρει μεγάλες τιμές.

Πιο συγκεκριμένα, όσον αφορά στην ***ερμηνεία των συντελεστών β***, αξίζουν να σημειωθούν τα εξής:

- Στην περίπτωση των ποσοτικών ανεξάρτητων μεταβλητών, θετικός συντελεστής σημαίνει ότι αύξηση της ανεξάρτητης μεταβλητής συνεπάγεται και αύξηση της πιθανότητας εμφάνισης υψηλότερων τιμών της εξαρτημένης διατάξιμης μεταβλητής, ενώ αρνητικός συντελεστής σημαίνει ότι αύξηση της ανεξάρτητης μεταβλητής συνεπάγεται και μείωση της πιθανότητας εμφάνισης υψηλότερων τιμών της εξαρτημένης διατάξιμης μεταβλητής.
- Στην περίπτωση των δίτιμων ανεξάρτητων μεταβλητών για να προχωρήσουμε στην ερμηνεία των συντελεστών θα πρέπει να γνωρίζουμε ποια κατηγορία της μεταβλητής χρησιμοποιείται ως κατηγορία αναφοράς. Για παράδειγμα, στο SPSS ως κατηγορία αναφοράς χρησιμοποιείται η κατηγορία που είναι κωδικοποιημένη με το μεγαλύτερο αριθμό. Συνεπώς, αν για το φύλο έχουμε άνδρας:1 και γυναίκα:0, η κατηγορία αναφοράς είναι ο άνδρας. Αν, λοιπόν, ο συντελεστής είναι θετικός σημαίνει ότι οι γυναίκες έχουν μεγαλύτερη πιθανότητα να έχουν υψηλότερες τιμές της εξαρτημένης μεταβλητής σε σχέση με του άνδρες.

- Ένα μειονέκτημα των συντελεστών β είναι ότι η ερμηνεία τους είναι δύσκολη. Το πλεονέκτημά τους, όμως, είναι ότι εύκολα μπορούν να εκφραστούν σε σχετικό λόγο (ΣΛ), όπως ακριβώς συμβαίνει και στην περίπτωση της κλασσικής λογαριθμιστικής παλινδρόμησης (ΣΛ = exp(β)). Σχετικός λόγος>1 εκφράζεται ως εξής: αύξηση της ποσοτικής μεταβλητής κατά 1 μονάδα συνεπάγεται αύξηση της πιθανότητας εμφάνισης υψηλότερων τιμών της εξαρτημένης διατάξιμης μεταβλητής κατά (1-exp(β))%. Αντίθετα, ΣΛ<1 εκφράζεται ως εξής: αύξηση της ποσοτικής μεταβλητής κατά 1 μονάδα συνεπάγεται μείωση της πιθανότητας εμφάνισης υψηλότερων τιμών της εξαρτημένης διατάξιμης μεταβλητής κατά (exp(β)-1)%. Στην περίπτωση που η ανεξάρτητη μεταβλητή είναι δίτιμη, ΣΛ>1 και ΣΛ<1 εκφράζει πόσο μεγαλύτερη ή μικρότερη, αντίστοιχα, είναι η πιθανότητα, τα άτομα που ανήκουν στην κατηγορία Α να έχουν υψηλότερες τιμές της εξαρτημένης μεταβλητής σε σχέση με τα άτομα που ανήκουν στην κατηγορία αναφοράς.

Τέλος, όσον αφορά στον έλεγχο των στατιστικών υποθέσεων και στην εκτίμηση των παραμέτρων του μοντέλου, ισχύουν όλα όσα παρουσιάστηκαν στην περίπτωση της λογαριθμιστικής παλινδρόμησης, ενώ δεν αναφέρονται στη βιβλιογραφία τρόποι ελέγχου της καλής προσαρμογής του μοντέλου και ελέγχου ύπαρξης έκτοπων παρατηρήσεων ή παρατηρήσεων επιρροής.

10.2.2 Έλεγχος για την ισχύ της υπόθεσης των αναλογικών λόγων συμπληρωματικών πιθανοτήτων (proportional odds assumption)

Μετά την επιλογή του βέλτιστου μοντέλου, πρέπει να ακολουθήσει ο έλεγχος για την ισχύ της προϋπόθεσης των αναλογικών λόγων συμπληρωματικών πιθανοτήτων για τα δεδομένα μας, βάσει της οποίας έγινε η εκτίμηση των παραμέτρων. Αυτό δύναται να επιτευχθεί συγκρίνοντας τους συντελεστές παλινδρόμησης που προκύπτουν για όλες τις κατηγορίες της εξαρτημένης μεταβλητής. Ουσιαστικά, σε αυτόν τον έλεγχο είναι σαν να εκτιμούνται διαφορετικά μοντέλα παλινδρόμησης για κάθε κατηγορία της εξαρτημένης μεταβλητής σε σχέση με κάποια κατηγορία αναφοράς (πολυωνυμική λογαριθμιστική παλινδρόμηση (βλ. ενότητα 10.3)) και στη συνέχεια να συγκρίνονται μεταξύ τους οι συντελεστές των μοντέλων για κάθε ανεξάρτητη μεταβλητή. Η μηδενική υπόθεση σε αυτό τον έλεγχο είναι ότι οι συντελεστές δεν διαφέρουν μεταξύ τους και συνεπώς ισχύει η προϋπόθεση της αναλογικότητας. Συνεπώς, λοιπόν, επιθυμούμε, ο συγκεκριμένος έλεγχος να δείξει ότι η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης είναι μεγαλύτερη από το 0,05 (το μέγιστο επιτρεπτό όριο) και συνεπώς δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση (άρα ισχύει η αναλογικότητα).

Στην περίπτωση που η μηδενική υπόθεση ή αλλιώς η προϋπόθεση της αναλογικότητας των συμπληρωματικών πιθανοτήτων απορριφθεί, υπάρχουν 2 πιθανές λύσεις:

- Να γίνει ομαδοποίηση κάποιων ομάδων έτσι ώστε να ισχύσει η προϋπόθεση της αναλογικότητας ή να μετατραπεί η εξαρτημένη μας μεταβλητή σε δίτιμη μεταβλητή, οπότε θα εφαρμόσουμε την κλασσική λογαριθμιστική παλινδρόμηση.
- Να εφαρμόσουμε πολυωνυμική λογαριθμιστική παλινδρόμηση (βλ. *ενότητα 10.3*).

10.3 Πολυωνυμική λογαριθμιστική παλινδρόμηση

Όταν η έκβαση/εξαρτημένη μεταβλητή (outcome ή dependent variable) αντί για δίτιμη είναι ονομαστική κατηγορική με περισσότερες κατηγορίες (π.χ συχνότητα κατανάλωσης φρούτων: ποτέ/σπάνια (0), μερικές φορές (1), συχνά/πολύ συχνά (2)) η κατάλληλη στατιστική ανάλυση είναι η πολυωνυμική λογαριθμιστική παλινδρόμηση (multinomial logistic regression). Επίσης, όπως, αναφέρεται και στην *ενότητα 10.2,* η πολυωνυμική λογαριθμιστική παλινδρόμηση εφαρμόζεται και στην περίπτωση που η εξαρτημένη μας μεταβλητή είναι διατάξιμη, αλλά δεν ισχύει η προϋπόθεση της αναλογικότητας των συμπληρωματικών πιθανοτήτων που είναι απαραίτητη για την ορθή εφαρμογή της διατάξιμης λογαριθμιστικής παλινδρόμησης. Ας υποθέσουμε, λοιπόν, ότι επιθυμούμε να διερευνήσουμε πως μία σειρά από άλλα χαρακτηριστικά (π.χ. ηλικία, φύλο, εισόδημα, οικογενειακή κατάσταση κτλ) συσχετίζονται με την παραπάνω ονομάστική μας εξαρτημένη μεταβλητή.

Ουσιαστικά, η πολυωνυμική λογαριθμιστική παλινδρόμηση είναι ένας συνδυασμός k-1 μοντέλων λογαριθμιστικής παλινδρόμησης για δίτιμες εξαρτημένες μεταβλητές (βλ. ενότητα 10.1). Κάθε μοντέλο εκτιμά την πιθανότητα η εξαρτημένη μεταβλητή Y να πάρει κάποια από τις πιθανές τιμές της έναντι του να πάρει την τιμή της κατηγορίας αναφοράς. Συνεπώς, λοιπόν, το πρώτο πράγμα που πρέπει να προσδιορίσουμε στην πολυωνυμική λογαριθμιστική παλινδρόμηση είναι η κατηγορία αναφοράς. Συνήθως, ορίζεται ως κατηγορία αναφοράς, η κατηγορία που έχει κωδικοποιηθεί με 0 κατά αντιστοιχία με αυτό που χρησιμοποιείται στην λογαριθμιστική παλινδρόμηση για δίτιμες εξαρτημένες μεταβλητές. Άρα, λοιπόν, στο συγκεκριμένο παράδειγμα, το ένα μοντέλο θα εκτιμά την πιθανότητα κατανάλωσης φρούτων μερικές φορές (Y=1) έναντι της σπάνιας/καθόλου (Y=0) κατανάλωσης και το άλλο μοντέλο θα εκτιμά την πιθανότητα συχνής/πολύ συχνής κατανάλωσης φρούτων (Y=2) έναντι της σπάνιας/καθόλου κατανάλωσης (Y=0).

Η αλγεβρική μορφή αυτών των μοντέλου θα είναι:

$$g_1(x) = \ln\left\{\frac{\pi_i(Y=1 \setminus x)}{\pi_i(Y=0 \setminus x)}\right\} = \beta_{10} + \beta_{11}X_{i1} + \beta_{12}X_{i2} + ... + \beta_{1p}X_{ip}$$

και

$$g_1(x) = \ln\left\{\frac{\pi_i(Y=2 \setminus x)}{\pi_i(Y=0 \setminus x)}\right\} = \beta_{20} + \beta_{21}X_{i1} + \beta_{22}X_{i2} + ... + \beta_{2p}X_{ip}$$

όπου $X_{i1}, X_{i2} \ldots, X_{ip}$: οι τιμές των ανεξάρτητων μεταβλητών (j = 1, 2, . . . , p) της i – οστής παρατήρησης

$\beta_{11}, \beta_{12}, ..., \beta_{1p}$: οι συντελεστές παλινδρόμησης για το μοντέλο που εκτιμά την πιθανότητα η εξαρτημένη μας μεταβλητή να πάρει την τιμή 1 έναντι της τιμής 0.

$\beta_{21}, \beta_{22}, ..., \beta_{2p}$: οι συντελεστές παλινδρόμησης για το μοντέλο που εκτιμά την πιθανότητα η εξαρτημένη μας μεταβλητή να πάρει την τιμή 2 έναντι της τιμής 0.

$\pi_i(Y_i = 1 \setminus x)$, η πιθανότητα η Y να πάρει την τιμή 1 όταν οι ανεξάρτητες μεταβλητές παίρνουν συγκεκριμένες τιμές (για συγκεκριμένο covariate pattern).

$\pi_i(Y_i = 2 \setminus x)$, η πιθανότητα η Y να πάρει την τιμή 2 όταν οι ανεξάρτητες μεταβλητές παίρνουν συγκεκριμένες τιμές (για συγκεκριμένο covariate pattern).
$\pi_i(Y_i = 0 \setminus x)$, η πιθανότητα η Y να πάρει την τιμή 0 όταν οι ανεξάρτητες μεταβλητές παίρνουν συγκεκριμένες τιμές (για συγκεκριμένο covariate pattern).
Συνεπώς, λοιπόν, για κάθε covariate pattern (δηλαδή τις τιμές των ανεξάρτητων μεταβλητών του κάθε συμμετέχοντα στη μελέτη), υπολογίζονται με την βοήθεια των παραπάνω μοντέλων, 3 πιθανότητες οι οποίες εκφράζουν την πιθανότητα το άτομο με τα συγκεκριμένα στοιχεία να ανήκει σε κάθε μία από τις τρεις κατηγορίες της εξαρτημένης μεταβλητής, αντίστοιχα. Το άτομο, με τη βοήθεια της πολυωνυμικής λογαριθμιστικής παλινδρόμησης κατατάσσεται στην κατηγορία της εξαρτημένης μεταβλητής για την οποία η πιθανότητα είναι υψηλότερη.
Εδώ πρέπει να σημειωθεί ότι όπως και στην κλασσική λογαριθμιστική παλινδρόμηση για δίτιμες εξαρτημένες μεταβλητές, έτσι και εδώ, ως ανεξάρτητες μεταβλητές μπορούν να χρησιμοποιηθούν είτε ποσοτικά είτε ποιοτικά χαρακτηριστικά.

10.3.1 Εκτίμηση των β-συντελεστών παλινδρόμησης και ερμηνεία

Η εκτίμηση των συντελεστών παλινδρόμησης πραγματοποιείται χρησιμοποιώντας τη *μέθοδο της μέγιστης πιθανοφάνειας (maximum likelihood estimation),* η οποία δίνει τέτοιες τιμές για τις άγνωστες παραμέτρους που να μεγιστοποιούν την πιθανότητα εμφάνισης των δεδομένων που υπάρχουν ως παρατηρήσεις.

Όσον αφορά στην ***ερμηνεία των συντελεστών***, αξίζουν να σημειωθούν τα εξής:

- Στην περίπτωση των ποσοτικών ανεξάρτητων μεταβλητών, θετικός συντελεστής (ή Exp(B)>1) σημαίνει ότι αύξηση της ανεξάρτητης μεταβλητής συνεπάγεται και αύξηση της πιθανότητας εμφάνισης της κατηγορίας 1 ή 2 (ανάλογα με το μοντέλο που ερμηνεύουμε) της εξαρτημένης μεταβλητής σε σχέση με την κατηγορία αναφοράς 0, ενώ αρνητικός συντελεστής (ή Exp(B)<1) σημαίνει ότι αύξηση της ανεξάρτητης μεταβλητής συνεπάγεται και μείωση της πιθανότητας εμφάνισης της κατηγορίας 1 ή 2 σε σχέση με την κατηγορία αναφοράς. Βέβαια, πρέπει να σημειωθεί ότι η ερμηνεία των συντελεστών «β» των μοντέλων δεν είναι εύκολη. Γι'αυτό είναι προτιμότερο να χρησιμοποιούμε τον σχετικό λόγο που είναι ίσος με το exp(B) προκειμένου να ερμηνεύσουμε τα αποτελέσματα. Συγκεκριμένα, αν το exp(B)>1 τότε ερμηνεύεται ως εξής: αύξηση της ανεξάρτητης μεταβλητής X κατά μία μονάδα συνεπάγεται αύξηση της πιθανότητα η Y να πάρει την τιμή 1 ή 2 (ανάλογα με το μοντέλο που ερμηνεύουμε) κατά (1-exp(B))% έναντι του να πάρει η Y την τιμή 0. Αντίθετα, αν το exp(B)<1 τότε ερμηνεύεται ως εξής: αύξηση της ανεξάρτητης μεταβλητής X κατά μία μονάδα συνεπάγεται μείωση της πιθανότητα η Y να πάρει την τιμή 1 ή 2 (ανάλογα με το μοντέλο που ερμηνεύουμε) κατά (exp(B)-1)% έναντι του να πάρει η Y την τιμή 0.

- Στην περίπτωση των δίτιμων ανεξάρτητων μεταβλητών για να προχωρήσουμε στην ερμηνεία των συντελεστών θα πρέπει να γνωρίζουμε την κατηγορία που χρησιμοποείται ως κατηγορία αναφοράς. Στο SPSS, για παράδειγμα, ως κατηγορία αναφοράς χρησιμοποιείται η κατηγορία που είναι κωδικοποιημένη με το μεγαλύτερο αριθμό. Συνεπώς, αν για το φύλο έχουμε άνδρας:1 και γυναίκα:0, η κατηγορία αναφοράς είναι ο άνδρας.

Αν, λοιπόν, ο συντελεστής είναι θετικός (ή Exp(B)>1) σημαίνει ότι οι γυναίκες έχουν μεγαλύτερη πιθανότητα να ανήκουν στην κατηγορία 1 ή 2 (ανάλογα με το μοντέλο που ερμηνεύουμε) έναντι της κατηγορίας 0 σε σχέση με του άνδρες, ενώ αν είναι αρνητικός έχουν μικρότερη πιθανότητα. Όσον αφορά στην ερμηνεία του σχετικού λόγου exp(B) ισχύουν αυτά που αναφέρονται παραπάνω.

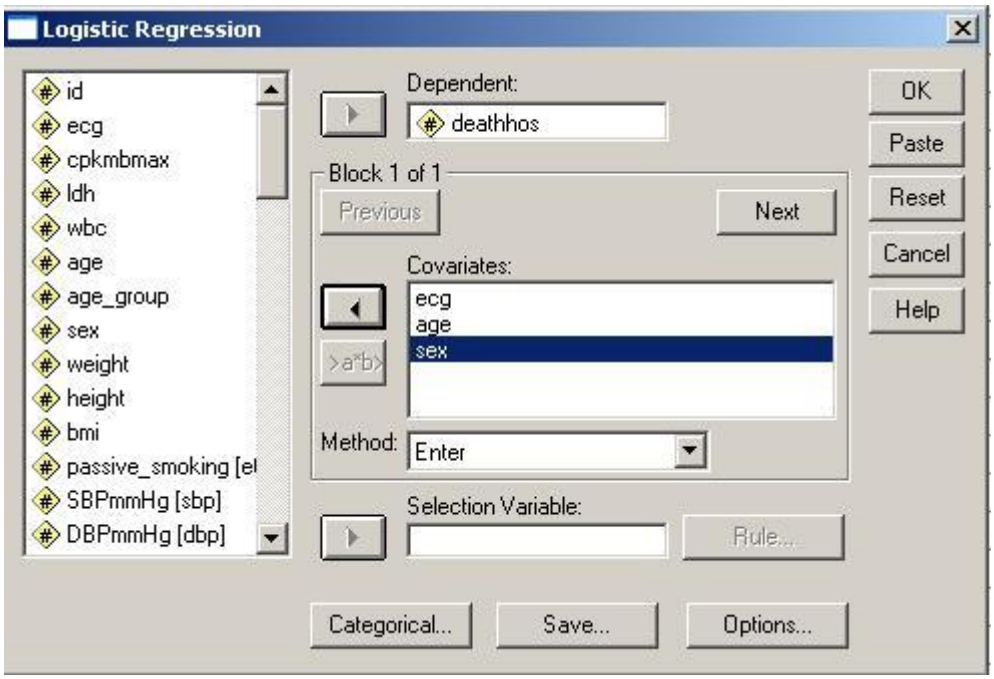
Όσον αφορά στον έλεγχο καλής προσαρμοστικότητας του μοντέλου ισχύει ότι ισχύει και στην περίπτωση της κλασσικής λογαριθμιστικής παλινδρόμησης. Επίσης, οι έλεγχοι υποθέσεων σε αυτή την μορφή παλινδρόμησης είναι αντίστοιχοι με αυτούς που πραγματοποιούνται στην κλασσική λογαριθμιστική παλινδρόμηση και πραγματοποιούνται εφαρμόζοντας τα ίδια στατιστικά κριτήρια.

10.4 Πολλαπλή λογαριθμιστική παλινδρόμηση με το SPSS

Ας υποθέσουμε ότι σκοπός της μελέτης μας είναι να διερευνήσουμε κατά πόσο το φύλο (sex), η ηλικία (age) και ο τύπος διάγνωσης του Οξέος Στεφανιαίου Συνδρόμου (ΟΣΣ) συσχετίζονται με την θνητότητα κατά τη διάρκεια της νοσηλείας (deathhos: όχι:0 και ναι:1). Η κατάλληλη στατιστική ανάλυση είναι η λογαριθμιστική παλινδρόμηση η οποία πραγματοποιείται ακολουθώντας τα εξής ***βήματα***:

Analyse → Regression → Binary Logistic

Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνα 10.1*.



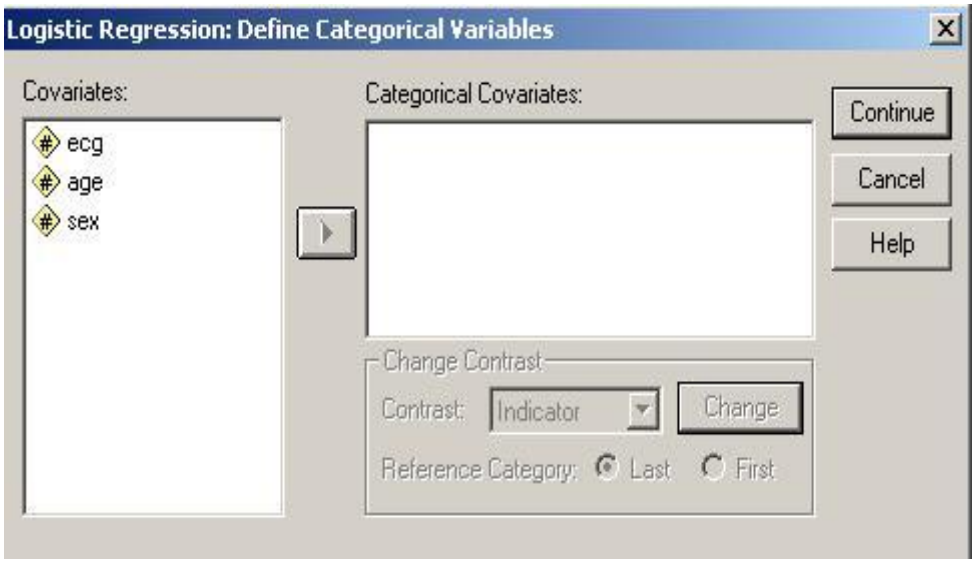
Εικόνα 10.1. Πραγματοποίηση λογαριθμικής παλινδρόμησης

i. ορίζουμε ως ***Dependent*** το δίτιμο χαρακτηριστικό/έκβαση (π.χ. ενδονοσοκομειακός θάνατος (Ναι/Όχι)), και

ii. ως ***Covariates*** ορίζουμε τους ανεξάρτητους παράγοντες (π.χ. ηλικία, φύλο και διάγνωση εισόδου).

iii. Στην περίπτωση που κάποιος από τους ανεξάρτητους παράγοντες είναι κατηγορική μεταβλητή, οφείλουμε να το δηλώσουμε πατώντας το κουμπί επιλογών ***Categorical*** και ανοίγει ένα νέο πλαίσιο διαλόγου (*Εικόνα 10.2*) δηλώνοντας τις κατηγορικές ανεξάρτητες μεταβλητές ως ***Categorical Covariates***. Στη συνέχεια πατάμε "***Continue***" και επιστρέφουμε στο παράθυρο διαλόγου της *Εικόνας 10.1*.

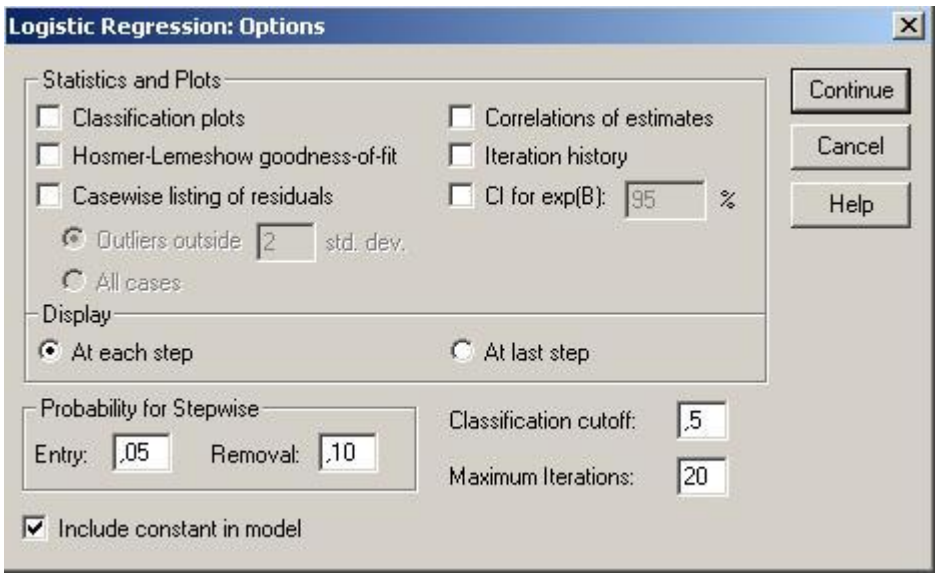
(***Σημείωση***: Στο SPSS είναι προεπιλεγμένο να χρησιμοποιείται η τελευταία κατηγορία της κάθε μεταβλητής ως κατηγορία αναφοράς.)



Εικόνα 10.2: Προσδιορισμός των ανεξάρτητων μεταβλητών που είναι κατηγορικές

iv. Επίσης, μέσω του κουμπιού επιλογών ***Options***, όπου ανοίγει το πλαίσιο διαλόγου της *Εικόνας 10.3* μπορούμε να επιλέξουμε να εμφανιστεί στα αποτελέσματα

- το διάστημα εμπιστοσύνης για το σχετικό λόγο (Odds Ratio = exp(B)), επιλέγοντας το «***CI for exp(B)***»,
- ο έλεγχος καλής προσαρμοστικότητας του μοντέλου επιλέγοντας την επιλογή «***Hosmer-Lemeshow goodness-of-fit***», και
- ο πίνακας με τα αποτελέσματα των ποσοστών σωστής ταξινόμησης των ατόμων του δείγματος με τη βοήθεια του παραπάνω μοντέλου , επιλέγοντας την επιλογή «***Classification Plots***» **.**



Εικόνα 10.3: Επιλογή εμφάνισης των διαστημάτων εμπιστοσύνης για τα OR και πραγματοποίηση του ελέγχου καλής προσαρμοστικότητας του μοντέλου (Κουμπί επιλογών «Options»)

v. Οι *Πίνακες 10.1, 10.2 & 10.3* παρουσιάζουν τα αποτελέσματα που προέκυψαν από την παραπάνω ανάλυση. Πιο συγκεκριμένα, από τον *Πίνακα 10.1* παρατηρούμε ότι το φύλο δεν συσχετίζεται στατιστικά σημαντικά με την πιθανότητα ενδονοσοκομειακού θανάτου, αφού η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης (p) είναι μεγαλύτερη από το α = 0,05 (sig. = 0,157). Αντίθετα, παρατηρούμε ότι η ηλικία συσχετίζεται στατιστικά σημαντικά με την πιθανότητα ενδονοσοκομειακού θανάτου, αφού sig = 0,000….1< α = 0,05. Συγκεκριμένα, φαίνεται ότι αύξηση της ηλικίας κατά ένα έτος οδηγεί σε αύξηση της πιθανότητας για θάνατο κατά περίπου 7% (exp(B) για το age=1,066), για συγκεκριμένο φύλο και συγκεκριμένη διάγνωση εισόδου. Επίσης, βλέπουμε ότι τα άτομα με διάγνωση εισόδου "STEMI" (ecg(1)) έχουν 4,669 φορές μεγαλύτερη πιθανότητα να πεθάνουν μέσα στο νοσοκομείο συγκριτικά με αυτούς με διάγνωση εισόδου "other" (κατηγορία αναφοράς), αφού exp(B) για ecg(1)= 4,669, ενώ αυτοί με διάγνωση εισόδου "NSTEMI" δεν φαίνεται να έχουν στατιστικά σημαντικά μεγαλύτερη πιθανότητα θανάτου συγκριτικά με αυτούς με διάγνωση "other" αφού sig. = 0,158 > α = 0,05.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1[a]	ecg			26,450	2	,000			
	ecg(1)	1,541	,332	21,488	1	,000	4,669	2,434	8,958
	ecg(2)	,550	,390	1,989	1	,158	1,733	,807	3,719
	age	,064	,011	33,294	1	,000	1,066	1,043	1,090
	sex	-,370	,261	2,007	1	,157	,690	,414	1,153
	Constant	-8,377	,928	81,480	1	,000	,000		

a. Variable(s) entered on step 1: ecg, age, sex.

Πίνακας 10.1: Αποτελέσματα πολλαπλής λογαριθμικής παλινδρόμησης με εξαρτημένη μεταβλητή τον ενδονοσοκομειακό θάνατο και ανεξάρτητες μεταβλητές το φύλο (males), την ηλικία (age) και την διάγνωση εισόδου (ecg).

Ο *Πίνακας 10.2* δίνει τα αποτελέσματα του ελέγχου Hosmer-Lemeshow. Ο συγκεκριμένος έλεγχος δίνει ένα μέτρο του βαθμού συμφωνίας μεταξύ των παρατηρούμενων τιμών της εξαρτημένης μας μεταβλητής στο δείγμα μας και των προβλεπόμενων με τη βοήθεια του συγκεκριμένου μοντέλου. Η μηδενική υπόθεση γι' αυτό τον έλεγχο είναι ότι το μοντέλο είναι καλό, δηλαδή ότι οι παρατηρούμενες τιμές συμπίπτουν με τις προβλεπόμενες. Άρα, από τον *Πίνακα* διαπιστώνουμε ότι το παρόν μοντέλο είναι ικανοποιητικό αφού sig.=0,170>α=0,05.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	11,608	8	,170

Πίνακας 10.2: Έλεγχος καλής προσαρμοστικότητας του μοντέλου

Τέλος, ο *Πίνακας 10.3* συνοψίζει τα αποτελέσματα της πρόβλεψης των τιμών της εξαρτημένης μεταβλητής χρησιμοποιώντας το συγκεκριμένο μοντέλο. Πιο συγκεκριμένα, παρατηρούμε ότι το ποσοστό σωστής ταξινόμησης είναι 96,1%. Αν και μεταξύ των ατόμων που δεν πέθαναν μέσα στο νοσοκομείο το ποσοστό ορθής πρόβλεψης χρησιμοποιώντας το συγκεκριμένο μοντέλο είναι 100%, το ποσοστό πρόβλεψης μεταξύ αυτών που πέθαναν είναι 0%. Συνεπώς, το συγκεκριμένο μοντέλο είναι ένα πολύ κακό μοντέλο όσον αφορά στον προσδιορισμό των ατόμων υψηλού κινδύνου για ενδο-νοσοκομειακό θάνατο μεταξύ αυτών που εισήχθησαν με ΟΣΣ.

Classification Table[a]

			Predicted		
			deathhos		Percentage
	Observed		0	1	Correct
Step 1	deathhos	0	1865	0	100,0
		1	75	0	,0
	Overall Percentage				96,1

a. The cut value is ,500

Πίνακας 10.3: Πίνακας ταξινόμησης

10.5 Πολλαπλή λογαριθμιστική παλινδρόμηση με το STATA

Η λογαριθμιστική παλινδρόμηση στο STATA πραγματοποιείται ακολουθώντας τα εξής ***βήματα***:

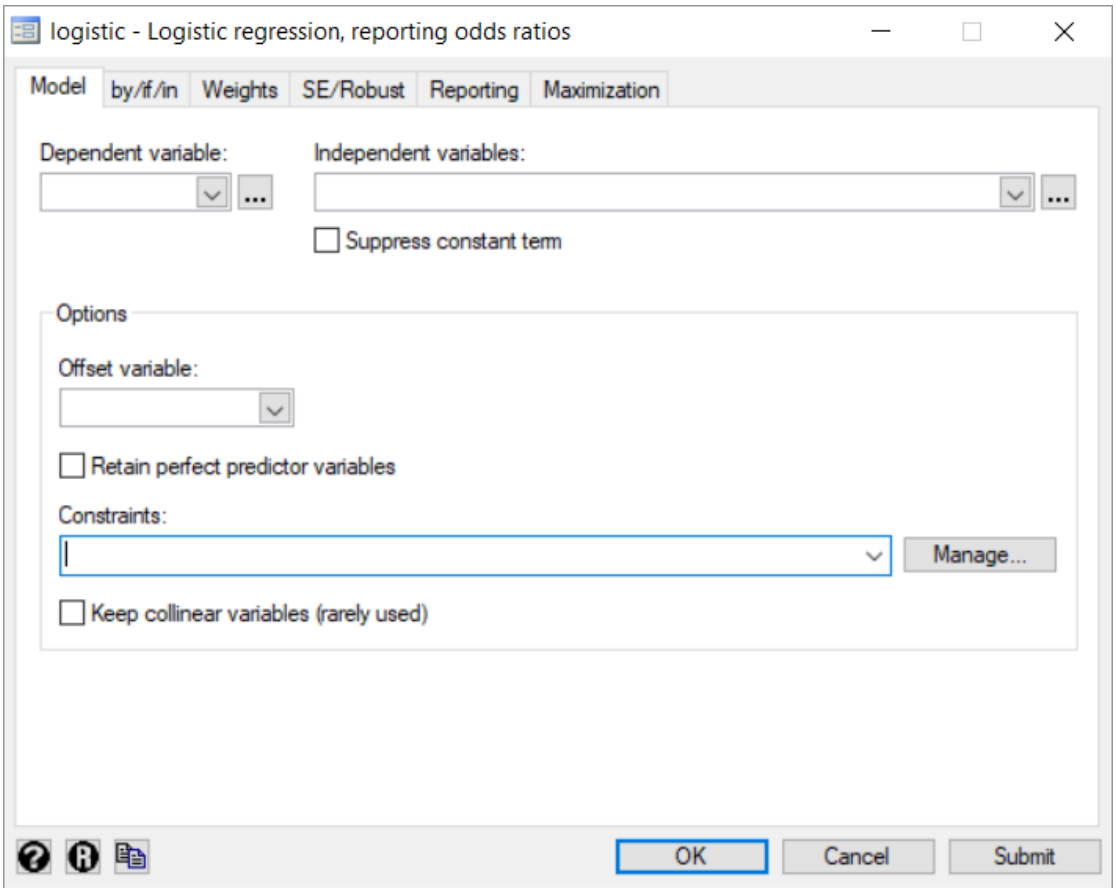
Statistics → Binary outcomes → Logistic regression (reporting odds ratios)

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνα 10.4.*

ii. Ορίζουμε ως ***Dependent variable*** το δίτιμο χαρακτηριστικό/έκβαση, και

iii. ως ***Independent variables*** ορίζουμε τους ανεξάρτητους παράγοντες.

iv. *«**OK**»*



Εικόνα 10.4. Πραγματοποίηση λογαριθμικής παλινδρόμησης

Στο STATA χρησιμοποιείται η εντολή ***logistic*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 10.5*:

Syntax

```
logistic depvar indepvars [if] [in] [weight] [, options]
```

options	Description
Model	
noconstant	suppress constant term
offset(*varname*)	include *varname* in model with coefficient constrained to 1
asis	retain perfect predictor variables
constraints(*constraints*)	apply specified linear constraints
collinear	keep collinear variables
SE/Robust	
vce(*vcetype*)	*vcetype* may be oim, robust, cluster *clustvar*, bootstrap, or jackknife
Reporting	
level(#)	set confidence level; default is level(95)
coef	report estimated coefficients
nocnsreport	do not display constraints
display_options	control column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
Maximization	
maximize_options	control the maximization process; seldom used
coeflegend	display legend instead of statistics

Εικόνα 10.5: Βασική σύνταξη της εντολής logistic

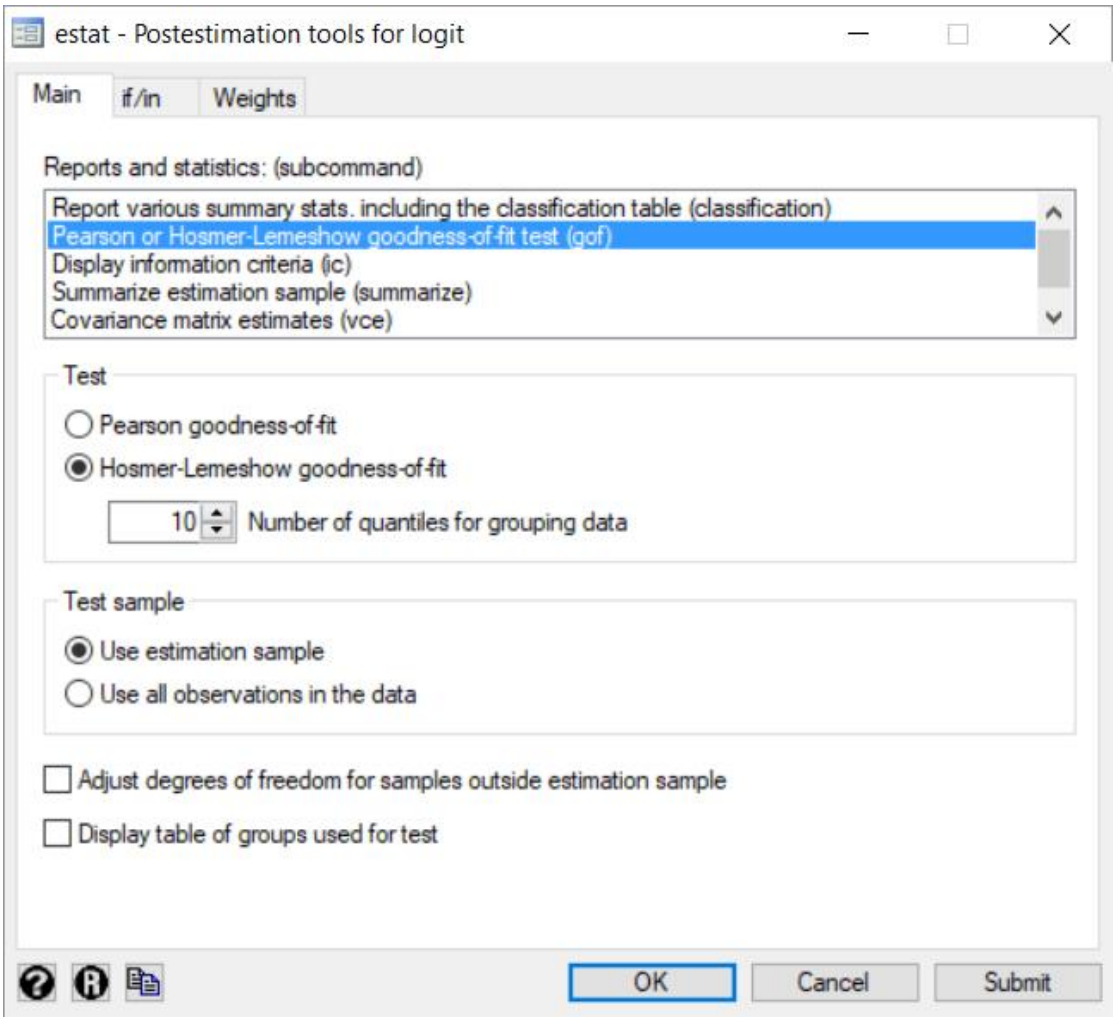
Ο έλεγχος καλής προσαρμοστικότητας ***Hosmer-Lemeshow*** του μοντέλου καθώς και άλλοι διαγνωστικοί έλεγχοι γίνονται ακολουθώντας τα εξής ***βήματα***:

Statistics → Binary outcomes → Postestimation → Goodness - of – fit after logistic/logit/probit

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνα 10.6.*

ii. Τσεκάρουμε την επιλογή «***Hosmer – Lemeshow goodness of fit***»

iii. «***OK***»



Εικόνα 10.6: Παράθυρο επιλογών διαγνωστικών εργαλείων

Στο STATA χρησιμοποιείται η εντολή ***estat gof*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 10.7*:

Syntax

```
estat gof [if] [in] [weight] [, options]
```

options	Description
Main	
`group(#)`	perform Hosmer–Lemeshow goodness-of-fit test using # quantiles
`all`	execute test for all observations in the data
`outsample`	adjust degrees of freedom for samples outside estimation sample
`table`	display table of groups used for test

`fweights` are allowed; see [U] **11.1.6 weight**.
`estat gof` is not appropriate after the `svy` prefix.

Menu for estat

Statistics > Postestimation > Reports and statistics

Description

`estat gof` reports the Pearson goodness-of-fit test or the Hosmer–Lemeshow goodness-of-fit test.

`estat gof` requires that the current estimation results be from `logistic`, `logit`, or `probit`; see [R] **logistic**, [R] **logit**, or [R] **probit**. For `estat gof` after `poisson`, see [R] **poisson postestimation**. For `estat gof` after `sem`, see [SEM] **estat gof**.

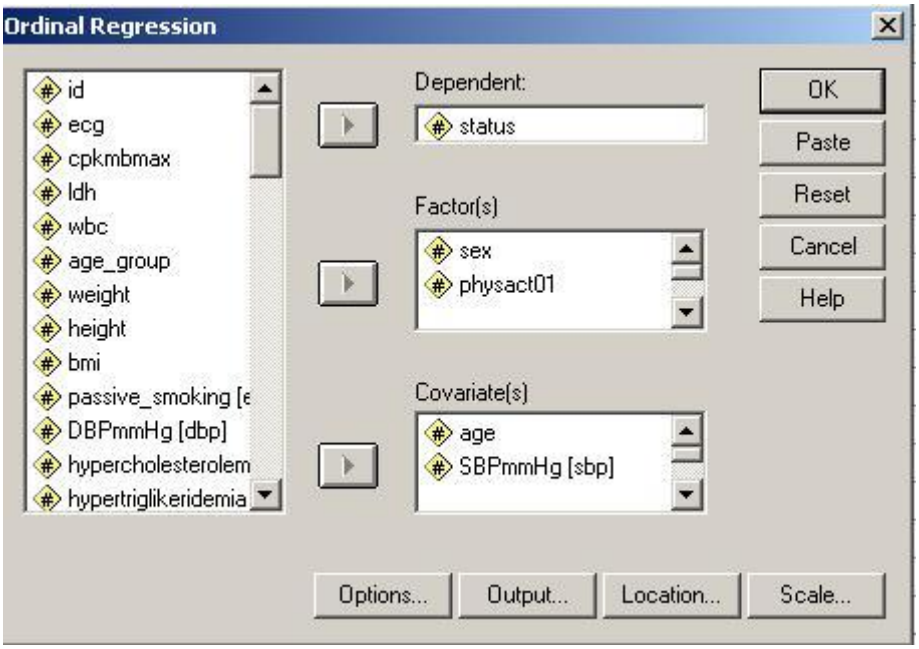
Εικόνα 10.7: Βασική σύνταξη της εντολής estat gof

10.6 Διατάξιμη λογαριθμιστική παλινδρόμηση με το SPSS

Ας υποθέσουμε ότι στόχος μας είναι να διερευνήσουμε κατά πόσο το φύλο (sex), η ηλικία (age), η συστολική αρτηριακή πίεση (SBP) και η σωματική δραστηριότητα (physact01) συσχετίζονται με την βαρύτητα του οξέος στεφανιαίου συνδρόμου (ΟΣΣ). Η βαρύτητα του ΟΣΣ είναι μία διατάξιμη κατηγορική μεταβλητή με 3 κατηγορίες (0: Ασταθής στηθάγχη, 1: non-Q έμφραγμα μυοκαρδίου (ΕΜ) και 2: Q ΕΜ). Συνεπώς, η κατάλληλη στατιστική ανάλυση είναι η διατάξιμη λογαριθμιστική παλινδρόμηση, η οποία πραγματοποιείται εφαρμόζοντας τα εξής ***βήματα***:

Analyse → Regression → Ordinal …

ii. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνα 10.8.*



Εικόνα 10.8: Πραγματοποίηση διατάξιμης λογαριθμιστικής παλινδρόμησης

iii. Ορίζουμε ως «***Dependent***» την διατάξιμη κατηγορική μεταβλητή/έκβαση (π.χ. status),

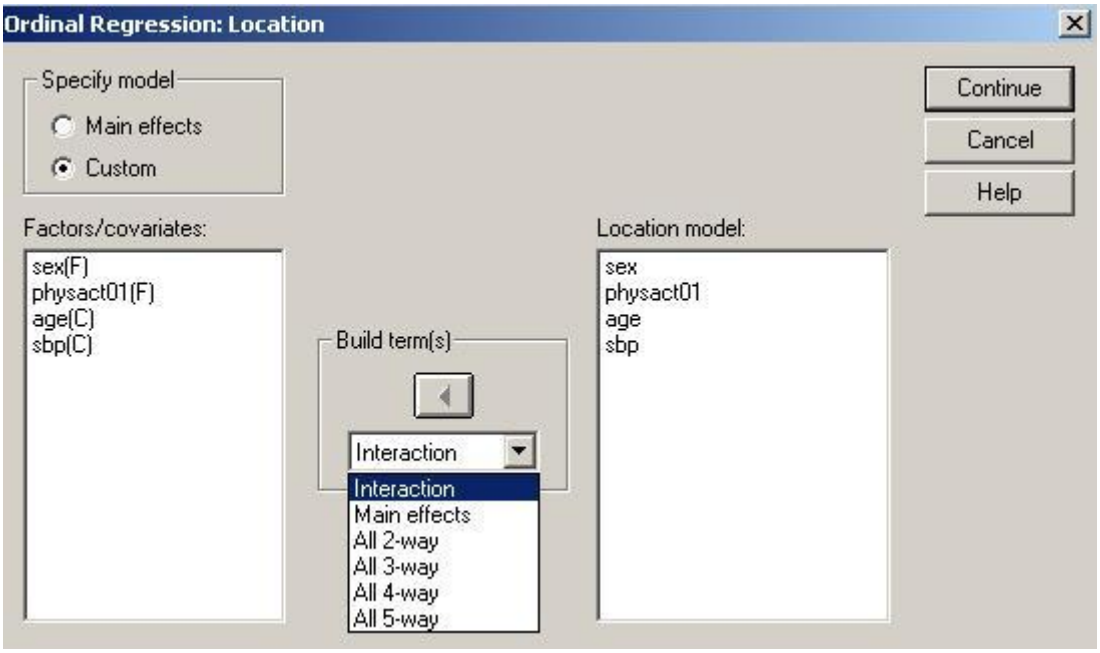
iv. Ως «***Covariates***» ορίζουμε τις ανεξάρτητες ποσοτικές μεταβλητές.

v. Ως «***Factors***» ορίζουμε τις ανεξάρτητες κατηγορικές μεταβλητές.

vi. Στη συνέχεια πατάμε το κουμπί επιλογών «***Location***» και εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 10.9.*

vii. Σε αυτό το πλαίσιο διαλόγου μπορούμε να δηλώσουμε τη μορφή του μοντέλου μας. Προεπιλεγμένο είναι το «Main effects», δηλαδή η «κυρίως επιδράσεις» των ανεξάρτητων μεταβλητών. Αν δεν αλλάξουμε αυτή την επιλογή, το μοντέλο που θα προκύψει θα περιέχει μόνο τις κυρίως επιδράσεις των ανεξάρτητων μεταβλητών που δηλώσαμε στο Covariates και Factors της *Εικόνας 10.4*. Μπορούμε, όμως, να επιλέξουμε το «Custom» όπου και ενεργοποιούνται τα υπόλοιπα κουμπιά του πλαισίου. Έτσι, λοιπόν, μπορούμε να επιλέξουμε να συμπεριληφθούν στο μοντέλο και διάφοροι όροι αλληλεπίδρασης («interactions») μεταξύ των ανεξάρτητων μεταβλητών μας, χρησιμοποιώντας το κουμπί πολλαπλών επιλογών που υπάρχει κάτω από το

«Build terms». (Στο συγκεκριμένο παράδειγμα επιλέξαμε να εμφανιστούν μόνο τα main effects των ανεξάρτητων μεταβλητών).



Εικόνα 10.9: Επιλογή της δομής του μοντέλου (π.χ. μόνο main effects, main effects και interaction)

viii. Πατώντας το κουμπί επιλογών «***Output***» ανοίγει το πλαίσιο διαλόγου της *Εικόνας 10.10*. Σε αυτό το πλαίσιο διαλόγου μπορούμε να επιλέξουμε τα στοιχεία που επιθυμούμε να εμφανιστούν στο output του SPSS (***Display***) και τα στοιχεία που επιθυμούμε να αποθηκευτούν στο αρχείο μας ως νέες μεταβλητές (***Saved variables***). (Στο παράδειγμά μας, επιλέξαμε τα «Goodness of fit statistics», «summary statistics», «Parameter estimates» & «Test of parallel lines»).

Display. Δίνει τους εξής Πίνακες:

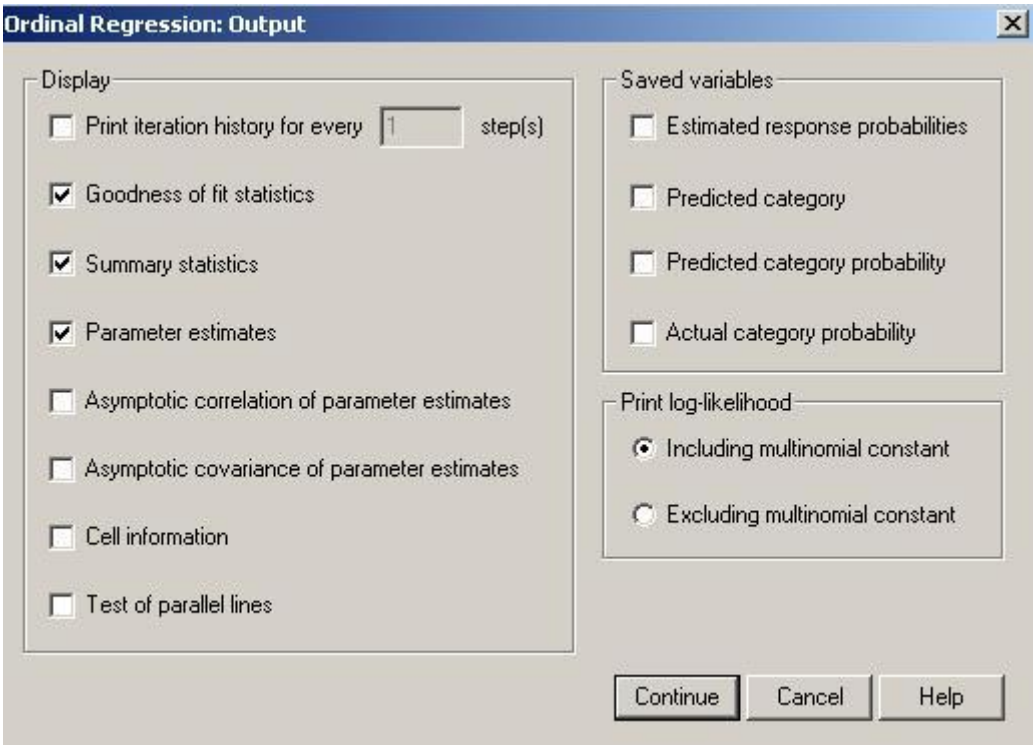
- **Print iteration history**. Η log-likelihood και οι εκτιμήσεις των παραμέτρων για την πρώτη και την τελευταία επανάληψη της διαδικασίας εκτίμησης των παραμέτρων του μοντέλου.
- **Goodness-of-fit statistics**. Δίνει τα Pearson και likelihood-ratio X^2 στατιστικά.
- **Summary statistics**. Cox and Snell's, Nagelkerke's, and McFadden's R2 στατιστικά.
- **Parameter estimates**. Εκτιμήσεις των συντελεστών του μοντέλου, τυπικό σφάλμα και διαστήματα εμπιστοσύνης.
- **Asymptotic correlation of parameter estimates**. Πίνακας συσχετίσεων των εκτιμήσεων των συντελεστών.
- **Asymptotic covariance of parameter estimates**. Πίνακας συν-διακυμάνσεων των εκτιμήσεων των συντελεστών.
- **Cell information**. Παρατηρούμενες και προσδοκόμενες συχνότητες και αθροιστικές συχνότητες, κατάλοιπα του Pearson για τις συχνότητες και τις αθροιστικές συχνότητες, παρατηρούμενες και αναμενόμενες πιθανότητες, και παρατηρούμενες και αναμενόμενες αθροιστικές πιθανότητες για κάθε

κατηγορία της εξαρτημένης μεταβλητής ανά συνδυασμό των ανεξάρτητων μεταβλητών. Πρέπει να σημειωθεί ότι για μοντέλα με πολλούς συνδυασμούς των ανεξάρτητων μεταβλητών (π.χ. μοντέλο με πολλές συνεχείς ανεξάρτητες μεταβλητές), αυτή η επιλογή μπορεί να οδηγήσει σε έναν πολύ μεγάλο πίνακα.

- **Test of parallel lines**. Τεστ για την προϋπόθεση των αναλογικών odds.

Saved variables. Αποθηκεύονται οι εξής μεταβλητές:

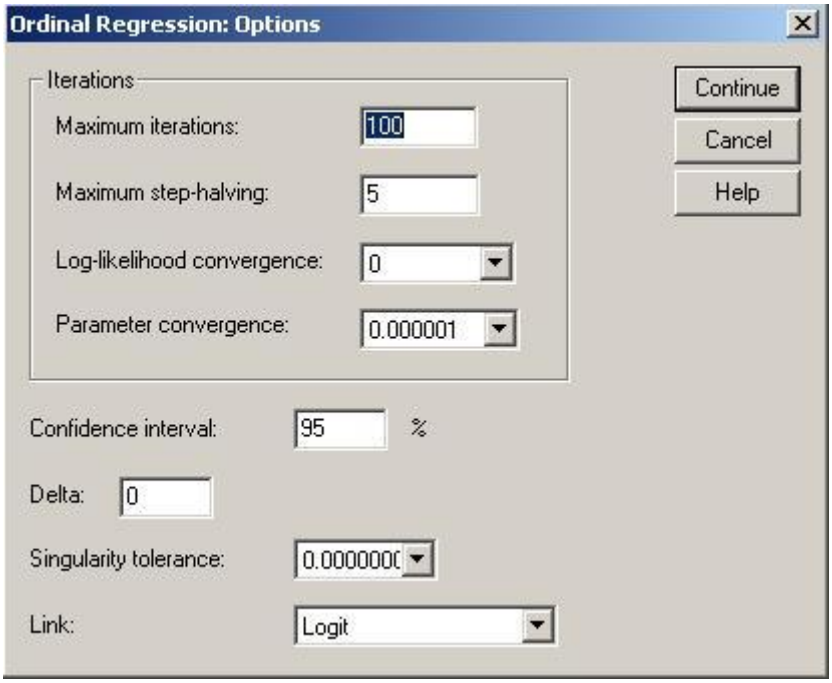
- **Estimated response probabilities**. Η εκτιμώμενες πιθανότητες βάσει του μοντέλου για την ταξινόμηση ενός συνδυασμού των ανεξάρτητων μεταβλητών στις κατηγορίες της εξαρτημένης μεταβλητής. Υπάρχουν τόσες πιθανότητες όσες και οι κατηγορίες της εξαρτημένης μεταβλητής.
- **Predicted category**. Η κατηγορία της εξαρτημένης μεταβλητής που έχει την μέγιστη εκτιμώμενη πιθανότητα για κάθε συνδυασμό των ανεξάρτητων μεταβλητών.
- **Predicted category probability**. Η εκτιμώμενη πιθανότητα ταξινόμησης ενός συνδυασμού ανεξάρτητων μεταβλητών στην προβλεπόμενη κατηγορία. **Actual category probability**. Η εκτιμώμενη πιθανότητα να ταξινομήσουμε έναν συνδυασμό των ανεξάρτητων μεταβλητών στην πραγματική κατηγορία.



Εικόνα 10.10: Επιλογή των στοιχείων που θα εμφανιστούν στο output και θα αποθηκευτούν ως νέες μεταβλητές.

ix. Πατώντας το κουμπί επιλογών «***Options***» εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 10.11*. Σε αυτό το πλαίσιο διαλόγου, μπορούμε να επιλέξουμε κάποια στοιχεία αναφορικά με τις επαναλήψεις που θα πραγματοποιηθουν προκειμένου να εκτιμηθούν οι συντελεστές του μοντέλου (επιλογές «Iterations»), το διάστημα εμπιστοσύνης των συντελεστών («Confidence interval»), το link function που θα χρησιμοποιήσουμε (προεπιλεγμένο είναι το

logit function) κτλ. Συνήθως, δεν χρειάζεται να τροποποιούμε κάποια από τις επιλογές αυτού του πλαισίου.



Εικόνα 10.11: Το πλαίσιο διαλόγου που ανοίγει πατώντας το κουμπί επιλογών «**Options**»

x. Τα αποτελέσματα της παραπάνω ανάλυσης για το παράδειγμά μας παρουσιάζονται στους *Πίνακες 10.8 – 10.11*.

Πιο συγκεκριμένα, στον *Πίνακα 10.4* παρουσιάζονται οι εκτιμήσεις των συντελεστών του μοντέλου. Οι γραμμές «***Threshold***» περιλαμβάνουν τις σταθερές του μοντέλου. Οι σταθερές του μοντέλου είναι τόσες όσα είναι και τα κατώφλια που χρησιμοποιούνται, δηλ. k-1, όπου k: ο αριθμός των κατηγοριών της εξαρτημένης μεταβλητής. Οι γραμμές «***Location***» περιλαμβάνουν τις εκτιμήσεις των συντελεστών για όλες τις ανεξάρτητες μεταβλητές και τα interactions (αν υπήρχαν). Για κάθε έναν από τους συντελεστές πραγματοποιείται και αναφέρεται και το p-value (sig.) από τον κατάλληλο έλεγχο χρησιμοποιώντας το Wald test. Επίσης, αναφέρεται και το αντίστοιχο διάστημα εμπιστοσύνης για κάθε συντελεστή (95% confidence interval). Αν το 0 συμπεριλαμβάνεται μέσα στο διάστημα εμπιστοσύνης, τότε δεν μπορούμε να υποθέσουμε ότι ο αντίστοιχος συντελεστής είναι διάφορος του 0 και το αντίστοιχο p-value θα είναι μη στατιστικά σημαντικό. Τέλος, στην περίπτωση των κατηγορικών μεταβλητών, συντελεστές εκτιμούνται για όλες τις κατηγορίες, εκτός από την κατηγορία αναφοράς. Όσον αφορά στην ***ερμηνεία των συντελεστών***, αξίζει να σημειωθεί ότι:

- Οι συντελεστές που παρουσιάζονται στον Πίνακα (estimate) δεν ερμηνεύονται ως σχετικός λόγος (Odds ratio) και το SPSS στην περίπτωση της διατάξιμης λογαριθμιστικής παλινδρόμησης δεν διαθέτει την επιλογή αυτόματου υπολογισμού του Odds Ratio όπως διαθέτει την λογαριθμιστική παλινδρόμηση για δίτιμες μεταβλητές. Συνεπώς, χρειάζεται προσοχή όσον αφορά στην ακριβή ερμηνεία των

αποτελεσμάτων. Στην περίπτωση που επιθυμούμε να αναφέρουμε πόσο ακριβώς αυξάνεται η πιθανότητα εμφάνισης υψηλότερων σκορ της διατάξιμης κατηγορικής μεταβλητής όταν μεταβάλλεται μία ανεξάρτητη θα πρέπει να υπολογίσουμε το αντίστοιχο Odds Ratio αντιλογαριθμίζοντας τον συντελεστή που εμφανίζεται στον Πίνακα «***Parameter estimates***».

Έτσι, λοιπόν, αναφορικά με το συγκεκριμένο παράδειγμα του *Πίνακα 10.4* διαπιστώνουμε ότι:

- οι γυναίκες (sex=0) έχουν στατιστικά σημαντικά μικρότερη πιθανότητα (p=0,001) εμφάνισης βαρύτερης μορφής ΟΣΣ (υψηλότερα σκορ έχουν χρησιμοποιηθεί για την κωδικοποίηση των βαρύτερων τύπων ΟΣΣ) σε σχέση με τους άνδρες (sex=1) αφού β=-0,389. Πιο συγκεκριμένα, οι γυναίκες έχουν περίπου 32% μικρότερη πιθανότητα έμφανισης βαρύτερης μορφής ΟΣΣ σε σχέση με τους άνδρες, αφού το Odds Ratio=exp(-0,389)=0,68.
- Αντίστοιχα, πραγματοποιείται και η ερμηνεία αναφορικά με την σωματική άσκηση. Πιο συγκεκριμένα, διαπιστώνουμε ότι αυτοί που δεν ασκούνται (physact01=0) είναι περίπου 33% πιο πιθανό να πάθουν βαρύτερη μορφή ΟΣΣ σε σχέση με αυτούς που ασκούνται (physact01=1), αφού Odds Ratio=exp(0,283)=1,33 και p=0,003.
- Όσον αφορά στη συστολική αρτηριακή πίεση, διαπιστώνουμε ότι αύξηση της πίεσης κατά 1 mm Hg συνεπάγεται μείωση 1% (Odds Ratio=exp(-0,007)=0,99) της πιθανότητας εμφάνισης βαρύτερης μορφής ΟΣΣ. Αν επιθυμούμε να περιγράψουμε το συγκεκριμένο αποτέλεσμα για αύξηση της πίεσης κατά 10 mm Hg, τότε η μείωση της πιθανότητας για βαρύτερη μορφή ΟΣΣ είναι $0,99^{10}$=0,90. Δηλαδή, αύξηση της πίεσης κατά 10 mm Hg συνεπάγεται μείωση της πιθανότητας εμφάνισης βαρύτερης μορφής ΟΣΣ κατά 10% περίπου.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[status = ,00]	-1,705	,342	24,914	1	,000	-2,375	-1,036
	[status = 1,00]	-,388	,339	1,313	1	,252	-1,053	,276
Location	[sex=0]	-,389	,112	12,087	1	,001	-,609	-,170
	[sex=1]	0[a]	.	.	0	.	.	.
	[physact01=0]	,283	,096	8,649	1	,003	,094	,471
	[physact01=1]	0[a]	.	.	0	.	.	.
	age	-,003	,004	,913	1	,339	-,011	,004
	sbp	-,007	,002	15,577	1	,000	-,011	-,004

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Πίνακας 10.4: Πίνακας με τις εκτιμήσεις των συντελεστών του μοντέλου.

Ο *Πίνακα 10.5* παρουσιάζει το αποτέλεσμα του ελέγχου υποθέσεων αναφορικά με το αν το συνολικό μοντέλο μας είναι καλύτερο από το μοντέλο που περιλαμβάνει μόνο την σταθερά (null model). Στην συγκεκριμένη περίπτωση, βλέπουμε ότι το μοντέλο μας είναι στατιστικά σημαντικά καλύτερο σε σχέση με το μοντέλο που δεν περιέχει καμία ανεξάρτητη μεταβλητή (sig.<0,001).

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	2962,206			
Final	2923,773	38,433	4	,000

Link function: Logit.

Πίνακας 10.5: Σύγκριση του μοντέλου με το μοντέλο που δεν περιλαμβάνει καμία ανεξάρτητη μεταβλητή.

Στον *Πίνακα 10.6* παρουσιάζονται κάποια περιληπτικά στατιστικά μέτρα του μοντέλου. Όλα αυτά τα μέτρα είναι ανάλογα του R-squared που υπολογίζεται στην γραμμική παλινδρόμηση, αλλά κανένα από αυτά δεν έχει την ερμηνεία του «ποσοστού της διακύμανσης που ερμηνεύεται» και δεν πρέπει να τα αναφέρουμε με αυτόν τον τρόπο. Αντίθετα, αυτά θα πρέπει να θεωρούνται ως επιπρόσθετα μέτρα του μεγέθους αποτελέσματος (effect size) του μοντέλου και συνεπώς όσο μεγαλύτερα τόσο το καλύτερο. Αυτά του *Πίνακα 10.6* εκφράζουν κακή προσαρμογή του μοντέλου. Όπως και σε άλλες τεχνικές, το μοντέλο μπορεί να είναι στατιστικά σημαντικά καλύτερο από το null μοντέλο, όμως, αυτό δεν σημαίνει απαραίτητα ότι και το effect size του μοντέλου είναι μεγάλο (π.χ. στην περίπτωση της γραμμικής παλινδρόμησης, μπορεί το μοντέλο να είναι στατιστικά σημαντικό, αλλά το R-squared να είναι πολύ μικρό). Συνεπώς, από κοινού η στατιστική σημαντικότητα του μοντέλου και αυτά τα μέτρα του effect size του μοντέλου θα πρέπει να ανφέρονται. Το Cox and Snell's R-Square είναι μία προσπάθεια να μυμηθούμε την ερμηνεία του R-square βάσει της πιθανοφάνειας, αλλά το μέγιστο (συνήθως) είναι μικρότερο της μονάδας και συνεπώς είναι δύσκολη η ερμηνεία. Το Nagelkerke's R-square είναι μία τροποποίηση του Cox and Snell συντελεστή έτσι ώστε να παίρνει τιμές μεταξύ 0 και 1. Συνεπώς, το Nagelkerke's R-square θα παίρνει φυσιολογικά υψηλότερες τιμές σε σχέση με το Cox and Snell συντελεστή. Το McFadden's R-square είναι ένα μέτρο της θεωρίας της πληροφορίας και ερμηνεύεται ως η μείωση της εντροπίας του μοντέλου σε σχέση με την εντροπία του μοντέλου με μόνο την σταθερά (null model).

Pseudo R-Square

Cox and Snell	,024
Nagelkerke	,027
McFadden	,011

Link function: Logit.

Πίνακας 10.6: Περιληπτικά στατιστικά μέτρα του μοντέλου (Summary statistics)

Στον *Πίνακα 10.7* παρουσιάζονται τα αποτελέσματα από τον έλεγχο που πραγματοποιείται προκειμένου να ελέγξουμε αν ισχύει η προϋπόθεση των αναλογικών odds. Η μηδενική υπόθεση είναι ότι ισχύει αυτή η προϋπόθεση. Συνεπώς, προκειμένου να συμπεράνουμε ότι ισχύει η προϋπόθεση θα πρέπει το αποτέλεσμα να είναι μη στατιστικά σημαντικό. Άρα, λοιπόν, όσον αφορά στο συγκεκριμένο παράδειγμα, διαπιστώνουμε ότι δεν ισχύει (sig<0,001). Σε αυτή την περίπτωση, η εναλλακτικές λύσεις που διαθέτει ο ερευνητής είναι να συνενώσει τις κατηγορίες της

διατάξιμης εξαρτημένης μεταβλητής μέχρι να πετύχει την αναλογικότητα των odds (αν υπάρχουν πολλές κατηγορίες) ή να εφαρμόσει τη λογαριθμιστική παλινδρόμηση για δίτιμες εξαρτημένες μεταβλητές (αν οι κατηγορίες είναι 3 οπότε με την συνένωση θα προκύψουν 2) ή τέλος, να εφαρμόσουν πολυωνυμική λογαριθμιστική παλινδρόμηση, με την οποία χάνουμε σε ισχύ των αποτελεσμάτων, αλλά τουλάχιστον αποφεύγουμε το πρόβλημα αυτής της σοβαρής προϋπόθεσης.

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	2923,773			
General	2883,195	40,578	4	,000

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

Πίνακας 10.7: Έλεγχος για την προϋπόθεση των αναλογικών odds

10.7 Διατάξιμη λογαριθμιστική παλινδρόμηση με το STATA

Για την διεξαγωγή μιας διατάξιμης λογαριθμιστικής παλινδρόμησης, τα ***βήματα*** που ακολουθούμε είναι τα εξής:

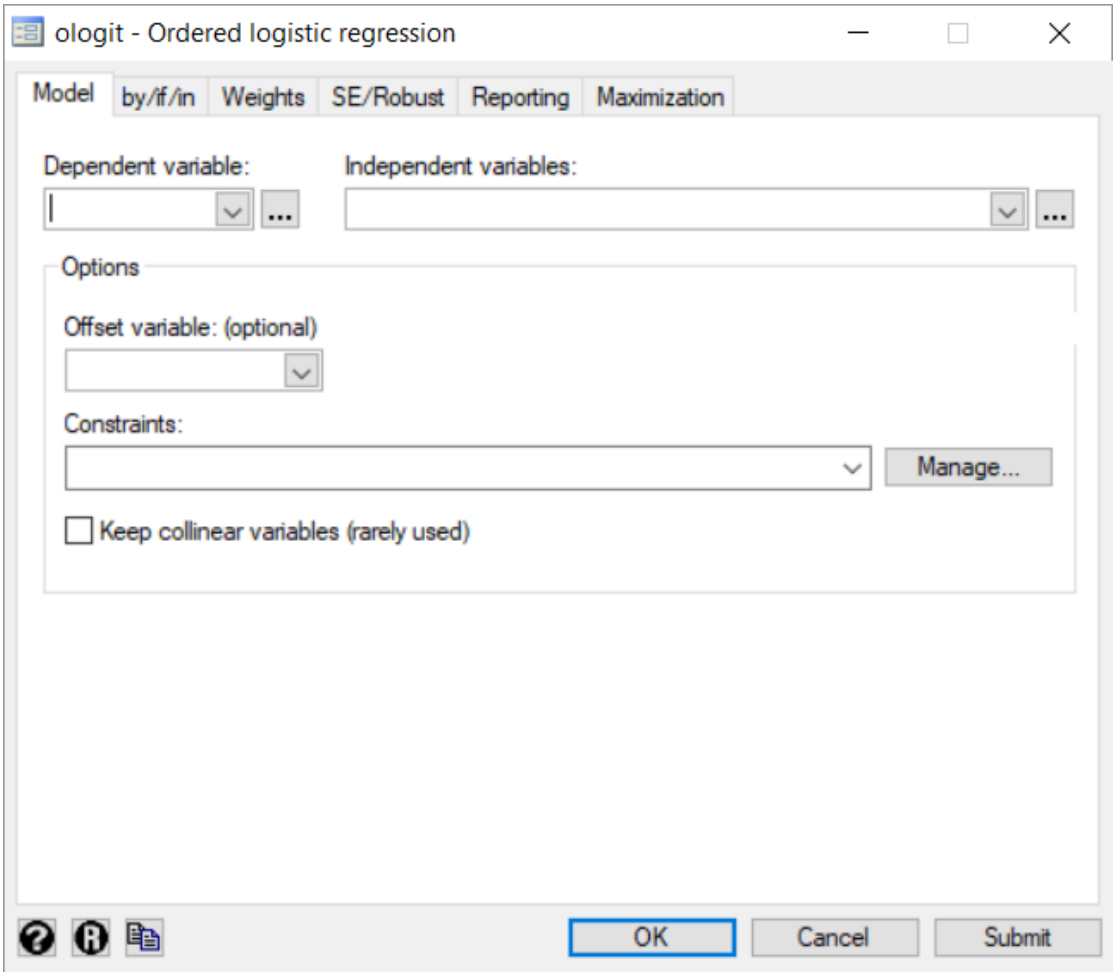
Statistics → Ordinal outcomes → Ordered logistic regression

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνα 10.12.*

ii. Ορίζουμε ως «***Dependent variable»*** την διατάξιμη κατηγορική μεταβλητή/έκβαση,

iii. Ως «***Independent variables»*** ορίζουμε τις ανεξάρτητες ποσοτικές μεταβλητές.

iv. «***OK***»



Εικόνα 10.12: Πραγματοποίηση διατάξιμης λογαριθμιστικής παλινδρόμησης

Στο STATA χρησιμοποιείται η εντολή ***ologit*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 10.13*:

Syntax

```
estat gof [if] [in] [weight] [, options]
```

options	Description
Main	
group(#)	perform Hosmer–Lemeshow goodness-of-fit test using # quantiles
all	execute test for all observations in the data
outsample	adjust degrees of freedom for samples outside estimation sample
table	display table of groups used for test

fweights are allowed; see [U] 11.1.6 weight.
estat gof is not appropriate after the svy prefix.

Menu for estat

Statistics > Postestimation > Reports and statistics

Description

estat gof reports the Pearson goodness-of-fit test or the Hosmer–Lemeshow goodness-of-fit test.

estat gof requires that the current estimation results be from logistic, logit, or probit; see [R] logistic, [R] logit, or [R] probit. For estat gof after poisson, see [R] poisson postestimation. For estat gof after sem, see [SEM] estat gof.

Syntax

```
ologit depvar [indepvars] [if] [in] [weight] [, options]
```

options	Description
Model	
offset(*varname*)	include *varname* in model with coefficient constrained to 1
constraints(*constraints*)	apply specified linear constraints
collinear	keep collinear variables
SE/Robust	
vce(*vcetype*)	*vcetype* may be oim, robust, cluster *clustvar*, bootstrap, or jackknife
Reporting	
level(#)	set confidence level; default is level(95)
or	report odds ratios
nocnsreport	do not display constraints
display_options	control column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
Maximization	
maximize_options	control the maximization process; seldom used
coeflegend	display legend instead of statistics

Εικόνα 10.13: Βασική σύνταξη της εντολής ologit

10.8 Πολυωνυμική λογαριθμιστική παλινδρόμηση με το SPSS

Ας θεωρήσουμε πάλι το παράδειγμα της ενότητας 10.4. Όπως είδαμε από τον *Πίνακα 10.7* η βασική προϋπόθεση για την ορθή εφαρμογή της διατάξιμης λογαριθμιστικής παλινδρόμησης δεν ισχύει και ένας από τους πιθανούς τρόπους να διαχειριστούμε τα παραπάνω δεδομένα χωρίς να παραβιάζεται καμία βασική προϋπόθεση, είναι να εφαρμόσουμε πολυωνυμική λογαριθμιστική παλινδρόμηση (multinomial logistic regression), η οποία πραγματοποιείται ακολουθώντας τα εξής ***βήματα***:

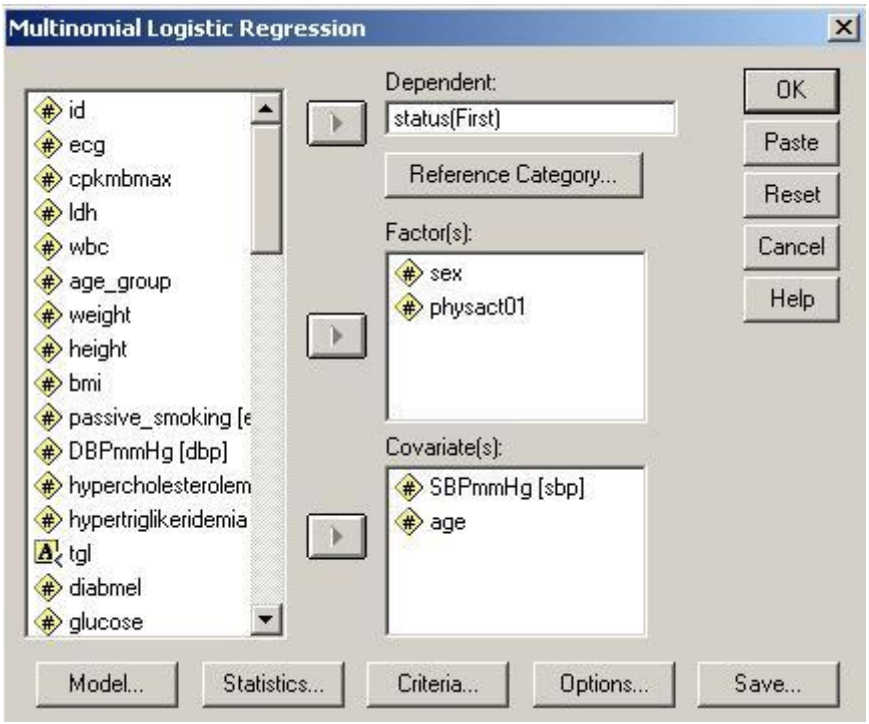
Analyse → Regression → Multinomial Logistic…

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνα 10.14.*

ii. Ορίζουμε ως «***Dependent***» την κατηγορική μεταβλητή/έκβαση (π.χ. status),

iii. Ως «***Covariates***» ορίζουμε τις ανεξάρτητες ποσοτικές μεταβλητές.

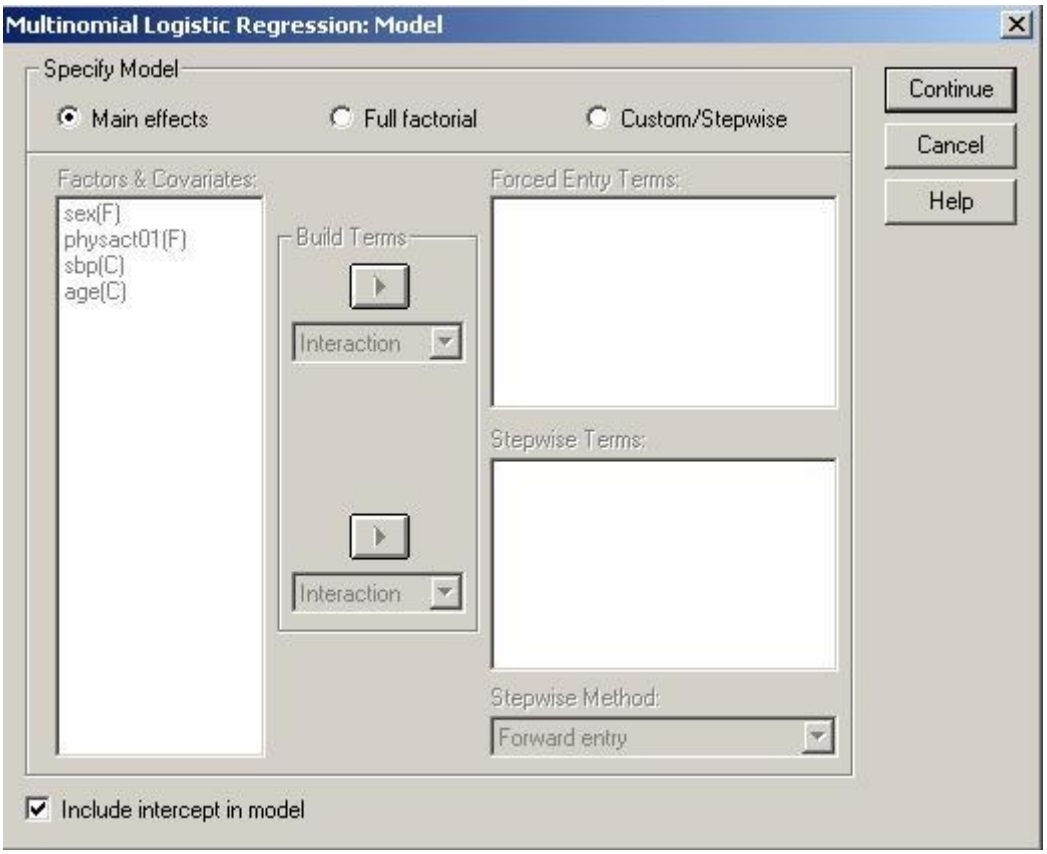
iv. Ως «***Factors***» ορίζουμε τις ανεξάρτητες κατηγορικές μεταβλητές.



Εικόνα 10.14: Πραγματοποίηση πολυωνυμικής λογαριθμιστικής παλινδρόμησης

v. Στη συνέχεια πατάμε το κουμπί επιλογών «***Model***» και εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 10.15*. Σε αυτό το πλαίσιο διαλόγου μπορούμε να ορίσουμε τι θα περιλαμβάνει το μοντέλο μας. Προεπιλεγμένο στο SPSS είναι το «main-effects model» που περιέχει μόνο τις κυρίως επιδράσεις των ανεξάρτητων μεταβλητών μας (συνεχείς και κατηγορικές) αλλά κανέναν όρο αλληλεπίδρασης μεταξύ των ανεξάρτητων μεταβλητών (interaction effect). Μπορούμε, επίσης, να επιλέξουμε το «full factorial model» που περιέχει όλα τις κυρίως επιδράσεις των ανεξάρτητων μεταβλητών μας αλλά και όλους τους πιθανούς όρους αλληλεπίδρασης μεταξύ των κατηγορικών μεταβλητών, αλλά όχι όρου αλληλεπίδρασης που να εμπλέκουν και τις συνεχείς μεταβλητές. Τέλος, μπορούμε να επιλέξουμε να δημιουργήσουμε ένα «custom model» στο

οποίο μπορούμε να προσδιορίσουμε μεταξύ ποιων παραγόντων (ποιοτικών ή συνεχών) τους όρους αλληλεπίδρασης επιθυμούμε να συμπεριλάβουμε στο μοντέλο ή να προσδιορίσουμε την βηματική διαδικασία (stepwise procedure) ως μέθοδος επιλογής βέλτιστου μοντέλου. (Στο συγκεκριμένο παράδειγμα θα έχουμε επιλέξει το μοντέλο να συμπεριλαμβάνει μόνο τις κυρίως επιδράσεις των ανεξάρτητων μεταβλητών μας).



Εικόνα 10.15: Προσδιορισμός των στοιχείων που θα περιλαμβάνει το μοντέλο.

vi. Στη συνέχεια πατάμε το κουμπί επιλογών «***Statistics***» και εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 10.16*. Σε αυτό το πλαίσιο διαλόγου, μπορούμε να ορίσουμε ποια στοιχεία επιθυμούμε να εμφανιστούν στο output του SPSS. Από τις πιο σημαντικές επιλογές που διαθέτει το συγκεκριμένο πλαίσιο είναι αυτές αναφορικά με το συνολικό μοντέλο και τις παραμέτρους που εκτιμούνται στο συγκεκριμένο μοντέλο και οι οποίες αναφέρονται παρακάτω:

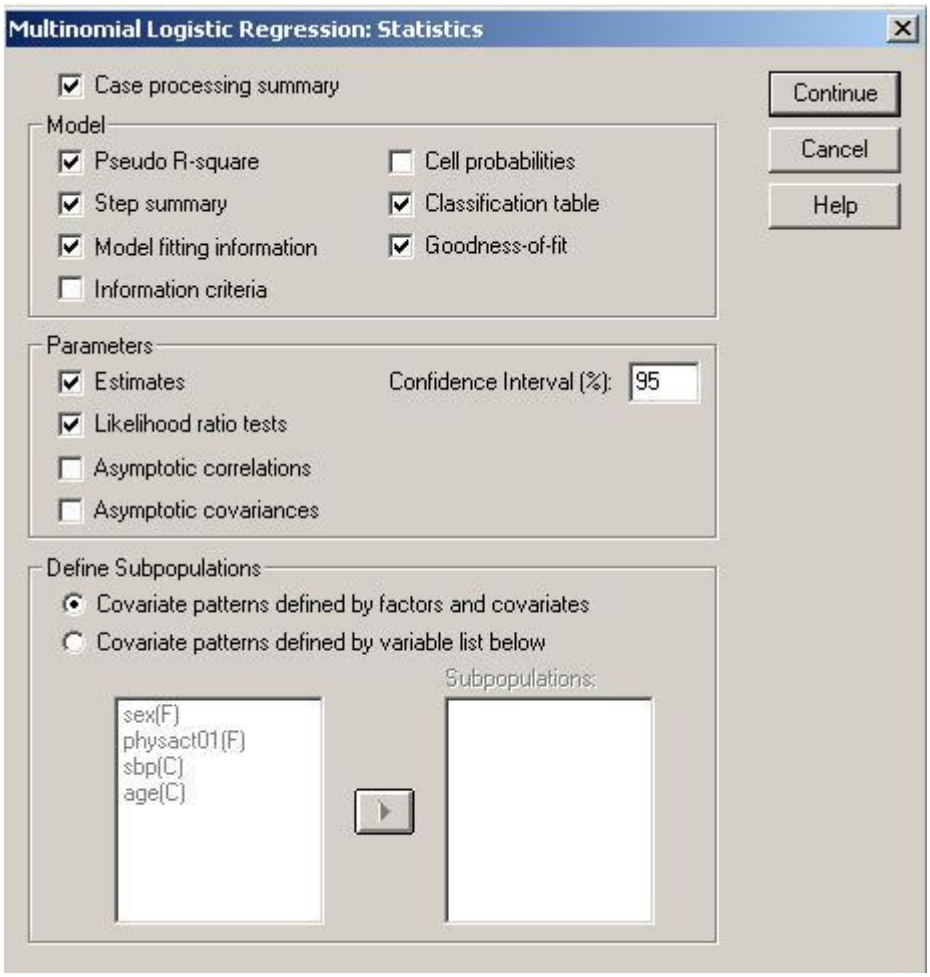
Model. Στατιστικά μέτρα που αφορούν ολόκληρο το μοντέλο.

- **Summary statistics**. Δίνει τα Cox and Snell, Nagelkerke, and McFadden R2 στατιστικά.
- **Step summary**. Αυτός ο Πίνακας περιγράφει το κάθε μοντέλο που προκύπτει κατά την βηματική διαδικασία. Αυτός ο Πίνακας δεν έχει νόημα και δεν δίνεται αν δεν έχει οριστεί από το πλαίσιο διαλόγου «***Model***» να δημιουργηθεί το μοντέλο ακολουθώντας την συγκεκριμένη διαδικασία.
- **Model fitting information**. Αυτός ο Πίνακας συγκρίνει το μοντέλο που προσαρμόσαμε με το μοντέλο που περιέχει μόνο την σταθερά (null model).

- **Information criteria**. Αυτός ο Πίνακας δίνει το Akaike's information criterion (AIC) και το Schwarz's Bayesian information criterion (BIC).
- **Cell probabilities**. Δίνει έναν Πίνακα με τις παρατηρούμενες και τις αναμενόμενες συχνότητες ανά συνδυασμό των ανεξάρτητων μεταβλητών και κατηγορία εξαρτημένης μεταβλητής.
- **Classification table**. Δίνει έναν Πίνακα με τις παρατηρούμενες έναντι των προβλεπόμενων τιμών.
- **Goodness of fit chi-square statistics**. Δίνει τα Pearson και likelihood-ratio chi-square στατιστικά.

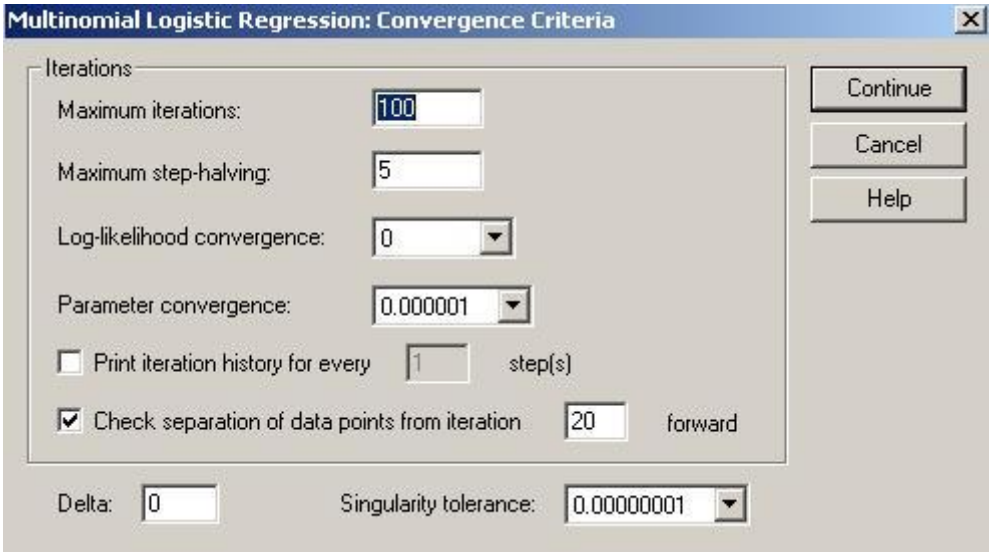
Parameters. Στατιστικά μέτρα που αφορούν την εκτίμηση των παραμέτρων.

- **Estimates**. Δίνει τον Πίνακα με τις εκτιμήσεις των παραμέτρων των συντελεστών του μοντέλου.
- **Likelihood ratio test**. Δίνει τα likelihood-ratio tests για τα μοντέλα με έναν έναν τον κάθε ανεξάρτητο παράγοντα του μοντέλου.
- **Asymptotic correlations**. Δίνει τον Πίνακα συσχετίσεων των παραμέτρων του μοντέλου που εκτιμήθηκαν.
- **Asymptotic covariances**. Δίνει τον Πίνακα συν-διακυμάνσεων των παραμέτρων του μοντέλου που εκτιμήθηκαν.



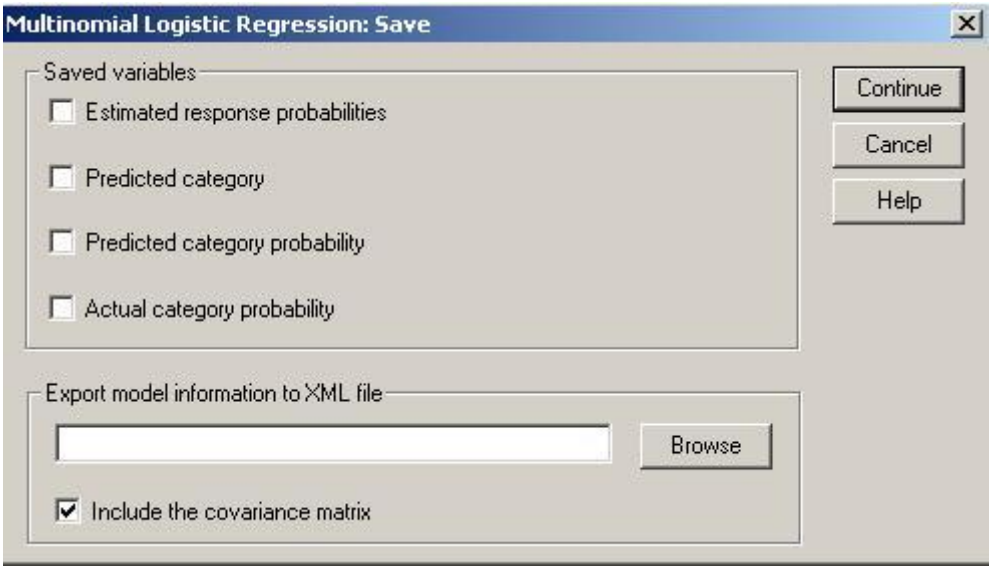
Εικόνα 10.16: Προσδιορισμός των στατιστικών μέτρων που επιθυμούμε να εμφανιστούν στο output του SPSS.

vii. Πατώντας το κουμπί επιλογών «***Criterion***» ανοίγει το πλαίσιο διαλόγου της *Εικόνας 10.17*. Σε αυτό το πλαίσιο μπορούμε να ορίσουμε διάφορα χαρακτηριστικά της επαναλληπτικής διαδικασίας που πραγματοποιείται προκειμένου να εκτιμηθούν οι συντελεστές του μοντέλου. Συνήθως, δεν χρειάζεται να αλλάξουμε κάτι στο συγκεκριμένο πλαίσιο διαλόγου.



Εικόνα 10.17: Προσδιορισμός των χαρακτηριστικών της επαναλληπτικής διαδικασίας που ακολουθείται για την εκτίμηση των παραμέτρων του μοντέλου.

viii. Αν επιθυμούμε να αποθηκεύσουμε κάποια από τα αποτελέσματα της παραπάνω ανάλυσης ως νέες μεταβλητές του αρχείου, πατάμε το κουμπί επιλογών «***Save***» και ανοίγει το πλαίσιο της *Εικόνας 10.12*.



Εικόνα 10.17: Επιλογές για την αποθήκευση κάποιων εκ των αποτελεσμάτων της πολυωνυμικής λογαριθμιστικής παλινδρόμησης ως νέες μεταβλητές.

ix. Τα αποτελέσματα της παραπάνω ανάλυσης παρουσιάζονται στους *Πίνακες 10.8-10.12*.

- Πιο συγκεκριμένα, στον *Πίνακα 10.8* παρουσιάζονται οι εκτιμήσεις των συντελεστών των 2 μοντέλων που προκύπτουν. Στις πρώτες γραμμές (status=1) παρουσιάζονται οι συντελεστές για το μοντέλο όπου η κατηγορία 1 της εξαρτημένης μεταβλητής συγκρίνεται με την κατηγορία αναφοράς (0) και στις γραμμές όπου status=2 παρουσιάζονται οι συντελεστές για το μοντέλο όπου η κατηγορία 2 της εξαρτημένης μεταβλητής συγκρίνεται με την κατηγορία αναφοράς (0) και πάλι. Για κάθε έναν από τους συντελεστές πραγματοποιείται και αναφέρεται και το p-value (sig.) από τον κατάλληλο έλεγχο χρησιμοποιώντας το Wald test. Επίσης, αναφέρεται και ο σχετικός λόγος (Odds Ratio) για κάθε μεταβλητή που προκύπτει αντιλογαριθμίζοντας τον συντελεστή B (Exp(B)) και το αντίστοιχο διάστημα εμπιστοσύνης για κάθε Odds Ratio (95% confidence interval). Αν το 1 συμπεριλαμβάνεται μέσα στο διάστημα εμπιστοσύνης, τότε δεν μπορούμε να υποθέσουμε ότι το αντίστοιχο Odds ratio είναι διάφορο του 1 και το αντίστοιχο p-value θα είναι μη στατιστικά σημαντικό. Τέλος, στην περίπτωση των κατηγορικών μεταβλητών, συντελεστές εκτιμούνται για όλες τις κατηγορίες, εκτός από την κατηγορία αναφοράς. Έτσι, λοιπόν, αναφορικά με το συγκεκριμένο παράδειγμα του *Πίνακα 10.8* διαπιστώνουμε ότι:

  - οι γυναίκες (sex=0) είναι περίπου 28% (1-Exp(B)=1-0,72) και 42% (1-Exp(B)=1-0,58) λιγότερο πιθανό να εμφανίσουν non-Q έμφραγμα (status=1) και Q έμφραγμα, (status=2) έναντι ασταθούς στηθάγχης (status=0), σε σχέση με τους άνδρες.

  - Αντίστοιχα, τα άτομα που δεν ασκούνται (physact01=0) είναι περίπου 59% (Exp(B)-1=1,589-1) και 41% (Exp(B)-1=11,41-1) πιο πιθανό να εμφανίσουν non-Q έμφραγμα (status=1) και Q έμφραγμα, (status=2) έναντι ασταθούς στηθάγχης (status=0), σε σχέση με τα άτομα που ασκούνται.

  - Όσον αφορά στην συστολική αρτηριακή πίεση, διαπιστώνουμε ότι αύξηση της πίεσης κατά 1 mm Hg συνεπάγεται μείωση 1% (1-Exp(B)=1-0,99) της πιθανότητας εμφάνισης Q εμφράγματος έναντι ασταθούς στηθάγχης, ενώ δεν υπάρχει στατιστικά σημαντική συσχέτιση ανάμεσα στις τιμές της συστολικής πίεσης και την εμφάνιση non-Q εμφράγματος έναντι ασταθούς στηθάγχης.

  - Αντίστοιχα, με την ερμηνεία για την συστολική αρτηριακή πίεση πραγματοποιείται και η ερμηνεία για την ηλικία. Σε αυτή την περίπτωση, όμως, παρατηρούμε ότι αύξηση της ηλικίας συνεπάγεται αύξηση της πιθανότητας εμφάνισης non-Q εμφράγματος έναντι της ασταθούς στηθάγχης (Odds Ratio=1,021), και μείωση της πιθανότητας εμφάνισης Q εμφράγματος έναντι της ασταθούς στηθάγχης (Odds Ratio=0,994)

Parameter Estimates

status[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
1,00	Intercept	-1,860	,470	15,636	1	,000			
	sbp	,001	,002	,130	1	,719	1,001	,996	1,005
	age	,021	,005	16,889	1	,000	1,021	1,011	1,031
	[sex=0]	-,328	,144	5,161	1	,023	,720	,543	,956
	[sex=1]	0[b]	.	.	0	.	.	.	.
	[physact01=0]	,463	,129	12,838	1	,000	1,589	1,233	2,047
	[physact01=1]	0[b]	.	.	0	.	.	.	.
2,00	Intercept	1,522	,447	11,590	1	,001			
	sbp	-,010	,002	16,718	1	,000	,990	,986	,995
	age	-,006	,005	1,361	1	,243	,994	,985	1,004
	[sex=0]	-,539	,152	12,602	1	,000	,584	,433	,786
	[sex=1]	0[b]	.	.	0	.	.	.	.
	[physact01=0]	,344	,126	7,477	1	,006	1,410	1,102	1,804
	[physact01=1]	0[b]	.	.	0	.	.	.	.

a. The reference category is: ,00.

b. This parameter is set to zero because it is redundant.

Πίνακας 10.8: Πίνακας με τις εκτιμήσεις των συντελεστών του μοντέλου.

Ο *Πίνακας 10.9* παρουσιάζει το αποτέλεσμα του ελέγχου υποθέσεων αναφορικά με το αν το συνολικό μοντέλο μας είναι καλύτερο από το μοντέλο που περιλαμβάνει μόνο την σταθερά (null model). Η μηδενική υπόθεση είναι ότι το πλήρες μοντέλο μας είναι ίδιο με το μοντέλο που περιλαμβάνει μόνο την σταθερά. Ο έλεγχος πραγματοποιείται συγκρίνοντας την πιθανοφάνεια (likelihood) των 2 μοντέλων μέσω του likelihood ratio test. Στην συγκεκριμένη περίπτωση, βλέπουμε ότι το μοντέλο μας είναι στατιστικά σημαντικά καλύτερο σε σχέση με το μοντέλο που δεν περιέχει καμία ανεξάρτητη μεταβλητή, αφού η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης είναι μικρότερη από το 0,05 (sig.<0,001).

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	2966,206	2976,970	2962,206			
Final	2895,820	2949,641	2875,820	86,386	8	,000

Πίνακας 10.9: Σύγκριση του μοντέλου μας με το μοντέλο που δεν περιλαμβάνει καμία ανεξάρτητη μεταβλητή.

Από τα παραπάνω διαπιστώνουμε ότι ο *Πίνακας 10.8* μας δίνει την απαραίτητη πληροφορία σχετικά με την επίδραση της κάθε ανεξάρτητης μεταβλητής σε κάθε ένα από τα 2 μοντέλα που εκτιμούνται στην συγκεκριμένη πολυωνυμική λογαριθμιστική παλινδρόμηση, ενώ ο *Πίνακας 9* μας πληροφορεί για το αν ολόκληρο το μοντέλο είναι στατιστικά σημαντικά καλύτερο σε σχέση με το μοντέλο που συμπεριλαμβάνει μόνο την σταθερά. Ο *Πίνακας 10* που παρουσιάζεται παρακάτω μας δείχνει αν γενικά κάθε μία από τις ανεξάρτητες μεταβλητές είναι σημαντική στο συνολικό μοντέλο της πολυωνυμικής λογαριθμιστικής παλινδρόμησης. Γι' αυτό το σκοπό χρησιμοποιείται

και πάλι το likelihood ratio test με το οποίο συγκρίνεται η συνολική πιθανοφάνεια του πλήρους μοντέλου με αυτή του μοντέλου που δεν περιέχει μία μία της ανεξάρτητες μεταβλητές. Η μηδενική υπόθεση λέει ότι η πιθανοφάνεια αυτών των 2 μοντέλων δεν διαφέρι στατιστικά σημαντικά και συνεπώς η ανεξάρτητη μεταβλητή που είχε εξαιρεθεί δεν είναι απαραίτητο να συμπεριληφθεί στο μοντέλο. Αν η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης είναι μικρότερη από το 0,05, τότε συμπεραίνουμε ότι η συγκεκριμένη μεταβλητή είναι απαραίτητο να συμπεριληφθεί στο μοντέλο. Αναφορικά με το συγκεκριμένο παράδειγμα, από τον παρακάτω Πίνακα διαπιστώνουμε ότι και οι 4 ανεξάρτητες μεταβλητές βελτιώνουν σημαντικά την πιθανοφάνεια του μοντέλου (αφού sig.<0,05 για όλες) και συνεπώς θα πρέπει να συμπεριληφθούν στο μοντέλο.

Likelihood Ratio Tests

	Model Fitting Criteria			Likelihood Ratio Tests		
Effect	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	2895,820	2949,641	2875,820[a]	,000	0	.
age	2921,593	2964,650	2905,593	29,773	2	,000
sbp	2915,457	2958,514	2899,457	23,637	2	,000
sex	2905,217	2948,274	2889,217	13,397	2	,001
physact01	2906,317	2949,374	2890,317	14,497	2	,001

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Πίνακας 10.10:

Στον *Πίνακα 10.11* παρουσιάζονται κάποια περιληπτικά στατιστικά μέτρα του μοντέλου. Όλα αυτά τα μέτρα είναι ανάλογα του R-squared που υπολογίζεται στην γραμμική παλινδρόμηση, αλλά κανένα από αυτά δεν έχει την ερμηνεία του «ποσοστού της διακύμανσης που ερμηνεύεται» και δεν πρέπει να τα αναφέρουμε με αυτόν τον τρόπο. Αντίθετα, αυτά θα πρέπει να θεωρούνται ως επιπρόσθετα μέτρα του μεγέθους αποτελέσματος (effect size) του μοντέλου και συνεπώς όσο μεγαλύτερα τόσο το καλύτερο. Αυτά του *Πίνακα 10.6* εκφράζουν κακή προσαρμογή του μοντέλου. Όπως και σε άλλες τεχνικές, το μοντέλο μπορεί να είναι στατιστικά σημαντικά καλύτερο από το null μοντέλο, όμως, αυτό δεν σημαίνει απαραίτητα ότι και το effect size του μοντέλου είναι μεγάλο (π.χ. στην περίπτωση της γραμμικής παλινδρόμησης, μπορεί το μοντέλο να είναι στατιστικά σημαντικό, αλλά το R-squared να είναι πολύ μικρό). Συνεπώς, από κοινού η στατιστική σημαντικότητα του μοντέλου και αυτά τα μέτρα του effect size του μοντέλου θα πρέπει να ανφέρονται. Το Cox and Snell's R-Square είναι μία προσπάθεια να μυμηθούμε την ερμηνεία του R-square βάσει της πιθανοφάνειας, αλλά το μέγιστο (συνήθως) είναι μικρότερο της μονάδας και συνεπώς είναι δύσκολη η ερμηνεία. Το Nagelkerke's R-square είναι μία τροποποίηση του Cox and Snell συντελεστή έτσι ώστε να παίρνει τιμές μεταξύ 0 και 1. Συνεπώς, το Nagelkerke's R-square θα παίρνει φυσιολογικά υψηλότερες τιμές σε

σχέση με το Cox and Snell συντελεστή. Το McFadden's R-square είναι ένα μέτρο της θεωρίας της πληροφορίας και ερμηνεύεται ως η μείωση της εντροπίας του μοντέλου σε σχέση με την εντροπία του μοντέλου με μόνο την σταθερά (null model).

Pseudo R-Square

Cox and Snell	,052
Nagelkerke	,059
McFadden	,025

Πίνακας 10.11: Περιληπτικά στατιστικά μέτρα του μοντέλου (Summary statistics)

Από τον *Πίνακα 10.12* ότι το ποσοστό των ατόμων που συμμετέχουν στην συγκεκριμένη μελέτη και ταξινομήθηκαν στην σωστή κατηγορία της εξαρτημένης μεταβλητής είναι 43,7%. Πιο συγκεκριμένα, παρατηρούμε ότι το 52,0% των ατόμων που στην πραγματικότητα ανήκει στην κατηγορία 0 (ασταθής στηθάγχη) έχει ταξινομηθεί σωστά με τη βοήθεια του συγκεκριμένου μοντέλου που προσαρμόσαμε. Τα αντίστοιχα ποσοστά για τις κατηγορίες 1 και 2 (non-Q και Q έμφραγμα, αντίστοιχα) είναι 36,5% και 41,3%, αντίστοιχα.

Classification

	Predicted			
Observed	,00	1,00	2,00	Percent Correct
,00	302	118	161	52,0%
1,00	189	183	129	36,5%
2,00	197	111	217	41,3%
Overall Percentage	42,8%	25,6%	31,5%	43,7%

Πίνακας 10.12: Πίνακας των παρατηρούμενων παρατηρήσεων έναντι των αναμενόμενων βάσει του μοντέλου.

10.9 Πολυωνυμική λογαριθμιστική παλινδρόμηση με το STATA

Η πολυωνυμική λογαριθμιστική παλινδρόμηση στο STATA (multinomial logistic regression), πραγματοποιείται ακολουθώντας τα εξής ***βήματα***:

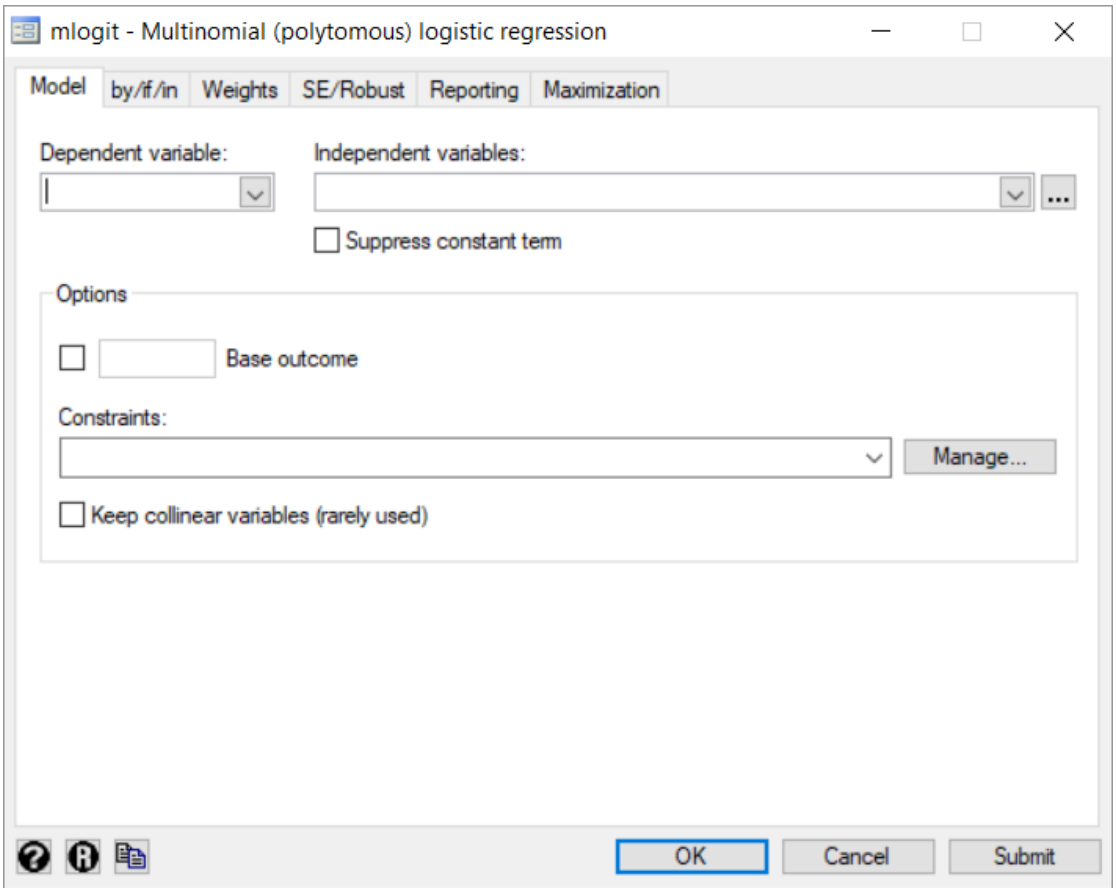
Statistics → Categorical outcomes → Multinomial logistic regression

x. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνα 10.18.*

v. Ορίζουμε ως «***Dependent variable»*** την κατηγορική μεταβλητή/έκβαση

vi. Ως «***Independent variables»*** ορίζουμε τις ανεξάρτητες ποσοτικές μεταβλητές.

vii. «***OK***»



Εικόνα 10.18: Πραγματοποίηση πολυωνυμικής λογαριθμιστικής παλινδρόμησης

Στο STATA χρησιμοποιείται η εντολή ***mlogit*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 10.19*:

Syntax

```
mlogit depvar [indepvars] [if] [in] [weight] [, options]
```

options	Description
Model	
noconstant	suppress constant term
baseoutcome(#)	value of *depvar* that will be the base outcome
constraints(*clist*)	apply specified linear constraints; *clist* has the form #[-#] [, #[-#] ...]
collinear	keep collinear variables
SE/Robust	
vce(*vcetype*)	*vcetype* may be oim, robust, cluster *clustvar*, bootstrap, or jackknife
Reporting	
level(#)	set confidence level; default is level(95)
rrr	report relative-risk ratios
nocnsreport	do not display constraints
display_options	control column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
Maximization	
maximize_options	control the maximization process; seldom used
coeflegend	display legend instead of statistics

Εικόνα 10.19: Βασική σύνταξη της εντολής mlogit

11. Ανάλυση για επαναλαμβανόμενες μετρήσεις.

11.1 Εισαγωγή

Το κοινό χαρακτηριστικό όλων των στατιστικών αναλύσεων που έχουν παρουσιαστεί μέχρι τώρα είναι ότι οι παρατηρήσεις είναι ανεξάρτητες μεταξύ τους. Συγκεκριμένα, για κάθε ένα χαρακτηριστικό (μεταβλητή) που συμπεριλαμβανόταν σε μια στατιστική ανάλυση υπήρχε μόνο μία μέτρηση για κάθε άτομο που συμμετείχε στη μελέτη. Όμως, υπάρχουν και μελέτες όπου για κάθε άτομο έχουν λάβει χώρα περισσότερες από μία μετρήσεις του ίδιου χαρακτηριστικού και αυτές οι μελέτες είναι γνωστές ως ***μελέτες επαναλαμβανόμενων μετρήσεων.***

Παρακάτω αναφέρονται μερικά ***παραδείγματα*** τέτοιου τύπου μελετών:

1. Θέλουμε να ελέγξουμε την αποτελεσματικότητα ενός αντι-υπερτασικού φάρμακου. Σε μία ομάδα υπερτασικών ασθενών μετράμε την ΑΠ πριν τη χορήγηση και μετά τη χορήγηση του φαρμάκου.
2. Σε 2 ομάδες ασθενών με HIV χορηγούνται δύο διαφορετικές θεραπείες στην αρχή της μελέτης. Προκειμένου να μελετηθεί & να συγκριθεί η αποτελεσματικότητα των φαρμάκων, οι ερευνητές μετρούν το ιικό φορτίο των ασθενών, σε μηνιαία βάση για τους επόμενους 12 μήνες.
3. Σε τρεις ομάδες νεογέννητων χορηγούνται τρεις διαφορετικοί τύποι γάλακτος. Μετράμε το ύψος και το βάρος αυτών σε μηνιαία βάση μέχρι να συμπληρώσουν τον πρώτο χρόνο ζωής, προκειμένου να συγκρίνουμε την επίδραση του γάλακτος στην ανάπτυξη των παιδιών.

Το ***κοινό χαρακτηριστικό*** των παραπάνω παραδειγμάτων είναι ότι σε κάθε άτομο που συμμετέχει στις μελέτες έχουν πραγματοποιηθεί περισσότερες από μία μετρήσεις για το χαρακτηριστικό ενδιαφέροντός μας (π.χ. ΑΠ, ιικό φορτίο, ύψος και βάρος παιδιών).

Η διαφορά των παραπάνω χαρακτηριστικών είναι ότι στην $1^{η}$ μελέτη έχουν λάβει χώρα μόνο 2 μετρήσεις του χαρακτηριστικού ενδιαφέροντος μας, ενώ στις άλλες 2 μελέτες έχουν λάβει χώρα περισσότερες μετρήσεις. Επίσης, στην $1^{η}$ μελέτη όλοι οι συμμετέχοντες ανήκουν στην ίδια ομάδα, σε αντίθεση με τις υπόλοιπες μελέτες όπου οι συμμετέχοντες ανήκουν σε περισσότερες ομάδες.

Οι κατάλληλες στατιστικές τεχνικές, λοιπόν, για την ανάλυση των δεδομένων του $1^{ου}$ παραδείγματος είναι οι αναλύσεις για 2 συσχετισμένα δείγματα που είναι οι εξής:

- ***paired-t-test,***
- ***Wilcoxon sign rank test***

Αντίθετα, η κατάλληλη στατιστική τεχνική για την ανάλυση των δεδομένων των άλλων 2 παραδειγμάτων είναι:

- η ***ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις*** (Repeated Measures ANOVA).

Σημείωση: Στην περίπτωση όπου έχουν λάβει χώρα περισσότερες από 2 μετρήσεις σε κάθε άτομο, όλοι οι συμμετέχοντες ανήκουν σε μία ομάδα και δεν πληρούνται οι απαραίτητες προϋποθέσεις για την εφαρμογή της ανάλυσης διακύμανσης για επαναλαμβανόμενες μετρήσεις, το κατάλληλο στατιστικό κριτήριο είναι το ***Friedman***.

11.1.1 Paired-t-test ή Wilcoxon/ sign rank test

Όπως προκύπτει και από τα παραπάνω και τα 2 κριτήρια εφαρμόζονται στις περιπτώσεις όπου ένα χαρακτηριστικό έχει μετρηθεί στα ίδια άτομα πριν και μετά από μία παρέμβαση, προκειμένου να διαπιστώσουμε αν η παρέμβαση είναι αποτελεσματική ή όχι. Η επιλογή του ενός ή του άλλου κριτηρίου εξαρτάται από τον έλεγχο της εξής ***προϋπόθεσης***:

- Το χαρακτηριστικό ακολουθεί την *κανονική κατανομή* και στις 2 μετρήσεις ή όχι?

Στην περίπτωση που το χαρακτηριστικό μας ακολουθεί την κανονική κατανομή τότε εφαρμόζεται το paired-t-test. Στην αντίθετη περίπτωση, εφαρμόζεται το Wilcoxon/sign rank test.

Και στις 2 περιπτώσεις, η μηδενική και εναλλακτική υπόθεση είναι:

H_0: Η θεραπεία δεν έχει καμία συσχέτιση με το χαρακτηριστικό που μετράμε (π.χ. ΑΠ) ή αλλιώς η μέση τιμή της ΑΠ είναι ίση πριν και μετά την θεραπεία (paired t test) ή η κατανομή της ΑΠ είναι ίδια πριν και μετά την θεραπεία (Wilcoxon ή sign rank test).

H_1: Η θεραπεία συσχετίζεται με το χαρακτηριστικό που μετράμε (π.χ. ΑΠ) ή αλλιώς η μέση τιμή της ΑΠ διαφέρει πριν και μετά την παρέμβαση (paired t test) ή η κατανομή της ΑΠ διαφέρει πριν και μετά την θεραπεία (Wilcoxon ή sign rank test).

Για να καταλήξουμε στο συμπέρασμα ότι υπάρχει συσχέτιση ανάμεσα στην παρέμβαση (π.χ. θεραπεία) και το χαρακτηριστικό που μετράμε (π.χ. ΑΠ), θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value) να είναι $< \alpha = 0,05$ ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).

11.1.2 Ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις

Η ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις είναι μία στατιστική μέθοδος η οποία βρίσκει εφαρμογή στην περίπτωση όπου οι μετρήσεις σε κάθε πειραματική μονάδα πραγματοποιούνται τις ίδιες n χρονικές στιγμές για όλες τις μονάδες, με καμία απόκλιση από αυτές τις χρονικές στιγμές και καμία ελλείπουσα τιμή για καμία μονάδα.

- Αυτό το μοντέλο βασίζεται σε μία πολύ ειδική παραδοχή αναφορικά με τον πίνακα συν-διακυμάνσεων των δεδομένων, που ίσως να μην ισχύει πάντα στις επαναλαμβανόμενες μετρήσεις.
- Συνεπώς, εξαιτίας αυτού, ενδέχεται να οδηγηθούμε σε λανθασμένα συμπεράσματα.
- Εξαιτίας της απλότητά του, αυτό το μοντέλο είναι αρκετά δημοφιλές και υιοθετείται εξ ορισμού, μερικές φορές χωρίς να γίνει έλεγχος της εγκυρότητας των προϋποθέσεων που απαιτούνται.

Ας θυμηθούμε τα παραδείγματα 2 και 3 της *παραγράφου 11.1*.

- Οι πειραματικές μονάδες τυχαιοποιούνται σε μία από τις $q > 1$ ομάδες. Στη βιβλιογραφία, αυτό αναφέρεται ως **παράγοντες μεταξύ των μονάδων**

(**between units factors or groups**). Στη μελέτη του παραδείγματος 2, q = 2 (2 θεραπείες).

- Το μέγεθος που μας ενδιαφέρει μετράται σε κάθε μία από n καταστάσεις. Αν και στις μελέτες επαναλαμβανόμενων μετρήσεων οι καταστάσεις που πραγματοποιούνται οι μετρήσεις είναι διάφορες χρονικές στιγμές, αυτό μπορεί να είναι και κάτι άλλο. Παρόλα αυτά, εμείς θα αναφερόμαστε σε αυτές τις καταστάσεις σαν «χρόνο» και στη βιβλιογραφία αυτό αναφέρεται ως **εντός των μονάδων παράγοντας** (**within-units factor**). Για παράδειγμα, στην μελέτη του παραδείγματος 2, το ιικό φορτίο μετρήθηκε 12 φορές (12 μήνες).

11.1.2.1 Πηγές διακύμανσης

Σε τέτοιες μελέτες υπάρχουν δύο πιθανές πηγές διακύμανσης που κάνουν τις παρατηρήσεις που έχουμε στις μονάδες της ίδιας ομάδας την ίδια χρονική στιγμή να διαφέρουν:

- Υπάρχει μία τυχαία διακύμανση στον πληθυσμό των πειραματικών μονάδων εξαιτίας της βιολογικής μεταβλητότητας. Για παράδειγμα, αν όλοι οι ασθενείς του παραδείγματος 2 έπαιρναν την ίδια θεραπεία, το ιικό τους φορτίο τον $1^{ο}$ μήνα θα ήταν διαφορετικό εξαιτίας των βιολογικών διαφορών, δηλαδή εξαιτίας του γεγονότος ότι αυτοί δεν είναι ίδιοι. Αυτή η πηγή διακύμανσης ορίζεται ως η **μεταξύ των πειραματικών μονάδων διακύμανση** (**among individuals random variation**).
- Υπάρχει επίσης μία τυχαία διακύμανση που προκαλείται από τη διαδικασία της μέτρησης σε μία πειραματική μονάδα σε κάποια συγκεκριμένη χρονική στιγμή. Για παράδειγμα, αν μετρήσουμε το ιικό φορτίο ενός ασθενούς σε κάποια συγκεκριμένη χρονική στιγμή, εξαιτίας του γεγονότος ότι το όργανο που χρησιμοποιείται δεν είναι τέλειο, οι παρατηρήσεις υπόκεινται στο σφάλμα μέτρησης. Ένα άλλο παράδειγμα είναι η μέτρηση της συγκέντρωσης ενός φαρμάκου στο αίμα σε έναν συγκεκριμένο άνδρα που ανήκει σε μία συγκεκριμένη ομάδα, ανάλογα με τη δόση φαρμάκου που του χορηγήθηκε. Τέτοιες μετρήσεις συγκέντρωσης μπορεί να διαφέρουν εξαιτίας α) σφάλματος στο όργανο που χρησιμοποιείται για τη μέτρηση της συγκέντρωσης στο δείγμα και β) του γεγονότος ότι ένα δείγμα έχει επιλεγεί από τον άντρα, εισάγοντας τη διακύμανση λόγω δειγματοληψίας. Αυτή η πηγή διακύμανσης ορίζεται ως την **εντός των μονάδων διακύμανση (within individuals random variation)**.

11.1.2.2 Προϋποθέσεις για την ορθή εφαρμογή της ανάλυσης διακύμανσης για επαναλαμβανόμενες μετρήσεις.

Προκειμένου να εφαρμοστεί ορθά η ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις θα πρέπει να ισχύουν οι εξής 2 προϋποθέσεις:

1. Κανονικότητα: Το χαρακτηριστικό που μελετάμε (π.χ. ιικό φορτίο) πρέπει να ακολουθεί την κανονική κατανομή σε όλες τις χρονικές στιγμές που έχουν πραγματοποιηθεί οι μετρήσεις.

2. Προϋπόθεση του Compound symmetry*:* Η διακύμανση του χαρακτηριστικού (π.χ. ιικό φορτίο) πρέπει να είναι ίδια σε όλες τις χρονικές στιγμές καθώς επίσης και όλες οι συν-διακυμάνσεις (και συνεπώς οι συσχετίσεις) των παρατηρήσεων μεταξύ

των διαφόρων χρονικών στιγμών να είναι ίσες άσχετα με το πόσο μακριά ή κοντά είναι αυτές στο χρόνο. Δηλαδή, ο πίνακας διακύμανσης – συνδιακύμανσης να είναι της μορφής *11.1*.

$$\text{var}(\mathbf{y}_{hl}) = \begin{pmatrix} \sigma_{\mathbf{b}}^2 + \sigma_e^2 & \sigma_{\mathbf{b}}^2 & \cdots & \sigma_{\mathbf{b}}^2 \\ \sigma_{\mathbf{b}}^2 & \sigma_{\mathbf{b}}^2 + \sigma_e^2 & \cdots & \sigma_{\mathbf{b}}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{\mathbf{b}}^2 & \sigma_{\mathbf{b}}^2 & \cdots & \sigma_{\mathbf{b}}^2 + \sigma_e^2 \end{pmatrix} \qquad (11.1)$$

11.1.2.3 Στατιστικές υποθέσεις.

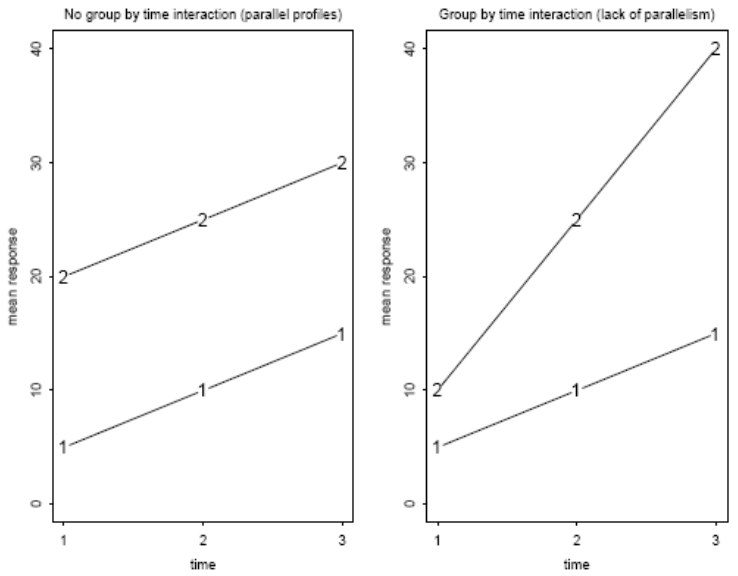
Αλληλεπίδραση χρόνου και ομάδας (Group by time interaction): Ο συνήθης στόχος της ανάλυσης των δεδομένων επαναλαμβανόμενων μετρήσεων είναι να διερευνηθεί αν ο τρόπος μεταβολής του μεγέθους που μετράτε κατά μήκος του χρόνου είναι διαφορετικός μεταξύ των διαφορετικών ομάδων. Για παράδειγμα, στη μελέτη του παραδείγματος 2, η ερώτηση είναι αν το ιικό φορτίο των ασθενών μεταβάλλεται με διαφορετικό τρόπο μεταξύ αυτών που έλαβαν τη θεραπεία Α και αυτών που έλαβαν τη θεραπεία Β.

Αυτό απεικονίζεται καλύτερα με μία εικόνα. Για την περίπτωση όπου έχουμε q = 2 ομάδες και n = 3 χρονικές στιγμές, η *Εικόνα 11.1* παρουσιάζει 2 πιθανά σενάρια. Σε κάθε σενάριο, οι γραμμές αντιπροσωπεύουν τις μέσες τιμές μ_{lj} για κάθε ομάδα. Και στα δύο σενάρια, η μέση τιμή κάθε χρονική στιγμή είναι μεγαλύτερη για την ομάδα 2 σε σχέση με την ομάδα 1 για όλες τις χρονικές στιγμές και η μορφή της μεταβολής των μέσων τιμών φαίνεται να είναι μία ευθεία γραμμή. Όμως, στο σενάριο που απεικονίζεται στο αριστερό πλαίσιο, ο ρυθμός μεταβολής των μέσων τιμών κατά μήκος του χρόνου είναι ο ίδιος και για τις δύο ομάδες, δηλαδή, τα προφίλ είναι παράλληλα. Στο σενάριο που παρουσιάζεται στο δεξιό πλαίσιο, ο ρυθμός μεταβολής είναι γρηγορότερος για την ομάδα 2 και συνεπώς τα προφίλ δεν είναι παράλληλα.
Στατιστικά, αυτός ο έλεγχος πραγματοποιείται εξετάζοντας αν ο όρος αλληλεπίδρασης (interaction term) ανάμεσα σε ομάδα και χρόνο είναι στατιστικά σημαντικός. Πιο συγκεκριμένα, η μηδενική και εναλλακτική υπόθεση διατυπώνεται ως εξής:

H_0: Δεν υπάρχει αλληλεπίδραση ανάμεσα στον χρόνο και την ομάδα και συνεπώς το χαρακτηριστικό που μετράμε μεταβάλλεται με τον ίδιο ρυθμό κατά μήκος του χρόνου σε όλες τις ομάδες.

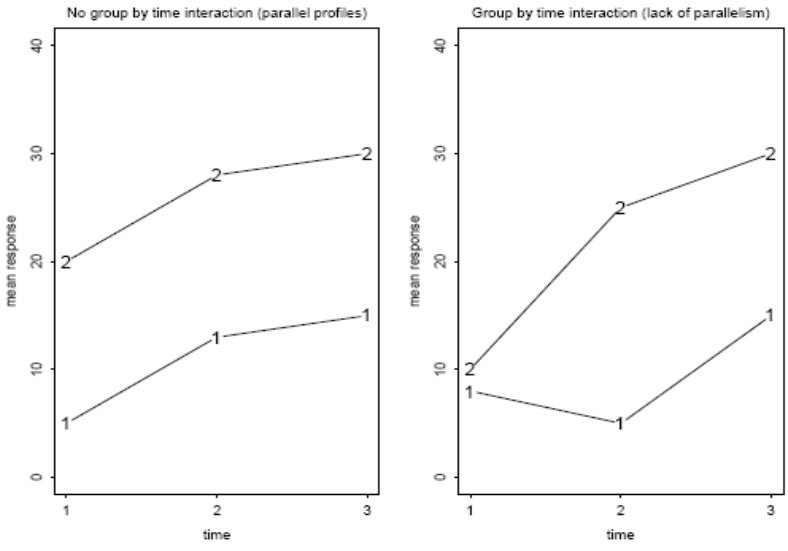
H_1: Υπάρχει αλληλεπίδραση ανάμεσα στον χρόνο και την ομάδα και συνεπώς ο ρυθμός μεταβολής του χαρακτηριστικού που μελετάμε κατά μήκος του χρόνου διαφέρει μεταξύ των διαφόρων ομάδων.

Για να καταλήξουμε στο συμπέρασμα ότι υπάρχει αλληλεπίδραση ανάμεσα στον χρόνο και το χαρακτηριστικό που μελετάμε (π.χ. ιικό φορτίο), θα πρέπει η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης H_o (δηλαδή το p-value) να είναι < α = 0,05 ή αλλιώς 5% (το οποίο έχει οριστεί αυθαίρετα, αλλά είναι ευρέως χρησιμοποιούμενο σε όλες τις ιατρό-βιολογικές έρευνες).



Εικόνα 11.1: Αλληλεπίδραση ομάδας και χρόνου. Τα σύμβολα υποδηλώνουν την ομάδα.

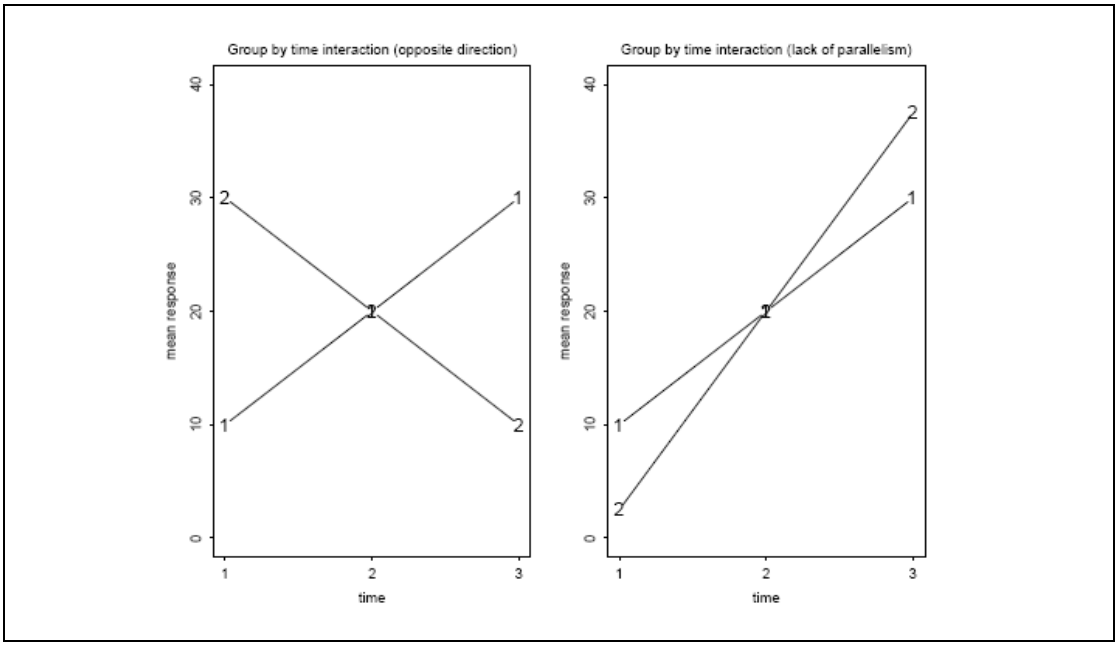
Στη γλώσσα που συνδέεται με τις επαναλαμβανόμενες μετρήσεις, ο έλεγχος για την αλληλεπίδραση μεταξύ ομάδας και χρόνου ονομάζεται **έλεγχος παραλληλότητας** (**test for parallelism**).
Είναι σημαντικό να σημειώσουμε ότι η παραλληλότητα δεν σημαίνει απαραίτητα ότι οι μέσες τιμές κατά μήκος του χρόνου πρέπει να μοιάζουν αυστηρά με ευθεία γραμμή. Στην *Εικόνα 11.2*, στο αριστερό πλαίσιο έχουμε παραλληλότητα, ενώ στο δεξί όχι.



Εικόνα 11.2: Αλληλεπίδραση ομάδας και χρόνου. Τα σύμβολα υποδηλώνουν την ομάδα.

Κυρίως επίδραση των ομάδων: Ξεκάθαρα, αν τα προφίλ είναι παράλληλα, τότε η αυτονόητη ερώτηση είναι αν αυτά **συμπίπτουν** (**coincident**), δηλαδή αν κάθε χρονική στιγμή οι μέσες τιμές των ομάδων είναι ίδιες. Η ερώτηση αν ο μέσος όρος των μέσων τιμών κατά μήκος του χρόνου είναι ίδιος για κάθε ομάδα αν τα προφίλ δεν είναι παράλληλα ίσως να μην έχει ενδιαφέρον.

- Για παράδειγμα αν η σωστή υπόθεση είναι αυτή που απεικονίζεται στο δεξί πλαίσιο των *Εικόνων 11.1 και 11.2*, τότε η ερώτηση αν ο μέσος όρος των μέσων τιμών κατά μήκος του χρόνου είναι διαφορετικός για τις δύο ομάδες ίσως είναι ενδιαφέρουσα.
- Από την άλλη πλευρά, ας θεωρήσουμε ότι το αριστερό πλαίσιο της *Εικόνας 11.3* απεικονίζει τη σωστή υπόθεση. Τότε ο έλεγχος αυτής της ερώτησης είναι μάλλον χωρίς νόημα, αφού η μεταβολή των μέσων τιμών κατά μήκος του χρόνου είναι στην αντίθετη κατεύθυνση για τις δύο ομάδες. Άρα ο μέσος όρος των μέσων τιμών κατά μήκος του χρόνου είναι μικρής σπουδαιότητας.
- Παρομοίως, αν το αντικείμενο μελέτης είναι κάτι σαν την ανάπτυξη, ο μέσος όρος των μέσων τιμών κατά μήκος του χρόνου ίσως έχει ελάχιστο νόημα. Για παράδειγμα, στο δεξί πλαίσιο της *Εικόνας 11.3*, οι μέσες τιμές κατά μήκος του χρόνου αυξάνουν για κάθε ομάδα με διαφορετικό ρυθμό, αλλά έχουν τον ίδιο μέσο όρο κατά μήκος του χρόνου. Ξεκάθαρα, η ομάδα με τον ταχύτερο ρυθμό μεταβολής θα έχει μεγαλύτερη μέση τιμή στο τέλος της χρονικής περιόδου.



Εικόνα 11.3: Αλληλεπίδραση ομάδας και χρόνου. Τα σύμβολα υποδηλώνουν την ομάδα.

Η μηδενική και εναλλακτική υπόθεση διατυπώνεται ως εξής:

H_0: Οι μέσες τιμές του χαρακτηριστικού ΕΙΝΑΙ ίσες μεταξύ των ομάδων ανεξαρτήτως χρονικής στιγμής που έχει πραγματοποιηθεί η μέτρηση.

H_1: Οι μέσες τιμές του χαρακτηριστικού ΔΕΝ είναι ίσες μεταξύ των ομάδων ανεξαρτήτως χρονικής στιγμής που έχει πραγματοποιηθεί η μέτρηση.

Κυρίως επίδραση του χρόνου: Μία άλλη ερώτηση είναι αν η μέση τιμή των παρατηρήσεων είναι σταθερή κατά μήκος του χρόνου. Αν τα προφίλ είναι παράλληλα, τότε είναι σαν να ρωτάμε αν ο μέσος όρος των παρατηρήσεων από όλες τις ομάδες είναι ο ίδιος σε κάθε χρονική στιγμή. Αν τα προφίλ δεν είναι παράλληλα, αυτή η ερώτηση μπορεί να μην έχει ενδιαφέρον. Για παράδειγμα, σημειώστε ότι στο αριστερό πλαίσιο της *Εικόνας 11.3*, ο μέσος όρος των παρατηρήσεων από τις ομάδες 1 και 2 είναι ίδιος για όλες τις χρονικές στιγμές. Όμως, η μέση τιμή των παρατηρήσεων μέσα σε κάθε ομάδα δεν είναι σταθερή.

Έτσι, αν θεωρήσουμε πως έχουμε q ομάδες και n χρονικές στιγμές, τότε η μηδενική και εναλλακτική υπόθεση διατυπώνεται ως εξής:

H_0: Οι μέσες τιμές του χαρακτηριστικού στις διάφορες χρονικές στιγμές ΕΙΝΑΙ ίσες ανεξαρτήτως ομάδας.

H_1: Οι μέσες τιμές του χαρακτηριστικού στις διάφορες χρονικές στιγμές ΔΕΝ είναι ίσες ανεξαρτήτως ομάδας.

11.1.2.4 Έλεγχος της υπόθεσης αναφορικά με τη δομή του πίνακα συν-διακύμανσης.

Test of sphericity: Ενδιαφέρον παρουσιάζει ο έλεγχος του αν η αληθινή δομή του πίνακα συν-διακύμανσης σε δεδομένα επαναλαμβανόμενων μετρήσεων είναι της μορφής Η. Ένα test με το οποίο μπορεί να πραγματοποιηθεί αυτός ο έλεγχος είναι το Mauchly's test of sphericity. Λεπτομέρειες σχετικά με αυτό το test είναι πέρα από το σκοπό των συγκεκριμένων σημειώσεων. Αυτό το test παρέχει έναν έλεγχο για την στατιστική υπόθεση

H_0 : Ο πίνακας διακύμανσης –συνδιακύμανσης είναι της μορφής Η, η αλλιώς ισχύει το compound symmetry

Αυτό το test έχει προσεγγιστικά την χ^2 κατανομή όταν ο αριθμός των μονάδων m είναι μεγάλος με βαθμούς ελευθερίας ίσους με $(n-2)(n-1)/2$. Έτσι το test πραγματοποιείται σε επίπεδο σημαντικότητας α συγκρίνοντας την τιμή του test με την κρίσιμη τιμή χ_a^2 με $(n-2)(n-1)/2$ βαθμούς ελευθερίας.

Το test έχει μερικούς περιορισμούς:

- Δεν έχει μεγάλη ισχύ όταν ο αριθμός των μονάδων σε κάθε ομάδα δεν είναι πολύ μεγάλος.
- Αυτό μπορεί να είναι παραπλανητικό αν το διάνυσμα δεδομένων δεν ακολουθεί την πολυδιάστατη κανονική κατανομή.

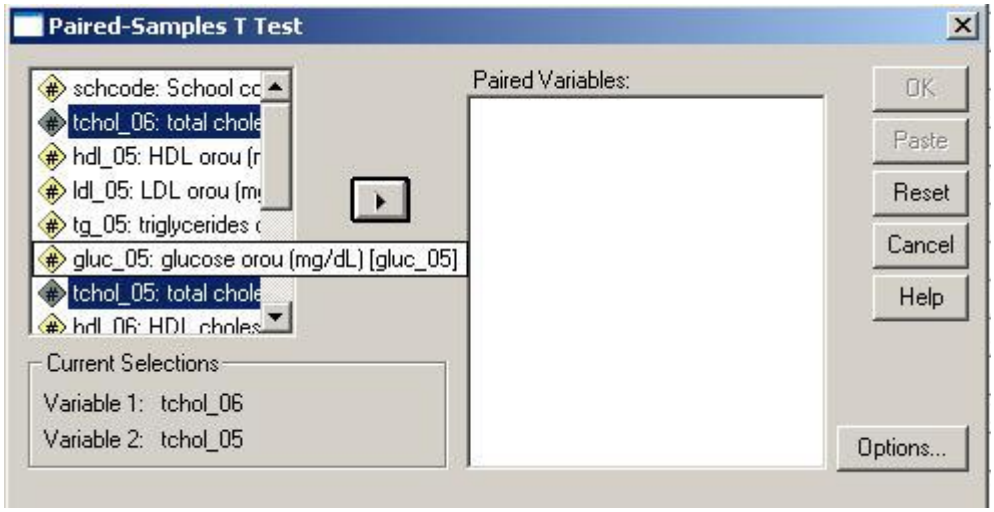
Στην περίπτωση που η προϋπόθεση για την δομή του πίνακα διακύμανσης συν-διακύμανσης παραβιάζεται, δεν συνεπάγεται αυτόματα ότι η ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις είναι εσφαλμένη. Σε αυτή την περίπτωση, όμως, είναι απαραίτητο τα αποτελέσματα της ανάλυσης να ερμηνεύονται αφού πρώτα πραγματοποιηθούν οι αντίστοιχες διορθώσεις.

11.2 Paired-t-test με τη χρήση του SPSS

Ας υποθέσουμε ότι σε μία ομάδα μαθητών πραγματοποιούμε μία διατροφική παρέμβαση για ένα χρόνο προκειμένου να ελέγξουμε αν αυτή η παρέμβαση είναι αποτελεσματική ή όχι όσον αφορά στο λιπιδαιμικό τους προφίλ. Μετράμε, ολική χοληστερόλη και τριγλυκερίδια πριν την έναρξη της παρέμβασης και αμέσως μετά το τέλος της παρέμβασης. Σε αυτή την περίπτωση έχουν πραγματοποιηθεί 2 μετρήσεις σε κάθε άτομο για το ίδιο χαρακτηριστικό και αυτές οι μελέτες είναι γνωστές ως «προ-μετά μελέτες». Ας υποθέσουμε ότι στόχος μας είναι να ελέγξουμε αν η παρέμβαση ήταν αποτελεσματική όσον αφορά στην μείωση των επιπέδων της ολικής χοληστερόλης και ας υποθέσουμε επιπλέον ότι η ολική χοληστερόλη ακολουθεί την κανονική κατανομή και στις 2 μετρήσεις. Σύμφωνα με όσα έχουν αναφερθεί παραπάνω είναι σαφές ότι ο κατάλληλος στατιστικός έλεγχος για να διερευνήσουμε αυτή την υπόθεση είναι το «***paired-t-test***».
Για να το εφαρμόσουμε πρέπει να ακολουθήσουμε τα εξής ***βήματα***:

Analyze ⟶ Compare means⟶ paired samples t test

i. Ανοίγει το παράθυρο διαλόγου που παρουσιάζεται στην *Εικόνα 11.4*



Εικόνα 11.4: Πραγματοποίηση του ελέγχου ***paired – t- test***.

ii. Επιλέγουμε τις μεταβλητές που είναι καταχωρημένες οι 2 μετρήσεις του χαρακτηριστικού (π.χ. tchol_05 & tchol_06)

iii. Με το βελάκι τις τοποθετούμε στο «***Paired Variables***»

iv. Πατάμε "***Ok***".

v. Στον *Πίνακα 11.1* παρουσιάζονται τα αποτελέσματα του ελέγχου. Από τον Πίνακα "***Paired Samples Test***" διαπιστώνουμε ότι η αλλαγή στα επίπεδα χοληστερόλης είναι στατιστικά σημαντική αφού sig.<0,001 (δηλ. η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης, ότι δηλαδή δεν υπάρχει συσχέτιση ανάμεσα στην παρέμβαση και τα επίπεδα χοληστερόλης είναι <α=0,05). Επίσης, από το «***Mean***» στον ίδιο πίνακα διαπιστώνουμε ότι η διαφορά tchol_06 – tchol_05 = -14,27 είναι αρνητική, δηλαδή βλέπουμε πως οι τιμές ολικής χοληστερόλης το 2005 ήταν υψηλότερες σε σχέση με το 2006.

Τέλος, ο *Πίνακας 11.2* παρουσιάζει αναλυτικά τα περιγραφικά στοιχεία των επιπέδων χοληστερόλης πριν και μετά την παρέμβαση.

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	tchol_06: total cholesterol (oliki xolisterini orou) (mg/dL) - tchol_05: total cholesterol (mg%)	-14,27114	34,03765	1,81939	-17,84949	-10,69280	-7,844	349	,000

Πίνακας 11.1: Αποτελέσματα του paired-t-test για την αποτελεσματικότητα ετήσιας διατροφικής παρέμβασης στα επίπεδα ολικής χοληστερόλης.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	tchol_06: total cholesterol (oliki xolisterini orou) (mg/dL)	158,9003	350	35,34525	1,88928
	tchol_05: total cholesterol (mg%)	173,171	350	27,3395	1,4614

Πίνακας 11.2: Περιγραφικά χαρακτηριστικά για τα επίπεδα ολικής χοληστερόλης πριν και μετά την παρέμβαση.

11.3 Paired-t-test με τη χρήση του STATA

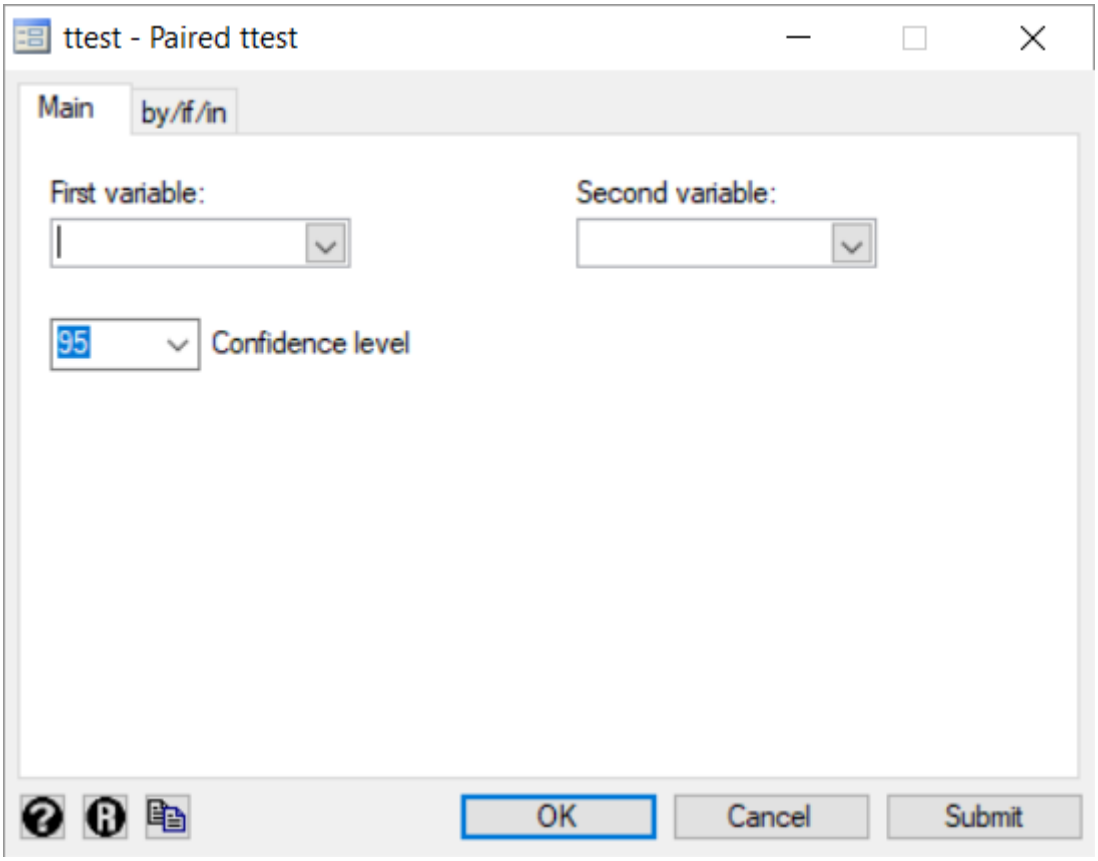
Για να το εφαρμόσουμε το paired – t- test στο STATA πρέπει να ακολουθήσουμε τα εξής ***βήματα***:

Statistics → Summaries, tables, and tests → Classical tests of hypotheses → Mean-comparison test, paired data

i. Ανοίγει το παράθυρο διαλόγου που παρουσιάζεται στην *Εικόνα 11.5*

ii. Τοποθετούμε την πρώτη μεταβλητή στο πλαίσιο «***First variable***» και την δεύτερη μεταβλητή στο πλαίσιο «***Second variable***».

iii. Πατάμε "***OK***".



Εικόνα 11.5: Πραγματοποίηση του ελέγχου ***paired – t- test***

Στο STATA χρησιμοποιείται η εντολή ***ttest*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 11.6*:

Syntax

One-sample t test

```
ttest varname == # [if] [in] [, level(#)]
```

Two-sample t test using groups

```
ttest varname [if] [in], by(groupvar) [options1]
```

Two-sample t test using variables

```
ttest varname1 == varname2 [if] [in], unpaired [unequal welch level(#)]
```

Paired t test

```
ttest varname1 == varname2 [if] [in] [, level(#)]
```

Immediate form of one-sample t test

```
ttesti #obs #mean #sd #val [, level(#)]
```

Immediate form of two-sample t test

```
ttesti #obs1 #mean1 #sd1 #obs2 #mean2 #sd2 [, options2]
```

options$_1$	Description
Main	
*by(*groupvar*)	variable defining the groups
unequal	unpaired data have unequal variances
welch	use Welch's approximation
level(#)	set confidence level; default is level(95)

*by(*groupvar*) is required.

options$_2$	Description
Main	
unequal	unpaired data have unequal variances
welch	use Welch's approximation
level(#)	set confidence level; default is level(95)

by is allowed with ttest; see [D] by.

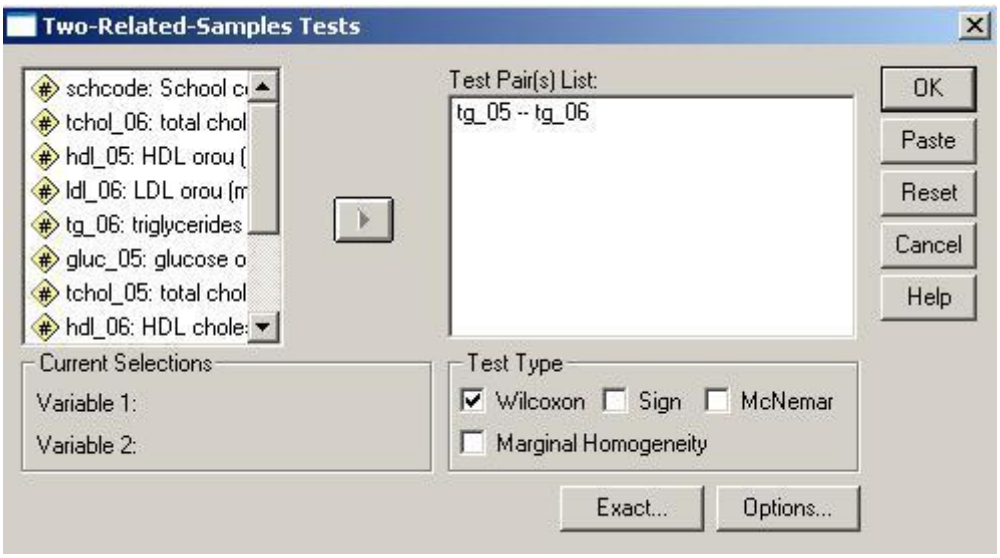
Εικόνα 11.6: Βασική σύνταξη της εντολής ttest

11.4 Wilcoxon Sign rank test με τη χρήση του SPSS

Ας επιστρέψουμε στο παραπάνω παράδειγμα όπου σε μία ομάδα μαθητών πραγματοποιούμε μία διατροφική παρέμβαση για ένα χρόνο προκειμένου να ελέγξουμε αν αυτή η παρέμβαση είναι αποτελεσματική ή όχι όσον αφορά στο λιπιδαιμικό τους προφίλ. Μετράμε, τριγλυκερίδια πριν την έναρξη της παρέμβασης και αμέσως μετά το τέλος της παρέμβασης. Τα τριγλυκερίδια, όμως, δεν ακολουθούν την κανονική κατανομή και συνεπώς ο κατάλληλος στατιστικός έλεγχος για να διερευνήσουμε αυτή την υπόθεση είναι το «***Wilcoxon Sign rank test***», όπως προκύπτει από τα παραπάνω.
Για να το εφαρμόσουμε πρέπει να ακολουθήσουμε τα εξής ***βήματα***:

Analyze ⟶ Nonparametric tests → 2 Related Samples

i. Ανοίγει το παράθυρο διαλόγου που παρουσιάζεται στην *Εικόνα 11.7*



Εικόνα 11.7: Πραγματοποίηση του ελέγχου Wilcoxon – Sign rank test

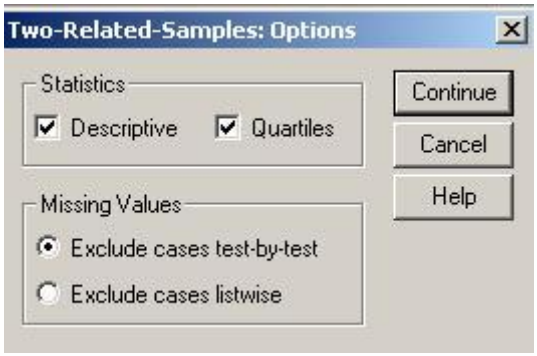
ii. Επιλέγουμε τις μεταβλητές που είναι καταχωρημένες οι 2 μετρήσεις του χαρακτηριστικού (π.χ. tg_05 & tg_06)

iii. Με το βελάκι τις τοποθετούμε στο «***test pair(s) List***»

iv. Στο παράθυρο διαλόγου της *Εικόνας 11.7* φαίνεται ότι είναι προεπιλεγμένο μόνο το Wilcoxon test. Αν επιθυμούμε, παράλληλα να εμφανιστούν και τα αποτελέσματα από το Sign rank test, πρέπει απλά να το επιλέξουμε.

v. Επίσης, πατώντας το κουμπί επιλογών «***Options***» ανοίγει το πλαίσιο διαλόγου της *Εικόνας 11.8*. Από αυτό το πλαίσιο μπορούμε να ζητήσουμε από το πρόγραμμα να μας εμφανίσει στο output και τα περιγραφικά χαρακτηριστικά των τριγλυκεριδίων στις 2 μετρήσεις, τσεκάροντας τις επιλογές «***Descriptive***» και «***Quartiles***».

vi. Πατάμε «***Continue***» & «***Ok***».



Εικόνα 11.8: Επιλογή για εμφάνιση των περιγραφικών χαρακτηριστικών των επιπέδων τριγλυκεριδίων πριν και μετά την παρέμβαση (Επιλογή «Options»)

vii. Στους *Πίνακες 11.3 & 11.4* παρουσιάζονται τα αποτελέσματα του Wilcoxon test. Από τον Πίνακα "***Test Statistics***" (*Πίνακας 11.3*) διαπιστώνουμε ότι η αλλαγή στα επίπεδα των τριγλυκεριδίων είναι στατιστικά σημαντική αφού Asymp. sig.<0,001 (δηλ. η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης, ότι δηλαδή δεν υπάρχει συσχέτιση ανάμεσα στην παρέμβαση και τα επίπεδα τριγλυκεριδιών) είναι <α=0,05. Από τον Πίνακα 10.4 διαπιστώνουμε ότι τόσο η διάμεσος, όσο και το 25^{ο} και 75^{ο} τεταρτημόριο της κατανομής των τριγλυκεριδίων αμέσως μετά την παρέμβαση είναι χαμηλότερα σε σχέση με τα αντίστοιχα πριν την παρέμβαση. Αντίθετα, η μέση τιμή των τριγλυκεριδίων μετά την παρέμβαση είναι υψηλότερη σε σχέση με πριν την παρέμβαση. Όμως, δεδομένου ότι η κατανομή των τριγλυκεριδίων δεν είναι κανονική, η μέση τιμή δεν είναι το κατάλληλο περιγραφικό στατιστικό μέτρο για την διεξαγωγή συμπερασμάτων.

Test Statistics[b]

	tg_06: triglycerides orou (mg/dL) - tg_05: triglycerides (mg%)
Z	-3,467[a]
Asymp. Sig. (2-tailed)	,001

a. Based on positive ranks.

b. Wilcoxon Signed Ranks Test

Πίνακας 11.3: Αποτελέσματα από τον έλεγχο Wilcoxon – Sign rank test για τον έλεγχο ύπαρξης συσχέτισης ανάμεσα στην πραγματοποίηση ετήσιας διατροφικής παρέμβασης και την αλλαγή στα επίπεδα τριγλυκεριδίων.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
tg_05: triglycerides (mg%)	398	75,540	31,7985	25,0	238,0	50,000	68,500	90,000
tg_06: triglycerides orou (mg/dL)	481	76,4978	61,94503	22,00	976,00	47,5000	62,0000	88,3500

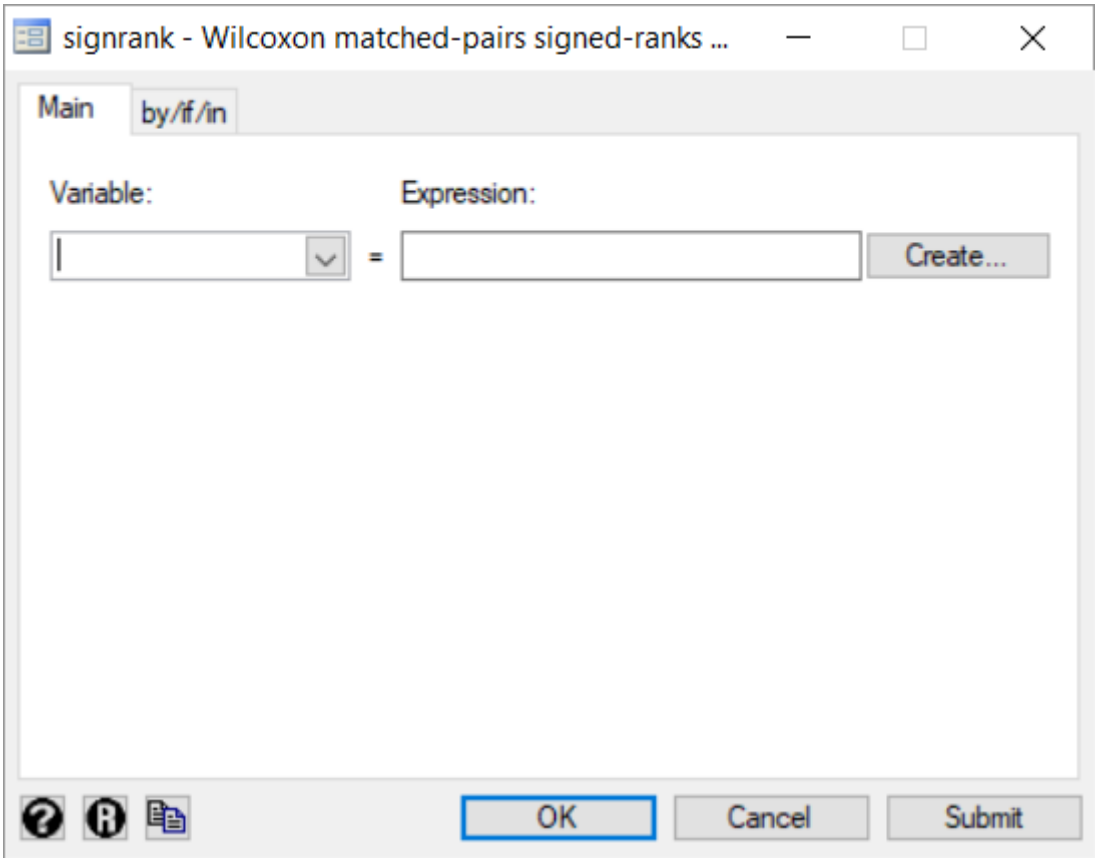
Πίνακας 11.4: Περιγραφικά χαρακτηριστικά για τα επίπεδα των τριγλυκεριδίων πριν και μετά την παρέμβαση.

11.5 Wilcoxon Sign rank test με τη χρήση του STATA

Για να το εφαρμόσουμε το Wilcoxon rank test στο STATA πρέπει να ακολουθήσουμε τα εξής ***βήματα***:

Statistics → Nonparametric analysis → Tests of hypotheses → Wilcoxon matched-pairs signed-rank test

Και ανοίγει το παράθυρο διαλόγου που παρουσιάζεται στην *Εικόνα 11.9*



Εικόνα 11.9: Πραγματοποίηση του ελέγχου Wilcoxon – Sign rank test

Στο STATA χρησιμοποιείται η εντολή ***signrank*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 11.10*:

Syntax

Wilcoxon matched-pairs signed-ranks test

`signrank` *varname* = *exp* [*if*] [*in*]

Sign test of matched pairs

`signtest` *varname* = *exp* [*if*] [*in*]

`by` is allowed with `signrank` and `signtest`; see [D] **by**.

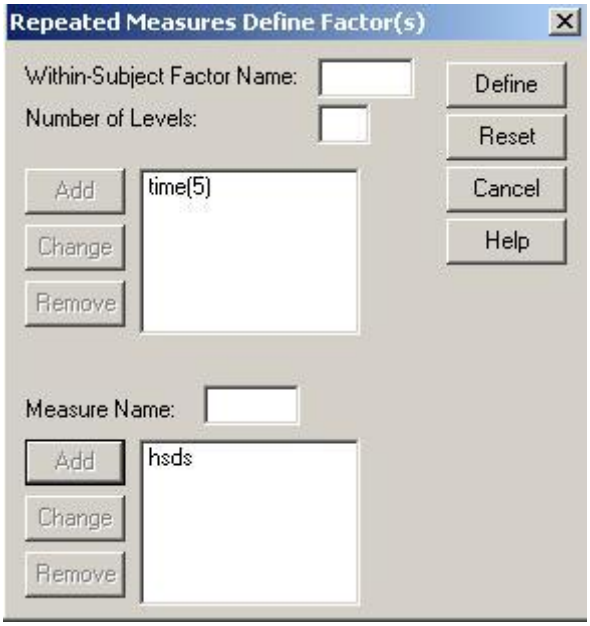
Εικόνα 11.10: Βασική σύνταξη της εντολής signrank

11.4 Ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις με τη χρήση του SPSS.

Ας υποθέσουμε σε κάποια παιδιά με άσθμα χορηγήθηκε το φάρμακο Α και στα υπόλοιπα χορηγήθηκε το φάρμακο Β. Σκοπός της μελέτης ήταν να μελετήσουμε αν το φάρμακο επηρεάζει την αύξηση του ύψους των παιδιών, δηλαδή να δούμε αν υπάρχει διαφορά στο ρυθμό αύξησης του ύψους μεταξύ των δύο ομάδων. Το ύψος των παιδιών εκφράστηκε ως height standard deviation score (hsds). Μετρήσεις πραγματοποιήθηκαν τη στιγμή που χορηγήθηκε η πρώτη δόση (χρονική στιγμή 0) καθώς επίσης και στους 6, 12, 24 και 36 μήνες μετά την πρώτη δόση.
Σε αυτή την μελέτη το χαρακτηριστικό που μας ενδιαφέρει (hsds) έχει μετρηθεί 5 φορές σε κάθε παιδί και η απάντηση στο συγκεκριμένο ερώτημα θα δοθεί πραγματοποιώντας ***ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις***, ακολουθώντας τα εξής ***βήματα***:

Analyze ⟶ General Linear Model ⟶ Repeated Measures…

i. Ανοίγει ένα παράθυρο διαλόγου (*Εικόνα 11.11*)



Εικόνα 11.11: Ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις

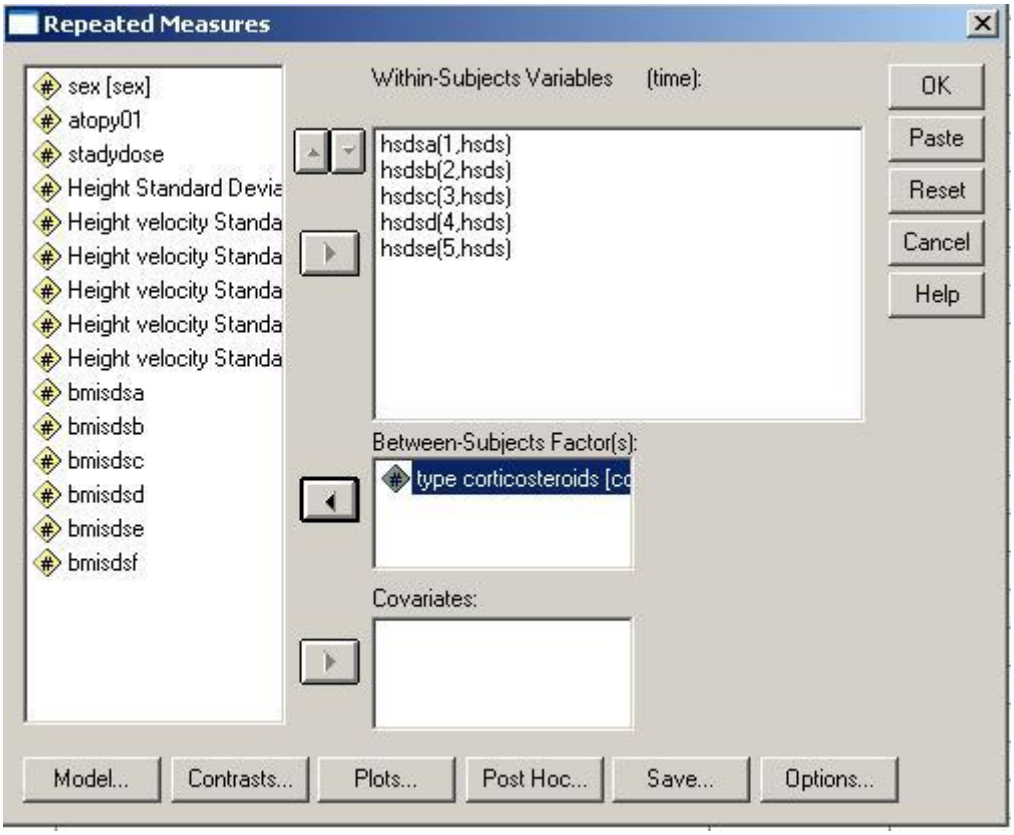
ii. Σε αυτό το παράθυρο διαλόγου πρέπει να συμπληρώσουμε:
- ***Within-Subject Factor Name*:** Ένα όνομα για τις καταστάσεις-χρονικές στιγμές στις οποίες πραγματοποιήθηκαν οι διάφορες μετρήσεις στο ίδιο υποκείμενο (π.χ. time).
- ***Number of levels*:** Τον αριθμό των μετρήσεων που πραγματοποιήθηκαν σε κάθε μονάδα (στο παράδειγμά μας: 5 μετρήσεις του ύψους σε κάθε παιδί).

iii. Αφού συμπληρώσουμε τα παραπάνω πατάμε το πλήκτρο «***Add***».

iv. Επίσης, πρέπει να συμπληρώσουμε το:
- ***Measure Name*:** Ένα όνομα για το μέγεθος που μετρήσαμε (π.χ. score ή distance, hsds (στο παράδειγμά μας)). Και πάλι πατάμε το πλήκτρο «***Add***».

Σημείωση: Εδώ πρέπει να σημειώσουμε ότι τα δεδομένα μας προκειμένου να είναι πραγματοποιήσιμη η συγκεκριμένη ανάλυση θα πρέπει να έχουν την εξής μορφή: Κάθε μέτρηση για κάθε μονάδα που συμμετέχει στη μελέτη θα πρέπει να είναι καταχωρημένη σε ξεχωριστή στήλη. Δηλαδή, πρέπει να έχουμε τόσες στήλες όσες είναι και οι επαναλαμβανόμενες μετρήσεις που έχουμε.

v. Στη συνέχεια, πατάμε το πλήκτρο «**Define**», και εμφανίζεται ένα άλλο πλαίσιο διαλόγου (*Εικόνα 11.12*).



Εικόνα 11.12: Προσδιορισμός των εντός των μονάδων και μεταξύ των μονάδων, παραγόντων.

Στο πλαίσιο της *Εικόνας 11.12* ορίζουμε ως:

i. ***Within-Subjects Variables:*** τις μεταβλητές στις οποίες είναι καταχωρημένες οι διάφορες μετρήσεις (π.χ. μεταβλητές hsdsa, hsdsb, hsdsc, hsdsd και hsdse που αντιστοιχούν στις μετρήσεις του height standard deviation score στους 0, 6, 12, 24 και 36 μήνες μετά τη χορήγηση της πρώτης δόσης, αντίστοιχα.).
ii. ***Between Subjects Factor(s):*** την ή τις κατηγορικές μεταβλητές, οι οποίες διαφέρουν μεταξύ των συμμετεχόντων ή αλλιώς χωρίζουν τους συμμετέχοντες σε 2 ή περισσότερες ομάδες (π.χ. τη μεταβλητή corticos (τύπος κορτικοστεροειδούς)).
iii. ***Covariates:*** Δηλώνουμε όσες συνεχείς μεταβλητές διαθέτουμε ως συν-μεταβλητές.

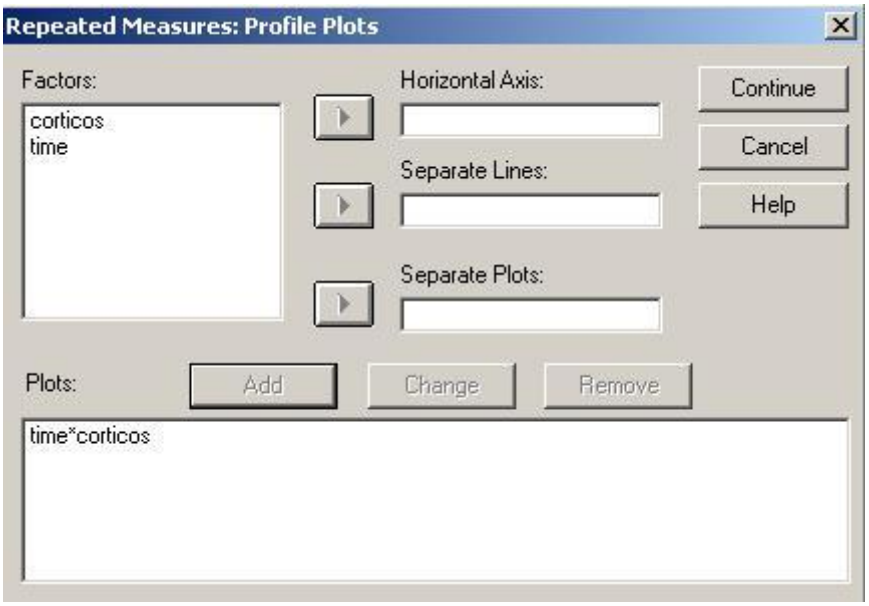
Στο κάτω μέρος του πλαισίου της *Εικόνας 11.8* υπάρχει μία σειρά από πλήκτρα επιλογών (***Model, Contrasts, Plots, Post Hoc, Save,*** και ***Options***). Οι επιλογές που

μας παρέχει κάθε ένα από αυτά τα πλήκτρα θα παρουσιαστούν παρακάτω αρκετά αναλυτικά.

Πατώντας το πλήκτρο ***Plots*** εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 11.13*. Μέσα από αυτό το πλαίσιο διαλόγου μας δίνεται η δυνατότητα να δημιουργήσουμε τα **profile plots (**γραφήματα που παριστάνουν τις μέσες τιμές του χαρακτηριστικού που μετράμε σε όλες τις χρονικές στιγμές και για κάθε ομάδα, ξεχωριστά**)**, ως εξής:

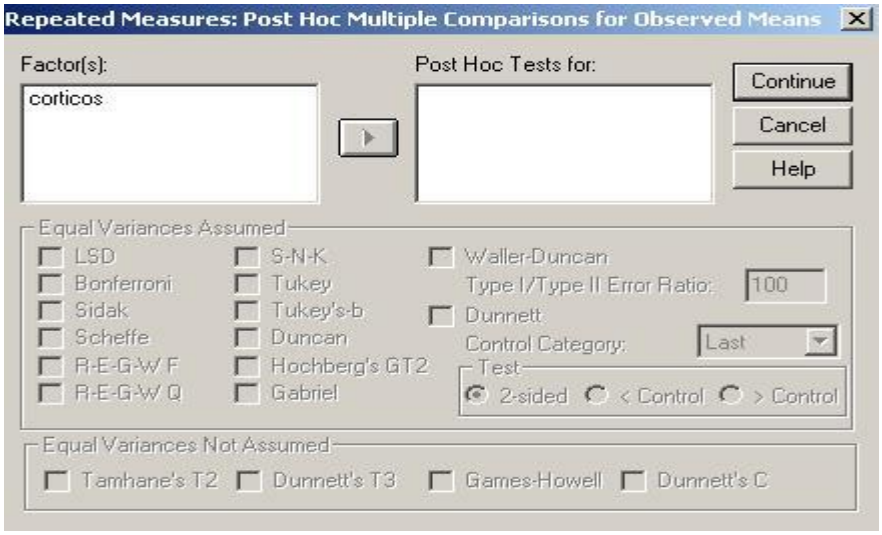
i. ***Separate lines:*** τον παράγοντα που διαφέρει μεταξύ των ατόμων (π.χ. τη μεταβλητή corticos (τύπος κορτικοστεροειδούς)).
ii. ***Horizontal Axis:*** Δηλώνουμε το όνομα που έχουμε ορίσει στο ***Within-Subject Factor Name,*** δηλ. τον «εντός των ατόμων» παράγοντα.
iii. Πατάμε το πλήκτρο «***Add***».

Από αυτό το γράφημα μπορούμε να δούμε αν υπάρχει αύξηση ή μείωση του μεγέθους κατά μήκος του «εντός των ατόμων» παράγοντα, στα διάφορα επίπεδα του παράγοντα που διαφέρει μεταξύ των ατόμων (ομάδες). Επίσης, μπορούμε να διαπιστώσουμε αν υπάρχει αλληλεπίδραση του «μεταξύ των ατόμων» παράγοντα με τον «εντός των ατόμων» παράγοντα, παρατηρώντας αν οι γραμμές που έχουν προκύψει ενώνοντας τις μέσες τιμές των μετρήσεών μας σε κάθε ομάδα και σε κάθε χρονική στιγμή διασταυρώνονται. Αν είναι παράλληλες, συνεπάγεται ότι δεν υπάρχει αλληλεπίδραση (interaction).



Εικόνα 11.13: Το πλαίσιο διαλόγου που εμφανίζεται πατώντας το πλήκτρο **Plots.**

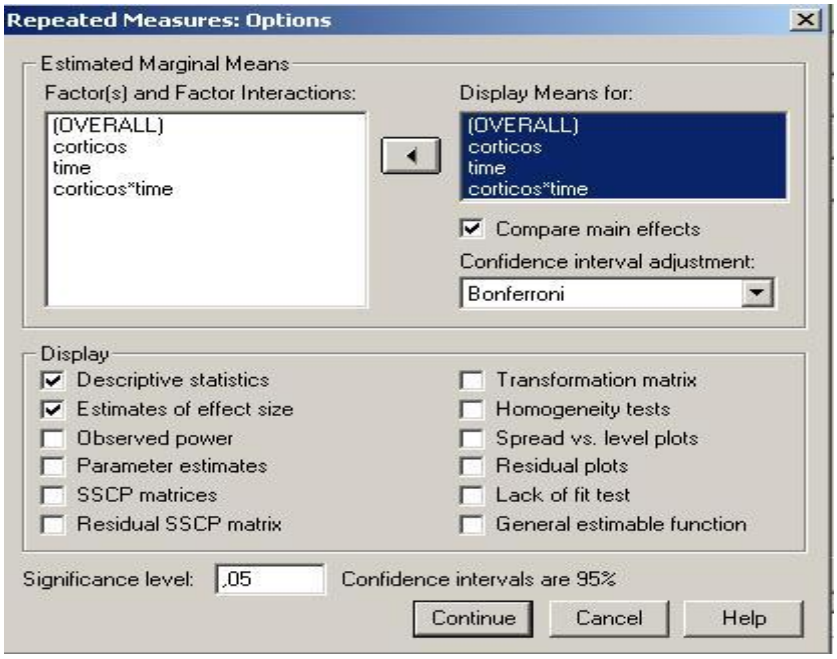
Πατώντας το πλήκτρο **Post Hoc** εμφανίζεται το πλαίσιο διαλόγου που απεικονίζεται στην *Εικόνα 11.14***.**



Εικόνα 11.14: Το πλαίσιο διαλόγου που εμφανίζεται πατώντας το πλήκτρο **Post Hoc.**

Εφόσον έχουμε διαπιστώσει ότι οι μέσες τιμές του μεγέθους που μας ενδιαφέρει διαφέρει μεταξύ των επιπέδων ενός ή περισσοτέρων παραγόντων, μπορούμε να διερευνήσουμε μεταξύ ποιων ακριβώς επιπέδων υπάρχει αυτή η διαφορά. Αυτό πραγματοποιείται επιλέγοντας τον παράγοντα από το πλαίσιο ***Factors*** και μετακινώντας το στο πλαίσιο ***Post Hoc test for.*** Στη συνέχεια θα πρέπει να επιλέξουμε ακριβώς από κάτω με ποιον από τους διαθέσιμους τρόπου επιθυμούμε να πραγματοποιηθεί η διόρθωση για τους πολλαπλούς ελέγχους. Αυτά τα test πραγματοποιούνται μόνο για τους παράγοντες μεταξύ των μονάδων που συμμετέχουν στη μελέτη και εφαρμόζονται στο μέσο όρο των παρατηρήσεων που προκύπτει αν συνυπολογίσουμε όλες τις μετρήσεις που πραγματοποιούνται κατά μήκος του χρόνου σε κάθε μονάδα.

Πατώντας το πλήκτρο ***Options*** εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 11.15.*



Εικόνα 11.15: Το πλαίσιο διαλόγου που εμφανίζεται πατώντας το πλήκτρο **Options.**

Από το πλαίσιο διαλόγου της *Εικόνας 11.15* μπορούμε να ζητήσουμε να εμφανιστούν στο output οι μέσες τιμές ανά επίπεδο παράγοντα μεταξύ των μονάδων ανεξαρτήτως χρονικής στιγμής που πραγματοποιήθηκε η μέτρηση, οι μέσες τιμές για τις διάφορες χρονικές στιγμές ανεξαρτήτως ομάδας, η συνολική μέση τιμή, μεταφέροντας τους παράγοντες από το πλαίσιο ***Factor(s) and Factor interactions*** στο πλαίσιο ***Display means for.*** Επίσης, μπορούμε να επιλέξουμε να γίνει και σύγκριση των παραπάνω μέσων τιμών τσεκάροντας την επιλογή ***Compare main effects***. Σε αυτή την περίπτωση θα πρέπει να επιλέξουμε και τη μέθοδο βάση της οποίας θα πραγματοποιηθεί η διόρθωση για τους πολλαπλούς ελέγχους (στο παράδειγμά μας επιλέξαμε τη μέθοδο Bonferroni). Τέλος, μπορούμε να ζητήσουμε να εμφανιστούν στο output και περιγραφικά χαρακτηριστικά (π.χ. μέση τιμή, τυπική απόκλιση και απόλυτος αριθμός) για κάθε χρονική στιγμή και κάθε επίπεδο του μεταξύ των μονάδων παράγοντα.

Πατώντας το πλήκτρο **Contrasts** εμφανίζεται το πλαίσιο διαλόγου που φαίνεται στην *Εικόνα 11.16*.

Εικόνα 11.16: Το πλαίσιο διαλόγου που εμφανίζεται πατώντας το πλήκτρο **Contrasts.**

Τα ***contrasts*** χρησιμοποιούνται για να ελέγξουμε τις διαφορές μεταξύ των επιπέδων διαφόρων παραγόντων. Μπορούμε να επιλέξουμε διαφορετικό contrast για κάθε παράγοντα. Αυτό πραγματοποιείται επιλέγοντας αρχικά τον παράγοντα που θέλουμε και στη συνέχεια επιλέγοντας το contrast που επιθυμούμε για να πραγματοποιηθεί η σύγκριση μεταξύ των επιπέδων του συγκεκριμένου παράγοντα και πατώντας το πλήκτρο **change.** Τα διαθέσιμα contrasts είναι None, Deviation, Simple, Difference, Helmet, Repeated, Polynomial.

- Deviation: Η τιμή του μεγέθους που μετράμε σε κάθε επίπεδο του παράγοντα συγκρίνεται με το συνολικό μέσο.
- Simple: Οι τιμές του μεγέθους σε κάθε επίπεδο συγκρίνονται με τις τιμές ενός επιπέδου που έχει οριστεί ως κατηγορία αναφοράς. Προεπιλεγμένο στο SPSS είναι να είναι κατηγορία αναφοράς η μεγαλύτερη κατηγορία. Υπάρχει όμως και η δυνατότητα να ορίσουμε ως κατηγορία αναφοράς την μικρότερη κατηγορία, επιλέγοντας το **First** και πατώντας **Change**.
- Difference: Συγκρίνουμε τις τιμές κάθε επιπέδου του παράγοντα με αυτές του αμέσως προηγούμενου επιπέδου (σύμφωνα με την κωδικοποίηση).
- Helmet: Συγκρίνουμε τις τιμές κάθε επιπέδου με το μέσο όρο των παρατηρήσεων όλων των επόμενων επιπέδων.
- Repeated: Οι τιμές κάθε επιπέδου συγκρίνονται με αυτές του αμέσως επόμενου επιπέδου. Δηλαδή, η σύγκριση γίνεται μεταξύ διαδοχικών επιπέδων.

- Polynomial: Ελέγχεται αν η σχέση που συνδέει τις τιμές του μεγέθους που μετράμε στα διάφορα επίπεδα του παράγοντα είναι γραμμική, πολυώνυμου $2^{ου}$ βαθμού ή πολυώνυμο $3^{ου}$ βαθμού.

Στους *Πίνακες 11.5-11.10* παρουσιάζονται τα αποτελέσματα από την ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις.
Το πρώτο πράγμα που πρέπει να ελέγξουμε είναι το αν ισχύει η προϋπόθεση «compound symmetry» για τον πίνακα διακυμάνσεων-συνδιακυμάνσεων. Από τον *Πίνακα 11.5* παρατηρούμε ότι απορρίπτεται η υπόθεση αυτή αφού από το Mauchly's Test of Sphericity έχουμε sig<0,001. Αφού η βασική προϋπόθεση της ανάλυσης διακύμανσης για επαναλαμβανόμενες μετρήσεις παραβιάζεται, τα αποτελέσματα θα ερμηνευτούν λαμβάνοντας υπόψη κάποιες από τις διορθώσεις (Greenhouse-Geisser, Huynh-Feldt, Lower-Bound).

Mauchly's Test of Sphericity[b]

Measure: hsds

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
timr	,470	75,753	9	,000	,733	,765	,250

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b.

Design: Intercept+corticos
Within Subjects Design: timr

Πίνακας 11.5: Έλεγχος για το αν ο πίνακας διακύμανσης-συνδιακύμανσης των παρατηρήσεων της ίδιας μονάδας έχει μία συγκεκριμένη δομή (compound symmetry).

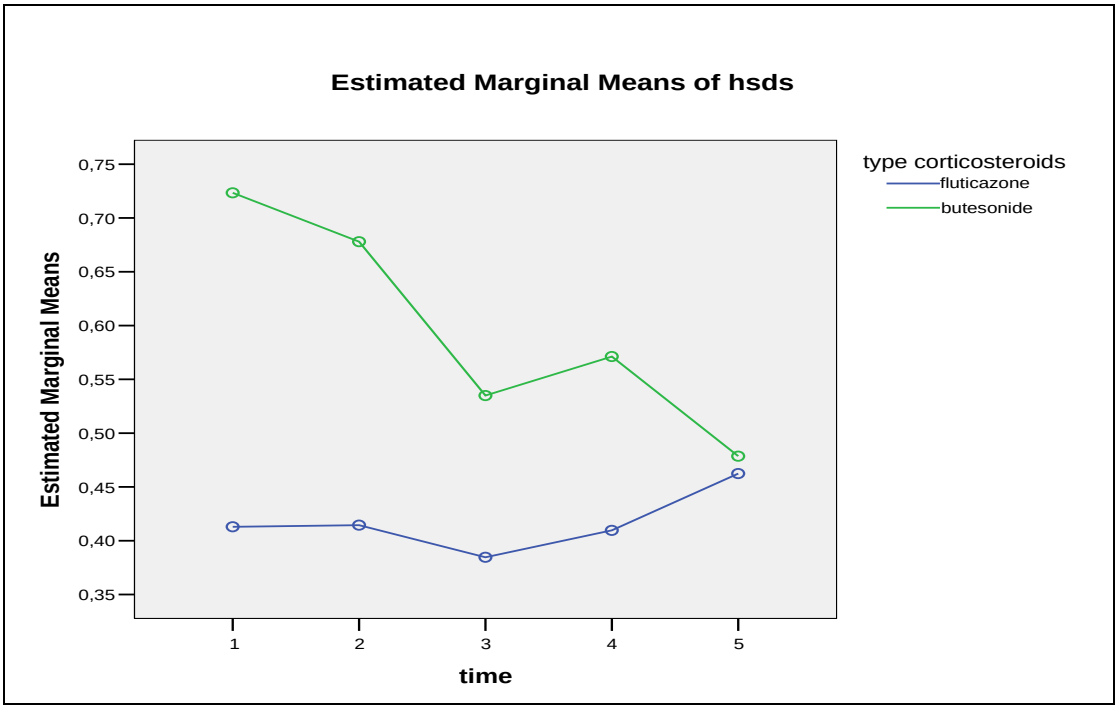
Το επόμενο πράγμα που πρέπει να ελέγξουμε είναι κατά πόσο η αλλαγή του μεγέθους που μετράμε (π.χ. height standard deviation score) κατά μήκος του χρόνου διαφέρει μεταξύ των ομάδων του «μεταξύ των ατόμων» παράγοντα (π.χ. 2 διαφορετικές θεραπείες με κορτικοστεροειδή). Δηλαδή, ουσιαστικά, πρέπει να διερευνήσουμε αν υπάρχει στατιστικά σημαντική αλληλεπίδραση. Η απάντηση σε αυτό το ερώτημα βρίσκεται στον *Πίνακας 11.6* & στην *Εικόνα 11.17*. Από τον *Πίνακα 11.6* παρατηρούμε ότι ο έλεγχος για την ύπαρξη ή όχι αλληλεπίδρασης ανάμεσα στην ομάδα (τύπος κορτικοστεροειδούς που χορηγήθηκε στα παιδιά) και το χρόνο οδηγεί σε στατιστικά σημαντικά αποτελέσματα, αφού sig. για το time*corticos <0,001. Αυτό σημαίνει ότι η τροχιά που προκύπτει συνδέοντας το μέσο όρο των height standard deviation scores όλων των παιδιών σε κάθε ομάδα για κάθε χρονική στιγμή είναι διαφορετική μεταξύ των δύο ομάδων (*Εικόνα 11.17*).

Tests of Within-Subjects Effects

Measure: hsds

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
time	Sphericity Assumed	,947	4	,237	7,160	,000
	Greenhouse-Geisser	,947	2,933	,323	7,160	,000
	Huynh-Feldt	,947	3,059	,310	7,160	,000
	Lower-bound	,947	1,000	,947	7,160	,009
time * corticos	Sphericity Assumed	1,351	4	,338	10,218	,000
	Greenhouse-Geisser	1,351	2,933	,461	10,218	,000
	Huynh-Feldt	1,351	3,059	,442	10,218	,000
	Lower-bound	1,351	1,000	1,351	10,218	,002
Error(time)	Sphericity Assumed	13,486	408	,033		
	Greenhouse-Geisser	13,486	299,139	,045		
	Huynh-Feldt	13,486	311,975	,043		
	Lower-bound	13,486	102,000	,132		

Πίνακας 11.6: Έλεγχος για ύπαρξη αλληλεπίδρασης ανάμεσα στον τύπο κορτικοστεροειδούς και το χρόνο. (έλεγχος παραλληλότητας)



Εικόνα 11.17: Profile plot

Στη συνέχεια, πρέπει να ελέγξουμε την κυρίως επίδραση του «μεταξύ των ατόμων» παράγοντα. Από τον *Πίνακα 11.7* προκύπτει ότι δεν υπάρχει καμία στατιστικά σημαντική διαφορά στις μέσες τιμές των height standard deviation scores μεταξύ των παιδιών που έλαβαν ως θεραπεία άσθματος τη θεραπεία Α και αυτών που έλαβαν τη θεραπεία Β (sig = 0,389). Βέβαια, αν θέλουμε να δούμε περισσότερες λεπτομέρειες αναφορικά με αυτό το θέμα μπορούμε να χρησιμοποιήσουμε τις επιλογές του πλήκτρου **Options**, όπου μπορούμε να δούμε μέση τιμή και τυπική απόκλιση του height standard deviation score για κάθε ομάδα (*Πίνακας 11.8*).

Σημείωση: Δεδομένου ότι έχουμε δείξει πως υπάρχει στατιστικά σημαντική αλληλεπίδραση ανάμεσα σε χρόνο και τύπο κορτικοστεροειδούς δε παρουσιάζει ιδιαίτερο ενδιαφέρον ο έλεγχος για την κυρίως επίδραση μεταξύ των ομάδων.

Tests of Between-Subjects Effects

Measure: hsds

Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	133,481	1	133,481	23,649	,000	,188
corticos	4,225	1	4,225	,749	,389	,007
Error	575,710	102	5,644			

Πίνακας 11.7: Έλεγχος για την κυρίως επίδραση του «μεταξύ των ατόμων» παράγοντα

Estimates

Measure: hsds

type corticosteroids	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
fluticazone	,417	,145	,130	,704
butesonide	,597	,150	,299	,895

Εικόνα 11.18: Περιγραφικά χαρακτηριστικά του height standard deviation score για κάθε ομάδα, ξεχωριστά.

Τέλος, μας ενδιαφέρει να ελέγξουμε την κυρίως επίδραση του «εντός των ατόμων» παράγοντα. Από τον *Πίνακα 11.6* βλέπουμε πως οι τιμές των height standard deviation scores δεν παραμένουν σταθερές κατά μήκος του χρόνου (sig<0,001 για το time). Από τις επιλογές του πλήκτρου ***Options*** μπορούμε να δούμε τις μέσες τιμές (τυπική απόκλιση) των height standard deviation scores σε κάθε χρονική στιγμή ανεξάρτητα από την ομάδα (τύπο κορτικοστεροειδούς) (*Πίνακας 11.9*). Επίσης, μπορούμε να δούμε όλους τους ανά δύο ελέγχους μεταξύ των χρονικών στιγμών (*Πίνακας 11.10*) έχοντας επιλέξει έναν τρόπο διόρθωσης για πολλαπλούς ελέγχους (π.χ. Bonferroni).

Estimates

Measure: hsds

time	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	,568	,104	,363	,774
2	,546	,099	,349	,743
3	,460	,107	,247	,673
4	,490	,109	,274	,707
5	,471	,108	,257	,684

Πίνακας 11.9: Περιγραφικά χαρακτηριστικά του height standard deviation score για κάθε χρονική στιγμή, ξεχωριστά.

Pairwise Comparisons

Measure: hsds

(I) time	(J) time	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	,022	,022	1,000	-,042	,086
	3	,108*	,026	,001	,033	,184
	4	,078	,028	,072	-,004	,159
	5	,098*	,031	,019	,010	,185
2	1	-,022	,022	1,000	-,086	,042
	3	,086*	,025	,007	,015	,158
	4	,056	,027	,438	-,023	,134
	5	,076	,029	,098	-,007	,158
3	1	-,108*	,026	,001	-,184	-,033
	2	-,086*	,025	,007	-,158	-,015
	4	-,031	,017	,686	-,078	,017
	5	-,011	,024	1,000	-,081	,060
4	1	-,078	,028	,072	-,159	,004
	2	-,056	,027	,438	-,134	,023
	3	,031	,017	,686	-,017	,078
	5	,020	,020	1,000	-,036	,076
5	1	-,098*	,031	,019	-,185	-,010
	2	-,076	,029	,098	-,158	,007
	3	,011	,024	1,000	-,060	,081
	4	-,020	,020	1,000	-,076	,036

Based on estimated marginal means

*. The mean difference is significant at the ,05 level.

a. Adjustment for multiple comparisons: Bonferroni.

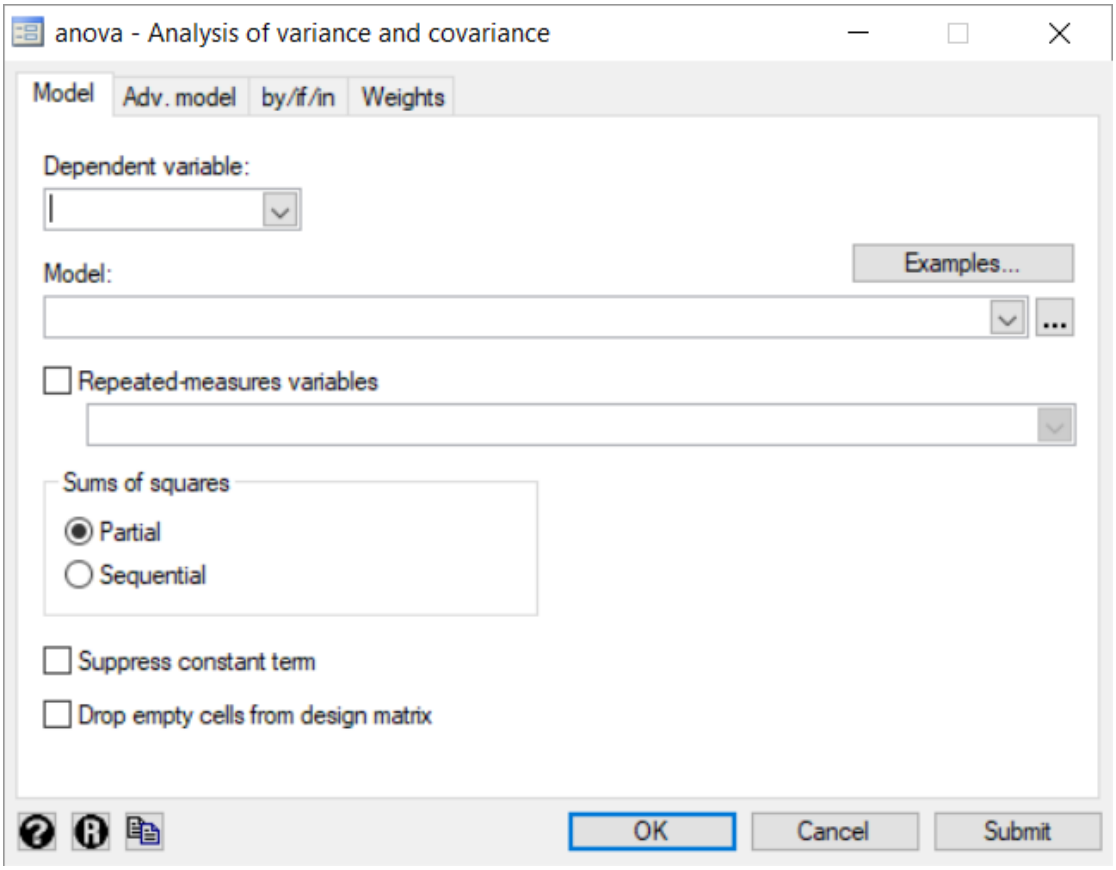
Πίνακας 11.10: Όλοι οι ανά δύο έλεγχοι μεταξύ των χρονικών στιγμών.

11.5 Ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις με τη χρήση του STATA.

Η ***ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις***, στο STATA πραγματοποιείται ακολουθώντας τα εξής ***βήματα***:

Statistics → Linear models and related → ANOVA/MANOVA → Analysis of variance and covariance

Και ανοίγει ένα παράθυρο διαλόγου (*Εικόνα 11.19*)



Εικόνα 11.19: Ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις

Στο STATA χρησιμοποιείται η εντολή ***anova*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 11.20*:

Syntax

anova *varname* [*termlist*] [*if*] [*in*] [*weight*] [, *options*]

where *termlist* is a factor-variable list (see [U] **11.4.3 Factor variables**) with the following additional features:

- Variables are assumed to be categorical; use the `c.` factor-variable operator to override this.
- The `|` symbol (indicating nesting) may be used in place of the `#` symbol (indicating interaction).
- The `/` symbol is allowed after a term and indicates that the following term is the error term for the preceding terms.

options	Description
Model	
`repeated(`*varlist*`)`	variables in *term*s that are repeated-measures variables
`partial`	use partial (or marginal) sums of squares
`sequential`	use sequential sums of squares
`noconstant`	suppress constant term
`dropemptycells`	drop empty cells from the design matrix
Adv. model	
`bse(`*term*`)`	between-subjects error term in repeated-measures ANOVA
`bseunit(`*varname*`)`	variable representing lowest unit in the between-subjects error term
`grouping(`*varname*`)`	grouping variable for computing pooled covariance matrix

`bootstrap`, `by`, `fp`, `jackknife`, and `statsby` are allowed; see [U] **11.1.10 Prefix commands**.
Weights are not allowed with the `bootstrap` prefix; see [R] **bootstrap**.
`aweights` are not allowed with the `jackknife` prefix; see [R] **jackknife**.
`aweights` and `fweights` are allowed; see [U] **11.1.6 weight**.
See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

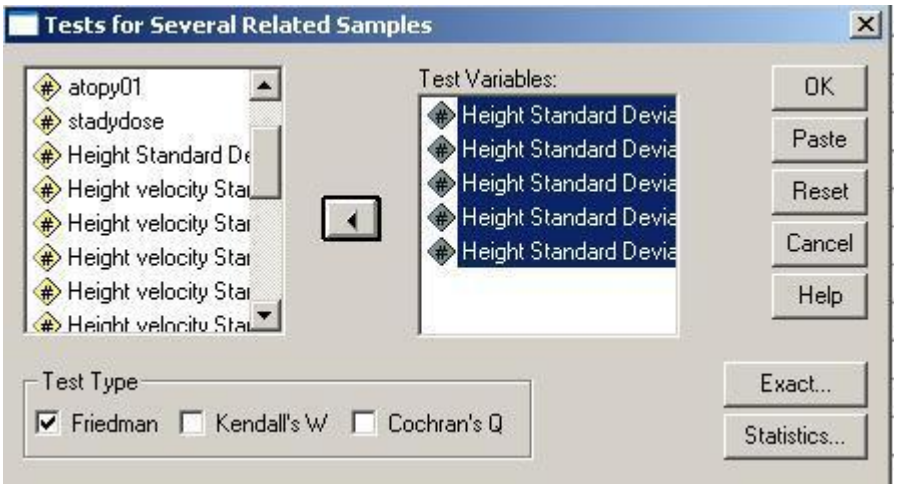
Εικόνα 11.20: Βασική σύνταξη της εντολής anova

11.6 Friedman test με τη χρήση του SPSS

Ο έλεγχος Friedman είναι ο αντίστοιχος μη παραμετρικός έλεγχος της ανάλυσης διακύμανσης για επαναλαμβανόμενες μετρήσεις, όταν δεν υπάρχει «μεταξύ των ατόμων» παράγοντας και οι μετρήσεις που έχουν πραγματοποιηθεί είναι περισσότερες από 2 για κάθε άτομο. Ας επιστρέψουμε, λοιπόν, στο παράδειγμα της *παραγράφου 11.4*, και ας θεωρήσουμε ότι σκοπός μας είναι να διερευνήσουμε αν υπάρχει μεταβολή σο ύψος των παιδιών στα οποία χορηγήθηκε η θεραπεία Α κατά μήκος του χρόνου. Ουσιαστικά, λοιπόν, σε αυτή την περίπτωση δεν υπάρχει «μεταξύ των ατόμων» παράγοντας. Επίσης, ας θεωρήσουμε ότι το ύψος των παιδιών σε κάθε χρονική στιγμή δεν ακολουθεί την κανονική κατανομή. Σε αυτή την περίπτωση, ο κατάλληλος στατιστικός έλεγχος είναι ο έλεγχος Friedman, ο οποίος πραγματοποιείται με τα εξής ***βήματα***:

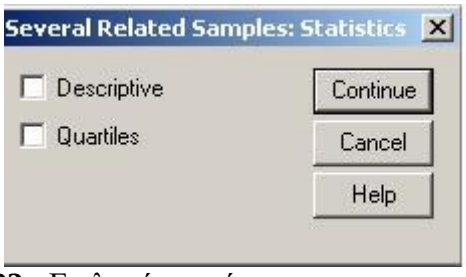
Analyze ⟶ Nonparametric tests → K Related Samples

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 11.21*.
ii. Στο «***Test Variables***» τοποθετούμε τις μεταβλητές στις οποίες είναι καταχωρημένες οι τιμές των μετρήσεων του ύψους στα παιδιά στις διάφορες χρονικές στιγμές.



Εικόνα 11.21: Πραγματοποίηση του ελέγχου «Friedman»

iii. Στη συνέχεια πατάμε το κουμπί επιλογών «***Statistics***» και εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 11.22*.
iv. Πατάμε «***Ok***»



Εικόνα 11.22: Επιλογή εμφάνισης των περιγραφικών χαρακτηριστικών για κάθε μέτρηση.

v. Τα αποτελέσματα παρουσιάζονται στους *Πίνακες 11.11 & 11.12*. Από τον πίνακα «***test Statistics***» του output, διαπιστώνουμε ότι η μεταβολή του ύψους των παιδιών κατά μήκος του χρόνου είναι στατιστικά σημαντική, αφού η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης (ότι δηλ. δεν υπάρχει διαφορά στο ύψος των παιδιών στις διάφορες χρονικές στιγμές που έγιναν οι μετρήσεις) είναι Asymp. Sig. <0,001<α=0,05.

Test Statistics[a]

N	104
Chi-Square	25,179
df	4
Asymp. Sig.	,000

a. Friedman Test

Εικόνα 11.23: Αποτελέσματα από τον έλεγχο Friedman για την αλλαγή του ύψους κατά μήκος του χρόνου σε παιδιά με άσθμα που του έχει χορηγηθεί η θεραπεία Α.

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Height Standard Deviation Score (0 m)	104	,5622	1,06286	-2,50	4,13	-,2875	,7100	1,2625
Height Standard Deviation Score (6 m)	104	,5412	1,01593	-2,12	3,82	-,1350	,7150	1,2725
Height Standard Deviation Score (12 m)	104	,4569	1,09059	-2,24	4,77	-,3100	,5750	1,1875
Height Standard Deviation Score (24 m)	104	,4873	1,11057	-2,47	4,76	-,3850	,6150	1,1775
Height Standard Deviation Score (36 m)	104	,4702	1,09086	-2,39	4,81	-,3475	,5100	1,0900

Εικόνα 11.24: Περιγραφικά στοιχεία της κάθε μέτρησης

12. Ανάλυση επιβίωσης

12.1 Εισαγωγή

Οι μελέτες επιβίωσης στηρίζονται στην ανάλυση του χρόνου από την έναρξη της παρατήρησης του ατόμου μέχρι τηη πραγματοποίηση του γεγονότος (time to event analysis). Είναι σαφές ότι τα δυο βασικά χαρακτηριστικά των μεθόδων αυτών είναι αφενός **ο χρόνος,** ο οποίος αποτελεί την μεταβλητή ενδιαφέροντος και αφετέρου ο προσδιορισμός **του γεγονότος,** η εμφάνιση του οποίου ποσοτικοποιεί τον χρόνο της παρατήρησης.

- **Γεγονός (event)** μπορεί να αποτελεί οποιοδήποτε συμβάν το οποίο μπορεί να οριστεί με απόλυτο και σαφή τρόπο. Ο όρος ανάλυση επιβίωσης είναι παραπλανητικός καθώς τα γεγονότα δεν αφορούν πάντοτε θάνατο ούτε έχουν εξ ορισμού αρνητική έννοια. Κατά αυτόν τον τρόπο γεγονός μπορεί να αποτελεί ένα θετικό συμβάν όπως η ίαση ή η βελτίωση ενός νοσήματος ή αντιθέτως ένα αρνητικό συμβάν όπως ο θάνατος ή η επιδείνωση. Παραδοσιακά το γεγονός στην ανάλυση επιβίωσης αναφέρεται ως αποτυχία (failure) καταχωρείται δε με 1 ενώ η απουσία του γεγονότος καταχωρείται με 0. Καθοριστική παράμετρο αποτελεί ο ορισμός του γεγονότος. Πρέπει να γίνεται πάντοτε προ της ενάρξεως της μελέτης και να ισχύει ομοιόμορφα και σταθερά καθ' όλη την διάρκειά της για όλους τους συμμετέχοντες. Το γεγονός μπορεί να αποτελεί κάτι το αντικειμενικά παρατηρούμενο και μετρήσιμο (όπως για παράδειγμα ο θάνατος, ή μια εργαστηριακή μέτρηση όπως ο αριθμός των λευκών αιμοσφαιρίων του αίματος) ή κάτι το οποίο εισάγει υποκειμενικότητα από τον ερευνητή (όπως η κλινική επιδείνωση ενός ασθενούς η εκτίμηση της οποίας βασίζεται στην εμπειρία, γνώση και ιατρική κουλτούρα του εξετάζοντος). Στην δεύτερη περίπτωση το γεγονός θα πρέπει να ορίζεται με αντικειμενικά κριτήρια, βάση της βιβλιογραφίας, χωρίς να αφήνονται περιθώρια για αμφισβητήσεις ή τροποποιήσεις. Ακόμη, όμως, και στην περίπτωση των αντικειμενικών συμβάντων όπως ο θάνατος απαιτείται πολλές φορές περαιτέρω αποσαφηνισμός των αιτιών αυτού καθώς είναι σύνηθες να ενδιαφέρει ένα είδος θανάτου μόνο. Για παράδειγμα, εάν κάποιος μελετά ως γεγονότα τους θανάτους μετά από μια χειρουργική επέμβαση θα πρέπει να ορίσει με σαφήνεια οτιδήποτε θεωρείται μετεγχειρητική επιπλοκή με βάση την βιβλιογραφία. Στην ίδια μελέτη όμως μπορεί ο ερευνητής να είναι επικεντρωμένος στους θανάτους από μετεγχειρητικές λοιμώξεις. Σε αυτή την περίπτωση όλοι οι υπόλοιποι θάνατοι δεν θα θεωρηθούν συμβάντα.
- **Χρόνος**. Ο χρόνος αποτελεί την μεταβλητή ενδιαφέροντος στην ανάλυση επιβίωσης. Η κλίμακα του χρόνου μπορεί να κυμαίνεται από λεπτά μέχρι χρόνια ανάλογα με την βιολογική διαδικασία του γεγονότος που μελετάται. Για την μέτρηση του χρόνου μέχρι το συμβάν είναι απαραίτητο να οριστεί το σημείο εισόδου στην μελέτη καθώς και το σημείο εξόδου.

Το σημείο εισόδου στην μελέτη είναι η χρονική στιγμή της έναρξης παρατήρησης του κάθε ατόμου. Η στιγμή αυτή μπορεί να προσδιορίζεται ηλικιακά, ημερολογιακά ή απλώς σε σχέση με κάποια διαδικασία. Κατά αυτό τον τρόπο το σημείο εισόδου μπορεί να αποτελεί η ημερομηνία γέννησης, μια συγκεκριμένη ημερολογιακή στιγμή η οποία να σηματοδοτεί την έναρξη της παρακολούθησης (για παράδειγμα η πιο πρόσφατη επίσκεψη στο Ιατρείο) ή τέλος σημείο εισόδου μπορεί να αποτελεί μια χειρουργική επέμβαση ή η έναρξη μιας θεραπευτικής αγωγής. Η είσοδος των ασθενών στην μελέτη μπορεί να γίνει ταυτόχρονα οπότε η ημερολογιακή έναρξή της είναι ταυτόσημη με την είσοδο για κάθε ασθενή, στις περισσότερες περιπτώσεις όμως

πραγματοποιείται σταδιακά σε μια χρονική περίοδο ή οποία μπορεί να κυμαίνεται από ώρες μέχρι έτη. Σε αυτή την περίπτωση οι συμμετέχοντες έχουν διαφορετική ημερολογιακή είσοδο γεγονός το οποίο είναι σημαντικό και σε ορισμένες περιπτώσεις πρέπει να λαμβάνεται υπόψιν.

Το σημείο εξόδου, σε μια ιδανική περίπτωση, θα πρέπει για όλους τους ασθενείς να είναι η χρονική στιγμή του συμβάντος. Αυτό όμως σε ελάχιστες περιπτώσεις είναι εφικτό για αρκετούς λόγους. Καταρχάς μπορεί να ολοκληρωθεί η περίοδος παρακολούθησης χωρίς να έχει συμβεί γεγονός σε όλους τους συμμετέχοντες. Εάν για παράδειγμα μελετάται η 5ετής επιβίωση καρκινοπαθών μετά από μια χειρουργική επέμβαση στο τέλος της παρακολούθησης ένα ποσοστό ασθενών θα είναι εν ζωή. Άλλο ενδεχόμενο είναι η αποχώρηση κάποιου ασθενούς από την μελέτη κατά την διάρκεια της παρακολούθησης (lost to follow up). Η αποχώρηση-έξοδος του ασθενούς από την μελέτη μπορεί να είναι προϊόν διαφόρων καταστάσεων: Αλλαγή Νοσοκομείου, επιδείνωση της υγείας, θάνατος από άλλα αίτια ή απλώς αλλαγή τόπου διαμονής. Όλες αυτές οι περιπτώσεις ανήκουν στις **αποκομμένες (censored)** παρατηρήσεις.

- **Αποκοπή (censoring)**: Όταν για ένα άτομο το οποίο συμμετέχει στην μελέτη δεν είναι γνωστό το χρονικό διάστημα από την έναρξη της μελέτης μέχρι το γεγονός, δηλαδή αυτή η παρατήρηση είναι ελλιπής, τότε ονομάζεται **αποκομένη (censored).**

12.1.1 Βασικές έννοιες

Τρεις βασικές συναρτήσεις στην ανάλυση επιβίωσης ανεξάρτητα από την μέθοδο που χρησιμοποιείται (απαραμετρική, ημι-παραμετρική, παραμετρική) είναι η **συνάρτηση επιβίωσης (survival function)**, η **συνάρτηση κινδύνου (hazard function)** και η **αθροιστική συνάρτηση κινδύνου (cumulative hazard function).**

Έστω T η τυχαία μεταβλητή (τ.μ) των χρόνων επιβίωσης του πληθυσμού που μελετάται. Ως τυχαία μεταβλητή θα έχει μια συγκεκριμένη κατανομή που θα περιγράφεται από την συνάρτηση πυκνότητας πιθανότητας $f_T(t)$ καθώς και την αθροιστική συνάρτηση πιθανότητας $F_T(t)$. Η αθροιστική συνάρτηση εκφράζει την πιθανότητα ο χρόνος επιβίωσης ενός ατόμου να είναι μικρότερος από t δηλαδή την πιθανότητα ένα άτομο να έχει υποστεί το γεγονός έως και την χρονική στιγμή t και δίνεται από την εξίσωση:

$$F(t) = P(T < t) = \int_0^t f(u)du$$

Η αθροιστική συνάρτηση πιθανότητα της μεταβλητής T έχει τιμή 0 κατά τον χρόνο t = 0 και ακολουθεί αύξουσα πορεία για να πάρει την τιμή 1 στο άπειρο t = ∞.

- **Η συνάρτηση επιβίωσης** $S(t)$ εκφράζει την πιθανότητα ένα άτομο να έχει χρόνο επιβίωσης μεγαλύτερο του t δηλαδή να του συμβεί το γεγονός μετά την χρονική στιγμή t. Αυτό ισοδυναμεί με την πιθανότητα το άτομο να ζει έως και την χρονική στιγμή t . Άρα η συνάρτηση επιβίωσης εκφράζει το ποσοστό των ατόμων στα οποία δεν έχει συμβεί τι γεγονός έως και την χρονική στιγμή t και δίνεται από την εξίσωση:

$$S(t) = P(T \geq t) = 1 - P(T < t) = 1 - F(t)$$

Η συνάρτηση επιβίωσης έχει τιμή 1 κατά τον χρόνο t = 0 και ακολουθεί φθίνουσα για να πάρει την τιμή 0 στο άπειρο t = ∞.

- **Η συνάρτηση κινδύνου** $h(t)$ εκφράζει τον στιγμιαίο κίνδυνο να συμβεί το γεγονός την χρονική στιγμή t δεδομένου ότι δεν έχει συμβεί το γεγονός έως

εκείνη την χρονική στιγμή. Αποτελεί μια έκφραση ρυθμού θανάτου την στιγμή t και στην βιβλιογραφία συχνά αναφέρεται ως στιγμιαίος ρυθμός θανάτου (instantaneous death rate) ή ροπή θνησιμότητας (force of mortality). Για τον μαθηματικό ορισμό της πρώτα σημειώνουμε την δεσμευμένη πιθανότητα το γεγονός να συμβεί μεταξύ t και $t + \delta t$ δεδομένου ότι δεν έχει συμβεί έως την στιγμή t . Δηλ.

$$P(t \le T < t + \delta t \mid T \ge t)$$

Διαιρώντας την πιθανότητα αυτή με δt ώστε να αποκτήσει χαρακτηριστικά ρυθμού (rate) και παίρνοντας το όριο όταν το δt τείνει στο 0 προκύπτει η συνάρτηση κινδύνου την χρονική στιγμή t:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \le T < t + \delta t \mid T \ge t)}{\delta t} \right\}$$

Χρησιμοποιώντας τις ιδιότητες των δεσμευμένων πιθανοτήτων και τον ορισμό της συνάρτησης πυκνότητας πιθανότητας $f(t)$ της μεταβλητής T προκύπτει η εξής σχέση μεταξύ συνάρτησης κινδύνου, συνάρτησης πυκνότητας πιθανότητας και συνάρτησης επιβίωσης:

$$h(t) = \frac{f(t)}{S(t)}$$

Η συνάρτηση κινδύνου παίρνει τιμές από το 0 έως το άπειρο.

- Η **αθροιστική συνάρτηση κινδύνου** $H(t)$ ορίζεται ως εξής:

$$H(t) = \int_0^t h(u)d(u)$$

Συνδέεται δε με τη συνάρτηση επιβίωσης με την εξής σχέση:

$$H(t) = -\log S(t) \quad \acute{\eta} \quad S(t) = e^{-H(t)}$$

Η αθροιστική συνάρτηση κινδύνου παίρνει τιμές από το 0 έως το άπειρο.

- **Διάμεσος επιβίωση** (median survival) ορίζεται ο χρόνος τ για τον οποίο ισχύει:

$S(\tau) = 0.5$

Καθώς στην συντριπτική πλειοψηφία των περιπτώσεων ο ακριβής χρόνος στον οποίο επιζεί το 50% των ατόμων δεν είναι δυνατό να εκτιμηθεί από το δείγμα (με εξαίρεση τις παραμετρικές μεθόδους) ως διάμεσος επιβίωση ορίζεται:

Το μικρότερο τ για το οποίο $S(\tau) \le 0.5$.

Είναι σαφές ότι με την μέθοδο αυτή κατά σύμβαση υπερεκτιμάται η διάμεσος επιβίωση.

- **Το p-οστό ποσοστημόριο** των χρόνων επιβίωσης ορίζεται ως ο χρόνος $t(p)$ τον οποίο ισχύει:

$$F\{t(p)\} = p/100$$

Εάν θέλουμε να μιλήσουμε με όρους συνάρτησης επιβίωσης το p-οστό ποσοστημόριο επιβίωσης δίδεται από την σχέση:

$$S\{t(p)\} = 1 - p/100$$

Και σε αυτή την περίπτωση επειδή στις απαραμετρικές μεθόδους στην πλειονότητα των περιπτώσεων ακριβής εκτίμηση των ποσοστημορίων από το δείγμα δεν θα είναι δυνατή, κατά σύμβαση ως p-οστό ποσοστημόριο ορίζεται:
Το μικρότερο t για το οποίο S {t(p)} ≤ 1 - p/100

12.1.2 Μη παραμετρικές μέθοδοι εκτίμησης

12.1.2.1 Kaplan-Meier

Έστω ότι έχουμε ένα δείγμα *n* ατόμων με *k* γεγονότα και *r* διακριτές χρονικές στιγμές γεγονότων. Δεδομένου πως ο χρόνος είναι συνεχής μεταβλητή είναι θεωρητικός αδύνατο να έχουμε δυο ταυτόχρονα γεγονότα. Στην πράξη όμως όλες οι μετρήσεις περιέχουν ένα βαθμό στρογγυλοποίησης με αποτέλεσμα δυο ή περισσότερα γεγονότα να μπορεί να καταγράφονται ότι συμβαίνουν ταυτόχρονα. Επιπλέον ορισμένες παρατηρήσεις μας είναι πολύ πιθανό να είναι δεξιά αποκομμένες. Έτσι συμβολίζουμε με d_j τον αριθμό των γεγονότων την χρονική στιγμή t_j όπου t_j οι διακριτές χρονικές στιγμές γεγονότων
(με j = 1,…,r και r ≤ k ≤ n).
Για την εφαρμογή της μεθόδου των καταρχάς διατάσσουμε σε αύξουσα σειρά τις χρονικές στιγμές t_j και θεωρούμε μια σειρά διαδοχικών χρονικών διαστημάτων τόσων όσες οι διακριτές χρονικές στιγμές *(r)* γεγονότων του δείγματος. Κάθε διάστημα εκτείνεται ακριβώς πριν από την χρονική στιγμή γεγονότος το οποίο περιλαμβάνει μέχρι ακριβώς πριν το επόμενο γεγονός χωρίς να το περιλαμβάνει. Θεωρούμε δε επίσης πως το γεγονός (ή γεγονότα αν είναι περισσότερα από ένα) συμβαίνει στην αρχή του διαστήματος. Εάν τώρα τα χρονικά διαστήματα τα ελαχιστοποιήσουμε ώστε να εκτείνονται από $t_j - \delta$ *(όπου* $\delta \to 0$*)* έως t_j , το κάθε διάστημα θα έχει απειροελάχιστο μήκος δ και θα περιέχει έστω d_j τον αριθμό των γεγονότων που παρατηρήθηκαν την χρονική στιγμή t_j χωρίς να περιλαμβάνει καμία αποκοπή. Εάν τυγχάνει κάποιος χρόνος αποκοπής να είναι ίδιος με ένα χρόνο γεγονότος τότε θεωρούμε ότι η αποκοπή συνέβη ακριβώς μετά την χρονική στιγμή t_j. Τελικά θα έχουμε *r* απειροελάχιστα χρονικά διαστήματα με n_j άτομα σε κίνδυνο στην αρχή του καθενός (συμπεριλαμβανομένων όσων υπέστησαν το γεγονός) και d_j γεγονότα σε αυτό το ελάχιστο διάστημα (ένα ή περισσότερα). Καμία αποκοπή δεν θα περιλαμβάνεται καθώς θα έχουν συμβεί στα μεσοδιαστήματα μεταξύ των απειροελάχιστων χρονικών διαστημάτων.
Ο κίνδυνος h_j για κάθε στιγμή γεγονότος μπορεί να εκτιμηθεί από τον λόγο d_j / n_j.
Ο παρονομαστής των κλασμάτων δυο συνεχόμενων στιγμών είναι μειωμένος κατά τα άτομα που υπέστησαν το γεγονός την προηγούμενη χρονική στιγμή καθώς και κατά αυτά των οποίων οι παρατηρήσεις αποκόπηκαν μεταξύ των δυο χρονικών στιγμών. Με αυτή την μέθοδο αποκτάται μία εκτίμηση κινδύνου μόνο στις χρονικές στιγμές γεγονότων προχωρώντας κατά βήματα.

- **Η συνάρηση επιβίωσης S(t)** για την χρονική στιγμή t_j μπορεί να γραφτεί ως εξής:

$$\mathbf{S(t)} = \Pr(T \geq t_j) = \prod_J \Pr(\text{επιβίωσης κατά το j-οστό διάστημα} \mid \text{επιβίωσε έως την αρχή του j-οστού διαστήματος})$$

$$= \prod_J \mathbf{(1 - d_j / n_j)}$$

Όπου:
d_j: Ο αριθμός των γεγονότων κατά το j-οστό διάστημα

n_j: Ο αριθμός των ατόμων σε κίνδυνο (at risk) κατά την αρχή του j-οστού διαστήματος.

Δηλαδή όσα άτομα υπέστησαν το γεγονός ή των οποίων οι παρατηρήσεις αποκόπηκαν **μετά** το j-οστό διάστημα.

Ο εκτιμητής **Kaplan-Meier** είναι το όριο του παραπάνω γινομένου όταν τα διαστήματα γίνονται άπειρα γι αυτό και ονομάζεται *product limit estimator*. [9,19]

Βασική προϋπόθεση για να ισχύει η παραπάνω σχέση είναι τα γεγονότα να είναι μεταξύ τους **ανεξάρτητα**. Επομένως ο **Kaplan-Meier** εκτιμητής γράφεται:

$$\mathrm{S}(\mathrm{t}) = \prod_{j:\tau_j < t}\left(1 - \frac{d_j}{n_j}\right) = \prod_{j:\tau_j < t}\left(\frac{n_j - d_j}{n_j}\right)$$

- **Ο κίνδυνος h_j** όπως είδαμε μπορεί να εκτιμηθεί με την μέθοδο Kaplan-Meier για κάθε χρονική στιγμή γεγονότος από τον **λόγο d_j / n_j.** Στην πράξη όμως πότε δεν έχουμε σημειακές χρονικές στιγμές με όση ακρίβεια και αν καταγράφεται ο χρόνος. Έτσι μπορεί να έχουμε ακρίβεια ημέρας (π.χ. ημερομηνία θανάτου). Λόγω αυτού ο κίνδυνος αποκτά χαρακτήρα ρυθμού προσθέτοντας στον παρονομαστή την μονάδα του χρόνου σύμφωνα με τον ορισμό του και γίνεται:

$$h_j = \frac{n_j}{\tau_j d_j}$$

Όπου τ_j το χρονικό διάστημα από t_j έως t_{j+1}

- Μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης για τη συνάρτηση επιβίωσης χρησιμοποιούμε το γεγονός ότι έχουμε στην ουσία μια σειρά από k διωνυμικές δοκιμές. Δηλαδή: $S(t_m) = \prod_{j=1}^{m} p_j \Rightarrow log S(t_m) = log \prod_{j=1}^{m} p_j$

$$\Rightarrow var\{logS(t_m)\} = \sum_{j=1}^{m} var\{log(p_j)\} \Rightarrow$$

$$se\{S(t_m)\} = S(t_m)\left\{\sum_{j=1}^{m} \frac{d_j}{n_j(n_j - d_j)}\right\}^{\frac{1}{2}}$$ *Greenwood's formula*.

12.1.2.2 Life table estimation (actuarial)

Σε αυτή την περίπτωση χωρίζουμε τον χρόνο παρακολούθησης σε j χρονικά διαστήματα (όχι υποχρεωτικά ίσα) και καταγράφουμε τον αριθμό των ατόμων σε κίνδυνο (n_j) στην αρχή κάθε διαστήματος, τον αριθμό των γεγονότων σε κάθε χρονικό διάστημα (d_j) καθώς και τον αριθμό των αποκομμένων παρατηρήσεων (c_j) σε κάθε χρονικό διάστημα. Θεωρώντας η αποκοπή των παρατηρήσεων συνέβη ομοιόμορφα σε κάθε χρονικό διάστημα ο αριθμός των ατόμων σε κίνδυνο για κάθε διάστημα τροποποιείται σε:

$n_j' = n_j - c_j/2$ (Actuarial assumption).

Η συνάρτηση επιβίωσης έως το m-οστό διάστημα θα δίνεται από τον τύπο:

$$S(t_m) = \prod_{j=1}^{m}\left(\frac{n'_j - d_j}{n'_j}\right)$$

12.1.3 Σύγκριση συναρτήσεων επιβίωσης δύο ομάδων

12.1.3.1 Log-rank test (Mantel-Haenszel)

Ας υποθέσουμε ότι θέλουμε να συγκρίνουμε την συνάρτηση επιβίωσης δυο ομάδων, group I & group II. Για την εφαρμογή του log rank test καταρχάς διατάσσουμε σε αύξουσα σειρά τους χρόνους γεγονότων των δύο ομάδων. Έστω ότι έχουμε *r* διακριτούς χρόνους γεγονότων συνολικά στις δυο ομάδες (t_j, $j = 1...r$ *και* $t_1 < t_2 < t_3 < ... < t_r$). Έστω τώρα ότι σε κάθε χρονική στιγμή γεγονότος t_j έχουμε n_{1j} άτομα σε κίνδυνο στο group I και n_{2j} άτομα σε κίνδυνο στο group II. Επίσης αντίστοιχα σε κάθε χρονική στιγμή γεγονότος d_{1j} ο αριθμός των γεγονότων στο group I και d_{2j} ο αριθμός των γεγονότων στο group II. Έτσι σε κάθε χρονική στιγμή γεγονότος t_j θα έχουμε $n_j = n_{1j} + n_{2j}$ άτομα σε κίνδυνο και $d_j = d_{1j} + d_{2j}$ γεγονότα. Τα δεδομένα αυτά τα διατάσσουμε στον παρακάτω 2x2 πίνακα.

Ομάδα	Αριθμός γεγονότων την χρονική στιγμή t_j	Αριθμός ατόμων που επιβιώνουν πέραν της χρονικής στιγμής t_j	Άτομα σε κίνδυνο την χρονική στιγμή t_j
I	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
II	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Σύνολο	d_j	$n_j - d_j$	n_j

Η μηδενική υπόθεση είναι ότι δεν υπάρχει καμία διαφορά μεταξύ των γεγονότων στις δυο ομάδες. Ένας τρόπος να το ελέγξουμε αυτό είναι συγκρίνοντας τον αριθμό των παρατηρούμενων γεγονότων σε κάθε χρονική στιγμή στις δυο ομάδες σε σχέση με τον αναμενόμενο υπό την μηδενική υπόθεση.
Αν θεωρήσουμε τα περιθώρια αθροίσματα του πίνακα δεδομένα (fixed) τότε το d_{1j} ακολουθεί την υπεργεωμετρική κατανομή με:

μέση τιμή $e_{1j} = n_{1j}\, d_j / n_j$

και διασπορά $u_{1j} = \dfrac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}$

Το στατιστικό που χρησιμοποιούμε είναι το: $U_L = \sum_{j=1}^{r} (d_{1j} - e_{1j})$

Η διασπορά του δίνεται από τον τύπο: $var(U_L) = \sum_{j=1}^{r} u_{1j} = V_L$

Η κατανομή του είναι: $\dfrac{U_L}{\sqrt{V_L}} \sim N(0,1)$ **ή** $\dfrac{U_L^2}{V_L} \sim X_1^2$

Το Log-rank test μπορεί να επεκταθεί και για συγκρίσεις μεταξύ περισσότερων των δύο ομάδων.

12.1.3.2 Cox proportional hazards model

Η μέθοδος του Cox δεν υποθέτει κάποια συγκεκριμένη κατανομή για τον χρόνο ή την συνάρτηση κινδύνου. Αντίθετα μοντελοποιεί τον λόγο των κινδύνων (hazard ratio) μεταξύ δύο ατόμων για τον οποίο υποθέτει ότι είναι σταθερός σε κάθε χρονική

στιγμή. Αυτή η παραδοχή αποτελεί βασικό χαρακτηριστικό της μεθόδου για αυτό και ονομάζεται αλλιώς *proportional hazards model*. Από μαθηματικής άποψης ανήκει στα log-linear μοντέλα και στην περίπτωση των δυο ομάδων περιγράφεται από την παρακάτω εξίσωση:

$$log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{1i}$$

Όπου x_{1i} η τιμή της μεταβλητής x_1 για το i-οστό άτομο στην μελέτη η οποία παίρνει την τιμή 1 ή 0 στην περίπτωση των δίτιμων μεταβλητών (π.χ. δυο ομάδες θεραπείας). Διαφορετικά η προηγούμενη εξίσωση γράφεται:

$$h_i(t) = h_0(t)e^{\beta x_i}$$

Ο όρος $h_0(t)$ αναφέρεται σαν baseline hazard function και αποτελεί την συνάρτηση κίνδυνου την χρονική στιγμή t για τα άτομα που ανήκουν στην ομάδα με μεταβλητή $x = 0$ (π.χ. $x = 0$ για τα άτομα στην κλασσική θεραπεία και $x = 1$ για τα άτομα στην νέα θεραπεία).

Η μέθοδος του Cox φυσικά μπορεί να επεκταθεί ώστε να συμπεριλάβει περισσότερες των δυο επεξηγηματικών μεταβλητών. Η γενική διατύπωση του μοντέλου είναι η εξής:

$$h_i(t) = h_0(t)e^{\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_\kappa x_{\kappa i}}$$

Σε αυτή την περίπτωση έχουμε *k* επεξηγηματικές μεταβλητές και η baseline hazard function $h_0(t)$ είναι η συνάρτηση κινδύνου την χρονική στιγμή *t* για τα άτομα που έχουν όλες τους τις μεταβλητές ίσες με 0. Για δυο οιαδήποτε άτομα *i* και *j* ο λόγος των κινδύνων τους **για κάθε χρονική στιγμή *t*** θα δίνεται από τον τύπο:

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)e^{\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_\kappa x_{\kappa i}}}{h_0(t)e^{\beta_1 x_{1j} + \beta_2 x_{2j} + ... + \beta_\kappa x_{\kappa j}}} \Rightarrow$$

$$\boldsymbol{log}\left\{\frac{\boldsymbol{h_i(t)}}{\boldsymbol{h_j(t)}}\right\} = \beta_1(\boldsymbol{x_{1i}} - \boldsymbol{x_{1j}}) + ... + \beta_k(\boldsymbol{x_{ki}} - \boldsymbol{x_{kj}})$$

Η πιθανοφάνεια βάση της οποίας προκύπτει η συμπερασματολογία του μοντέλου παρέχεται από τον τύπο:

$$\boldsymbol{L}(\beta) = \prod_{j=1}^{r} \frac{exp(\boldsymbol{\beta' x_{(j)}})}{\sum_{l \in R(t_{(j)})} exp(\boldsymbol{\beta' x_l})}$$

Όπου:

$\boldsymbol{x_{(j)}}$ το διάνυσμα των μεταβλητών για το άτομο που υφίσταται το γεγονός την χρονική στιγμή $t_{(j)}$

$\sum_{l \in R(t_{(j)})} exp(\boldsymbol{\beta' x_l})$ το άθροισμα όλων των τιμών των $exp(\boldsymbol{\beta' x_l})$ για όλα τα άτομα που βρίσκονται σε κίνδυνο την χρονική στιγμή $t_{(j)}$.

Βασική προϋπόθεση για την εφαρμογή του παραπάνω τύπου της πιθανοφάνειας είναι η παραδοχή ότι σε κάθε χρονική στιγμή γεγονότος **συμβαίνει ένα γεγονός και όχι περισσότερα**. Στις περισσότερες, όμως, περιπτώσεις ανάλυσης επιβίωσης συμβαίνει

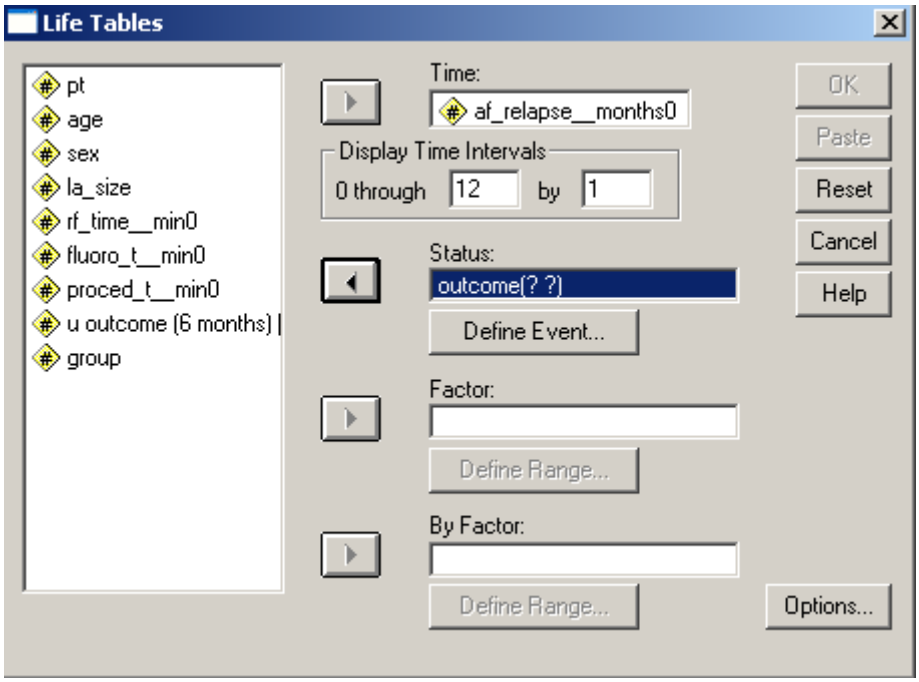
να έχουμε σε μια ή περισσότερες χρονικές στιγμές ταυτόχρονα γεγονότα σε δυο ή περισσότερα άτομα. Αυτές η περιπτώσεις ονομάζονται **"ισοπαλίες" (ties)** και πρέπει να υπάρξει κατάλληλη τροποποίηση της πιθανοφάνειας ώστε να εξακολουθεί να ισχύει. Διάφορες τέτοιες τροποποιήσεις έχουν προταθεί (Breslow, Efron, Cox, Kalbfleisch & Prentice).

12.2 Πίνακες επιβίωσης

Ας υποθέσουμε ότι υπάρχουν 2 ομάδες ασθενών με καρδιαγγειακή νόσο, όπου η μία ομάδα έχει υποβληθεί στην επέμβαση Α και η άλλη στην επέμβαση Β (group 0 & 1, αντίστοιχα). Όλοι οι ασθενείς παρακολουθήθηκαν μέχρι την εμφάνιση υποτροπής της νόσου, και η μελέτη διήρκησε 12 μήνες μετά την επέμβαση. Για να πραγματοποιήσουμε του «πίνακες ανάλυσης επιβίωσης», ακολουθούμε τα εξής ***βήματα***:

Analyze ⟶ Survival → Life Tables …

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 12.1*.



Εικόνα 12.1: Πραγματοποίηση των πινάκων επιβίωσης (Life Tables)

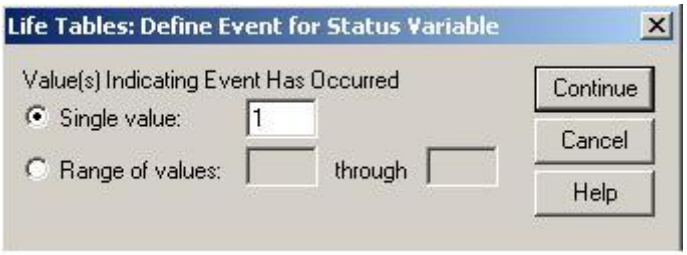
ii. Στο «***Time***» ορίζουμε την μεταβλητή που υποδηλώνει το χρόνο που ο κάθε ασθενής παρακολουθήθηκε, δηλ. το χρόνο μέχρι την εμφάνιση υποτροπής ή το χρόνο μέχρι να χαθεί ο ασθενής από τη μελέτη για άσχετους λόγους (lost to follow-up) ή μέχρι να ολοκληρωθεί η μελέτη.

iii. Στο "***Display Time Intervals***" δηλώνουμε το μέγιστο χρονικό διάστημα παρακολούθησης και τα διαστήματα στα οποία υποδιαιρείται αυτό το χρονικό διάστημα προκειμένου να δημιουργηθούν οι πίνακες επιβίωσης.

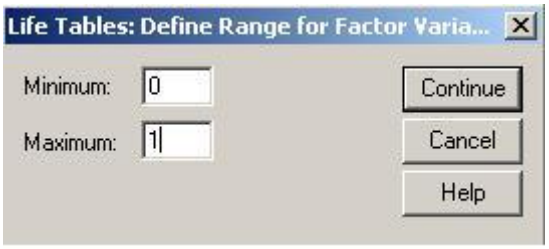
iv. Στο "***Status***" δηλώνουμε την μεταβλητή που υποδηλώνει αν ο ασθενής εμφάνισε υποτροπή ή αν βγήκε από τη μελέτη χωρίς υποτροπή (lost to follow up ή ολοκληρώθηκε η μελέτη). Στη συνέχεια πιέζουμε το κουμπί "***Define Event***" έτσι ώστε να ορίσουμε με ποια ή ποιες τιμές υποδηλώνεται η εμφάνιση υποτροπής (*Εικόνα 12.2*) και πιέζουμε "***Continue***".

v. Στο "***Factor***" δηλώνουμε τη μεταβλητή που υποδηλώνει το αν ο κάθε ασθενής έχει υποβληθεί στην επέμβαση Α ή στην επέμβαση Β (*Σημείωση*: Ενδέχεται να μην αποτελεί στόχο της μελέτης να γίνει ξεχωριστός πίνακας επιβίωσης για

κάθε κατηγορία ενός ποιοτικού χαρακτηριστικού). Και εδώ θα πρέπει να πιέσουμε το κουμπί "***Define Range***", όπου εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 12.3*, προκειμένου να δηλώσουμε τα νούμερα με τα οποία έχουν κωδικοποιηθεί οι επεμβάσεις και πιέζουμε "***Continue***".



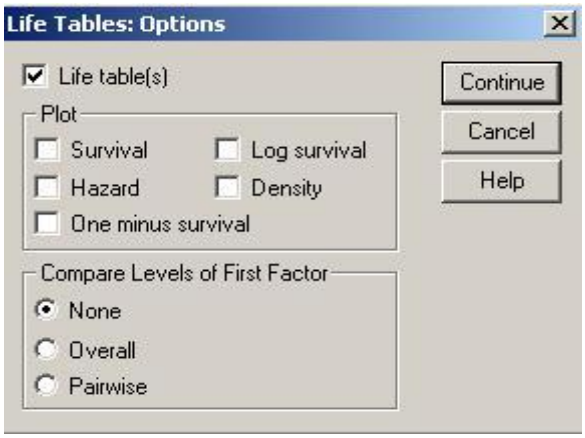
Εικόνα 12.2: Προσδιορισμός της τιμής με την οποία έχει κωδικοποιηθεί η εμφάνιση του γεγονότος (π.χ. υποτροπή νόσου).



Εικόνα 12.3: Προσδιορισμός των αριθμών με τους οποίους έχουν κωδικοποιηθεί οι 2 κατηγορίες του παράγοντα βάσει του οποίου θα υπολογιστούν οι πίνακες επιβίωσης (π.χ., επέμβαση Α:0 & επέμβαση Β: 1)

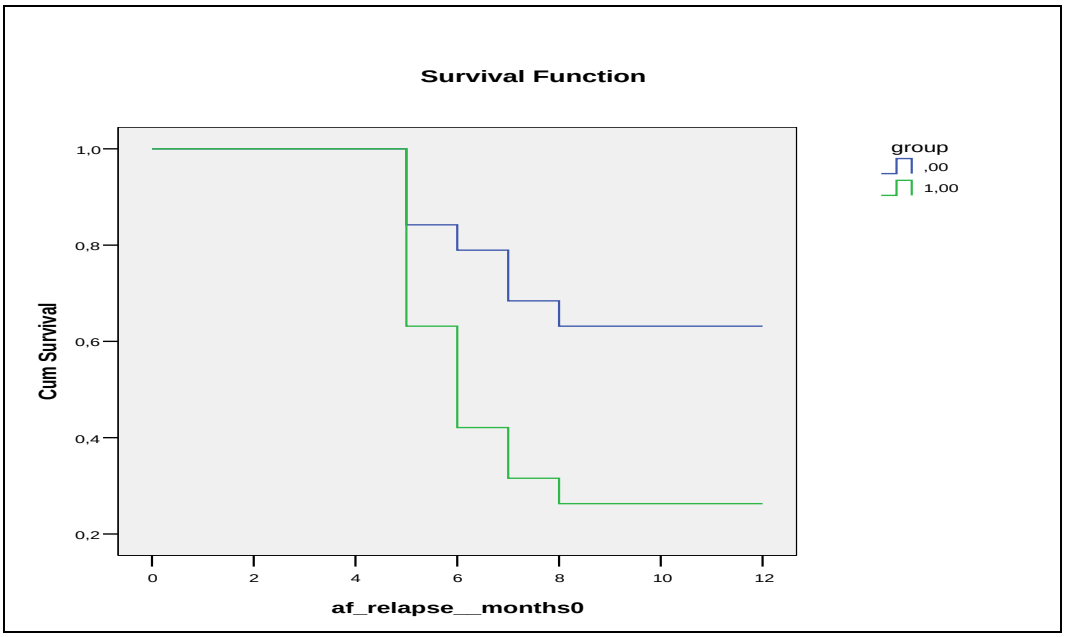
vi. Πιέζοντας το κουμπί επιλογών "***Options***" εμφανίζεται ένα νέο πλαίσιο διαλόγου το οποίο φαίνεται στην *Εικόνα 12.4* και στο οποίο μπορούμε να επιλέξουμε:

- **Plots:** Να εμφανιστεί κάποιο γράφημα
- **Compare levels of first factor:** Να πραγματοποιηθεί σύγκριση της επιβίωσης ανάμεσα στα διάφορα επίπεδα του παράγοντα (π.χ. επέμβαση).



Εικόνα 12.4: Μενού «Options» στους πίνακες επιβίωσης.

vii. Στους *Πίνακες 12.1 & 12.2* καθώς επίσης και στο *Εικόνα 12.5* παρουσιάζονται τα αποτελέσματα.



Εικόνα 12.5: Καμπύλες επιβίωσης γι' αυτούς που υπεβλήθησαν στην Α και Β επέμβαση ξεχωριστά.

Η *Εικόνα 12.5* απεικονίζει τις καμπύλες επιβίωσης για τα άτομα που υπεβλήθησαν στην επέμβαση Α (group 0) και την επέμβαση Β (group 1), ξεχωριστά. Διαπιστώνουμε, λοιπόν, ότι η επιβίωση των ατόμων που υπεβλήθησαν στην επέμβαση Α είναι υψηλότερη σε σχέση με αυτή των ατόμων που υπεβλήθησαν στην επέμβαση Β.

Life Table

First-order Controls		Interval Start Time	Number Entering Interval	Number Withdrawing during Interval	Number Exposed to Risk	Number of Terminal Events	Proportion Terminating	Proportion Surviving	Cumulative Proportion Surviving at End of Interval	Std. Error of Cumulative Proportion Surviving at End of Interval	Probability Density	Std. Error of Probability Density	Hazard Rate	Std. Error of Hazard Rate
group	0	0	19	0	19,000	0	,00	1,00	1,00	,00	,000	,000	,00	,00
		1	19	0	19,000	0	,00	1,00	1,00	,00	,000	,000	,00	,00
		2	19	0	19,000	0	,00	1,00	1,00	,00	,000	,000	,00	,00
		3	19	0	19,000	0	,00	1,00	1,00	,00	,000	,000	,00	,00
		4	19	0	19,000	3	,16	,84	,84	,08	,158	,084	,17	,10
		5	16	0	16,000	1	,06	,94	,79	,09	,053	,051	,06	,06
		6	15	0	15,000	2	,13	,87	,68	,11	,105	,070	,14	,10
		7	13	0	13,000	1	,08	,92	,63	,11	,053	,051	,08	,08
		8	12	0	12,000	0	,00	1,00	,63	,11	,000	,000	,00	,00
		9	12	0	12,000	0	,00	1,00	,63	,11	,000	,000	,00	,00
		10	12	0	12,000	0	,00	1,00	,63	,11	,000	,000	,00	,00
		11	12	0	12,000	0	,00	1,00	,63	,11	,000	,000	,00	,00
	1	0	19	0	19,000	0	,00	1,00	1,00	,00	,000	,000	,00	,00
		1	19	0	19,000	0	,00	1,00	1,00	,00	,000	,000	,00	,00
		2	19	0	19,000	0	,00	1,00	1,00	,00	,000	,000	,00	,00
		3	19	0	19,000	0	,00	1,00	1,00	,00	,000	,000	,00	,00
		4	19	0	19,000	7	,37	,63	,63	,11	,368	,111	,45	,17
		5	12	0	12,000	4	,33	,67	,42	,11	,211	,094	,40	,20
		6	8	0	8,000	2	,25	,75	,32	,11	,105	,070	,29	,20
		7	6	0	6,000	1	,17	,83	,26	,10	,053	,051	,18	,18
		8	5	0	5,000	0	,00	1,00	,26	,10	,000	,000	,00	,00
		9	5	0	5,000	0	,00	1,00	,26	,10	,000	,000	,00	,00
		10	5	0	5,000	0	,00	1,00	,26	,10	,000	,000	,00	,00
		11	5	0	5,000	0	,00	1,00	,26	,10	,000	,000	,00	,00

Πίνακας 12.1: Πίνακας επιβίωσης γι' αυτούς που υπεβλήθησαν στην επέμβαση Α & Β, ξεχωριστά.

Στον *Πίνακα 12.1* παρατηρούμε τον αριθμό των ατόμων που εισήχθησαν χωρίς επιπλοκή σε κάθε χρονικό διάστημα (Number entering), τον αριθμό των ατόμων που χάθηκαν (number withdrawing during interval), τον αριθμό των ατόμων που είναι σε κίνδυνο να εμφανίσουν επιπλοκή σε κάθε χρονικό διάστημα (number exposed to risk), τον αριθμό των ατόμων που εμφάνισαν επιπλοκή (number of terminal events), το ποσοστό των επιπλοκών (proportion terminating), το ποσοστό επιβίωσης (proportion surviving), το αθροιστικό ποσοστό επιβίωσης (και το τυπικό σφάλμα) σε κάθε διάστημα (cumulative proportion surviving at end of interval), το ποσοστό κινδύνου (και τυπικό σφάλμα) (hazard rate). Για παράδειγμα, όσον αφορά στα άτομα που υπεβλήθησαν στην επέμβαση Α, παρατηρούμε ότι κατά το χρονικό διάστημα 0-1, κανένα άτομο δεν χάθηκε (number withdrawing during interval), κανένα άτομο δεν εμφάνισε υποτροπή (number of terminal events), άρα το ποσοστό επιβίωσης σε αυτό το χρονικό διάστημα είναι 100% (proportion surviving). Στο χρονικό διάστημα, όμως, 4-5, 3 άτομα εμφάνισαν υποτροπή, άρα το ποσοστό επιβίωσης σε αυτό το διάστημα είναι 84% (16/19).

Σημείωση: Ο αριθμός τον ατόμων που εισάγονται σε κάθε χρονικό διάστημα χωρίς επιπλοκή και ο αριθμός των ατόμων που είναι σε κίνδυνο να εμφανίσουν επιπλοκή σε κάθε χρονικό διάστημα, δεν είναι ίδιος στην περίπτωση που υπάρχουν άτομα που χάθηκαν από τη μελέτη στο συγκεκριμένο χρονικό διάστημα. Σε αυτή την περίπτωση, επειδή δεν γνωρίζουμε αν τα άτομα χάθηκαν στην αρχή ή στο τέλος του χρονικού διαστήματος, θεωρούμε ότι χάθηκαν στην μέση οπότε τα άτομα που είναι σε κίνδυνο για εμφάνιση επιπλοκής είναι τόσα όσα εισήχθησαν στο διάστημα μείον (αυτών που χάθηκαν στο συγκεκριμένο χρονικό διάστημα)/2.

Overall Comparisons[a]

Wilcoxon (Gehan) Statistic	df	Sig.
5,320	1	,021

a. Comparisons are exact.

Πίνακας 12.2: Σύγκριση των καμπυλών επιβίωσης μεταξύ των αυτών που υπεβλήθησαν στην Α & Β επέμβαση, ξεχωριστά.

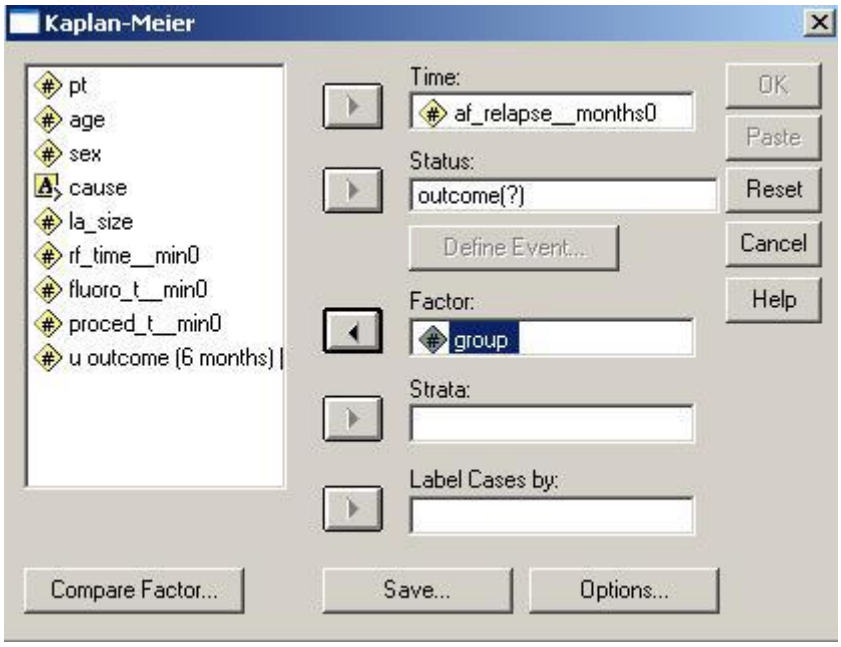
Από τον *Πίνακα 12.2* διαπιστώνουμε ότι οι επιβιώσεις των ατόμων που υπεβλήθησαν στις 2 διαφορετικές επεμβάσεις διαφέρουν στατιστικά σημαντικά αφού η πιθανότητα εσφαλμένης απόρριψης της μηδενικής υπόθεσης (ότι δηλαδή οι συναρτήσεις επιβίωσης των ατόμων που υπεβλήθησαν στις 2 διαφορετικές μονάδες είναι ίσες) είναι μικρότερη από το 5% αφού sig.=0,021.

12.3 Καμπύλες επιβίωσης Kaplan-Meier

Εναλλακτικά, μπορούμε να πραγματοποιήσουμε τις καμπύλες επιβίωσης «Kaplan Meier», οι οποίες πραγματοποιούνται ακολουθώντας τα εξής ***βήματα***:

Analyze → Survival → Kaplan Meier …

i. Ανοίγει το πλαίσιο διαλόγου της *Εικόνας 12.6*.



Εικόνα 12.6: Πραγματοποίηση της «Kaplan – Meier» ανάλυσης.

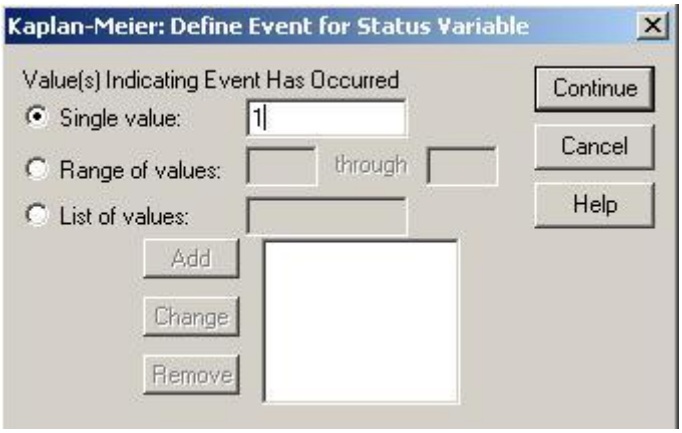
ii. Στο «***Time***» ορίζουμε την μεταβλητή που υποδηλώνει το χρόνο που ο κάθε ασθενής παρακολουθήθηκε, δηλ. το χρόνο μέχρι την εμφάνιση υποτροπής ή το χρόνο μέχρι να χαθεί ο ασθενής από τη μελέτη για άσχετους λόγους (lost to follow-up) ή μέχρι να ολοκληρωθεί η μελέτη.

iii. Στο "***Status***" δηλώνουμε την μεταβλητή που υποδηλώνει αν ο ασθενής εμφάνισε υποτροπή ή αν βγήκε από τη μελέτη χωρίς υποτροπή (lost to follow up ή ολοκληρώθηκε η μελέτη). Στη συνέχεια πιέζουμε το κουμπί "***Define Event***", όπου ανοίγει το πλαίσιο διαλόγου της *Εικόνας 12.7*, έτσι ώστε να ορίσουμε με ποια ή ποιες τιμές υποδηλώνεται η εμφάνιση υποτροπής και πιέζουμε "***Continue***".

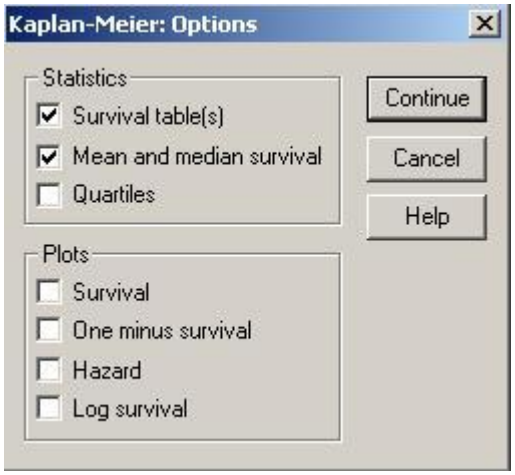
iv. Στο "***Factor***" δηλώνουμε τη μεταβλητή που υποδηλώνει το αν ο κάθε ασθενής έχει υποβληθεί στην παρέμβαση Α ή στην παρέμβαση Β.

v. Πιέζοντας το κουμπί επιλογών "***Options***" εμφανίζεται ένα νέο πλαίσιο διαλόγου το οποίο φαίνεται στην *Εικόνα 12.8* και στο οποίο μπορούμε να επιλέξουμε:

- **Plots:** Να εμφανιστεί κάποιο γράφημα
- **Statistics:** Εκτός από τα προεπιλεγμένα (Mean and median survival and survival table(s)) μπορούμε να επιλέξουμε και Quartiles.



Εικόνα 12.7: Προσδιορισμός της τιμής με την οποία έχει κωδικοποιηθεί η εμφάνιση του γεγονότος (π.χ. υποτροπή νόσου).

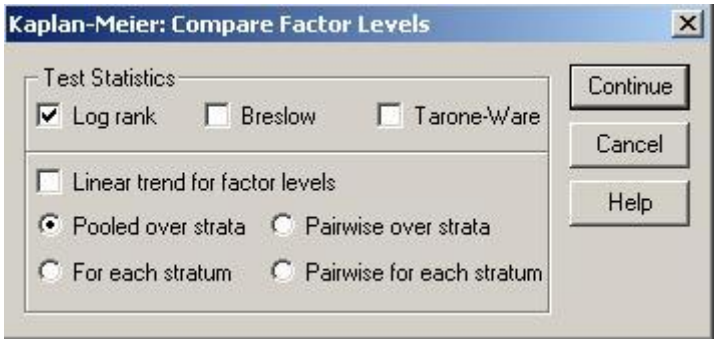


Εικόνα 12.8: Επιλογή «Options» της «Kaplan – Meier» ανάλυσης

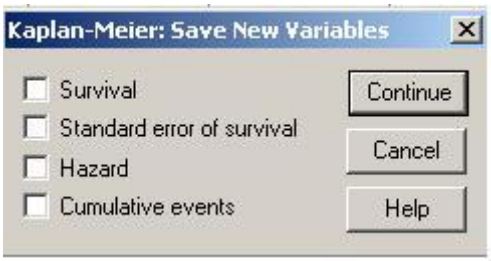
vi. Πιέζοντας το κουμπί επιλογών "***Compare factor***" εμφανίζεται ένα νέο πλαίσιο διαλόγου το οποίο φαίνεται στην *Εικόνα 12.9* και στο οποίο μπορούμε να επιλέξουμε το στατιστικό κριτήριο με το οποίο θα πραγματοποιηθεί η σύγκριση. Διαθέσιμα στατιστικά είναι τα: log rank, Breslow, and Tarone-Ware.

vii. Πιέζοντας το κουμπί επιλογών "***Save***" εμφανίζεται ένα νέο πλαίσιο διαλόγου (*Εικόνα 12.10*) από το οποίο μπορούμε να επιλέξουμε τις πληροφορίες από το Kaplan-Meier πίνακα που επιθυμούμε να σώσουμε στο αρχείο μας ως νέες μεταβλητές προκειμένου να χρησιμοποιηθούν σε άλλες αναλύσεις. Μπορούμε να επιλέξουμε να σωθούν κάποιο από τα εξής: survival, standard error of survival, hazard, και cumulative events.

viii. Στους *Πίνακες 12.3. 12.4 & 12.5* και στην *Εικόνα 12.11* παρουσιάζονται τα αποτελέσματα.



Εικόνα 12.9: Επιλογή «Compare factors» της «Kaplan – Meier» ανάλυσης



Εικόνα 12.10: Επιλογές αποθήκευσης μεταβλητών

Από τον *Πίνακα 12.3* παρατηρούμε για κάθε ένα άτομο που συμμετέχει στην μελέτη (ξεχωριστά για την επέμβαση A & B) τη χρονική στιγμή που εξήλθε από τη μελέτη και την έκβαση του (π.χ. αν χάθηκε: Status = 0 ή αν εμφάνισε υποτροπή: Status = 1). Αυτά τα στοιχεία εμφανίζονται με χρονική σειρά εξόδου από τη μελέτη, έτσι ώστε να μπορεί να υπολογιστεί η αθροιστική συνάρτηση επιβίωσης (και το τυπικό σφάλμα αυτή) για κάθε χρονική στιγμή (cumulative proportion surviving at the time: estimate and std.error). Για παράδειγμα, παρατηρούμε ότι μεταξύ των ασθενών που υπεβλήθησαν στην επέμβαση A (group=0), τη χρονική στιγμή 5 μήνες (time) είχαν εξέλθει από τη μελέτη μόνο 4 άτομα (N of cumulative events) έχοντας όλα εμφανίσει υποτροπή (status=1). Αυτή τη χρονική στιγμή, λοιπόν, η επιβίωση για την συγκεκριμένη ομάδα ασθενών είναι 78,9% (estimate).

Survival Table

group		Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
				Estimate	Std. Error		
,00	1	4,000	1,00	.	.	1	18
	2	4,000	1,00	.	.	2	17
	3	4,000	1,00	,842	,084	3	16
	4	5,000	1,00	,789	,094	4	15
	5	6,000	1,00	.	.	5	14
	6	6,000	1,00	,684	,107	6	13
	7	7,000	1,00	,632	,111	7	12
	8	12,000	,00	.	.	7	11
	9	12,000	,00	.	.	7	10
	10	12,000	,00	.	.	7	9
	11	12,000	,00	.	.	7	8
	12	12,000	,00	.	.	7	7
	13	12,000	,00	.	.	7	6
	14	12,000	,00	.	.	7	5
	15	12,000	,00	.	.	7	4
	16	12,000	,00	.	.	7	3
	17	12,000	,00	.	.	7	2
	18	12,000	,00	.	.	7	1
	19	12,000	,00	.	.	7	0
1,00	1	4,000	1,00	.	.	1	18
	2	4,000	1,00	.	.	2	17
	3	4,000	1,00	.	.	3	16
	4	4,000	1,00	.	.	4	15
	5	4,000	1,00	.	.	5	14
	6	4,000	1,00	.	.	6	13
	7	4,000	1,00	,632	,111	7	12
	8	5,000	1,00	.	.	8	11
	9	5,000	1,00	.	.	9	10
	10	5,000	1,00	.	.	10	9
	11	5,000	1,00	,421	,113	11	8
	12	6,000	1,00	.	.	12	7
	13	6,000	1,00	,316	,107	13	6
	14	7,000	1,00	,263	,101	14	5
	15	12,000	,00	.	.	14	4
	16	12,000	,00	.	.	14	3
	17	12,000	,00	.	.	14	2
	18	12,000	,00	.	.	14	1
	19	12,000	,00	.	.	14	0

Πίνακας 12.3: Πίνακας συνάρτησης επιβίωσης βάσει της «Kaplan – Meier» ανάλυσης.

Means and Medians for Survival Time

group	Mean[a]				Median			
			95% Confidence Interval				95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
,00	9,474	,775	7,955	10,992	.	.	.	.
1,00	6,684	,752	5,210	8,158	5,000	,538	3,945	6,055
Overall	8,079	,585	6,932	9,226	6,000	1,027	3,986	8,014

a. Estimation is limited to the largest survival time if it is censored.

Πίνακας 12.4: Μέση τιμή και διάμεσος του χρόνου επιβίωσης

Από τον *Πίνακα 12.4* παρατηρούμε ότι ο μέσος χρόνος επιβίωσης των ασθενών που υπεβλήθησαν στην επέμβαση Α (group 0) είναι περίπου 9,5 μήνες, ενώ ο μέσος χρόνος επιβίωσης των ασθενών που υπεβλήθησαν στην επέμβαση Β (group 1) είναι περίπου 6,7 μήνες. Επίσης, από τον ίδιο Πίνακα διαπιστώνουμε ότι το 50% των ασθενών που υπεβλήθησαν στην επέμβαση Β είχε εμφανίσει υποτροπή στους 5 μήνες μετά την πραγματοποίηση της επέμβασης (διάμεσος χρόνος επιβίωσης), ενώ αντίθετα μέχρι και την ολοκλήρωση της μελέτης (12 μήνες) το ποσοστό των ασθενών που εμφάνισαν υποτροπή μεταξύ αυτών που υπεβλήθησαν στην θεραπεία Β είναι αρκετά υψηλότερο από το 50% (συνεπώς, δεν μπορεί να εκτιμηθεί ο διάμεσος χρόνος επιβίωσης).

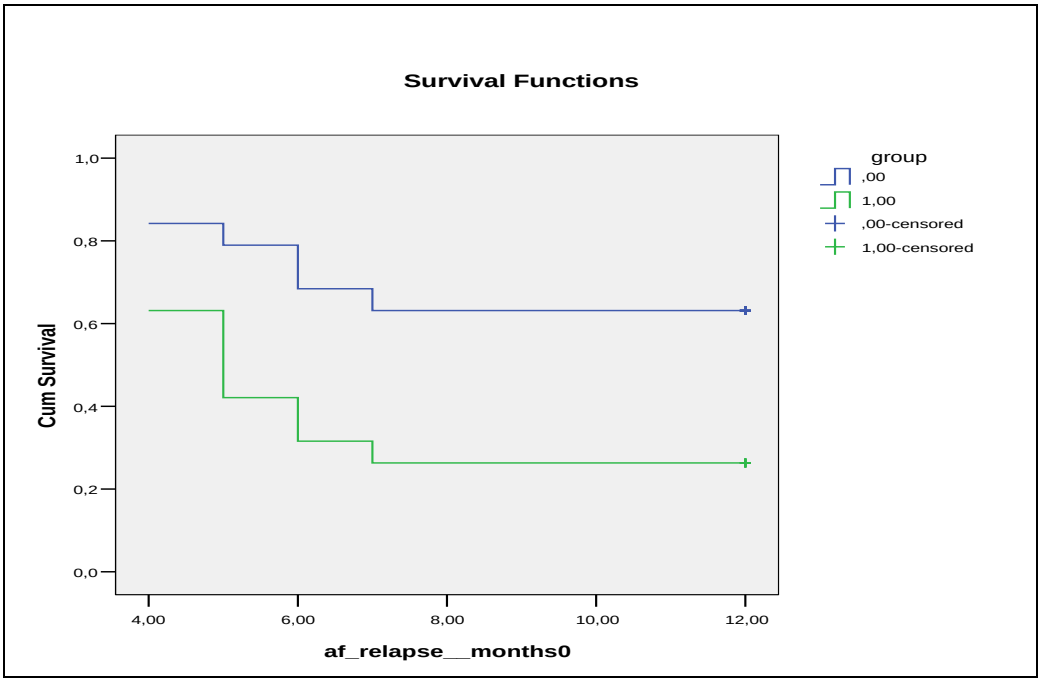
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	5,663	1	,017

Test of equality of survival distributions for the different levels of group.

Πίνακας 12.5: Log-rank test: Σύγκριση της συνάρτησης επιβίωσης μεταξύ των ατόμων που υπεβλήθησαν στην επέμβαση Α και αυτών που υπεβλήθησαν στην επέμβαση Β.

Από τον *Πίνακα 12.5* διαπιστώνουμε ότι η επιβίωση μεταξύ των ατόμων που υπεβλήθησαν στην επέμβαση Α και αυτών που υπεβλήθησαν στην επέμβαση Β διαφέρει στατιστικά σημαντικά αφού sig.=0,017<0,05. Από την *Εικόνα 12.11* διαπιστώνουμε ότι η επιβίωση των ασθενών που υπεβλήθησαν στην επέμβαση Α είναι υψηλότερη σε σχέση με αυτή των ασθενών που υπεβλήθησαν στην επέμβαση Β.



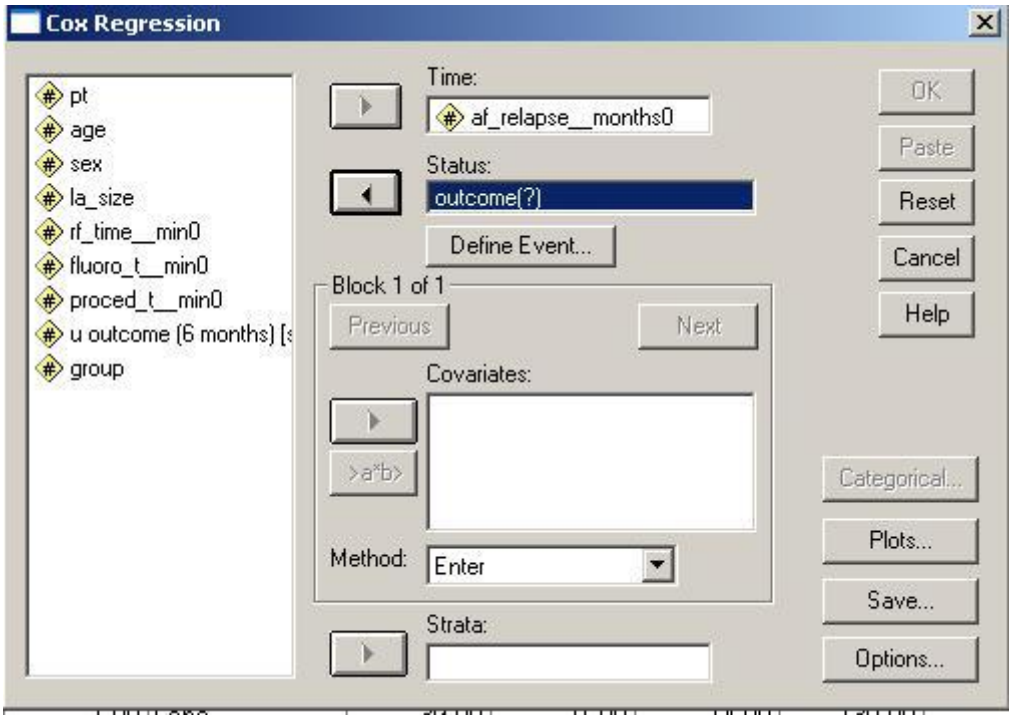
Εικόνα 12.11: Kaplan-Meier καμπύλες επιβίωσης ξεχωριστά για τους ασθενείς που υπεβλήθησαν στην επέμβαση Α και αυτούς που υπεβλήθησαν στην επέμβαση Β, ξεχωριστά.

12.4 Μοντέλα παλινδρόμησης Cox

Τα μοντέλα παλινδρόμησης Cox είναι μία τεχνική αντίστοιχη με αυτή της γραμμικής παλινδρόμησης, η οποία, όμως, είναι κατάλληλη να μοντελοποιεί το χρόνο μέχρι να συμβεί ένα γεγονός διαχειρίζοντας κατάλληλα τις λογοκριμένες τιμές (censored cases). Επίσης, η συγκεκριμένη μέθοδος ανάλυσης δίνει τη δυνατότητα να συμπεριληφθούν στο ίδιο μοντέλο πολλοί παράγοντες (ποιοτικά ή ποσοτικά χαρακτηριστικά). Ας χρησιμοποιήσουμε και πάλι το παραπάνω παράδειγμα και ας υποθέσουμε ότι επιθυμούμε να ελέγξουμε ταυτόχρονα το αν η ηλικία, το φύλο και η επέμβαση στην οποία υποβλήθηκε το κάθε άτομο, συσχετίζονται με τον χρόνο που μεσολαβεί από την πραγματοποίηση της επέμβασης μέχρι την εμφάνιση κάποιας επιπλοκής ή την ολοκλήρωση της μελέτης. Για να πραγματοποιήσουμε το μοντέλο παλινδρόμησης Cox ακολουθούμε τα εξής ***βήματα***:

Analyze ⟶ Survival → Cox regression …

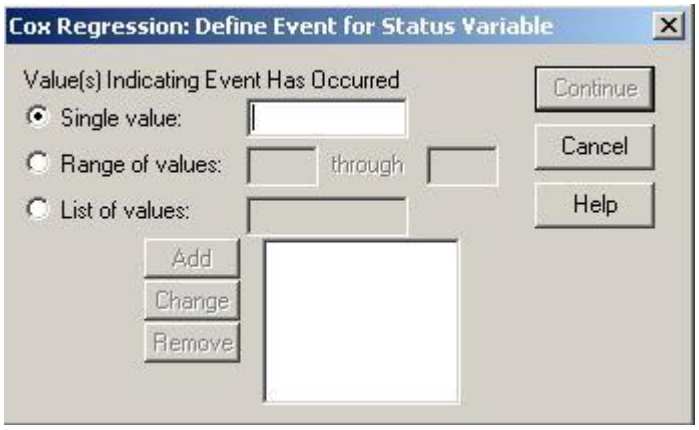
i. Ανοίγει το παράθυρο διαλόγου της *Εικόνας 12.12*.



Εικόνα 12.12: Πραγματοποίηση της ανάλυσης παλινδρόμησης Cox.

ii. Στο «***Time***» ορίζουμε την μεταβλητή που υποδηλώνει το χρόνο που ο κάθε ασθενής παρακολουθήθηκε, δηλ. το χρόνο μέχρι την εμφάνιση υποτροπής ή το χρόνο μέχρι να χαθεί ο ασθενής από τη μελέτη για άσχετους λόγους (lost to follow-up) ή μέχρι να ολοκληρωθεί η μελέτη.

iii. Στο "***Status***" δηλώνουμε την μεταβλητή που υποδηλώνει αν ο ασθενής εμφάνισε υποτροπή ή αν βγήκε από τη μελέτη χωρίς υποτροπή (lost to follow up ή ολοκληρώθηκε η μελέτη). Στη συνέχεια πιέζουμε το κουμπί "***Define Event***" έτσι ώστε να ορίσουμε με ποια ή ποιες τιμές υποδηλώνεται η εμφάνιση υποτροπής (*Εικόνα 12.13*) και πιέζουμε "***Continue***".

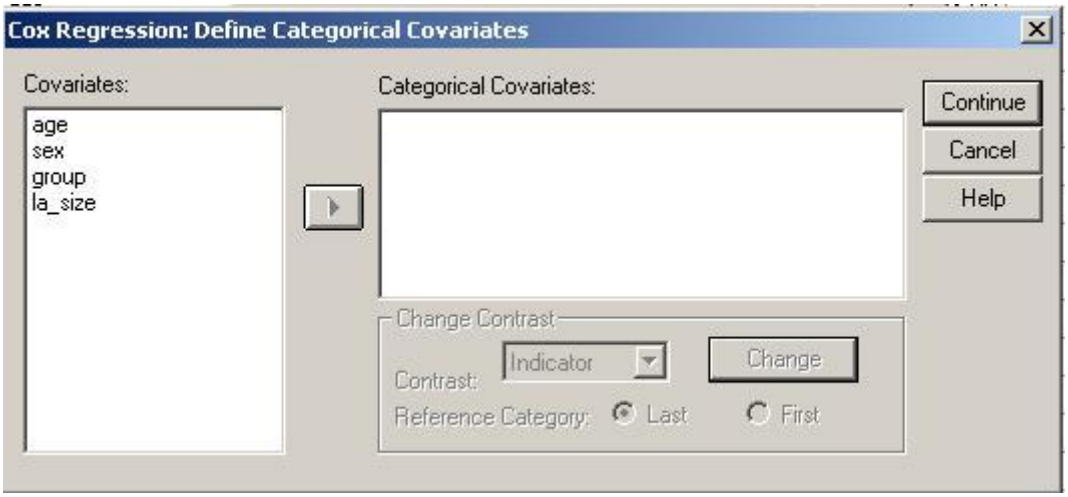


Εικόνα 12.13: Προσδιοσρισμός της τιμής με την οποία είναι κωδικοποιημένη η εμφάνιση ενός γεγονότος.

iv. Στο "***Covariates***" δηλώνουμε όλους τους παράγοντες που επιθυμούμε να ελέγξουμε αν επηρεάζουν σημαντικά το χρόνο μέχρι την εμφάνιση της υποτροπής.

v. Στο ***"Strata"*** μπορούμε να δηλώσουμε την κατηγορική μεταβλητή βάση της οποίας επιθυμούμε να πραγματοποιηθεί στρωματοποιημένη ανάλυση παλινδρόμησης Cox.

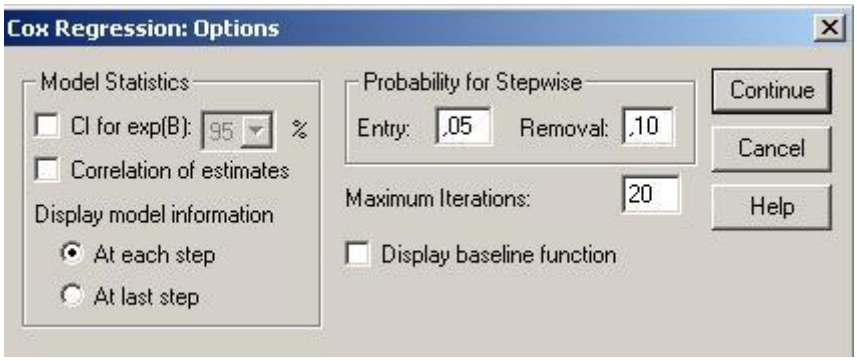
vi. Αν υπάρχουν κατηγορικές μεταβλητές μεταξύ των ***Covariates***, θα πρέπει να το δηλώσουμε και αυτό επιτυγχάνεται πατώντας το κουμπί επιλογών "***Categorical***" όπου ανοίγει ένα νέο παράθυρο διαλόγου (*Εικόνα 12.14*). Σε αυτό το παράθυρο, από το πλαίσιο "***Covariates***" τοποθετούμε στο πλαίσιο "***Categorical covariates***" τις κατηγορικές μεταβλητές και στη συνέχεια μπορούμε να ορίσουμε αν επιθυμούμε να είναι κατηγορία αναφοράς η πρώτη ή η τελευταία κατηγορία αυτών των μεταβλητών. Προεπιλεγμένο είναι να είναι κατηγορία αναφοράς η τελευταία κατηγορία. Αν επιθυμούμε να είναι η πρώτη απλά τσεκάρουμε το "***first***" και πιέζουμε "***change***" και μετά "***continue***".



Εικόνα 12.14: Προσδιορισμός των παραγόντων μεταξύ των ***Covariates*** που είναι κατηγορικές μεταβλητές.

vii. Πατώντας το κουμπί επιλογών "***Options***" ανοίγει ένα νέο παράθυρο διαλόγου (*Εικόνα 12.15*) στο οποίο ενδιαφέρον έχει να επιλέξουμε να εμφανιστεί στο output του SPSS

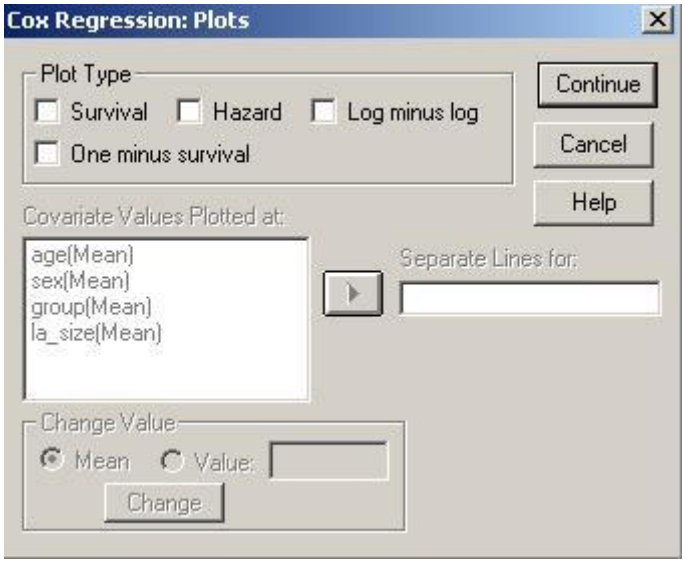
- **CI for exp(β):** το διάστημα εμπιστοσύνης των εκτιμήσεων των συντελεστών
- "**Display baseline function**": μας δίνει το ***Hazard rate*** για κάθε χρονική στιγμή όταν όλες οι ***covariates*** είναι 0, καθώς επίσης το ***survival function*** και το ***hazard rate*** όταν όλες οι ***covariates*** παίρνουν τιμή ίση με την μέση τους τιμή.



Εικόνα 12.15: Κουμπί επιλογών «*Options*» στην ανάλυση παλινδρόμησης Cox.

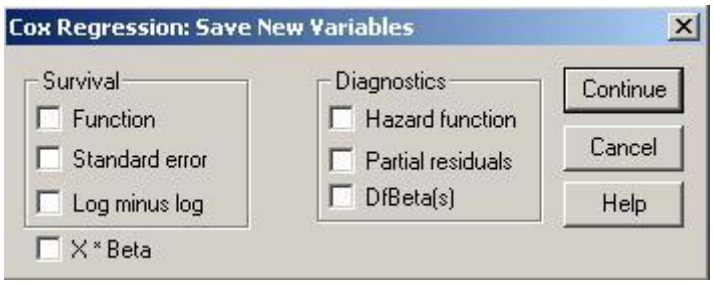
viii. Στη συνέχεια, πατάμε το κουμπί επιλογών «***Plots***» όπου ανοίγει ένα νέο παράθυρο διαλόγου (*Εικόνα 12.16*). Σε αυτό το παράθυρο μπορούμε να επιλέξουμε κάποιο από τα εξής γραφήματα, τα οποία εξυπηρετούν στον έλεγχο της βασικής προϋπόθεσης για την εφαρμογή των μοντέλων παλινδρόμησης Cox, που είναι ότι οι κίνδυνοι είναι αναλογικοί κατά μήκος του χρόνου μεταξύ των επιπέδων μιας κατηγορικής μεταβλητής:

- **Survival Plot**: Σε αυτό το γράφημα παριστάνεται στον κάθετο άξονα το cumulative survival function και στον οριζόντιο άξονα ο χρόνος. Κάθε γραμμή του γραφήματος αντιστοιχεί σε κάθε επίπεδο της κατηγορικής μεταβλητής. Αν οι γραμμές είναι παράλληλες, τότε ισχύει η αναλογικότητα των κινδύνων.
- **Hazard Plot.** Παρομοίως με το ***Survival plot***, αλλά στον κάθετο άξονα παριστάνεται το cumulative hazard function.
- **Log-Minus-Log**: Σε αυτό το γράφημα επιλέγουμε την μεταβλητή μεταξύ των covariates που επιθυμούμε να πραγματοποιηθεί ο έλεγχος. Στον κάθετο άξονα παριστάνεται το log-minus-log της survival function και στον οριζόντιο άξονα παριστάνεται η survival function. Αν οι γραμμές που θα προκύψουν στο γράφημα είναι παράλληλες τότε μπορούμε να ισχυριστούμε ότι ισχύει η προϋπόθεση της αναλογικότητας.
- **One Minus Survival**. Παρομοίως με το ***Survival plot***, αλλά στον κάθετο άξονα παριστάνεται το 1-cumulative Survival function.



Εικόνα 12.16: Μενού επιλογών «***plots***» στην ανάλυση παλινδρόμησης Cox.

ix. Τέλος, πατώντας το κουμπί επιλογών «***Save***», ανοίγει το πλαίσιο διαλόγου της *Εικόνας 12.17*.



Εικόνα 12.17: Επιλογές για την αποθήκευση νέων μεταβλητών.

x. Τα αποτελέσματα παρουσιάζονται στους *Πίνακες 12.6, 12.7, 12.8 & 12.9* καθώς επίσης και στις *Εικόνες 12.18 & 12.19*.

Omnibus Tests of Model Coefficients

-2 Log Likelihood
142,935

Πίνακας 12.6: -2log likelihood για το Cox μοντέλο χωρίς covariates (null model)

Omnibus Tests of Model Coefficients[a,b]

-2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
135,175	7,763	3	,051	7,761	3	,051	7,761	3	,051

a. Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 142,935

b. Beginning Block Number 1. Method = Enter

Πίνακας 12.7: -2log likelihood για το Cox μοντέλο συμπεριλαμβανόμενων όλων των covariates (age, sex, group).

Από τους *Πίνακες 12.6 & 12.7*, ελέγχουμε αν το μοντέλο είναι καλό. Συγκεκριμένα, μπορούμε να ελέγξουμε αν το μοντέλο που συμπεριλαμβάνει όλα τα covariates είναι καλύτερο σε σχέση με το μοντέλο που δεν περιλαμβάνει κανένα covariate στο να ερμηνεύσει το χρονικό διάστημα μέχρι την εμφάνιση ενός γεγονότος. Αυτό επιτυγχάνεται συγκρίνοντας το -2log likelihood του null μοντέλου = 142,935 (*Πίνακας 12.6*) με το -2log likelihood του πλήρους μοντέλου = 135,175 (*Πίνακας 12.7*), με το likelihood ratio test του οποίου το p-value είναι 0,051 (sig. = 0,051, *Πίνακας 12.7*). Αυτό σημαίνει ότι οριακά το πλήρες μοντέλο προσφέρει σημαντική πληροφορία στην ερμηνεία του χρόνου μέχρι την εμφάνιση της υποτροπής σε σχέση με το μοντέλο που δεν περιέχει καμία covariate.

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
group	1,026	,471	4,743	1	,029	2,791	1,108	7,029
age	,042	,025	2,884	1	,089	1,043	,994	1,095
sex	-,273	,559	,239	1	,625	,761	,254	2,275

Πίνακας 12.8: Αποτελέσματα αναφορικά με την επίδραση της κάθε covariate στον χρόνο που μεσολαβεί μέχρι την εμφάνιση της υποτροπής.

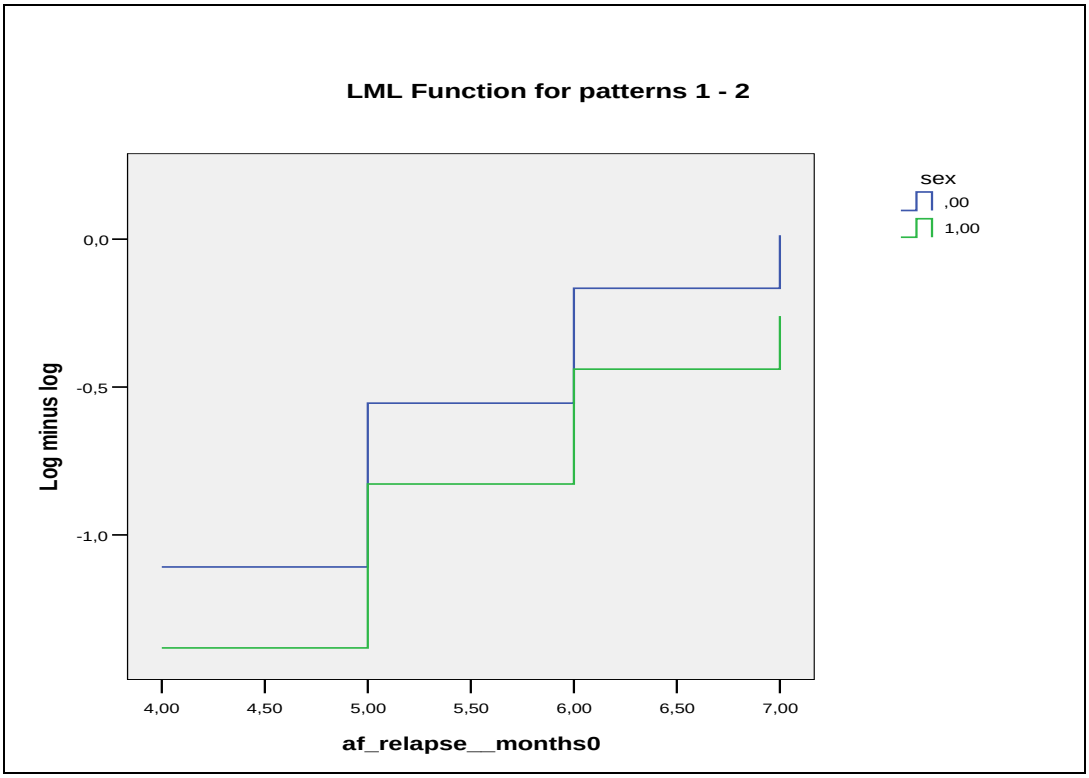
Από τον *Πίνακα 12.8* διαπιστώνουμε ότι ο τύπος της επέμβασης στην οποία υπεβλήθησαν οι ασθενείς διαδραματίζει στατιστικά σημαντικό ρόλο στην ερμηνεία του χρόνου εμφάνισης υποτροπής (sig.=0,029), ενώ αντίθετα η ηλικία και το φύλο, όχι (sig. = 0,089 & 0,625, αντίστοιχα). Επίσης, από τον ίδιο *Πίνακα* διαπιστώνουμε ότι ο κίνδυνος εμφάνισης υποτροπής στους ασθενείς που υπεβλήθησαν στην θεραπεία Β είναι 2,8 φορές μεγαλύτερος σε σχέση με αυτούς που υπεβλήθησαν στην θεραπεία Α, αφού το hazard ratio (exp(B)) = 2,791.

Survival Table

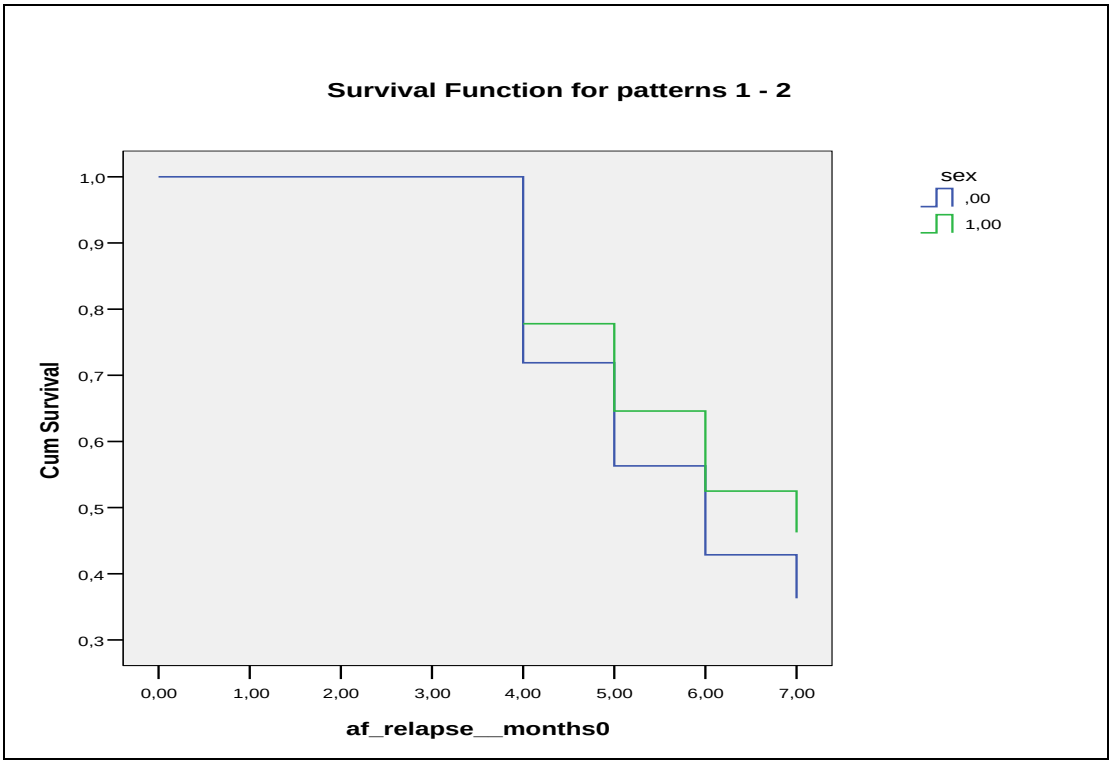
Time	Baseline Cum Hazard	At mean of covariates		
		Survival	SE	Cum Hazard
4,00	,022	,769	,058	,262
5,00	,038	,634	,070	,456
6,00	,056	,510	,075	,673
7,00	,067	,447	,077	,805

Πίνακας 12.9: Cumulative hazard όταν όλες οι covariates πάρουν τιμή ίσον με το 0 και όταν όλες οι covariates πάρουν τιμές ίσες με τη μέση τιμή.

Από τον *Πίνακα 12.9* παρατηρούμε για κάθε χρονική στιγμή μεταξύ αυτών που εμφανίστηκαν υποτροπές, ποιος είναι ο cumulative hazard όταν όλες οι covariates πάρουν τιμή ίση με το μηδέν και ποιος είναι όταν όλες οι covariates πάρουν τιμές ίσες με τη μέση τιμή.



Εικόνα 12.18: log minus log της αθροιστικής συνάρτησης επιβίωσης



Εικόνα 12.19: Αθροιστικής συνάρτησης επιβίωσης

Από τις *Εικόνες 12.18 & 12.19* παρατηρούμε ότι αναφορικά με το φύλο ισχύει η προϋπόθεση της αναλογικότητας των κινδύνων.

13. Ανάλυση σε Κύριες Συνιστώσες

13.1 Εισαγωγή

Η Ανάλυση σε Κύριες Συνιστώσες είναι μια πολυμεταβλητή στατιστική μέθοδος η οποία αποσκοπεί στην δημιουργία ενός συνόλου νέων μεταβλητών χρησιμοποιώντας τις πιθανές συσχετίσεις μεταξύ ενός συνόλου χαρακτηριστικών (μεταβλητών) ενός δείγματος. Οι νέες αυτές μεταβλητές είναι ασυσχέτιστες μεταξύ τους και αποτελούν γραμμικούς συνδυασμούς των αρχικών μεταβλητών από τις οποίες προέκυψαν. Με τον τρόπο αυτό προκύπτει στην πραγματικότητα ισάριθμο πλήθος γραμμικών μετασχηματισμών με το πλήθος των αρχικών μεταβλητών.

Το κέρδος του ερευνητή από τη διαδικασία της Ανάλυσης σε Κύριες Συνιστώσες είναι πολλαπλό.

1. Από ένα σύνολο συσχετισμένων μεταβλητών προκύπτει ένα σύνολο νέων μεταβλητών, γραμμικών μετασχηματισμών των αρχικών, οι οποίες φέρουν την εξαιρετική ιδιότητα να είναι ασυσχέτιστες μεταξύ τους. Πρόκειται για μια πολλή χρήσιμη ιδιότητα η οποία αποτελεί μάλιστα απαραίτητη προϋπόθεση για πλήθος άλλων στατιστικών τεχνικών. Ως παράδειγμα αναφέρεται η μέθοδος της γραμμικής παλινδρόμησης όπου, αν μεταξύ των ανεξάρτητων μεταβλητών, που χρησιμοποιούνται στο μοντέλο, υφίστανται ισχυρές συσχετίσεις τότε υπάρχει πρόβλημα πολυσυγγραμμικότητας γεγονός που καθιστά το μοντέλο εξαιρετικά ασταθές και τις αντίστοιχες εκτιμήσεις, που προκύπτουν απ' αυτό, μάλλον αναξιόπιστες.
2. Οι νέες μεταβλητές-γραμμικοί συνδυασμοί των αρχικών μεταβλητών- οι οποίες προκύπτουν από την εφαρμογή της Ανάλυσης σε Κύριες Συνιστώσες, εμφανίζονται σε φθίνουσα σειρά με βάση το ποσοστό της αρχικής μεταβλητότητας του συστήματος που η κάθε μια ερμηνεύει. Αν η εσωτερική δομή των δεδομένων το επιτρέπει, δηλαδή οι συσχετίσεις μεταξύ των αρχικών μεταβλητών της έρευνας είναι αρκετά ισχυρές και σαφείς, είναι δυνατόν ένας πολύ μικρός αριθμός νέων μεταβλητών, σε σχέση με τον αριθμό των αρχικών μεταβλητών, να ερμηνεύει μεγάλο (ικανό) ποσοστό της αρχικής μεταβλητότητας των δεδομένων. Αν κάτι τέτοιο παρατηρηθεί τότε ο ερευνητής βρίσκεται στην ευχάριστη θέση να έχει μειώσει σημαντικά τις διαστάσεις του προβλήματος που μελετά έχοντας όσο το δυνατόν μικρότερο κόστος όσον αφορά το ποσοστό της συνολικής μεταβλητότητας που τελικά αντιπροσωπεύεται από το σύνολο των τελικών μεταβλητών που αποφασίζει να διατηρήσει.

   Με τον τρόπο αυτό μπορεί να χάνεται μικρό μέρος της αρχικής πληροφορίας, παράλληλα όμως εξοικονομείται σημαντικός χώρος (στην πολύ συνηθισμένη περίπτωση των τεράστιων βάσεων δεδομένων) και υπολογιστική ισχύς, ενώ παράλληλα το πρόβλημα ανάγεται σε μελέτη των σχέσεων μεταξύ ενός πολύ μικρότερου και άρα πιο εύκολα παρατηρήσιμου και ερμηνεύσιμου πλήθους μεταβλητών.

   Η μείωση των διαστάσεων ενός προβλήματος είναι πολύ χρήσιμη και σε μελέτες που για τα υπό μελέτη φαινόμενα υπάρχει πλήθος μεταβλητών / χαρακτηριστικών και δυσανάλογα μικρός αριθμός παρατηρήσεων (π.χ. στην ιατρική στην περίπτωση που εξετάζονται εξαιρετικά σπάνιες ασθένειες-ασθένειες με πολύ μικρό επιπολασμό). Στην περίπτωση αυτή καμιά περαιτέρω στατιστική ανάλυση δεν μπορεί να πραγματοποιηθεί αν προηγουμένως δε

μειωθούν οι διαστάσεις του προβλήματος ώστε για κάθε υπό μελέτη χαρακτηριστικό να υφίσταται ένας ικανός αριθμός μετρήσεων.

3. Η μέθοδος της Ανάλυσης σε Κύριες Συνιστώσες βασίζεται επίσης στη μελέτη των συσχετίσεων μεταξύ των μεταβλητών και τη διαπίστωση τυχόν ομοιοτήτων μεταξύ τους. Είναι πιθανόν κατά το σχεδιασμό μιας έρευνας να συλλεγεί πληροφορία για πλήθος μεταβλητών κάποιες από τις οποίες μπορεί ουσιαστικά να μετρούν το ίδιο ή παρόμοιο χαρακτηριστικό. Στην περίπτωση αυτή η συνεισφορά της μιας εκ των δύο μεταβλητών είναι επουσιώδης στη μελέτη. Με την Ανάλυση σε Κύριες Συνιστώσες υπάρχει η δυνατότητα διάκρισης μεταξύ των μεταβλητών που αντιπροσωπεύουν παρόμοια χαρακτηριστικά και δημιουργίας νέων μεταβλητών οι οποίες θα συνοψίσουν την πληροφορία κατά τέτοιο τρόπο ώστε κάθε συνδυασμός μεταβλητών με ισχυρές μεταξύ τους συσχετίσεις να αντιπροσωπεύεται μοναδικά (μονοσήμαντα) από μία (την πιο χαρακτηριστική) μεταβλητή στο γραμμικό συνδυασμό.
4. Ένα επίσης πολύ σημαντικό πλεονέκτημα της Ανάλυσης σε Κύριες Συνιστώσες, είναι η δυνατότητα που παρέχει η μέθοδος για ποσοτικοποίηση μη μετρήσιμων εννοιών (χαρακτηριστικών-ποσοτήτων). Πιο συγκεκριμένα, από ερωτηματολόγια που συμπεριλαμβάνουν μία σειρά ερωτήσεων σχετικών με τις γνώσεις των παιδιών και την όποια κοινωνική υποστήριξη που αυτά έχουν από το περιβάλλον τους, αναφορικά με διάφορα θέματα διατροφής, ο συνδυασμός της πληροφορίας με τη βοήθεια των συγκεκριμένων ερωτηματολογίων και η σύνθεσή της με τη βοήθεια της μεθόδου της Ανάλυσης σε Κύριες Συνιστώσες, έχει ως αποτέλεσμα τη δημιουργία νέων μεταβλητών, συνισταμένων των αρχικών μεταβλητών, οι οποίες υποδηλώνουν μη μετρήσιμες έννοιες όπως π.χ. την αντίδραση του παιδιού σε προσπάθεια επιβολής κατανάλωσης κάποιων ειδών τροφίμων από το οικογενειακό του περιβάλλον. Βέβαια ο χαρακτηρισμός των νέων μεταβλητών που προκύπτουν με τη μέθοδο αυτή και η απόδοση σ' αυτές συγκεκριμένης υποδηλούσας έννοιας είναι διαδικασίες που ενέχουν σε σημαντικό βαθμό το υποκειμενικό στοιχείο. Ως εκ τούτου πολλοί θεωρούν το σημαντικό αυτό πλεονέκτημα της μεθόδου ως ένα από τα βασικότερα μειονεκτήματά της.

13.1.1 Διαδικασία εύρεσης Κύριων Συνιστωσών

13.1.1.1 Έλεγχος Συσχετίσεων

Ανεξάρτητα από το αν θα χρησιμοποιήσουμε τον πίνακα διακύμανσης ή τον πίνακα συσχετίσεων, σκόπιμο είναι να εξετάσουμε τον πίνακα συσχετίσεων για να διαπιστώσουμε αν μεταξύ των αρχικών μεταβλητών μας υφίστανται ικανές συσχετίσεις ώστε να προχωρήσουμε σε εφαρμογή της μεθόδου της ανάλυσης σε κύριες συνιστώσες. Σε περίπτωση που ανιχνεύσουμε μεταβλητές οι οποίες φαίνεται να είναι ασυσχέτιστες από τις υπόλοιπες εξαιρούμε τις μεταβλητές αυτές από την ανάλυση σε κύριες συνιστώσες που σκοπεύουμε να εφαρμόσουμε.

Όταν μιλάμε για ικανές συσχετίσεις, εννοούμε συσχετίσεις μεγάλες σε απόλυτη τιμή, είτε θετικές είτε αρνητικές. Μιλώντας για μεγάλες συσχετίσεις δεν εννοούμε απαραίτητα για στατιστικά σημαντικές σύμφωνα με το αποτέλεσμα κάποιου σχετικού στατιστικού ελέγχου. Η πράξη έχει δείξει ότι ακόμα και συσχετίσεις τις τάξεως του 0,1 είναι στατιστικά σημαντικές για δείγματα μετρίου μεγέθους (π.χ. όταν διαθέτουμε δείγματα περίπου 300 παρατηρήσεων). Για να είναι όμως οι συσχετίσεις ικανοποιητικές προκειμένου να προχωρήσουμε σε ανάλυση σε

κύριες συνιστώσες, θα πρέπει να είναι, σε απόλυτη τιμή, της τάξεως του 0,4 ή και μεγαλύτερες.

Προκειμένου να εξακριβώσουμε αν οι υφιστάμενες συσχετίσεις μεταξύ των αρχικών μεταβλητών του υπό εξέταση προβλήματος είναι ικανοποιητικές για την εφαρμογή της μεθόδου της ανάλυσης σε κύριες συνιστώσες, έχει προταθεί η ακόλουθη στατιστική συνάρτηση:

$$\varphi = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{p} r_{ij}^2 - p}{p(p-1)}}$$

όπου, r_{ij} είναι το ij στοιχείο του πίνακα συσχετίσεων, δηλαδή η συσχέτιση μεταξύ των μεταβλητών X_i και X_j. Το στατιστικό φ παίρνει τιμές μεταξύ 0 και 1. Τιμές κοντά στο 1 υποδηλώνουν μεγάλες συσχετίσεις μεταξύ των μεταβλητών, καθώς σ' αυτή την περίπτωση όλα τα r_{ij} θα πλησιάζουν τη μονάδα και επομένως το άθροισμά των τετραγώνων τους θα είναι κοντά στο p^2, δηλαδή ο αριθμητής θα τείνει να είναι ίσος με τον παρονομαστή. Αν, αντίθετα, οι συσχετίσεις μεταξύ των μεταβλητών είναι ασθενείς ή ανύπαρκτες, οι τιμές του στατιστικού φ θα είναι κοντά στο 0, καθώς μόνο τα p διαγώνια στοιχεία θα είναι 1, άρα το άθροισμα τετραγώνων τους θα είναι περίπου p. Αυτό σημαίνει ότι ο αριθμητής του κλάσματος, και κατά συνέπεια και το στατιστικό φ, θα τείνει να μηδενιστεί. Στην πράξη τιμές συσχετίσεων μεταξύ των μεταβλητών γύρω στο 0,4 θεωρούνται ικανοποιητικές για την εφαρμογή της μεθόδου της ανάλυσης σε κύριες συνιστώσες.

13.1.1.2 Επιλογή πίνακα

Όπως έχει ήδη αναφερθεί, είναι δυνατό στην ανάλυσή μας να χρησιμοποιήσουμε είτε τον πίνακα διακυμάνσεων είτε τον πίνακα συσχετίσεων. Τα αποτελέσματα στα οποία θα καταλήξουμε θα είναι διαφορετικά ανάλογα με την επιλογή του πίνακα τον οποίο θα χρησιμοποιήσουμε. Για το λόγο αυτό είναι πολύ σημαντικό, πριν προχωρήσουμε στην εφαρμογή της συγκεκριμένης μεθόδου, να εξετάσουμε τα δεδομένα μας και να αποφασίσουμε, βασισμένοι σε σωστά κριτήρια, με ποιον από τους δύο πίνακες θα δουλέψουμε.

13.1.1.3 Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων

Ανάλογα με τον πίνακα τον οποίο επιλέξαμε για να στηρίξουμε την ανάλυσης μας, υπολογίζουμε τις ιδιοτιμές και τα ιδιοδιανύσματα, με τον τρόπο που έχει ήδη περιγραφή στην παράγραφο **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.** του κεφαλαίου. Πρέπει να σημειωθεί ότι τα ιδιοδιανύσματα που δίνουν τα στατιστικά πακέτα είναι κανονικοποιημένα, δηλαδή έχουν άθροισμα τετραγώνων ίσο με τη μονάδα. Επίσης δεν είναι μοναδικά με την έννοια ότι μπορούμε να αλλάξουμε το πρόσημο σε όλα τα στοιχεία τους. Κατά συνέπεια η λύση που προκύπτει από την εφαρμογή της ανάλυσης σε κύριες συνιστώσες μπορεί να διαφέρει από στατιστικό πακέτο σε στατιστικό πακέτο ως προς τα πρόσημα.

13.1.1.4 Απόφαση του αριθμού των κύριων συνιστωσών που θα διατηρήσουμε

Όπως έχει ήδη αναφερθεί, η εφαρμογή της ανάλυσης σε κύριες συνιστώσες δίνει ως λύση τόσες νέες μεταβλητές, τις κύριες συνιστώσες, όσες και οι αρχικές μεταβλητές του υπό εξέταση προβλήματος. Το σύνολο των κύριων συνιστωσών που προκύπτουν ερμηνεύει το σύνολο της αρχικής μεταβλητότητας του συστήματος. Επιλέγοντας να

διατηρήσουμε λιγότερες κύριες συνιστώσες από όσες μεταβλητές είχαμε αρχικά, χάνουμε αναπόφευκτα ένα μέρος της ερμηνευόμενης μεταβλητότητας και κατά συνέπεια της αρχικής πληροφορίας που διαθέταμε. Αυτό όμως είναι ουσιαστικά και το «κόστος» που καλούμαστε να «πληρώσουμε» προκειμένου να μειώσουμε τις διαστάσεις του προβλήματος που εξετάζουμε.

Ένα από τα πιο σημαντικά ερωτήματα στο οποίο καλούμαστε να δώσουμε μια σαφή και ικανοποιητική απάντηση κατά την εφαρμογή της συγκεκριμένης μεθόδου, είναι το πόσες τελικά από τις προκύπτουσες νέες συνιστώσες είναι σκόπιμο να διατηρήσουμε στη μελέτη μας. Στην βιβλιογραφία παρουσιάζεται ένας μεγάλος αριθμός κριτηρίων τα οποία απαντούν κατά κάποιο τρόπο στο ερώτημα αυτό. Μερικά από αυτά παρουσιάζονται συνοπτικά στις επόμενες παραγράφους.

1. Ποσοστό συνολικής διακύμανσης που ερμηνεύουν οι συνιστώσες.

Σύμφωνα με το κριτήριο αυτό εξετάζουμε το ποσοστό της συνολικής διακύμανσης του συστήματος που ερμηνεύουν οι κύριες συνιστώσες. Θέτουμε ένα (αυθαίρετο) όριο (π.χ. γύρω στο 80%) και διαλέγουμε τις πρώτες κύριες συνιστώσες που προκύπτουν από την εφαρμογή της μεθόδου και οι οποίες ερμηνεύουν αθροιστικά ποσοστό διακύμανσης ίσο ή ανώτερο από το όριο αυτό. Πρόκειται για ένα εξαιρετικά απλό, άρα και εύχρηστο, κριτήριο, ωστόσο δε δίνει ιδιαίτερα καλά αποτελέσματα ιδίως αν ο στόχος (το όριο) που θέτουμε είναι ιδιαίτερα ψηλός. Σημαντικό επίσης μειονέκτημά του αποτελεί και το αυθαίρετο του τρόπου με τον οποίο καθορίζουμε το ποσοστό της συνολικής διακύμανσης του συστήματος το οποίο θέλουμε τελικά να ερμηνεύεται από τις κύριες συνιστώσες που διατηρούμε στο σύστημα.

Πολλές φορές η πράξη δείχνει ότι μέτρια αποτελέσματα μπορούμε να επιτύχουμε αν θέσουμε ως όριο ένα 55 με 60% της αρχικής μεταβλητότητας του συστήματος. Εφαρμογή της χρήσης ενός ανάλογου ορίου θα περιγράψουμε στο ειδικό μέρος της παρούσας διπλωματικής.

2. Το κριτήριο του Kaiser

Το κριτήριο του Kaiser προτείνει να διατηρήσουμε εκείνες τις κύριες συνιστώσες των οποίων οι αντίστοιχες ιδιοτιμές έχουν τιμή μεγαλύτερη από τη μέση τιμή του συνόλου των ιδιοτιμών που προκύπτουν από την εφαρμογή της μεθόδου. Τα πράγματα είναι απλούστερα στην περίπτωση που δουλεύουμε με τον πίνακα συσχετίσεων. Στην περίπτωση αυτή η μέση τιμή των ιδιοτιμών θα είναι ίση με τη μονάδα, δηλαδή θα ισχύει ότι $\bar{\lambda}=1$. Αυτό σημαίνει ότι, σύμφωνα με το κριτήριο του Kaiser, καλούμαστε να διατηρήσουμε όσες από τις κύριες συνιστώσες που προκύπτουν έχουν ιδιοτιμές μεγαλύτερες της μονάδας.

Επίσης πρόκειται για ένα κριτήριο απόφασης για τις κύριες συνιστώσες που θα διατηρήσουμε, εξαιρετικά απλό και κατά συνέπεια εύχρηστο. Ωστόσο η υπόθεση στην οποία βασίζεται είναι ιδιαίτερα απλοϊκή καθώς, στην πράξη είναι ανεδαφικό να υποστηρίζουμε τελεία απουσία δομής στα δεδομένα μας και κατά συνέπεια εξίσωση όλων των ιδιοτιμών με τη μονάδα. Στην πραγματικότητα, όταν δουλέψουμε με ένα δείγμα, κάποιες από τις ιδιοτιμές θα είναι ούτως ή άλλως μεγαλύτερες της μονάδας αφού το άθροισμά τους πρέπει να ισούται με των αριθμό των αρχικών μεταβλητών και, κατά συνέπεια, και των κύριων συνιστωσών που προκύπτουν με την εφαρμογή της συγκεκριμένης μεθόδου. Ως εκ τούτου, το κριτήριο του Kaiser συνήθως υπερεκτιμά των αριθμών των κύριων συνιστωσών που χρειάζεται να κρατήσουμε στη μελέτη μας.

3. Ποσοστό διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται.

Όπως έχει ήδη ειπωθεί, αν διατηρήσουμε μόνο μερικές από τις συνιστώσες που προκύπτουν από την εφαρμογή της μεθόδου αυτής, χάνουμε ένα μέρος της αρχικής πληροφορίας που έχουμε συλλέξει. Είναι δυνατόν να υπολογίσουμε το ποσοστό της διακύμανσης των αρχικών μεταβλητών που ερμηνεύεται όταν κρατήσουμε ένα συγκεκριμένο αριθμό συνιστωσών. Το συγκεκριμένο κριτήριο υποδεικνύει να κρατήσουμε τόσες κύριες συνιστώσες όσες χρειάζονται ώστε να ερμηνεύεται για κάθε αρχική μεταβλητή ένα υψηλό, σχετικά, ποσοστό διακύμανσης. Η επιλογή του ποσοστού, που τίθεται ως όριο, στην περίπτωση αυτή είναι καθαρά υποκειμενική.

4. Scree Plot

Πρόκειται για ένα επίσης υποκειμενικό και σχετικά εύχρηστο κριτήριο το οποίο στηρίζει την επιλογή του αριθμού κύριων συνιστωσών, που πρέπει να διατηρηθούν, σε μια γραφική απεικόνιση των κύριων αυτών συνιστωσών και των αντίστοιχων ιδιοτιμών τους. Πιο συγκεκριμένα, το scree plot είναι ένα γράφημα το οποίο παριστά στον οριζόντιο άξονα των x τη σειρά και στον κάθετο άξονα των y την τιμή της κάθε ιδιοτιμής που αντιστοιχεί σε κάθε κύρια συνιστώσα. Η ένωση των σημείων δίνει ως γραφική απεικόνιση μια τεθλασμένη γραμμή με φθίνουσα κλίση.

Το κριτήριο προτείνει να διατηρήσουμε τόσες συνιστώσες μέχρι το σημείο που η τεθλασμένη γραμμή του γραφήματος να αλλάζει απότομα κλίση. Ωστόσο το συγκεκριμένο κριτήριο πρέπει να χρησιμοποιείται με μεγάλη προσοχή, αφενός μεν επειδή ενέχει έντονα το υποκειμενικό στοιχείο, αφετέρου δε επειδή βασίζεται σε γραφική απεικόνιση των αποτελεσμάτων στην οποία πολλές φορές μπορεί να μην είναι απολύτως ξεκάθαρη η υποδηλούμενη δομή των αποτελεσμάτων.

13.1.1.5 Ερμηνεία κύριων συνιστωσών

Έχοντας καταλήξει στον τελικό αριθμό κύριων συνιστωσών που θα διατηρήσουμε στη μελέτη μας, το επόμενο βήμα είναι να προσπαθήσουμε να ερμηνεύσουμε τις συνιστώσες αυτές. Αν η ανάλυση έχει γίνει σωστά και οι υφιστάμενες συσχετίσεις μεταξύ των αρχικών μεταβλητών το επιτρέπουν, μπορούμε, με την εφαρμογή της συγκεκριμένης μεθόδου, αφενός να μειώσουμε σημαντικά τις διαστάσεις του προβλήματος διατηρώντας σχετικά μικρό αριθμό κύριων συνιστωσών στη μελέτη μας, και αφετέρου να ποσοτικοποιήσουμε μη μετρήσιμες έννοιες.

Η ερμηνεία των κύριων συνιστωσών βασίζεται κυρίως στους συντελεστές των γραμμικών σχέσεων που κάθε μια από αυτές έχει με τις αρχικές μεταβλητές πάνω στις οποίες βασίστηκε η συγκεκριμένη μέθοδος. Εξετάζουμε τους συντελεστές αυτούς τόσο ως προς τις απόλυτες τιμές τους όσο και ως προς τα πρόσημά τους. Επίσης, σε συνεργασία με τον επιστήμονα, στο γνωστικό αντικείμενο του οποίου πραγματοποιείται η συγκεκριμένη μελέτη, είναι δυνατόν να ανιχνεύσουμε και να αποδώσουμε στις κύριες συνιστώσες κάποια επιπλέον χαρακτηριστικά. Τα χαρακτηριστικά αυτά είναι πιθανόν να είναι λίγο πιο εξειδικευμένα και όχι εύκολα ανιχνεύσιμα με απλή παρατήρηση των τιμών και των πρόσημων των συντελεστών που έχουν οι αρχικές μεταβλητές σε κάθε μια από τις προκύπτουσες συνιστώσες.

Στο πλαίσιο της προσπάθειας ερμηνείας των κύριων συνιστωσών ανήκει και η λεγόμενη «περιστροφή των αξόνων». Πρόκειται για μια τεχνική στην οποία μπορεί να καταφύγει ο ερευνητής προκειμένου να διευκολύνει τη διαδικασία ερμηνείας των κύριων συνιστωσών. Η συγκεκριμένη τεχνική είναι ουσιαστικά πολλαπλασιασμός

του πίνακα συντελεστών των κύριων συνιστωσών που βρήκαμε με έναν ορθογώνιο πίνακα. Από τους άπειρους ορθογώνιους πίνακες μπορούμε να διαλέξουμε κάποιον με βάση συγκεκριμένα κριτήρια βελτιστοποίησης, όπως για παράδειγμα την απαίτηση κάθε κύρια συνιστώσα να έχει όσο το δυνατόν λιγότερες μεταβλητές με μεγάλους συντελεστές. Η περιστροφή των αξόνων καταλήγει στο αποτέλεσμα κάθε συνιστώσα να έχει λίγους συντελεστές με μεγάλες απόλυτες τιμές ενώ όλοι οι υπόλοιποι να είναι κοντά στο μηδέν. Αυτό συντελεί στην πιο εύκολη αναγνώριση και ερμηνεία της κάθε συνιστώσας καθώς οι μεταβλητές και, κατά συνέπεια, η συνεισφορά τους στους διάφορους συνδυασμούς με τους οποίους εμφανίζονται στις κύριες συνιστώσες, ξεχωρίζουν πιο έντονα με βάση τα πρόσημά τους.

13.1.1.7 Δημιουργία νέων μεταβλητών

Οι κύριες συνιστώσες, οι οποίες προκύπτουν από τη συγκεκριμένη ανάλυση, αποτελούν τις νέες μεταβλητές της μελέτης μας. Για κάθε μονάδα του δείγματός μας μπορούμε να αποκτήσουμε μια τιμή για κάθε μια από τις κύριες συνιστώσες που αποφασίσαμε να διατηρήσουμε στη μελέτη. Αυτό μπορεί να γίνει πολλαπλασιάζοντας, σε κάθε συνδυασμό αρχικών μεταβλητών σε μια κύρια συνιστώσα, την τιμή κάθε συντελεστή με την τιμή που έχει η συγκεκριμένη μονάδα για την αντίστοιχη μεταβλητή στο δείγμα.

13.2 Ανάλυση σε κύριες συνιστώσες με τη χρήση του SPSS

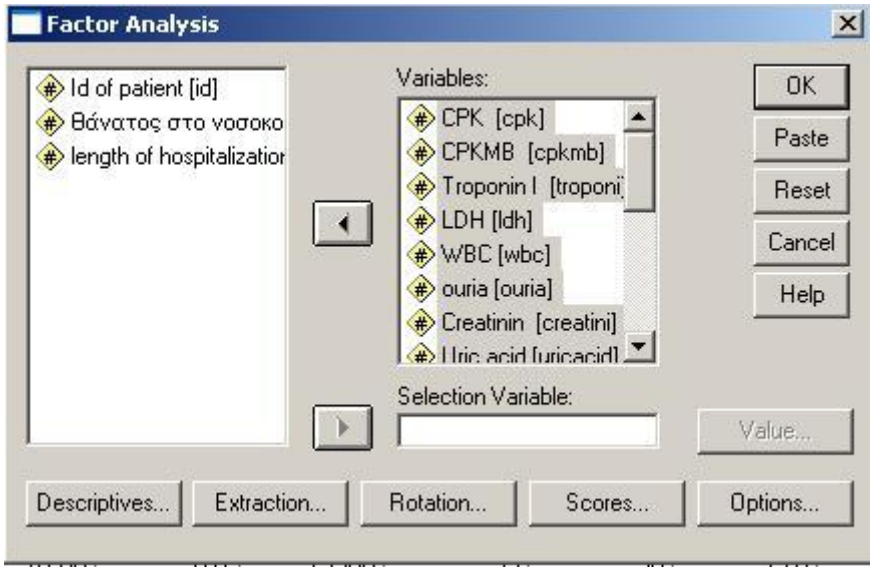
Έστω ότι διαθέτουμε μια σειρά από βιοχημικά, αιματολογικά, κοινωνικό-δημογραφικά και διατροφικά χαρακτηριστικά ασθενών που εισήχθησαν στο νοσοκομείο με Οξύ Στεφανιαίο Σύνδρομο (ΟΣΣ). Συγκεκριμένα, ας θεωρήσουμε ότι διαθέτουμε τα εξής στοιχεία:
cpk: Επίπεδα της CPK κατά την εισαγωγή του ασθενή (ng/ml)
cpkmb: Επίπεδα του ισοενζύμου MB της CPK κατά την εισαγωγή του ασθενή (ng/ml)
troponi: Επίπεδα τροπονίνης I κατά την εισαγωγή του ασθενή (ng/ml)
LDH: Γαλακτική αφυδρογονάση (mg/dL)
WBC: Αριθμός λευκών αιμοσφαιρίων κατά την εισαγωγή (αριθμός κυττάρων/dL)
ouria: Επίπεδα ουρίας κατά την εισαγωγή (mg/dL)
creatinine: Επίπεδα κρεατινίνης κατά την εισαγωγή (mg/dL)
uric acid: Ουρικό οξύ (mg/dL)
age: Ηλικία σε έτη
sex: Φύλο (Άνδρας: 1 & Γυναίκες: 0)
weight: Βάρος σε kgr
height: Ύψος σε cm
legumes: Κατανάλωση οσπρίων (φορές / εβδομάδα, 0 - 5)
vegetabl: Κατανάλωση λαχανικών (φορές / εβδομάδα, 0 - 5)
salads: Κατανάλωση σαλάτας (φορές / εβδομάδα, 0 - 5)

Επειδή κάποιες από τις παραπάνω μεταβλητές μετρούν το ίδιο χαρακτηριστικό (π.χ. CPK, CPKMB και troponi, είναι βιοχημικοί δείκτες νέκρωσης του μυοκαρδίου), ενδέχεται να συσχετίζονται ισχυρά μεταξύ τους και γι' αυτό είναι δόκιμο να μη χρησιμοποιηθούν ως έχουν σε περαιτέρω αναλύσεις, αλλά να πραγματοποιηθεί ανάλυση σε κύριες συνιστώσες, αρχικά.
Για να πραγματοποιήσουμε ανάλυση σε κύριες συνιστώσες ακολουθούμε τα εξής ***βήματα***:

Analyse → **Data Reduction** → **Factor**

Τότε εμφανίζεται το πλαίσιο διαλόγου που φαίνεται στην *Εικόνα 13.1*. Σε αυτό το πλαίσιο διαλόγου δηλώνουμε:

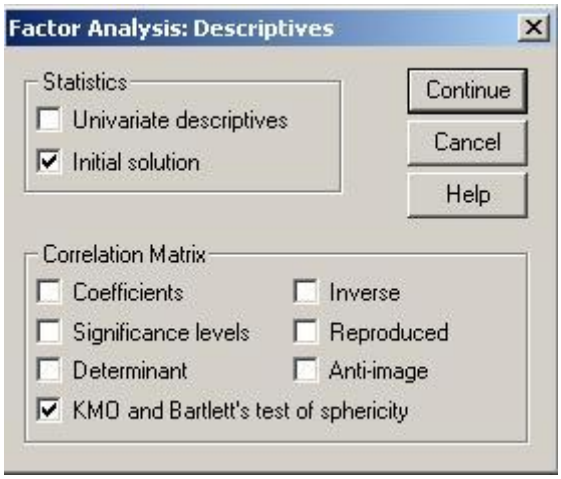
i. τις μεταβλητές «***Variables***» που θέλουμε να χρησιμοποιηθούν στην ανάλυση (π.χ. όλες τις παραπάνω).



Εικόνα 13.1: Το αρχικό παράθυρο της ανάλυσης σε κύριες συνιστώσες

ii. Επιλέγοντας το πλήκτρο «***Descriptives»*** εμφανίζεται το παράθυρο της *Εικόνας 13.2*. Οι επιλογές αφορούν διάφορα περιγραφικά μέτρα που είναι χρήσιμα για την ανάλυση μας. Μπορούμε να επιλέξουμε όποια από αυτά θέλουμε. Τα πιο σημαντικά μέτρα από αυτά που προκύπτουν είναι:

- ***Coefficients***: Ο πίνακας συσχετίσεων, αφού είναι γνωστό ότι απαραίτητη προϋπόθεση για την ορθή εφαρμογή της ανάλυσης σε κύριες συνιστώσες είναι η ύπαρξη ισχυρών συσχετίσεων ανάμεσα στις αρχικές μεταβλητές (βλ. *Πίνακα 13.1* για τα αποτελέσματα από το παράδειγμά μας). Παρατηρούμε ότι υπάρχουν αρκετοί συντελεστές συσχέτισης > |0,3|.
- ***Significance levels***: Ο πίνακας με τη στατιστική σημαντικότητα κάθε συσχέτισης ξεχωριστά (βλ. *Πίνακα 13.1* για τα αποτελέσματα από το παράδειγμά μας).
- ***Determinant***: Η ορίζουσα του πίνακα συσχετίσεων. Τιμές κοντά στο 0 υποδηλώνουν την ύπαρξη συσχετίσεων (βλ. *Πίνακα 13.2* για τα αποτελέσματα από το παράδειγμά μας). Η *Determinant* είναι 0,046.
- ***KMO and Bartlett's test of sphericity***: Ο έλεγχος σφαιρικότητας του Bartlett και η Kaiser-Meyer-Olkin στατιστική συνάρτηση για την καταλληλότητα των δεδομένων (βλ. *Πίνακα 13.2* για τα αποτελέσματα από το παράδειγμά μας). Παρατηρούμε ότι το KMO είναι 0,645>0,5, που σημαίνει ότι υπάρχουν αρκετά ισχυρές συσχετίσεις.



Εικόνα 13.2: Επιλογές από το παράθυρο Descriptives

Correlation Matrix[a]

		CPK	CPKMB	Troponin I	LDH	WBC	ouria	Creatinin	Uric acid	age	sex	Weight (Kg)	Height (cm)	Legume consum	Vegetables consum	salads
Correlation	CPK	1,000	,748	,129	,563	,205	-,016	-,001	,008	-,064	,092	,000	,037	,037	-,045	-,049
	CPKMB	,748	1,000	,194	,575	,221	,012	,009	,020	-,012	,060	-,040	-,011	,090	-,054	-,061
	Troponin I	,129	,194	1,000	,007	,101	-,003	,010	,003	-,007	,024	-,011	,013	-,046	-,049	,012
	LDH	,563	,575	,007	1,000	,189	,057	,021	,038	-,009	,033	,005	-,006	,169	,017	-,026
	WBC	,205	,221	,101	,189	1,000	,112	,084	,040	-,016	,082	,016	,049	,048	-,036	-,013
	ouria	-,016	,012	-,003	,057	,112	1,000	,492	,131	,264	-,080	-,100	-,077	,070	,042	,055
	Creatinin	-,001	,009	,010	,021	,084	,492	1,000	,063	,084	,048	-,027	,020	-,014	-,016	,001
	Uric acid	,008	,020	,003	,038	,040	,131	,063	1,000	,070	,058	,025	,061	,059	-,002	,043
	age	-,064	-,012	-,007	-,009	-,016	,264	,084	,070	1,000	-,234	-,305	-,310	,105	,111	,056
	sex	,092	,060	,024	,033	,082	-,080	,048	,058	-,234	1,000	,295	,599	-,049	-,033	-,047
	Weight (Kg)	,000	-,040	-,011	,005	,016	-,100	-,027	,025	-,305	,295	1,000	,529	-,029	-,038	,022
	Height (cm)	,037	-,011	,013	-,006	,049	-,077	,020	,061	-,310	,599	,529	1,000	-,055	-,020	,023
	Legume consum	,037	,090	-,046	,169	,048	,070	-,014	,059	,105	-,049	-,029	-,055	1,000	,242	,109
	Vegetables consum	-,045	-,054	-,049	,017	-,036	,042	-,016	-,002	,111	-,033	-,038	-,020	,242	1,000	,409
	salads	-,049	-,061	,012	-,026	-,013	,055	,001	,043	,056	-,047	,022	,023	,109	,409	1,000
Sig. (1-tailed)	CPK		,000	,000	,000	,000	,269	,485	,381	,008	,000	,494	,079	,083	,045	,033
	CPKMB	,000		,000	,000	,000	,326	,361	,228	,331	,012	,067	,339	,000	,022	,011
	Troponin I	,000	,000		,397	,000	,457	,349	,462	,399	,187	,339	,319	,042	,033	,330
	LDH	,000	,000	,397		,000	,016	,211	,075	,369	,109	,426	,409	,000	,259	,168
	WBC	,000	,000	,000	,000		,000	,001	,064	,276	,001	,270	,032	,036	,089	,318
	ouria	,269	,326	,457	,016	,000		,000	,000	,000	,001	,000	,002	,004	,058	,019
	Creatinin	,485	,361	,349	,211	,001	,000		,008	,001	,037	,154	,228	,294	,279	,489
	Uric acid	,381	,228	,462	,075	,064	,000	,008		,004	,015	,177	,011	,014	,473	,052
	age	,008	,331	,399	,369	,276	,000	,001	,004		,000	,000	,000	,000	,000	,017
	sex	,000	,012	,187	,109	,001	,001	,037	,015	,000		,000	,000	,033	,106	,038
	Weight (Kg)	,494	,067	,339	,426	,270	,000	,154	,177	,000	,000		,000	,141	,076	,208
	Height (cm)	,079	,339	,319	,409	,032	,002	,228	,011	,000	,000	,000		,019	,224	,194
	Legume consum	,083	,000	,042	,000	,036	,004	,294	,014	,000	,033	,141	,019		,000	,000
	Vegetables consum	,045	,022	,033	,259	,089	,058	,279	,473	,000	,106	,076	,224	,000		,000
	salads	,033	,011	,330	,168	,318	,019	,489	,052	,017	,038	,208	,194	,000	,000	

a. Determinant = ,046

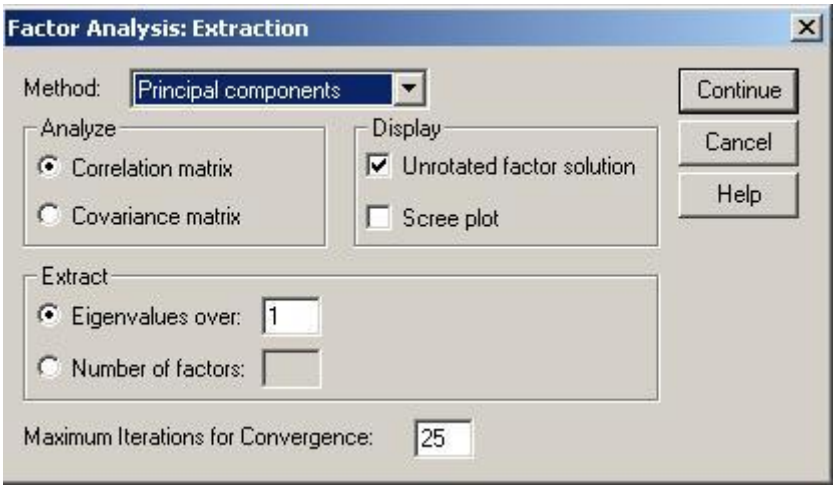
Πίνακας 13.1: Πίνακας συσχετίσεων μεταξύ των αρχικών μεταβλητών

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,653
Bartlett's Test of Sphericity	Approx. Chi-Square	4362,150
	df	105
	Sig.	,000

Πίνακας 13.2: KMO και Bartlett's Test of sphericity: Καταλληλότητα των δεδομένων για την πραγματοποίηση της ανάλυσης σε κύριες συνιστώσες.

iii. Επιλέγοντας το πλήκτρο «***Extraction***» ανοίγει το πλαίσιο διαλόγου που παρουσιάζεται στην *Εικόνα 13.3* και σε αυτό μπορούμε να ορίσουμε τον αριθμό των συνιστωσών που θα κρατήσουμε. Θεωρητικά, δημιουργούνται τόσες συνιστώσες όσες και οι αρχικές μεταβλητές που χρησιμοποιούνται στην ανάλυση, δηλαδή στο συγκεκριμένο παράδειγμα θα δημιουργηθούν 15 συνιστώσες (ΠΡΟΣΟΧΗ: Οι συνιστώσες δεν έχουν μονάδες μέτρησης, γιατί ουσιαστικά προσθέτουμε μεταβλητές με διαφορετικές μονάδες μέτρησης). Όσον αφορά στην επιλογή του αριθμού των συνιστωσών, το SPSS χρησιμοποιεί αυτόματα το κριτήριο του Kaiser. Όμως, δίνει και τη δυνατότητα να επιλέξουμε εμείς τον αριθμό των συνιστωσών που επιθυμούμε, χρησιμοποιώντας είτε το scree plot είτε τον κανόνα της συνολικής μεταβλητότητας των αρχικών μεταβλητών που ερμηνεύουν οι συνιστώσες. Στο παράδειγμά μας χρησιμοποιώντας το κριτήριο του Kaiser προκύπτουν 5 συνιστώσες (*Πίνακας 13.3*).



Εικόνα 13.3: Επιλογή μεθόδου και αριθμού συνιστωσών

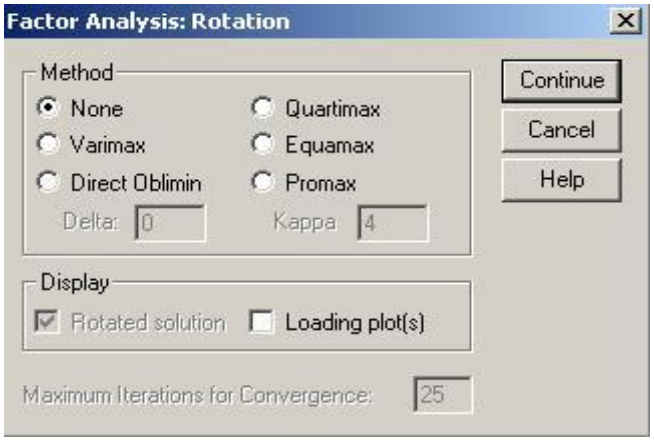
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2,451	16,343	16,343	2,451	16,343	16,343
2	2,242	14,949	31,292	2,242	14,949	31,292
3	1,620	10,801	42,092	1,620	10,801	42,092
4	1,491	9,940	52,033	1,491	9,940	52,033
5	1,044	6,963	58,996	1,044	6,963	58,996
6	,973	6,490	65,486			
7	,883	5,887	71,373			
8	,826	5,508	76,881			
9	,765	5,098	81,978			
10	,677	4,515	86,493			
11	,548	3,652	90,145			
12	,466	3,106	93,251			
13	,440	2,936	96,187			
14	,328	2,190	98,377			
15	,243	1,623	100,000			

Extraction Method: Principal Component Analysis.

Πίνακας 13.3: Ιδιοτιμές κάθε συνιστώσας και ποσοστό συνολικής μεταβλητότητας που ερμηνεύει η κάθε μία.

iv. Επιλέγοντας το πλήκτρο «***Rotation***» ανοίγει το πλαίσιο διαλόγου που φαίνεται στην *Εικόνα 13.4* και μπορούμε να επιλέξουμε τη μέθοδο περιστροφής που θέλουμε. Αν δεν επιλέξουμε καμία, δε θα γίνει περιστροφή και θα πάρουμε απλά τη λύση χωρίς περιστροφή. Πιο συνηθισμένη είναι η ορθογώνια επιλογή «Varimax».



Εικόνα 13.4: Το μενού επιλογών «***Rotation***»

v. Επιλέγοντας το κουμπί επιλογών «***Options***» ανοίγει το πλαίσιο διαλόγου που φαίνεται στην *Εικόνα 13.5* και μπορούμε να επιλέξουμε κάποιες επιλογές σχετικά με το χειρισμό των ελλειπουσών τιμών (***Missing values***) και με τον τρόπο που θέλουμε να εμφανίζονται τα αποτελέσματα των εκτιμημένων συνιστωσών (***Coefficient Display Format***). Πιο συγκεκριμένα:

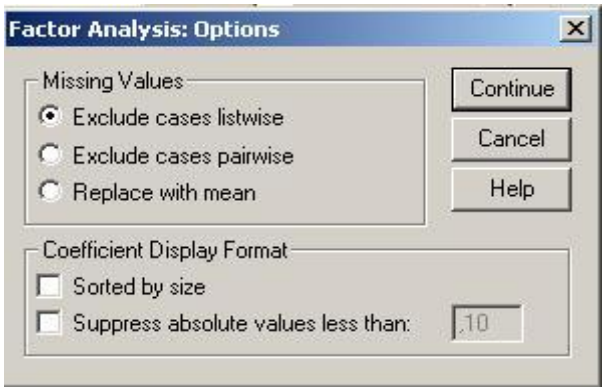
Ελλείπουσες τιμές

- **Exclude cases listwise:** Εξαιρούνται από την ανάλυση όλα τα άτομα που έχουν ελλείπουσες τιμές έστω και σε μία από τις μεταβλητές που συμμετέχουν στην ανάλυση.

- **Exclude cases pairwise:** Εξαιρούνται τα άτομα που έχουν ελλείπουσες τιμές σε μία η και τις 2 μεταβλητές που εμπλέκονται στον υπολογισμό ενός συγκεκριμένου στατιστικού.
- **Replace with mean:** Να αντικατασταθούν οι ελλείπουσες τιμές με τη μέση τιμή της μεταβλητής.

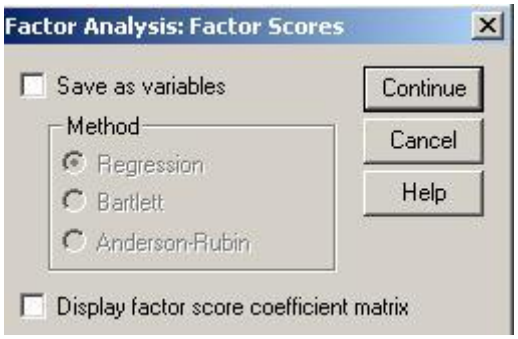
Παρουσίαση των εκτιμούμενων συνιστωσών

- **Sorted by size:** Οι αρχικές μεταβλητές παρουσιάζονται με σειρά απόλυτης τιμής (από τη μεγαλύτερη και συνεπώς την πιο σημαντική για την κάθε συνιστώσα)
- **Suppress absolute values less than:** Δεν εμφανίζονται κάποιοι συντελεστές που είναι μικρότεροι από το όριο που θα δηλώσουμε.



Εικόνα 13.5: Το μενού επιλογών «***Options***»

vi. Από το πλήκτρο «***Scores*»,** μπορούμε να δημιουργήσουμε τόσες νέες μεταβλητές όσες και οι συνιστώσες που επιλέξαμε (Εικόνα 13.6).



Εικόνα 13.6: Αποθήκευση των συνιστωσών που επιλέξαμε

vii. Ερμηνεύουμε τα αποτελέσματα χρησιμοποιώντας τα αποτελέσματα που παρουσιάζονται στον *Πίνακα 13.4*.

Rotated Component Matrix[a]

	Component				
	1	2	3	4	5
CPKMB	,883	-,021	-,016	-,036	,128
CPK	,868	,045	-,045	-,048	,086
LDH	,809	-,003	,046	,040	-,177
WBC	,352	,084	,240	,004	,265
Height (cm)	-,006	,870	,072	,035	,017
sex	,079	,748	,101	-,031	,032
Weight (Kg)	-,034	,730	-,028	,030	-,051
age	-,031	-,544	,313	,155	-,049
ouria	,010	-,171	,830	,045	,025
Creatinin	-,018	,012	,792	-,079	,101
Uric acid	,050	,096	,345	,062	-,171
Vegetables consum	-,031	-,043	-,004	,816	-,069
salads	-,085	,022	,021	,782	,155
Legume consum	,223	-,084	,075	,460	-,407
Troponin I	,139	-,015	-,034	,062	,850

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.

Πίνακας 13.4. Κύριες συνιστώσες που εξάγονται χρησιμοποιώντας το κριτήριο του Kaiser.

Λαμβάνοντας υπόψη, το υποκειμενικό κριτήριο, ότι κάθε συνιστώσα χαρακτηρίζεται κυρίως από εκείνες τις μεταβλητές των οποίων οι συντελεστές είναι μεγαλύτεροι (μεγαλύτερο το 0,3), διαπιστώνουμε ότι οι συνιστώσες ερμηνεύονται ως εξής:

a. Συνιστώσα 1: Βιοχημικοί δείκτες νέκρωσης του μυοκαρδίου
b. Συνιστώσα 2: Δημογραφικά και σωματομετρικά χαρακτηριστικά
c. Συνιστώσα 3: Νεφρική λειτουργία
d. Συνιστώσα 4: Διατροφικές συνήθειες
e. Συνιστώσα 5: Επίπεδα Τροπονίνης Ι στο αίμα.

Παρατηρούμε πως αν και η Τροπονίνη Ι, ως γνωστό, είναι ο πιο ευαίσθητος βιοχημικός δείκτης νέκρωσης του μυοκαρδίου, δεν σχετίζεται αρκετά με την 1η συνιστώσα, όπου την χαρακτηρίζουν όλοι οι υπόλοιποι βιοχημικοί δείκτες νέκρωσης του μυοκαρδίου. Αυτό οφείλεται στο γεγονός ότι η Τροπονίνη Ι δεν σχετίζεται ισχυρά με κάποια από τις υπόλοιπες αρχικές μεταβλητές, όπως φαίνεται και στον πίνακα συσχετίσεων. Συνεπώς, κάποια μεταβλητή που δεν έχει ισχυρές συσχετίσεις με τις υπόλοιπες αρχικές μεταβλητές, δεν έχει νόημα να συμπεριλαμβάνεται στην ανάλυση σε κύριες συνιστώσες, γιατί θα χαρακτηρίζει μόνη της κάποια από αυτές τις συνιστώσες.

14. K-means ανάλυση κατά συστάδες

14.1. Εισαγωγή

Η ανάλυση κατά συστάδες είναι μια πολυμεταβλητή στατιστική μέθοδο η οποία σκοπό έχει να κατατάξει τις υπάρχουσες παρατηρήσεις σε ομάδες χρησιμοποιώντας την πληροφορία που υπάρχει σε διάφορες μεταβλητές. Δύο βασικές έννοιες της συγκεκριμένης μεθόδου είναι η ***έννοια της απόστασης*** και η ***έννοια της ομοιότητας***. Πρόκειται για δύο αντίθετες, μεταξύ τους, έννοιες οι οποίες ουσιαστικά ποσοτικοποιούν αυτό που στην καθομιλουμένη εννοούν. Παρατηρήσεις που μοιάζουν πολύ μεταξύ τους έχουν, δηλαδή, σχετικά όμοιες τιμές, θα έχουν μεγάλες τιμές αν καθορίσουμε κάποιο μέτρο ομοιότητας και μικρές τιμές αν καθορίσουμε κάποιο μέτρο το οποίο θα δηλώνει τη μεταξύ τους απόσταση.

Η ανάλυση κατά συστάδες έχει σκοπό, χρησιμοποιώντας τις προαναφερθείσες έννοιες της ομοιότητας και της απόστασης, να δημιουργήσει ομάδες από όμοιες παρατηρήσεις. Μια επιτυχημένη ανάλυση κατά συστάδες θα έχει δημιουργήσει ομάδες με μεγάλη εσωτερική ομοιογένεια και έντονη εξωτερική ανομοιογένεια. Δηλαδή, οι παρατηρήσεις μέσα στις ομάδες θα είναι όσο το δυνατόν πιο όμοιες μεταξύ τους, την ίδια στιγμή που παρατηρήσεις μεταξύ οποιωνδήποτε δύο διαφορετικών ομάδων θα διαφέρουν πολύ μεταξύ τους.

Η μέθοδος της ανάλυση κατά συστάδες μπορεί να χρησιμοποιηθεί σε πλήθος περιπτώσεων, μερικές από τις οποίες αναφέρονται ακολούθως:

- Για τη διερεύνηση των διαθέσιμων δεδομένων και την προσπάθεια ανίχνευσης τυχόν μεταξύ τους ομοιοτήτων, σχέσεων, μεταβλητών με διακριτική ικανότητα κ.λ.π.
- Για τη μείωση των διαστάσεων του προβλήματος. Σε τεράστιες βάσεις δεδομένων συχνά περιέχεται μικρότερη από την αναμενόμενη πληροφορία, καθώς μπορεί να υπάρχουν επικαλύψεις μεταξύ της πληροφορίας που περιέχεται σε διάφορες μεταβλητές, στοιχεία χωρίς ιδιαίτερο ενδιαφέρον κ.λ.π. Ομαδοποιώντας τα δεδομένα, μας δίνεται η δυνατότητα να εστιάσουμε στις μεταβλητές οι οποίες παίζουν σημαντικό ρόλο στην ερμηνεία τους και να επικεντρωθούμε σ' αυτές.
- Για δημιουργία και έλεγχο υποθέσεων σχετικά με τα δεδομένα μας. Σε πολλές περιπτώσεις είναι πιθανόν να υποψιαζόμαστε μια δομή για τα δεδομένα μας με βάση κάποιο θεωρητικό μοντέλο. Η ανάλυση κατά συστάδες δίνει τη δυνατότητα στον ερευνητή να εξετάσει τις υποθέσεις αυτές ομαδοποιώντας κατάλληλα τα διαθέσιμα δεδομένα.
- Για προβλέψεις νέων τιμών. Έχοντας ήδη ομαδοποιήσει τα δεδομένα μας με την εφαρμογή της ανάλυσης κατά συστάδες, μας δίνεται η δυνατότητα να κατατάξουμε καινούριες διαθέσιμες παρατηρήσεις. Χαρακτηριστικό παράδειγμα της συγκεκριμένης εφαρμογής έχουμε στις τράπεζες οι οποίες κατατάσσουν τους πελάτες τους σε διάφορες κατηγορίες ανάλογα με κάποια διαθέσιμα χαρακτηριστικά και σύμφωνα με τη δυνατότητα που έχουν για αποπληρωμή κάποιου δανείου που έχουν πάρει. Το ενδιαφέρον της τράπεζας εστιάζεται, εκτός των άλλων, και στη σωστή κατάταξη νέων πελατών σε κάποια από τις ομάδες αυτές ώστε να αποφασίσει, με το λιγότερο δυνατό ρίσκο αν θα τους δώσει δάνειο ή όχι.

14.1.1 Η μέθοδος των κ μέσων (K-Means)

Η μέθοδος των κ μέσων ανήκει σε μια μεγάλη κατηγορία αλγορίθμων ομαδοποίησης γνωστών ως αλγόριθμοι διαμέρισης (partitioning algorithms). Πρόκειται για

αλγορίθμους κατάλληλα φτιαγμένους ώστε να διαμερίζουν το πολυεπίπεδο που δημιουργούν πολυμεταβλητά δεδομένα σε περιοχές και να αντιστοιχεί μια περιοχή σε κάθε ομάδα.

14.1.1.1 Ο αλγόριθμος

Η μέθοδος προαπαιτεί ορισμό των ομάδων στις οποίες θέλουμε να χωρίσουμε τα δεδομένα μας. Η μέθοδος των κ μέσων πρόκειται για μια επαναληπτική μέθοδο η οποία χρησιμοποιεί την έννοια του κέντρου της ομάδας (centroid) και κατατάσσει τις παρατηρήσεις ανάλογα με την απόστασή τους από τα κέντρα όλων των ομάδων. Ως κέντρο της ομάδας θεωρεί το διάνυσμα των μέσων τιμών, δηλαδή το διάνυσμα το οποίο αποτελείται από τις μέσες τιμές όλων των μεταβλητών, με βάση τις οποίες γίνεται η ομαδοποίηση. Αρχικά ορίζονται τα κέντρα των κ ομάδων που θέλουμε να δημιουργηθούν από την εφαρμογή της συγκεκριμένης μεθόδου. Ως αρχικά κέντρα, των ομάδων που θέλουμε να προκύψουν, λαμβάνονται από τον αλγόριθμο ισάριθμες, με τον αριθμό των ομάδων, παρατηρήσεις οι οποίες βρίσκονται όσο το δυνατόν σε μεγαλύτερη απόσταση μεταξύ τους.

Για κάθε παρατήρηση υπολογίζεται η απόστασή της (συνήθως ευκλείδεια) από τα κέντρα των ομάδων και κατατάσσεται στην ομάδα με την οποία είναι πιο κοντά. Αφού ομαδοποιηθούν με τον τρόπο αυτό όλες οι παρατηρήσεις τότε υπολογίζονται εκ νέου τα κέντρα των ομάδων ως τα διανύσματα των μέσων των μεταβλητών για τις παρατηρήσεις που ανήκουν στην κάθε ομάδα. Με βάση τα νέα αυτά κέντρα επαναλαμβάνουμε τη μέθοδο και ομαδοποιούμε τις παρατηρήσεις εκ νέου με κριτήριο πάλι την απόστασή τους από τα νέα κέντρα των ομάδων που δημιουργήθηκαν από το προηγούμενο βήμα. Η διαδικασία αυτή επαναλαμβάνεται έως ότου τα νέα κέντρα που προκύπτουν να μη διαφέρουν από τα κέντρα της τελευταίας εφαρμογής της μεθόδου. Αυτό δηλαδή που προσπαθεί να επιτευχθεί μέσω του συγκεκριμένου αλγορίθμου είναι μείωση των τετραγωνικών αποστάσεων των παρατηρήσεων από τα κέντρα των ομάδων που ανήκουν.

Όπως ήδη αναφέρθηκε, συνήθως χρησιμοποιείται η ευκλείδεια απόσταση κατά τη διαδικασία ομαδοποίησης των παρατηρήσεων. Ωστόσο είναι δυνατόν, για το σκοπό αυτό, να χρησιμοποιηθεί και οποιαδήποτε άλλη απόσταση επιθυμούμε αρκεί πρώτα να κάνουμε κατάλληλους μετασχηματισμούς των δεδομένων μας. Για μη συνεχή δεδομένα υπάρχει το πρόβλημα ότι δεν είναι δυνατόν να ορίσουμε το μέσο της ομάδας, μπορούμε ωστόσο να χρησιμοποιήσουμε εναλλακτικά κάποιο άλλο αντίστοιχο αντιπροσωπευτικό μέτρο για την ομάδα. Έτσι, για παράδειγμα, σε κατηγορικά δεδομένα σε διατεταγμένη κλίμακα (ordinal scale) μπορούμε να χρησιμοποιήσουμε το διάνυσμα των διαμέσων (metroid) ως κέντρο κάθε ομάδας. Αντίστοιχα, για δεδομένα σε ονομαστική κλίμακα (nominal scale) είναι δυνατόν να χρησιμοποιηθεί το διάνυσμα των κορυφών, δηλαδή οι τιμές με τη μεγαλύτερη συχνότητα. Βέβαια όλα αυτά τα μέτρα είναι πολύ κατώτερα ως επιλογή λόγω των ιδιοτήτων τους, ωστόσο μας δίνουν τη δυνατότητα χρήσης του αλγορίθμου σε κάθε τύπου δεδομένα. Στην περίπτωση μικτού τύπου δεδομένων το κέντρο κάθε ομάδας μπορεί να προσδιοριστεί από τις κορυφές των ονομαστικών μεταβλητών, τις διαμέσους των διατεταγμένων και τις μέσες τιμές των συνεχών μεταβλητών. κατώτερα

Συνοπτικά τα ***βήματα*** της μεθόδου είναι τα εξής:

$1^{ο}$ Βήμα: Ορισμός των κ αρχικών κέντρων των ομάδων.

2ο Βήμα: Ομαδοποίηση των παρατηρήσεων στις κ αρχικές ομάδες με κριτήριο την απόστασή τους από τα κέντρα των ομάδων, όπως αυτά ορίστηκαν στο 1ο βήμα.
3ο Βήμα: Υπολογισμός των νέων κέντρων των ομάδων με βάση τις μέσες τιμές των μεταβλητών για τις παρατηρήσεις οι οποίες ανήκουν σε κάθε ομάδα.
4ο Βήμα: Αν τα νέα κέντρα που προέκυψαν δε διαφέρουν από τα κέντρα που ορίστηκαν στο 1ο βήμα, ο αλγόριθμος τερματίζεται και η ομαδοποίηση των παρατηρήσεων είναι τελική. Διαφορετικά το 2ο και 3ο βήμα του αλγορίθμου επαναλαμβάνονται έως ότου τα νέα κέντρα των ομάδων που θα προκύψουν στο βήμα 4 να μη διαφέρουν από αυτά που ορίστηκαν στη αρχή της επανάληψης.

14.1.1.2 Αξιολόγηση της μεθόδου (Μειονεκτήματα - Πλεονεκτήματα)

Ένα από τα βασικά πλεονεκτήματα του συγκεκριμένου αλγορίθμου είναι ότι δουλεύει ικανοποιητικά για μεγάλα σετ δεδομένων. Όταν διαθέτουμε μεγάλες βάσεις δεδομένων δουλεύει, για παράδειγμα, πολύ πιο γρήγορα από την ιεραρχική ομαδοποίηση.

Επιπλέον ο συγκεκριμένος αλγόριθμος είναι ιδιαίτερα χρήσιμος καθώς δε χρειάζεται να κρατά στη μνήμη πολλά στοιχεία και συνεπώς δεν απαιτεί πολύ μεγάλη υπολογιστική ισχύ. Στο σημείο αυτό θα πρέπει να τονισθεί πως η δυναμική του αλγορίθμου είναι ότι με τις πρώτες λίγες επαναλήψεις πλησιάζει πολύ κοντά στην τελική λύση. Οι υπόλοιπες επαναλήψεις χρειάζονται λόγω διαφορών οι οποίες οφείλονται σε μικρή μετακίνηση κάποιων λίγων παρατηρήσεων που βρίσκονται ουσιαστικά ανάμεσα σε δύο ομάδες. Συνεπώς δεν είναι απαραίτητος ένας μεγάλος αριθμός επαναλήψεων καθώς η τελική δομή θα σχηματιστεί πολύ γρήγορα. Πολύ σημαντικό, και επιθυμητό πολλές φορές, είναι επίσης το γεγονός ότι από την εφαρμογή της μεθόδου των κ μέσων προκύπτουν ομάδες με παρόμοιο αριθμό παρατηρήσεων.

Ένα από τα βασικά μειονεκτήματα της μεθόδου είναι ότι ο αριθμός των ομάδων είναι προκαθορισμένος. Υπάρχουν διάφοροι τρόποι με τους οποίους μπορούμε να αποφασίσουμε σε πόσες ομάδες θέλουμε να χωρίσουμε τα δεδομένα μας. Μία επιλογή είναι να τρέξουμε τον αλγόριθμο αρκετές φορές με διαφορετικό κάθε φορά αριθμό ομάδων και να συγκρίνουμε τα αποτελέσματα. Εναλλακτικά θα μπορούσε να χρησιμοποιηθεί πιθανή διαθέσιμη πρότερη γνώση για το αντικείμενο που μελετούμε, με βάση την οποία είναι γνωστός εκ των προτέρων ο αριθμός των ομάδων που θέλουμε να προκύψουν. Στο σημείο αυτό μπαίνει πολλές φορές ο υποκειμενικός παράγοντας. Έχοντας επαναλάβει τον αλγόριθμο αρκετές φορές ο ερευνητής μπορεί να εξετάσει τα αποτελέσματα και να συνενώσει ομάδες που θεωρεί ότι δε διαφέρουν σημαντικά (όχι απαραίτητα με κάποιο στατιστικό έλεγχο) μεταξύ τους. Μπορεί ακόμα να χρησιμοποιήσει τη διαίσθησή του καθορίζοντας εκ των προτέρων τον αριθμό των ομάδων με βάση το πώς περίπου περιμένει να ομαδοποιηθούν τα δεδομένα του.

Σημαντικό επίσης μειονέκτημα της μεθόδου των κ μέσων αποτελεί το γεγονός ότι ο αλγόριθμος εξαρτάται σε μεγάλο βαθμό από τις αρχικές τιμές οι οποίες θεωρούνται ως μέσα των ομάδων και αν δεν οριστούν καλά είναι δυνατόν να οδηγήσουν σε τελείως διαφορετική ομαδοποίηση. Για να ξεπεραστεί το πρόβλημα αυτό μια λύση είναι να επαναλαμβάνουμε το συγκεκριμένο αλγόριθμο με διαφορετικές κάθε φορά αρχικές τιμές, ώστε να είμαστε σίγουροι ότι δεν παγιδεύτηκε σε κάποια βέλτιστη λύση.

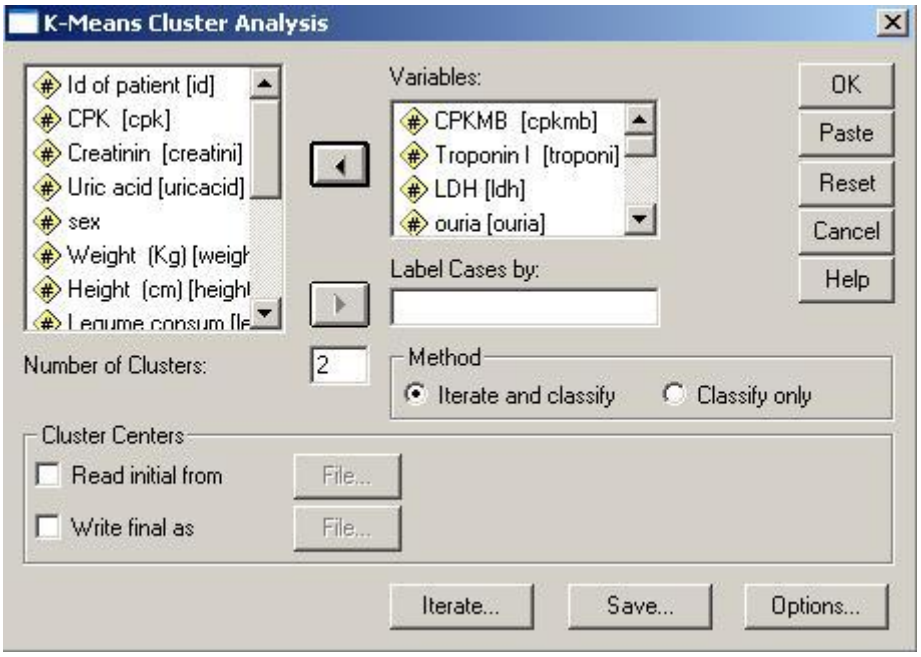
14.2 Ανάλυση κατά συστάδες με τη χρήση του SPSS

Ας υποθέσουμε πως επιθυμούμε να ομαδοποιήσουμε τους ασθενείς που έχουν εισαχθεί στο νοσοκομείο με Οξύ Στεφανιαίο Σύνδρομο χρησιμοποιώντας τα επίπεδα CPKMB, Τροπονίνης, LDH, λευκών αιμοσφαιρίων, ουρίας και την ηλικία των ασθενών την ώρα της εισαγωγής τους.
Η συσταδική ανάλυση πραγματοποιείται ακολουθώντας τα εξής ***βήματα***.

Analyse → Classify → K-means Cluster

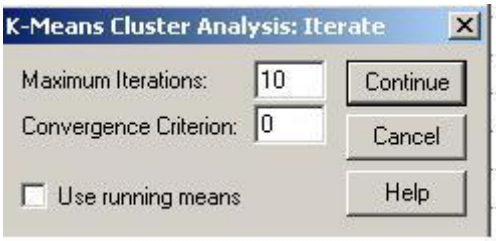
Στη συνέχεια ανοίγει ένα πλαίσιο διαλόγου (*Εικόνα 14.1*), στο οποίο πρέπει να δηλώσουμε:

i. τις μεταβλητές βάση των οποίων θα πραγματοποιηθεί η συσταδική ανάλυση (***Variables***) και τον αριθμό των ομάδων που θέλουμε να σχηματιστούν (***Number of Clusters***). Επίσης, υπάρχουν ακόμα τρεις επιλογές (**Iterate, Save** και **Options**), που ενεργοποιούν τα παράθυρα που θα δούμε παρακάτω. Τα αρχικά κέντρα των ομάδων μπορούν είτε να τεθούν από τον χρήστη είτε να βρεθούν από το πακέτο με βάση κάποιον αλγόριθμο.



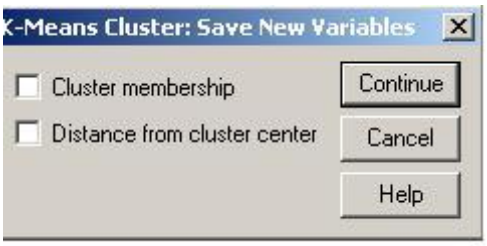
Εικόνα 14.1: Το βασικό παράθυρο της K-means ομαδοποίησης

ii. Πατώντας το κουμπί επιλογών «**Iterate**» ανοίγει το πλαίσιο διαλόγου που φαίνεται στην *Εικόνα 14.2* και επιλέγουμε τα κριτήρια τερματισμού του αλγορίθμου. Μπορούμε να επιλέξουμε να σταματήσει έπειτα από συγκεκριμένο αριθμό επαναλήψεων (10 στην περίπτωσή μας) ($1^{ο}$ κριτήριο) είτε όταν η μεγαλύτερη απόσταση ανάμεσα σε διαδοχικά κέντρα όλων των ομάδων γίνει 0 ($2^{ο}$ κριτήριο). Το $2^{ο}$ κριτήριο αντιστοιχεί στην περίπτωση που οι ομάδες δεν αλλάξουν καθόλου μετά από μία επανάληψη. Αν το $2^{ο}$ κριτήριο επιτευχθεί πριν το $1^{ο}$ τότε η επαναληπτική διαδικασία σταματά πριν ολοκληρωθούν όλες οι επαναλήψεις, διαφορετικά σταματά όταν πραγματοποιηθεί ο μέγιστος αριθμός επαναλήψεων.



Εικόνα 14.2: Το παράθυρο για τα κριτήρια τερματισμού του επαναληπτικού αλγόριθμου.

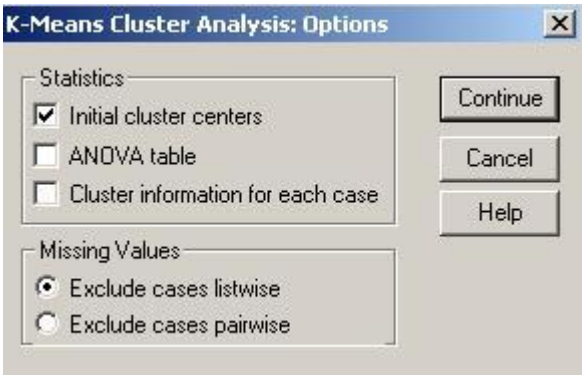
iii. Πατώντας το κουμπί επιλογών «***Save***» ανοίγει το πλαίσιο διαλόγου της *Εικόνας 14.3* και επιλέγουμε τις καινούριες μεταβλητές που θέλουμε να δημιουργήσουμε. Η επιλογή ***Cluster membership*** θα μας δημιουργήσει μια καινούρια στήλη όπου σε κάθε παρατήρηση θα δίνεται η τιμή της ομάδας που την κατατάξαμε. Αυτή η μεταβλητή είναι ιδιαίτερα χρήσιμη, όταν μετά την ανάλυση προσπαθήσουμε να δούμε τα χαρακτηριστικά κάθε ομάδας, αλλά και το αν κάποιες μεταβλητές προσφέρουν πληροφορία σχετικά με την ομαδοποίηση που κάναμε.



Εικόνα 14.3: Το παράθυρο για τη δημιουργία νέων μεταβλητών με τα αποτελέσματα της ανάλυσης (κουμπί «Save»).

iv. Τέλος, πατώντας το κουμπί επιλογών «***Options***» ανοίγει το παράθυρο της *Εικόνας 14.4* και μπορούμε να επιλέξουμε ποια αποτελέσματα θα εμφανιστούν στην οθόνη. Μπορούμε να επιλέξουμε να μας δώσει ο υπολογιστής:

- ***Initial cluster centers:*** τα αρχικά κέντρα των ομάδων, η χρησιμότητά των οποίων έγκειται στο να ελέγξουμε κατά πόσο αλλάζουν τα αποτελέσματά σε διαδοχικές εκτελέσεις του αλγορίθμου,
- ***Cluster information for each case:*** τα τελικά κέντρα των ομάδων και τις Ευκλείδειες αποστάσεις μεταξύ των ατόμων και των κέντρων των συστάδων που χρησιμοποιήθηκαν για την ταξινόμηση του κάθε ατόμου.
- ***ANOVA table:*** τους πίνακες ανάλυσης διακύμανσης για τις μεταβλητές που χρησιμοποιήσαμε, ώστε να βρούμε ποιες περιέχουν όντως πληροφορία για την ομαδοποίηση που κάναμε.



Εικόνα 14.4: Διάφορες άλλες επιλογές (κουμπί «Options»)

Για το παράδειγμά μας, τα αρχικά και τελικά κέντρα των ομάδων (3 ομάδες), καθώς ο πίνακας με την ανάλυση διακύμανσης για όλες τις αρχικές μεταβλητές και ο αριθμός των ατόμων που συμπεριλαμβάνονται σε κάθε τελική ομάδα, φαίνονται στους *Πίνακες 14.1, 14.2 και 14.3*, αντίστοιχα.

Initial Cluster Centers

	Cluster		
	1	2	3
CPKMB	22	1	40
Troponin I	15,20	2,00	11,50
LDH	148	168	334
WBC	1040	48100	26800
ouria	42	35	32
age	79	76	70

Πίνακας 14.1: Τα αρχικά κέντρα των 3 ομάδων.

Final Cluster Centers

	Cluster		
	1	2	3
CPKMB	32	5	70
Troponin I	8,22	1,00	39,21
LDH	351	175	533
WBC	7579	41720	13081
ouria	45	54	51
age	66	74	66

Πίνακας 14.2: Τα τελικά κέντρα των 3 ομάδων

Από τον *Πίνακα 14.2* παρατηρούμε ότι η τρίτη ομάδα χαρακτηρίζεται από υψηλές τιμές των δεικτών σχετικών με τη νέκρωση του μυοκαρδίου, η πρώτη ομάδα από χαμηλότερες τιμές όλων αυτών των δεικτών συγκριτικά με την τρίτη ομάδα, ενώ η δεύτερη ομάδα χαρακτηρίζεται από ακραίες υψηλές ή ακραίες χαμηλές τιμές των διαφόρων δεικτών. Συνεπώς, η τρίτη ομάδα συμπεριλαμβάνει τους ασθενείς που πάσχουν από βαρύ στεφανιαίο σύνδρομο, η πρώτη, αυτούς που πάσχουν από ελαφριά μορφή στεφανιαίας νόσου και η δεύτερη αυτούς που έχουν ακραίες τιμές σχεδόν σε όλες τις μεταβλητές που συμμετάσχουν στην ανάλυση. Λαμβάνοντας υπόψη και τον μικρό αριθμό των ατόμων που ανήκουν στη δεύτερη ομάδα (*Πίνακας 14.3*), πιθανολογούμε πως πρόκειται για περιπτώσεις λάθος καταγραφής των τιμών κάποιων ή όλων των μεταβλητών.

Number of Cases in each Cluster

Cluster	1	1387,000
	2	2,000
	3	539,000
Valid		1928,000
Missing		244,000

Πίνακας 14.3: Αριθμός παρατηρήσεων που περιέχονται σε κάθε ομάδα τελικά.

Από τον *Πίνακα 143.4* προκύπτει ότι όλες οι μεταβλητές εκτός από την ηλικία συνεισφέρουν στατιστικά σημαντικά στο διαχωρισμό των ατόμων σε συστάδες.

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
CPKMB	282686,922	2	5458,279	1925	51,790	,000
Troponin I	186663,623	2	15516,179	1925	12,030	,000
LDH	6514680,512	2	143791,524	1925	45,306	,000
WBC	6936866508	2	3897879,760	1925	1779,651	,000
ouria	5814,222	2	694,032	1925	8,377	,000
age	66,525	2	169,912	1925	,392	,676

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Πίνακας 14.4: Ο πίνακας που περιέχει την ανάλυση διακύμανσης για το αν διαφέρουν οι μέσες τιμές των μεταβλητών που χρησιμοποιήθηκαν στην ανάλυσή μας ανάμεσα στις ομάδες.

15. Διαχωριστική Ανάλυση

15.1. Εισαγωγή

Η βασική ιδέα της διαχωριστικής ανάλυσης (διακριτική ή διακρινούσα ανάλυση κατά άλλους και discriminant analysis στα αγγλικά) είναι να ταξινομήσει τις παρατηρήσεις στις κατηγορίες μιας κατηγορικής μεταβλητής (***criterion variable***) με βάση μια σειρά άλλων χαρακτηριστικών (***discriminating variables***). Συνεπώς, η διαχωριστική ανάλυση αποτελεί μία μέθοδο με πλήθος εφαρμογών σε πολλές επιστήμες. Ένα παράδειγμα εφαρμογών της στην Ιατρική είναι:

- Διάγνωση της ασθένεια κάποιου ασθενή με βάση τα συμπτώματα που έχει αυτός. Δεδομένου πως τα συμπτώματα κάθε ασθένειας είναι γνωστά, σκοπός μας είναι η δημιουργία ενός κανόνα που θα κάνει τη διάγνωση για έναν καινούριο ασθενή, λαμβάνοντας υπόψη τα συμπτώματά της ασθένειας αλλά και τη γνώση μας για τα συμπτώματα ενός συνόλου ασθενειών.

Επιπλέον, η διαχωριστική ανάλυση συμβάλει στο να διακρίνουμε ποια χαρακτηριστικά είναι αυτά που συμβάλλουν περισσότερο στην ταξινόμηση των παρατηρήσεων μεταξύ των κατηγοριών μιας μεταβλητής.

Οφείλει να παρατηρήσει κανείς ότι η κατάταξη γίνεται είτε σε δύο (παράδειγμα τράπεζας) είτε σε περισσότερες ομάδες (παράδειγμα διάγνωσης), και επίσης, πρέπει να υπογραμμιστεί πως αν και η διαχωριστική ανάλυση μοιάζει με την ανάλυση σε συστάδες, υπάρχουν σημαντικές ***διαφορές*** μεταξύ τους, όπως:

- Στη διαχωριστική ανάλυση, οι ομάδες είναι γνωστές, ενώ στην ανάλυση σε ομάδες δεν είναι και σκοπός μας να βρούμε αυτές τις ομάδες.
- Στη διαχωριστική ανάλυση κύριο μέλημα μας είναι η κατασκευή ενός κανόνα που θα μας βοηθήσει να λάβουμε αποφάσεις στο μέλλον, ενώ στην ανάλυση κατά συστάδες ο κύριος στόχος μας είναι να δημιουργήσουμε ομοειδείς ομάδες με σκοπό την κατανόηση των ήδη υπαρχόντων στοιχείων και τη μείωση της διασποράς σε επιμέρους ομάδες.

15.1.1 Βασικές έννοιες ης διαχωριστικής ανάλυσης

Διαχωριστική συνάρτηση (Discriminant function): Μία διαχωριστική συνάρτηση είναι μία ψευδομεταβλητή που δημιουργείται ως ένας γραμμικός συνδυασμός των ανεξάρτητων μεταβλητών, όπως

$$L = b_1x_1 + b_2x_2 + ... + b_nx_n + c, \quad (1)$$

Όπου,
τα b's είναι συντελεστές (discriminant coefficients),
τα x's είναι οι ανεξάρτητες μεταβλητές, και
η c είναι μία σταθερά.
Δηλαδή, η διαχωριστική συνάρτηση είναι κάτι ανάλογο με την πολλαπλή παλινδρόμηση, αλλά τα b's είναι συντελεστές που μεγιστοποιούν την απόσταση μεταξύ των μέσων τιμών των ανεξάρτητων μεταβλητών στις κατηγορίες της κατηγορικής μεταβλητής ή αλλιώς μεγιστοποιούν την διαφορά μεταξύ των μέσων τιμών του σκορ που υπολογίζεται από την παραπάνω συνάρτηση. Ο αριθμός των διαχωριστικών συναρτήσεων που δημιουργούνται είναι μόνο μία αν ο διαχωρισμός

πρόκειται να γίνει μεταξύ 2 ομάδων (κατηγορική με 2 κατηγορίες). Αν όμως, ο διαχωρισμός πρόκειται να γίνει μεταξύ περισσότερων κατηγοριών, ο αριθμός των διαχωριστικών συναρτήσεων που δημιουργούνται είναι ή (g - 1), όπου g είναι ο αριθμός των κατηγοριών της κατηγορικής μεταβλητής, ή p: ο αριθμός των ανεξάρτητων μεταβλητών. Πιο συγκεκριμένα, αν g-1<p τότε θα δημιουργηθούν g-1 συναρτήσεις, αλλιώς θα δημιουργηθούν p συναρτήσεις. Στην περίπτωση που δημιουργούνται περισσότερες από 1 συναρτήσεις, όλες είναι ασυσχέτιστες μεταξύ τους. Η πρώτη συνάρτηση έχει την πιο ισχυρή διαχωριστική ικανότητα.

Βέβεια, εκτός από την παραπάνω διαχωριστική συνάρτηση, μπορεί να υπολογιστεί και η τυποποιημένη διαχωριστική συνάρτηση η οποία προκύπτει αντικαθιστώντας τους συντελεστές (b_i) της παραπάνω συνάρτησης με τους αντίστοιχους τυποποιημένους ($beta_i$) και τις ανεξάρτητες μεταβλητές (x_i) με τις αντίστοιχες τυποποιημένες (z_i). Οπότε η συγκεκριμένη διαχωριστική συνάρτηση παίρνει τη μορφή:

$$D = beta_1z_1 + beta_2z_2 + ... + beta_nz_n \ , \ (2)$$

Το **διαχωριστικό σκορ** (**discriminant score)**, είναι η τιμή της διαχωριστικής συνάρτησης για κάθε ένα άτομο που συμμετέχει στην μελέτη. Αν το διαχωριστικό σκορ της συνάρτησης είναι μικρότερο ή ίσο με το κατώφλι που ορίζουμε, τότε το άτομο ταξινομείται στην ομάδα «0», ενώ αν είναι μεγαλύτερο, ταξινομείται στην ομάδα «1». Αυτό το κατώφλι είναι ίσο με το μέσο όρο των 2 κεντροειδών (μέση τιμή του διαχωριστικού σκορ σε κάθε κατηγορία της εξαρτημένης μεταβλητής) αν ο αριθμός των ατόμων και στι 2 ομάδες της εξαρτημένης μεταβλητής είναι ίδιος, ενώ με τον σταθμισμένο μέσο όρο αν δεν είναι ίδιος ο αριθμός των ατόμων. Το *Z score* είναι το διαχωριστικό σκορ που προκύπτει από τα τυποποιημένα δεδομένα. Η μέση τιμή του *Z score* είναι μηδέν επειδή η μέση τιμή όλων των ανεξάρτητων μεταβλητών είναι μηδέν όταν τυποποιηθούν. Έτσι, ένα άτομο μπορεί να ταξινομηθεί στην ομάδα «1» αν το *Z score* είναι μεγαλύτερο από το 0 και στην ομάδα «0» αν είναι μικρότερο από το μηδέν.

Συναρτήσεις ταξινόμησης (Classification functions): Οι συναρτήσεις ταξινόμησης μοιάζουν με τις διαχωριστικές συναρτήσεις, με τη διαφορά ότι δημιουργούνται τόσες συναρτήσεις ταξινόμησης όσες και οι κατηγορίες της κατηγορικής εξαρτημένης μεταβλητής. Οι πιο γνωστές συναρτήσεις ταξινόμησεις είναι αυτές του Fisher οι συντελεστές των οποίων προκύπτουν από τους μη τυποποιημένους συντελεστές της διαχωρριστικής συνάρτησης. Το κάθε άτομο ταξινομείται στην κατηγορία όπου προκύπτει το υψηλότερο σκορ.

Το **canonical correlation, R***, είναι ένα μέτρο της σχέσης ανάμεσα στο σκορ της διαχωριστικής συνάρτησης και την εξαρτημένη κατηγορική μεταβλητή. Όταν το R* είναι μηδέν, δεν υπάρχει καμία συσχέτιση ανάμεσα στην εξαρτημένη μεταβλητή και τη συνάρτηση. Όταν το canonical correlation είναι υψηλό, υπάρχει μία υψηλή συσχέτιση ανάμεσα στην εξαρτημένη μεταβλητή και τη συνάρτηση. R* είναι χρήσιμο στο να μας πει πόσο πολύ κάθε συνάρτηση είναι χρήσιμη στο να προσδιορίζει διαφορές μεταξύ των ομάδων. Αν R* είναι 1.0 υποδηλώνει ότι όλη η μεταβλητότητα του διαχωριστικού σκορ ερμηνεύεται από την εξαρτημένη κατηγορική μεταβλητή. Εδώ αξίζει να σημειωθεί ότι όταν η εξαρτημένη μεταβλητή αποτελείται από 2 ομάδες μόνο, το canonical correlation είναι ίσο με το συντελεστή συσχέτισης του Pearson ανάμεσα στο διαχωριστικό σκορ και την εξαρτημένη μεταβλητή.

Μη τυποποιημένη συντελεστές (Unstandardized discriminant coefficients) είναι οι συντελσετές της συνάρτησης 1. Είναι κάτι αντίστοιχο με τους partial coefficients, και αντανακλούν την μοναδική συνεισφορά κάθε ανεξάρτητης μεταβλητής στην ταξινόμηση των ατόμων μεταξύ των καητγοριών της εξαρτημένης μεταβλητής. Οι **τυποποιημένοι συντελεστές (*standardized discriminant coefficients)***, είναι κα΄τγι αντίστοιχο με τα beta weights στην παλινδρόμηση, και χρησιμοποιούνται για να αξιολογήσουμε την σχετική συνεισφορά της κάθε ανεξάρτητης μεταβλητής στην διαδικασία της ταξινόμησης. Ουσιαστικά, λοιπόν, οι τυποποιημένοι συντελεστές είναι χρήσιμοι στο να κατατάξουμε τις ανεξάρτητες μεταβλητές από αυτήν που ασκεί την μεγαλύτερη επίδραση στην διαδικασία της ταξινόμησης (πιο σημαντική) προς την λιγότερο σημαντική. Γι' αυτό το σκοπό χρησιμοποιούμε τις απόλυτες τιμές αυτών των συντελσετών. Όμως, αν η εξαρτημένη μεταβλητή αποτελείται από περισσότερες από 2 κατηγορίες, αυτοί οι συντελεστές δεν είναι αρκετοί για να δείξουν μεταξύ ποιων κατηγοριών οι μεταβλητές έχουν μεγαλύτερη διαχωριστική ικανότητα. Σε αυτή την περίπτωση, εξετάζονται τα κεντρωειδή.

Structure coefficients: καλούνται και **structure correlations** ή **discriminant loadings** και είναι ο συντελεστής συσχέτισης ανάμεσα στην ανεξάρτητη μεταβλητή και το διαχωριστικό σκορ.

15.1.2 Έλεγχοι υποθέσεων

Έλεγχος για την συνεισφορά ολόκληρου του μοντέλου: Ουσιαστικά, στην συγκεκριμένη περίπτωση αυτό που ελέγχουμε είναι αν η μέση τιμή του διαχωριστικού σκορ είναι ίση στις κατηγορίες της εξαρτημένης μεταβλητής, οπότε συμπεραίνουμε ότι το μοντέλο μας δεν έχει καμία διαχωριστική ικανότητα. Η μηδενική και εναλλακτική υπόθεση, λοιπόν, διατυπώνεται ως εξής:

H_0: Οι μέσες τιμές του διαχωριστικού σκορ είναι ίσες στις κατηγορίες της εξαρτημένης μεταβλητής και συνεπώς το μοντέλο μας ΔΕΝ έχει καλή διαχωριστική ικανότητα.

H_1: Οι μέσες τιμές του διαχωριστικού σκορ ΔΕΝ είναι ίσες μεταξύ των κατηγοριών της εξαρτημένης μεταβλητής και συνεπώς το μοντέλο μας ΕΧΕΙ καλή διαχωριστική ικανότητα.

Για να ελεγχθεί η παραπάνω υπόθεση, υπολογίζεται το **Wilks' lambda** του μοντέλου. Όσο πιο μεγάλο είναι το λάμδα, τόσο πιο πιθανό είναι να είναι σημαντική η διαχωριστική ικανότητα του μοντέλου μας.

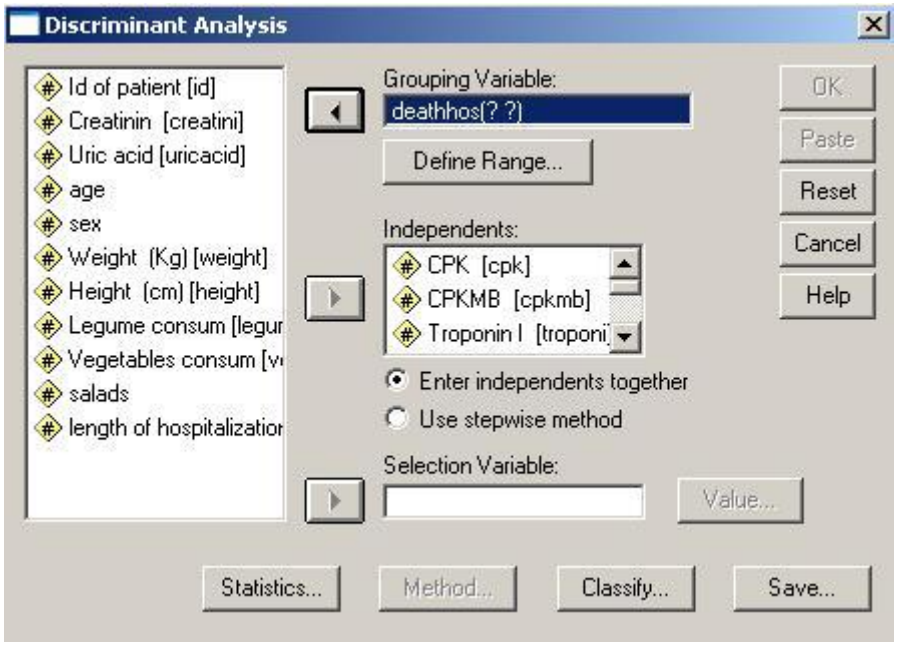
Έλεγχος για την συνεισφορά κάθε ανεξάρτητης μεταβλητής: Υπολογίζεται το **Wilks' lambda** για κάθε μεταβλητή, το οποίο υποδηλώνει ποια ανεξάρτητη μεταβλητή συνεισφέρει σημαντικά στην διαχωριστική συνάρτηση. Όσο μικρότερο είναι το Wilks' lambda για μία ανεξάρτητη μεταβλητή, τόσο περισσότερο αυτή η συγκεκριμένη ανεξάρτητη μεταβλητή συνεισφέρει στη διαχωριστική συνάρτηση. Το λάμδα του Wilks' διακυμαίνεται μεταξύ 0 και 1, με το 0 να σημαίνει ότι οι μέσες τιμές της συγκεκριμένης μεταβλητής διαφέρουν μεταξύ των κατηγοριών, και το 1 να σημαίνει ότι δεν διαφέρουν. Το F test του λάμδα του Wilks's δείχνει ποιων ανεξάρτητων μεταβλητών η συνεισφορά είναι σημαντική.

15.2 Διαχωριστική ανάλυση με τη χρήση του SPSS

Ας υποθέσουμε ότι σκοπός μας είναι να φτιάξουμε μία συνάρτηση χρησιμοποιώντας μια σειρά από αιματολογικά, βιοχημικά και κοινωνικό-δημογραφικά χαρακτηριστικά που συλλέγονται κατά την εισαγωγή των ασθενών με Οξύ Στεφανιαίο Σύνδρομο, με τη βοήθεια της οποίας να διαχωρίσουμε τους ασθενείς (τη στιγμή της εισαγωγής τους στο νοσοκομείο) σε αυτούς που θα έχουν αυξημένη πιθανότητα να πεθάνουν κατά τη διάρκεια της νοσηλείας τους και σε αυτούς που θα έχουν αυξημένη πιθανότητα να επιβιώσουν. Συγκεκριμένα, θα χρησιμοποιήσουμε τα εξής στοιχεία: ηλικία, φύλο, βάρος, ύψος, επίπεδα κρεατινίνης, ουρίας, ουρικού οξέος, λευκών αιμοσφαιρίων & LDH. Αυτό μπορεί να επιτευχθεί πραγματοποιώντας διαχωριστική ανάλυση ακολουθώντας τα εξής ***βήματα***:

Analyse → **Classify** → **Discriminant.**

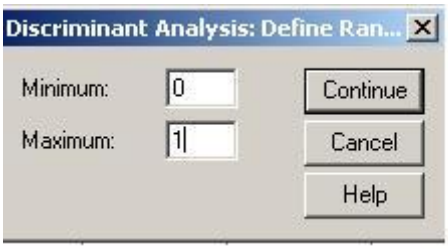
Από την παραπάνω διαδρομή προκύπτει το παράθυρο διαλόγου που φαίνεται στην *Εικόνα 15.1*.



Εικόνα 15.1. Πραγματοποίηση διαχωριστικής ανάλυσης

Σε αυτό το παράθυρο διαλόγου δηλώνουμε τα εξής:

i. **Grouping Variable:** Η μεταβλητή που καθορίζει τις ομάδες. Στο παράδειγμά μας τοποθετούμε τη μεταβλητή «deathhos», η οποία μας υποδεικνύει εάν ο κάθε ασθενής του δείγματος μας έχει πεθάνει ή όχι. Επιπλέον πρέπει να ορίσουμε το εύρος των ομάδων στην επιλογή «***Define Range***». Πατώντας αυτό το κουμπί ανοίγει το πλαίσιο διαλόγου της *Εικόνας 15.2*, όπου ορίζουμε σαν ελάχιστη τιμή (minimum) το 0 και μέγιστη το 1 και πατάμε «***Continue***».



Εικόνα 15.2: Προσδιορισμός των ομάδων

ii. **Independents:** Εδώ τοποθετούμε τις ανεξάρτητες μεταβλητές μας με βάση τις οποίες θα γίνει η ταξινόμηση της κάθε παρατήρησης. Στο παράδειγμά μας οι μεταβλητές αυτές είναι: LDH, WBC, ouria, Creatinine, uric acid, age, sex, weight, height.

iii. Οι υπόλοιπες επιλογές περιλαμβάνουν τη χρήση όλων των ανεξάρτητων μεταβλητών (***Enter independent together***) ή εναλλακτικά τη χρήση κλιμακωτών μεθόδων επιλογής των ανεξάρτητων μεταβλητών (***Use stepwise Method***). Η δεύτερη επιλογή είναι πολύ χρήσιμη στην πράξη και εντοπίζει βήμα-βήμα τις ασήμαντες μεταβλητές για το διαχωρισμό και τις αφαιρεί από τη διαχωριστική συνάρτηση.

iv. Πατώντας το κουμπί επιλογών «***Statistics***» ανοίγει ένα νέο παράθυρο διαλόγου (*Εικόνα 15.3*) από το οποίο μπορούμε να επιλέξουμε να εμφανιστούν στο output του SPSS, κάποιοι περιγραφικοί δείκτες (***Descriptives***), συντελεστές της διαχωριστικής συνάρτησης (***Function Coefficients***) και Μήτρες-Πίνακες (***Matrices***). Οι επιλογές που έχουμε σε κάθε μία από τις παραπάνω κατηγορίες του υπο-μενού **Statistics** παρουσιάζονται παρακάτω αναλυτικά.

Περιγραφικοί Δείκτες (Descriptives):

- ***Means***: όπου δίνει μέση τιμή και τυπική απόκλιση για όλες τις ανεξάρτητες μεταβλητές τόσο για κάθε μία από τις ομάδες της grouping variable, ξεχωριστά όσο και συνολικά στο δείγμα μας
- ***Univariate ANOVAs***: Δίνει αποτελέσματα της ανάλυσης διακύμανσης ή του t–test (όταν έχουμε 2 ομάδες), για κάθε μία από τις ανεξάρτητες μεταβλητές πριν αυτές εισαχθούν στο μοντέλο. Δηλαδή, δίνει τα αποτελέσματα του ελέγχου ισότητας των μέσων τιμών για όλες τις ανεξάρτητες μεταβλητές, δίνοντας μία εικόνα του πόσο πολύ συμβάλλει η κάθε μεταβλητή στο διαχωρισμό των παρατηρήσεων στις ομάδες.
- ***Box's M***: Είναι ένας έλεγχος της ισότητας των πινάκων διακύμανσης-συνδιακύμανσης των ανεξάρτητων μεταβλητών στις διάφορες ομάδες.

Συντελεστές της διαχωριστικής συνάρτησης (Function Coefficients):

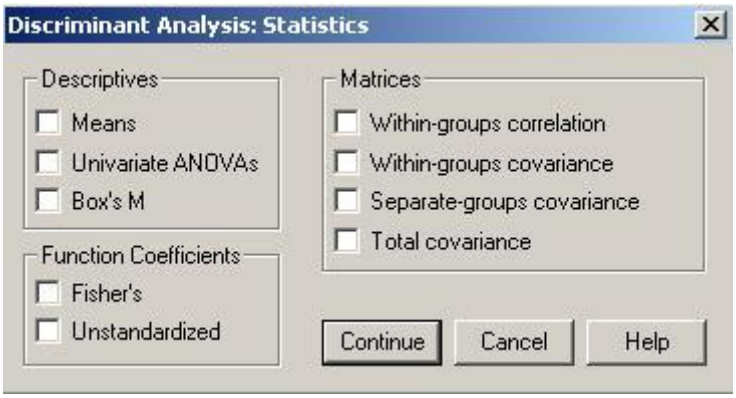
- ***Fisher's*:** Δίνει τους τυποποιημένους συντελεστές για κάθε μία από τις ανεξάρτητες μεταβλητές για τη διαχωριστική συνάρτηση του Fisher.
- ***Unstandardized*:** Δίνει τους μη τυποποιημένους συντελεστές.

Μήτρες-Πίνακες (Matrices):

- ***Within-groups correlation***: Δίνει τον συνδυασμένο πίνακα συσχετίσεων, υποθέτοντας ότι οι συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών είναι ίσες σε όλες τις ομάδες
- ***Within-groups covariance***: Δίνει τον συνδυασμένο πίνακα διακυμάνσεων - συνδιακυμάνσεων, υποθέτοντας ότι οι διακυμάνσεις

και οι συνδιακυμάνσεις μεταξύ των ανεξάρτητων μεταβλητών είναι ίσες σε όλες τις ομάδες

- ***Separate-groups covariance*:** Ξεχωριστοί πίνακες διακυμάνσεων – συνδιακυμάνσεων για κάθε μία ομάδα.
- ***Total covariance***: Πίνακας διακυμάνσεων–συνδιακυμάνσεων, συνολικά.



Εικόνα 15.3: Το μενού Statistics

Στο ***παράδειγμά μας*** επιλέξαμε να εμφανιστούν όλα τα περιγραφικά χαρακτηριστικά και οι συντελεστές της διαχωριστικής συνάρτησης και τα αποτελέσματα παρουσιάζονται στους παρακάτω *Πίνακες 15.1 -15.6.*

Η επιλογή **Means** μας δίνει μέση τιμή και τυπική απόκλιση για όλες τις μεταβλητές που συμπεριλήφθηκαν στην ανάλυση, ξεχωριστά για τα άτομα που πέθαναν μέσα στο νοσοκομείο και γι' αυτά που δεν πέθαναν (*Πίνακας 15.1*).

Η επιλογή ***Univariate ANOVAs*** δίνει τον *Πίνακα 15.2*, από τον οποίο βλέπουμε ότι για την LDH, λευκά αιμοσφαίρια (WBC), ουρία (ouria), κρεατινίνη (creatinin), ηλικία (age), φύλο (sex) και ύψος (height) οι μέσες τιμές στις δύο ομάδες διαφοροποιούνται σημαντικά ($p<0,05$). Επιπλέον, ο δείκτης ***λάμδα του Wilks*** μας δίνει χρήσιμες πληροφορίες για τις διαφορές των ομάδων. Ο δείκτης αυτός είναι το ποσοστό της διακύμανσης το οποίο δεν εξηγείται από το μοντέλο της ανάλυσης διακύμανσης κατά ένα παράγοντα. Κυμαίνεται από το μηδέν (0) έως το ένα (1). Τιμές κοντά στο μηδέν υποδεικνύουν ισχυρές διαφορές, ενώ τιμές κοντά στο ένα (1) υποδεικνύουν ότι δεν υπάρχουν διαφορές. Ο δείκτης αυτός δεν είναι παρά η ποσότητα $1- R^2$ όπου R^2 είναι ο συντελεστής προσδιορισμού από την παλινδρόμηση με εξαρτημένη τη μεταβλητή που μας ενδιαφέρει και ανεξάρτητες ψευδομεταβλητές για τις ομάδες.

Η επιλογή του τεστ για την ισότητα των πινάκων συνδιακύμανσης (***Box's M***) δίνει τον *Πίνακα 15.3*. Το test αυτό ελέγχει την υπόθεση H_0: $\Sigma_1 = \Sigma_2 = ... = \Sigma_K$, άρα για p-value (sig)<0,05 , όπως στο παράδειγμά μας, απορρίπτεται η υπόθεση της ισότητας των πινάκων διακύμανσης. Το τεστ αυτό είναι πολύ ευαίσθητο σε αποκλίσεις από την κανονική κατανομή.

Group Statistics

Θάνατος στο νοσοκομείο		Mean	Std. Deviation	Valid N (listwise) Unweighted	Weighted
No	LDH	398,9028	374,34923	1414	1414,000
	WBC	9039,6450	3242,75039	1414	1414,000
	ouria	45,4583	24,11216	1414	1414,000
	Creatinin	1,1666	,90867	1414	1414,000
	Uric acid	5,9652	4,75614	1414	1414,000
	age	65,7313	12,68090	1414	1414,000
	sex	,7702	,42088	1414	1414,000
	Weight (Kg)	79,1036	13,02514	1414	1414,000
	Height (cm)	169,0566	7,96105	1414	1414,000
Yes	LDH	624,7619	575,78315	42	42,000
	WBC	10982,38	3926,93501	42	42,000
	ouria	73,2857	39,31566	42	42,000
	Creatinin	1,6619	,89114	42	42,000
	Uric acid	7,1714	3,18068	42	42,000
	age	75,1429	10,02575	42	42,000
	sex	,5714	,50087	42	42,000
	Weight (Kg)	76,6905	14,72262	42	42,000
	Height (cm)	166,0952	8,38061	42	42,000
Total	LDH	405,4179	383,22860	1456	1456,000
	WBC	9095,6854	3279,05999	1456	1456,000
	ouria	46,2610	25,09738	1456	1456,000
	Creatinin	1,1809	,91165	1456	1456,000
	Uric acid	6,0000	4,72163	1456	1456,000
	age	66,0027	12,70744	1456	1456,000
	sex	,7644	,42450	1456	1456,000
	Weight (Kg)	79,0340	13,07778	1456	1456,000
	Height (cm)	168,9712	7,98585	1456	1456,000

Πίνακας 15.1: Περιγραφικά μέτρα για τις 2 ομάδες

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
LDH	,990	14,297	1	1454	,000
WBC	,990	14,450	1	1454	,000
ouria	,966	51,899	1	1454	,000
Creatinin	,992	12,131	1	1454	,001
Uric acid	,998	2,665	1	1454	,103
age	,985	22,708	1	1454	,000
sex	,994	8,988	1	1454	,003
Weight (Kg)	,999	1,389	1	1454	,239
Height (cm)	,996	5,627	1	1454	,018

Πίνακας 15.2: Έλεγχος διακυμάνσεων

Test Results

Box's M		118,715
F	Approx.	2,429
	df1	45
	df2	16345,702
	Sig.	,000

Tests null hypothesis of equal population covariance matrices.

Πίνακας 15.3: Έλεγχος ισότητας πινάκων διακύμανσης

Όσον αφορά στους συντελεστές διαχωριστικής συνάρτησης του Fisher (***Function coefficients***), τα αποτελέσματα του παραδείγματός μας παρουσιάζονται στον *Πίνακα 15.4* .

Classification Function Coefficients

	Θάνατος στο νοσοκομείο	
	No	Yes
LDH	,004	,005
WBC	,001	,001
ouria	-,014	,023
Creatinin	,343	,399
Uric acid	-,159	-,139
age	,908	,947
sex	-39,778	-40,705
Weight (Kg)	-,412	-,398
Height (cm)	4,707	4,701
(Constant)	-400,452	-407,013

Fisher's linear discriminant functions

Πίνακας 15.4: Οι διαχωριστικές συναρτήσεις του Fisher

Για κάθε νέο ασθενή υπολογίζονται 2 διαφορετικά σκορ χρησιμοποιώντας τους παραπάνω συντελεστές ως εξής:

*W1=-400,452+0,004*LDH + 0,001*WBC - 0,014*ouria + 0,343*Creatinin - 0,159*Uric acid + 0,908*age -39,778*sex - 0,412*weight +4,707*Height.*

*W1=-407,013+0,005*LDH + 0,001*WBC + 0,023*ouria + 0,399*Creatinin - 0,139*Uric acid + 0,947*age -40,705*sex - 0,398*weight +4,701*Height.*

Ο ασθενής κατατάσσεται στην ομάδα όπου παρατηρείται το μέγιστο σκορ (π.χ. σε αυτούς που θα πεθάνουν μέσα στο νοσοκομείο ή όχι).
Οι παραπάνω συντελεστές είναι ανάλογοι των μη τυποποιημένων συντελεστών (***unstandardized function coefficients***). Αυτοί οι συντελεστές όσον αφορά στο παράδειγμά μας φαίνονται στον *Πίνακα 14.5* και η διαχωριστική συνάρτηση μπορεί να γραφτεί ως εξής:

*Z=-3,768+0,001*LDH + 0,000*WBC + 0,025*ouria + 0,038*Creatinin + 0,013*Uric acid + 0,026*age -0,630*sex + 0,010*weight -0,004*Height.*

Canonical Discriminant Function Coefficients

	Function
	1
LDH	,001
WBC	,000
ouria	,025
Creatinin	,038
Uric acid	,013
age	,026
sex	-,630
Weight (Kg)	,010
Height (cm)	-,004
(Constant)	-3,768

Unstandardized coefficients

Πίνακας 15.5: Μη τυποποιημένοι συντελεστές

Οι αντίστοιχοι τυποποιημένοι συντελεστές της κανονικοποιημένης διαχωριστικής συνάρτησης (***standardized canonical discrimination function coefficients***) δίνουν μια ένδειξη της συνεισφοράς της κάθε μεταβλητής στη διαχωριστική συνάρτηση (*Πίνακας 15.6*). Πιο συγκεκριμένα, όσο πιο μεγάλος ο συντελεστής τόσο πιο μεγάλη και η συνεισφορά της συγκεκριμένης μεταβλητής στη διαχωριστική συνάρτηση. Από τον *Πίνακα 15.6* διαπιστώνουμε ότι τη μεγαλύτερη συνεισφορά έχει η «ουρία» με συντελεστή 0,610.

Standardized Canonical Discriminant Function Coefficients

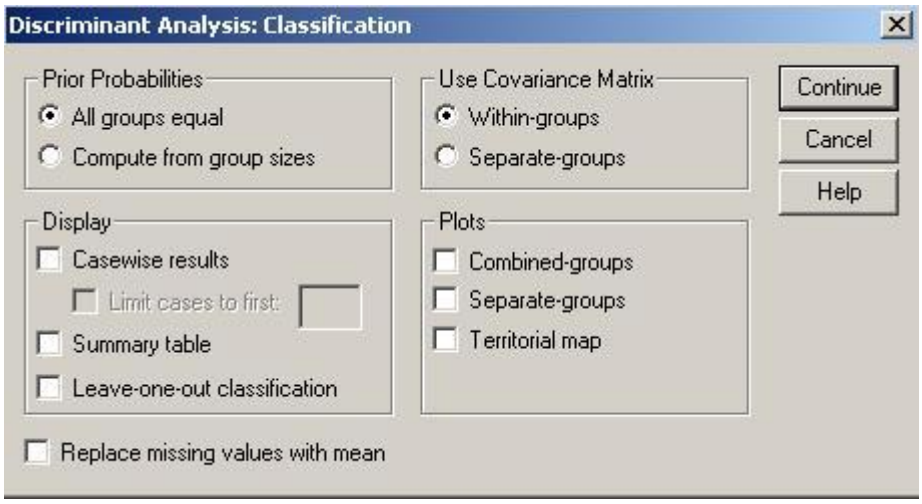
	Function
	1
LDH	,338
WBC	,310
ouria	,610
Creatinin	,034
Uric acid	,064
age	,333
sex	-,267
Weight (Kg)	,126
Height (cm)	-,035

Πίνακας 15.6: Τυποποιημένοι συντελεστές της κανονικοποιημένης διαχωριστικής συνάρτησης.

Σημείωση 1: Η κατάταξη των ανεξάρτητων μεταβλητών από αυτήν που συνεισφέρει περισσότερο στο διαχωρισμό των ατόμων σε αυτήν που συνεισφέρει λιγότερο πραγματοποιείται χρησιμοποιώντας την απόλυτη τιμή των τυποποιημένων συντελεστών.

Σημείωση 2: Οι τυποποιημένοι συντελεστές είναι προτιμότερο να χρησιμοποιούνται για την κατάταξη των ανεξάρτητων μεταβλητών από την πιο σημαντική στη λιγότερο σημαντική, σε σχέση με τους μη τυποποιημένους συντελεστές, ειδικά στην περίπτωση που οι μονάδες μέτρησης των ανεξάρτητων μεταβλητών διαφέρουν.

v. Πατώντας το κουμπί επιλογών «***Classify***» ανοίγει ένα νέο παράθυρο διαλόγου (*Εικόνα 15.4*) από το οποίο μπορούμε να επιλέξουμε:



Εικόνα 15.4: Επιλογή **«Clasify»**

Prior probabilities, όπου μπορούμε να επιλέξουμε ανάμεσα σε:

- ***All groups equal***: ίσες εκ των προτέρων πιθανότητες για τις 2 ομάδες ή
- ***Compute from group sizes***: υπολογισμό από το μέγεθος των δειγμάτων

Display, όπου μπορούμε να επιλέξουμε:

- ***Casewise results***: όπου παρουσιάζεται η ομάδα που ανήκει πραγματικά το κάθε άτομο καθώς και η εκ των υστέρων πιθανότητα, η τιμή του σκορ της διαχωριστικής συνάρτησης,
- ***Summary table***: όπου προκύπτει ένας πίνακας που να παρουσιάζει τον αριθμό των ατόμων που σωστά ή λάθος έχουν καταταχθεί σε κάποια ομάδα (βλ. *Πίνακα 15.7* για αποτελέσματα από το παράδειγμά μας). Παρατηρούμε ότι το 71,4% αυτών που πέθαναν μέσα στο νοσοκομείο και το 82,5% αυτών που δεν πέθαναν, κατατάσσονται ορθώς βάση της διαχωριστικής ανάλυσης.
- ***Leave-one-out classification***: όπου προκύπτει ένας πίνακας που να παρουσιάζει την κατάταξη των παρατηρήσεων όταν κάθε μία κατατάχθηκε χρησιμοποιώντας συνάρτηση που δημιουργήθηκε χωρίς να λαμβάνεται υπόψη η συγκεκριμένη παρατήρηση.

Use Covariance Matrix, όπου υπάρχει η δυνατότητα η διαχωριστική ανάλυση να πραγματοποιηθεί χρησιμοποιώντας:

- ***Within-groups***: το συνδυασμένο πίνακα διακυμάνσεων – συνδιακυμάνσεων ή
- ***Separate-groups:*** διαφορετικούς πίνακες για κάθε ομάδα.

Γι' αυτήν την επιλογή είναι απαραίτητη η εφαρμογή του Box's m test, το οποίο θα μας δείξει αν ισχύει η ισότητα των διασπορών μεταξύ των ομάδων ή όχι.

Plots, όπου μπορούμε να πάρουμε:

- ***Combined-groups:*** ένα στικτόγραμμα των τιμών των δύο πρώτων διαχωριστικών συναρτήσεων για όλες τις ομάδες. Αν υπάρχει μόνο μία συνάρτηση θα πάρουμε ένα ιστόγραμμα,
- ***Separate-groups:*** στικτογράμματα για τις τιμές των δύο πρώτων διαχωριστικών συναρτήσεων για κάθε ομάδα ξεχωριστά. Αν υπάρχει μόνο μία συνάρτηση θα πάρουμε ένα ιστόγραμμα, και τέλος
- ***Territorial map:*** Ένα γράφημα των ορίων που χρησιμοποιούνται για την ταξινόμηση των παρατηρήσεων στις ομάδες βάση των τιμών των συναρτήσεων. Οι αριθμοί αντιστοιχούν στις ομάδες που ταξινομούνται οι παρατηρήσεις. Ο μέσος κάθε ομάδας υποδηλώνεται με έναν αστερίσκο. Αυτό το γράφημα δεν παράγεται αν υπάρχει μόνο μία διαχωριστική συνάρτηση.

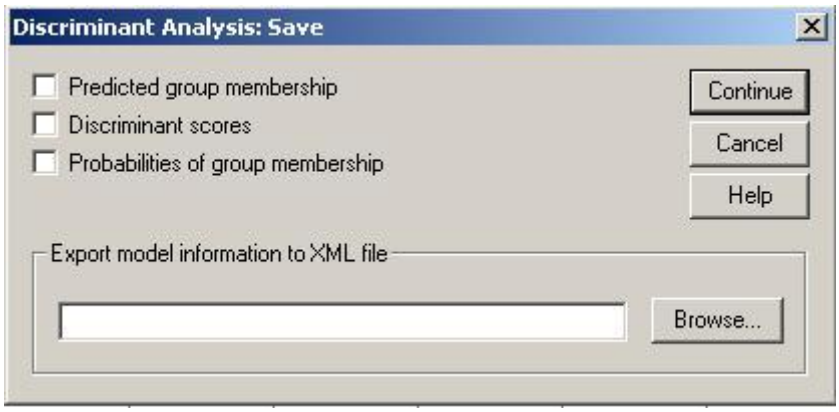
Classification Results[a]

			Predicted Group Membership		
		Θάνατος στο νοσοκομείο	No	Yes	Total
Original	Count	No	1167	247	1414
		Yes	12	30	42
	%	No	82,5	17,5	100,0
		Yes	28,6	71,4	100,0

a. 82,2% of original grouped cases correctly classified.

Πίνακας 15.7: Αποτελέσματα της κατάταξης των παρατηρήσεων

vi. Πατώντας το κουμπί επιλογών «***Save***» ανοίγει ένα νέο παράθυρο διαλόγου (*Εικόνα 15.5*) από το οποίο μπορούμε να επιλέξουμε:



Εικόνα 15.5 : Το μενού «Save»

- ***Predicted group membership***: την ομάδα στην οποία προβλέπεται να ανήκει το κάθε άτομο σύμφωνα με το διαχωριστικό κανόνα που δημιουργήθηκε (μία μεταβλητή),
- ***Discriminant scores***: την τιμή που παίρνει κάθε μία από τις διαχωριστικές συναρτήσεις για κάθε άτομο (δημιουργούνται τόσες μεταβλητές όσες και οι διαχωριστικές συναρτήσεις), και
- ***Probabilities of group-membership:*** την πιθανότητα κάθε άτομο να ανήκει σε κάθε ομάδα δεδομένης της τιμής της διαχωριστικής συνάρτησης (δημιουργούνται τόσες μεταβλητές όσες και οι ομάδες).

16. Έλεγχος αξιοπιστίας

16.1 Εισαγωγή

Η αξιοπιστία ενός εργαλείου εκφράζει τη συνέπεια με την οποία ένα εργαλείο (π.χ. ερωτηματολόγιο) μετράει ένα συγκεκριμένο χαρακτηριστικό. Πιο συγκεκριμένα ένα εργαλείο χαρακτηρίζεται ως αξιόπιστο όταν: όσες φορές και αν εφαρμοστεί στα ίδια άτομα θα δώσει παρόμοια αποτελέσματα.

Ο έλεγχος αξιοπιστίας ενός ερωτηματολογίου είναι **απαραίτητος** όταν:

- Συνθέτουμε ένα καινούριο ερωτηματολόγιο ή χρησιμοποιούμε για πρώτη φορά ένα νέο όργανο ή μηχάνημα για την πραγματοποίηση οποιουδήποτε τύπου μετρήσεων.
- Χρησιμοποιούμε ερωτηματολόγια που έχουν χρησιμοποιηθεί ευρέως και με επιτυχία σε πληθυσμούς με διαφορετικά χαρακτηριστικά από αυτά που θα έχει ο δικός μας πληθυσμός (π.χ. αν ένα ερωτηματολόγιο έχει χρησιμοποιηθεί σε πληθυσμό ενηλίκων και εμείς θα το χρησιμοποιήσουμε σε πληθυσμό εφήβων).
- Μεταφράζουμε ένα ερωτηματολόγιο.

Οι **μορφές αξιοπιστίας** είναι οι εξής:

1. Εσωτερική συνοχή (Internal consistency)
2. Επαναληψιμότητα (Test-retest reliability)
3. Μεταξύ των παρατηρητών/κριτών αξιοπιστία (interobserver or interrater reliability)

16.1.1 Εσωτερική συνοχή

Αυτή η μορφή αξιοπιστίας είναι ένα μέτρο που παρουσιάζει ιδιαίτερο ενδιαφέρον στην περίπτωση που ένα ερωτηματολόγιο χρησιμοποιείται για την ποσοτικοποίηση ενός μη μετρήσιμου χαρακτηριστικού (π.χ. κατάθλιψη, ποιότητα ζωής, επιθετικότητα κτλ). Τέτοιου τύπου ερωτηματολόγια αποτελούνται από μία σειρά ερωτήσεων που αποσκοπούν στην αξιολόγηση των διαφόρων πτυχών του ίδιου χαρακτηριστικού. Κάθε ερώτηση διαθέτει μία σειρά πιθανών απαντήσεων (π.χ. καθόλου, λίγο, μέτρια, πολύ, πάρα πολύ) στις οποίες αποδίδουμε ένα συγκεριμένο σκορ (π.χ. 1, 2, 3, 4 και 5, αντίστοιχα / κλίμακα Likert). Αθροίζοντας τα σκορ όλων των ερωτήσεων, προκύπτει το συνολικό σκορ.

Η ***εσωτερική συνοχή*** είναι ένας δείκτης που εκφράζει το πόσο καλά οι διάφορες ερωτήσεις του ερωτηματολογίου μετρούν το ίδιο χαρακτηριστικό. Είναι σαφές πως για να χαρακτηρίζεται ένα ερωτηματολόγιο από μεγάλη εσωτερική συνοχή είναι απαραίτητο όλες οι ερωτήσεις από τις οποίες αποτελείται να αξιολογούν το ίδιο χαρακτηριστικό ή αλλιώς να υπάρχει ομοιογένεια μεταξύ των ερωτήσεων.

Το στατιστικό μέτρο που χρησιμοποιείται για την αξιολόγηση της εσωτερικής συνοχής είναι ο συντελεστής ***Cronbach's a*** (Cronbach's a coefficient). Το εύρος τιμών αυτού του συντελεστή διακυμαίνεται μεταξύ 0 και 1. Το μεγάλο μειονέκτημα αυτού του κριτηρίου είναι ότι δεν υπάρχουν συγκεκριμένα, εδραιωμένα κατώφλια με τα οποία να μπορούμε να συγκρίνουμε την τιμή του συντελεστή ***Cronbach's a*** που προκύπτει από τα δεδομένα μας προκειμένου να αξιολογήσουμε την εσωτερική συνοχή του ερωτηματολογίου μας ως άριστη, καλή, μέτρια ή κακή. Συνεπώς, ο τρόπος αξιολόγησης αυτού του κριτηρίου είναι αυθαίρετος και υποκειμενικός (ο κάθε

ερευνητής αξιολογεί και ερμηνεύει τα αποτελέσματα του συγκεκριμένου κριτηρίου αυθαίρετα). Ένας από τους κανόνες που έχει προταθεί για την αξιολόγηση του συγκεκριμένου κριτηρίου είναι:
a>0,9: εξαιρετική αξιοπιστία,
a>0,8: καλή,
a>0,7: αποδεκτή,
a>0,6: αμφισβητήσιμη,
a>0,5: φτωχή και
a<0,5: μη αποδεκτή.
Γενικά, a>0,8 θεωρείτε ικανοποιητική και αποδεκτή αξιοπιστία.

ΠΡΟΣΟΧΗ!!!:
1. Στην περίπτωση που ο τρόπος βαθμολόγησης κάποιων ερωτήσεων του ερωτηματολογίου είναι αντίστροφος (π.χ. καθόλου: 5, λίγο: 4, μέτρια:3, πολύ:2 πάρα πολύ:1, ενώ για κάποιες άλλες ισχύει η βαθμολόγηση καθόλου: 1, λίγο:2, μέτρια:3, πολύ:4, πάρα πολύ:5), πρώτα πρέπει να πραγματοποιηθεί η αντιστροφή στα σκορ και μετά να υπολογιστεί ο συντελεστής Cronbach's a για τον έλεγχο εσωτερικής συνοχής του ερωτηματολογίου.
2. Υπάρχουν περιπτώσεις όπου από ένα ερωτηματολόγιο που αποσκοπεί στην αξιολόγηση ενός συγκεκριμένου χαρακτηριστικού (π.χ. ποιότητα ζωής) να υπολογίζενται υποκλίμακες. Σε αυτή την περίπτωση, το ιδανικότερο είναι να ελέγχεται η εσωτερική συνοχή (υπολογισμός Cronbach's a) για κάθε υποκλίμακα, ξεχωριστά.

Σε περίπτωση που η εσωτερική συνοχή ενός ερωτηματολογίου βρεθεί να είναι χαμηλή, οφείλουν να γίνουν προσπάθειες βελτίωσης αυτής με κάποιον από τους εξής τρόπους:

a. Προσθήκη επιπλέον ερωτήσεων που ενδεχομένως να είχαν παραλειφθεί και να είχαν σαν αποτέλεσμα την κακή αντιπροσώπευση κάποιων πτυχών του χαρακτηριστικού που μετράμε μέσα στο ερωτηματολόγιο.
b. Επαναδιατύπωση κάποιων εκ των ερωτήσεων.
c. Αφαίρεση κάποιων εκ των ερωτήσεων.

Αφού γίνουν οι απαραίτητες διορθώσεις στο ερωτηματολόγιο, τότε οφείλουμε να ελέγξουμε και πάλι την εσωτερική του συνοχή προκειμένου να βεβαιωθούμε ότι αυτές οι αλλαγές ήταν ικανοποιητικές (δηλαδή, όντως αυξήθηκε η εσωτερική συνοχή).

Σημείωση: Ο εντοπισμός των ερωτήσεων που οφείλουν να αφαιρεθούν από το ερωτηματολόγιο προκειμένου να βελτιωθεί η εσωτερική συνοχή του είναι εύκολος αν υπολογίσουμε το συντελεστή Cronbach's a όταν αφαιρέσουμε μία μία τις ερωτήσεις. Για παράδειγμα, σε ένα ερωτηματολόγιο με 15 ερωτήσεις, αφαιρούμε την πρώτη και υπολογίζουμε τον συντελεστή Cronbach's a χρησιμοποιώντας τις υπόλοιπες 14. Πραγματοποιούμε την ίδια διαδικασία μέχρι να υπολογίσουμε όλους τους πιθανούς συντελεστές Cronbach's a (15 στο συγκεκριμένο παράδειγμα). Ο υπολογισμός όλων αυτών των συντελεστών Cronbach's a πραγματοποιείται αυτόματα στο στατιστικό πρόγραμμα SPSS και χαρακτηρίζονται ως «Cronbach's a if item deleted». Η ερώτηση που οφείλει να απομακρυνθεί από το ερωτηματολόγιο προκειμένου να βελτιωθεί η εσωτερική του συνοχή είναι αυτή που ο συντελεστής Cronbach's a που υπολογίστηκε εν τη απουσία της ήταν αρκετά υψηλότερος σε σχέση με τον αρχικό.

16.1.2 Επαναληψιμότητα

Η ***επαναληψιμότητα*** είναι μια ευρέως χρησιμοποιούμενη μορφή αξιοπιστίας διαφόρων εργαλείων (π.χ. ερωτηματολογίων, μηχανημάτων κτλ). Αυτή η μορφή αξιοπιστίας ελέγχεται εφαρμόζοντας το ίδιο εργαλείο σε 2 διαφορετικές χρονικές στιγμές (τουλάχιστον) στην ίδια ομάδα ατόμων προκειμένου να ελέγξουμε αν τα αποτελέσματα των 2 μετρήσεων συμπίπτουν. Τα αποτελέσματα του εργαλείου στις 2 διαφορετικές χρονικές στιγμές συγκρίνονται με κάποιο κατάλληλο στατιστικό κριτήριο (π.χ. *intraclass correlation coefficient (ICC) ή kappa measure of agreement (κ)*).

Το μεγάλο μειονέκτημα αυτής της μορφής αξιοπιστίας είναι η επιλογή του χρονικού διαστήματος που θα μεσολαβήσει μεταξύ των 2 διαδοχικών μετρήσεων. Πιο συγκεκριμένα:

- Ενδέχεται ότι όσο μεγαλύτερο είναι το χρονικό διάστημα τόσο μικρότερη θα είναι η συσχέτιση μεταξύ των 2 μετρήσεων. Αυτό δεν είναι απαραίτητο να εκφράζει ότι το εργαλείο μας δεν είναι αξιόπιστο. Απλά, είναι πιθανό, όταν μεσολαβεί μεγάλο χρονικό διάστημα μεταξύ της πρώτης και της δεύτερης μέτρησης, το χαρακτηριστικό το οποίο μετράμε με το εργαλείο να έχει μεταβληθεί.
- Αν είναι πολύ μικρό το χρονικό διάστημα, ενδέχεται η δεύτερη μέτρηση να έχει επηρεαστεί από την πρώτη.

Συνεπώς, λοιπόν, στην αξιολόγηση της επαναληψιμότητας ενός εργαλείου μεγάλο ρόλο διαδραματίζει το χρονικό διάστημα που θα επιλέξουμε να επαναληφθεί η μέτρηση.

Σε αυτό το σημείο, πρέπει να σημειωθεί ότι κάποιες φορές τα δεδομένα δεν συγκεντρώνονται από τους ίδιους τους συμμετέχοντες (δηλαδή δεν χρησιμοποιείται κάποιο αυτό-απαντούμενο ερωτηματολόγιο). Αντίθετα, ενδέχεται κάποιος ερευνητής να απαντά το ερωτηματολόγιο για κάθε συμμετέχοντα (π.χ. ο ψυχίατρος απαντά ένα ερωτηματολόγιο που αποσκοπεί στην αξιολόγηση της επιθετικότητας των παιδιών). Σε αυτή την περίπτωση, οφείλουμε να ελέγξουμε αν ο ερευνητής είναι αξιόπιστος και αυτό επιτυγχάνεται ζητώντας από τον ερευνητή να πραγματοποιήσει τουλάχιστον 2 μετρήσεις (να συμπληρώσει τουλάχιστον 2 φορές το ερωτηματολόγιο) σε κάθε ερωτώμενο. Αυτή η μορφή αξιοπιστίας είναι μία άλλη μορφή της επαναληψιμότητας και ονομάζεται «***εντός των παρατηρητών/κριτών αξιοπιστία***» («intraobserver ή intrarater reliability»). Τα στατιστικά κριτήρια που χρησιμοποιούνται και για τον έλεγχο αυτής της μορφής αξιοπιστίας είναι:

- *Kappa measure of agreement (κ):* όταν το χαρακτηριστικό που μετράμε είναι κατηγορικό, π.χ. ο ακτινολόγος χαρακτηρίζει κάποια γυναίκα ως ασθενή με Ca μαστού ή ως υγιή βλέποντας τη μαστογραφία της. Οι τιμές αυτού του κριτηρίου είναι ≤ 1.
- *Intra-class correlation coefficient (ICC)*: όταν το χαρακτηριστικό μου μετράμε είναι συνεχές, π.χ. ψυχολόγοι παρατηρούν κάποια άτομα και βαθμολογούν με ένα σκορ την επιθετικότητά τους. Το εύρος τιμών αυτού του κριτηρίου διακυμαίνεται μεταξύ -1 και +1.

Το μεγάλο πρόβλημα που αντιμετωπίζουμε στην αξιολόγηση των 2 παραπάνω στατιστικών κριτηρίων, είναι ότι δεν υπάρχει προκαθορισμένο κατώφλι βάσει του οποίου να χαρακτηρίσουμε την αξιοπιστία που προκύπτει υψηλή, χαμηλή ή μέτρια. Γενικά, υπάρχει η τάση *κ ή ICC >0,8* να ερμηνεύονται ως υψηλή αξιοπιστία.

16.1.3 Αξιοπιστία μεταξύ των παρατηρητών/κριτών

Η «αξιοπιστία μεταξύ των κριτών» είναι ένα μέτρο συμφωνίας 2 ή περισσότερων κριτών αναφορικά με την αξιολόγηση ενός χαρακτηριστικού. Η συγκεκριμένη μορφή αξιοπιστίας δεν παρουσιάζει κανένα ενδιαφέρον στην περίπτωση που τα ερωτηματολόγια είναι αυτό-απαντούμενα από τους συμμετέχοντες. Αντίθετα, ο έλεγχος αυτής της μορφής αξιοπιστίας είναι *απαραίτητος* όταν:

- το ερωτηματολόγιο πρέπει να συμπληρώνεται από ερευνητές παρατηρώντας τους συμμετέχοντες στη μελέτη (π.χ. ερωτηματολόγιο αξιολόγησης συμπεριφοράς αυτιστικών παιδιών), και
- 2 ή περισσότεροι ερευνητές πρέπει να χρησιμοποιηθούν για την γρήγορη διεξαγωγή της μελέτης.

Σε αυτή την περίπτωση, πρέπει να εξασφαλίσουμε ότι ο τρόπος αξιολόγησης του χαρακτηριστικού με τη βοήθεια του ερωτηματολόγιου είναι ίδιος από όλους τους ερευνητές που θα χρησιμοποιηθούν στη μελέτη.

Τα απαραίτητα ***βήματα*** για την πραγματοποίηση του ελέγχου αυτής της μορφής αξιοπιστίας είναι:

- Επιλέγουμε ένα μικρό δείγμα από τον πληθυσμό αναφοράς
- 2 ή περισσότεροι ερευνητές εφαρμόζουν το ίδιο εργαλείο σε όλα τα άτομα του δείγματος (ο ένας ερευνητής ανεξάρτητα από τον άλλο).
- Συγκρίνεται ο βαθμός συμφωνίας των αποτελεσμάτων των 2 ερευνητών, χρησιμοποιώντας το *intraclass correlation coefficient ή kappa measure of agreement.*

Αν από τα παραπάνω στατιστικά κριτήρια βρεθεί ότι ο βαθμός συμφωνίας μεταξύ των κριτών είναι μεγάλος, τότε στο συνολικό δείγμα της μελέτης το εργαλείο μπορεί να εφαρμοστεί σε μερικούς συμμετέχοντες από τον ένα ερευνητή και στους υπόλοιπους συμμετέχοντες από τον άλλο ερευνητή.

Μία άλλη περίπτωση που ο έλεγχος αυτής της μορφής αξιοπιστίας είναι εξαιρετικά χρήσιμος είναι όταν επιθυμούμε να αξιολογήσουμε την αξιοπιστία του ίδιου του εργαλείου. Για παράδειγμα, όταν στόχος μας είναι να αξιολογήσουμε αν η μαστογραφία είναι αξιόπιστο εργαλείο για την διάγνωση καρκίνου του μαστού ή όχι.

Τα απαραίτητα ***βήματα*** για την πραγματοποίηση του ελέγχου αυτής της μορφής αξιοπιστίας είναι:

- Επιλέγουμε ένα μικρό δείγμα από τον πληθυσμό αναφοράς
- 2 ή περισσότεροι ερευνητές αξιολογούν τις μαστογραφίες των συμμετεχόντων (ο ένας ερευνητής ανεξάρτητα από τον άλλο).
- Συγκρίνεται ο βαθμός συμφωνίας των συμπερασμάτων των 2 ερευνητών, χρησιμοποιώντας το *kappa measure of agreement.*

Αν το συγκεκριμένο κριτήριο δείξει ότι ο βαθμός συμφωνίας των ερευνητών είναι μεγάλος, το συμπέρασμά μας είναι ότι η μαστογραφία είναι αξιόπιστο εργαλείο για την διάγνωση καρκίνου του μαστού.

***Σημείωση*:** Ο έλεγχος αξιοπιστίας εντός και μεταξύ των κριτών παρουσιάζει ιδιαίτερο ενδιαφέρον όχι μόνο στην περίπτωση όπου χρησιμοποιείται ερωτηματολόγιο που συμπληρώνουν ερευνητές, αλλά και σε οποιαδήποτε άλλη περίπτωση όπου υπεισέρχεται το υποκειμενικό στοιχείο του ερευνητή για την αξιολόγηση ενός χαρακτηριστικού (π.χ. αξιολόγηση μιας απεικονιστικής διαγνωστικής εξέτασης, μέτρηση αποστάσεων σε μία ακτινογραφία κτλ).

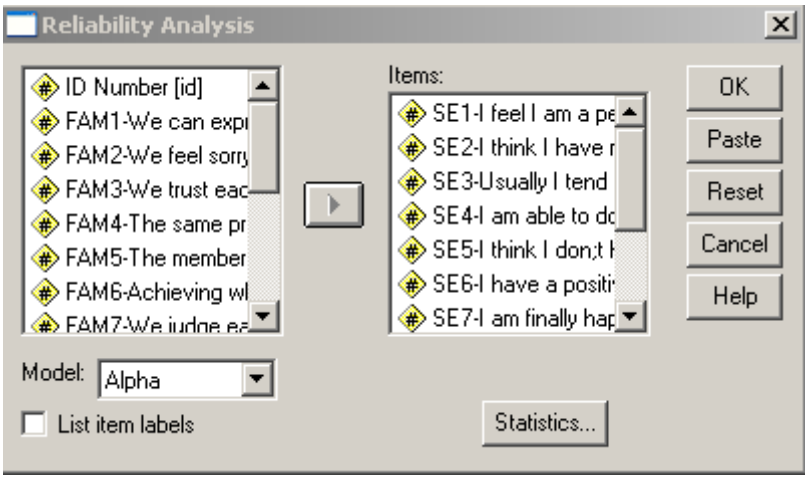
16.2 Έλεγχος εσωτερικής συνοχής με τη χρήση του SPSS

Ας υποθέσουμε ότι στόχος μας είναι να ελέγξουμε αν υπάρχει εσωτερική συνοχή μεταξύ των 10 ερωτήσεων ενός ερωτηματολογίου που έχει δημιουργηθεί για να ελεγχθεί ο βαθμός αυτοεκτίμησης μαθητών δευτεροβάθμιας εκπαίδευσης. Οι απαντήσεις σε όλες τις ερωτήσεις είναι της μορφής: Διαφωνώ απόλυτα, Διαφωνώ, Συμφωνώ, συμφωνώ απόλυτα και το εύρος τιμών του σκορ που αποδίδεται σε αυτές τις απαντήσεις διακυμαίνεται μεταξύ 1 και 4. Αθροίζοντας, τα σκορ και των 10 ερωτήσεων προκύπτει ένα συνολικό σκορ με εύρος τιμών 10-40, όπου μεγαλύτερες τιμές εκφράζουν υψηλότερη αυτοεκτίμηση. Το πρόβλημα, όμως, σε αυτό το ερωτηματολόγιο είναι ότι κάποιες ερωτήσεις είναι διατυπωμένες να εκφράζουν αισιοδοξία (π.χ. Νομίζω ότι έχω αρκετά καλές ιδιότητες) ενώ κάποιες άλλες είναι διατυπωμένες να εκφράζουν απαισιοδοξία (π.χ. έχω την τάση να σκέφτομαι ότι είμαι αποτυχημένο άτομο). Συνεπώς, λοιπόν, δεν πρέπει να ακολουθήσουμε τον ίδιο τρόπο απόδοσης βαθμών στις απαντήσεις όλων των ερωτήσεων. Πιο συγκεκριμένα, στις ερωτήσεις που εκφράζουν αισιοδοξία θα αποδώσουμε το μέγιστο σκορ (4) στην απάντηση «συμφωνώ απόλυτα», ενώ στις ερωτήσεις που εκφράζουν απαισιοδοξία το μέγιστο σκορ θα αποδοθεί στην απάντηση «διαφωνώ απόλυτα». Αν και μόνο αν, έχουν πραγματοποιηθεί σωστά όλα τα παραπάνω, ο έλεγχος εσωτερικής συνοχής θα μας δείξει την πραγματική εσωτερική συνοχή του ερωτηματολογίου.

Τα ***βήματα*** που πρέπει να ακολουθήσουμε για την πραγματοποίηση του ελέγχου εσωτερικής συνοχής είναι:

Analyse → Scale → Reliability analysis

Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 16.1*.



Εικόνα 16.1: Πραγματοποίηση ελέγχου εσωτερικής συνοχής.

I. Ως «***Items***» δηλώνουμε τις ερωτήσεις του ερωτηματολογίου, των οποίων την εσωτερική συνοχή επιθυμούμε να ελέγξουμε.
II. Πατώντας το κουμπί επιλογών «***Statistics***» εμφανίζεται ένα νέο πλαίσιο διαλόγου που απεικονίζεται στην *Εικόνα 16.2*. Σε αυτό το πλαίσιο διαλόγου μπορούμε να επιλέξουμε διάφορα στατιστικά μέτρα που να περιγράφουν τόσο τις μεμονωμένες ερωτήσεις του ερωτηματολογίου όσο και το σκορ που υπολογίζεται αθροίζοντας αυτές τις ερωτήσεις.

Descriptives for:

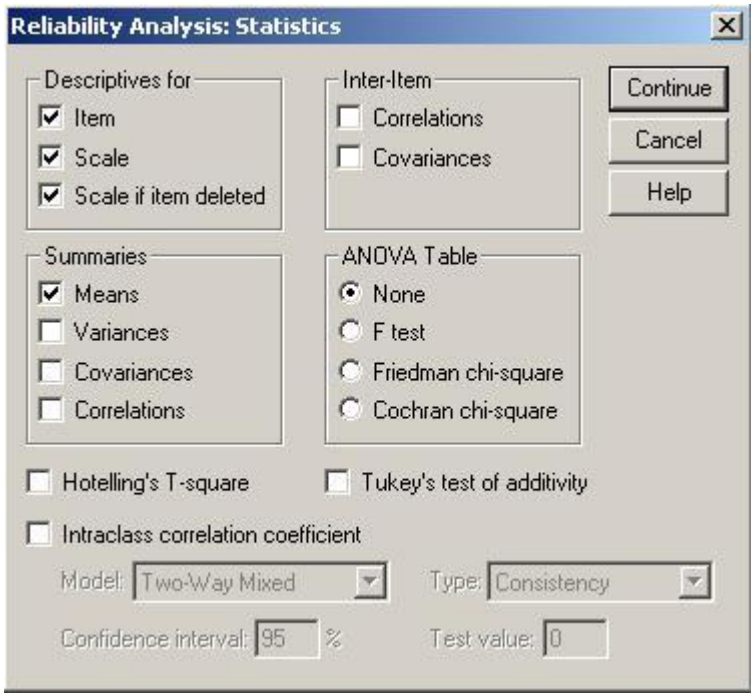
- ***Item:*** Δίνει μέση τιμή (mean) και τυπική απόκλιση (Std. Deviation) για κάθε ερώτηση (*Πίνακας 16.1*)
- ***Scale:*** Δίνει μέση τιμή (mean), διακύμανση (Variance) και τυπική απόκλιση (Std. Deviation) για το συνολικό σκορ που υπολογίζεται από αυτές τις ερωτήσεις (*Πίνακας 16.2*)
- ***Scale if item deleted:*** Δίνει τη μέση τιμή (scale mean in item deleted), και τη διακύμανση (scale variance if item deleted) για το συνολικό σκορ που υπολογίζεται αθροίζοντας όλες τις ερωτήσεις του ερωτηματολογίου με εξαίρεση μίας κάθε φορά (*Πίνακας 16.3*) καθώς επίσης και τον αντίστοιχο συντελεστή Cronbach's α (Cronbach's alpha if item deleted). Αν ο συντελεστής Cronbach α, αυξάνεται πολύ όταν αφαιρείται μία ερώτηση, τότε αυτή η ερώτηση δεν παρουσιάζει εσωτερική συνοχή με τις υπόλοιπες και πρέπει να αφαιρεθεί από τον υπολογισμό του τελικού σκορ της κλίμακας.

Summaries:

- ***Means*:** Περιγραφικά στατιστικά για τις μέσες τιμές των ερωτήσεων. Τη μεγαλύτερη, την μικρότερη, την μέση τιμή των μέσων τιμών των ερωτήσεων, το εύρος και τη διακύμανση των μέσων τιμών των ερωτήσεων (*Πίνακας 16.4*)
- ***Variances*:** Περιγραφικά στατιστικά για τις διακυμάνσεις των ερωτήσεων. Τη μεγαλύτερη, την μικρότερη, την μέση τιμή των διακυμάνσεων των ερωτήσεων, το εύρος και τη διακύμανση των διακυμάνσεων τιμών των ερωτήσεων (*Πίνακας 16.4*)
- ***Covariances*:** Περιγραφικά στατιστικά για τις συν-διακυμάνσεις ανάμεσα στις ερωτήσεις. Τη μεγαλύτερη, την μικρότερη, την μέση τιμή των συν-διακυμάνσεων των ερωτήσεων, το εύρος και τη διακύμανση των συν-διακυμάνσεων των ερωτήσεων (*Πίνακας 16.4*).
- ***Correlations*:** Περιγραφικά στατιστικά για τους συντελεστές συσχέτισης ανάμεσα στις ερωτήσεις. Τη μεγαλύτερη, την μικρότερη, την μέση τιμή των συντελεστών συσχέτισης των ερωτήσεων, το εύρος και τη διακύμανση των συντελεστών συσχέτισης των ερωτήσεων (*Πίνακας 16.4*).

Inter item: Δημιουργεί πίνακες συσχετίσεων ή συν-διακυμάνσεων ανάμεσα στις ερωτήσεις.

ANOVA table: Δίνει τεστ ισότητας των μέσων τιμών.



Εικόνα 16.2: Μενού επιλογών «Statistics»

Item Statistics

	Mean	Std. Deviation	N
SE1-I feel I am a person as worthy as others	3,47	,735	1020
SE2-I think I have many abilities	3,37	,687	1020
SE3-Usually I tend to think I am a failure	2,82	,979	1020
SE4-I am able to do things as good as others	3,35	,786	1020
SE5-I think I don;t have a lot of things that I am proud of	2,69	,996	1020
SE6-I have a positive view of myself	3,15	,818	1020
SE7-I am finally happy with myself	3,17	,817	1020
SE8-I wish I could have more respect for myself	2,12	,947	1020
SE9-Sometimes I feel useless	2,75	,992	1020
SE10-Sometimes I feel I am not good at all	2,67	,993	1020

Πίνακας 16.1: Περιγραφικά στατιστικά για τις ερωτήσεις από τις οποίες αποτελείται η κλίμακα

Scale Statistics

Mean	Variance	Std. Deviation	N of Items
29,56	27,768	5,270	10

Πίνακας 16.2: Περιγραφικά στατιστικά για το συνολικό σκορ που υπολογίζεται από το άθροισμα των ερωτήσεων.

Από τον *Πίνακα 16.3* διαπιστώνουμε ότι η εσωτερική συνοχή του ερωτηματολογίου δεν μεταβάλλεται σημαντικά με την αφαίρεση οποιασδήποτε ερώτησης, αφού οι τιμές του συντελεστή Cronbach's a if item deleted δεν διαφέρουν πολύ μεταξύ τους.

Item-Total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
SE1-I feel I am a person as worthy as others	26,09	24,462	,380	,366	,792
SE2-I think I have many abilities	26,19	24,167	,463	,428	,785
SE3-Usually I tend to think I am a failure	26,74	21,706	,560	,398	,771
SE4-I am able to do things as good as others	26,21	24,041	,403	,369	,790
SE5-I think I don;t have a lot of things that I am proud of	26,87	23,135	,380	,228	,795
SE6-I have a positive view of myself	26,41	22,499	,593	,598	,769
SE7-I am finally happy with myself	26,39	22,499	,594	,577	,769
SE8-I wish I could have more respect for myself	27,45	24,785	,221	,158	,812
SE9-Sometimes I feel useless	26,82	21,067	,627	,556	,762
SE10-Sometimes I feel I am not good at all	26,89	21,562	,566	,478	,770

Πίνακας 16.3: Περιγραφικά στατιστικά της κλίμακας αν κάποια ερώτηση αφαιρείται κάθε φορά.

Summary Item Statistics

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	2,956	2,116	3,474	1,358	1,642	,179	10
Item Variances	,778	,472	,991	,519	2,100	,042	10
Inter-Item Covariances	,222	-,041	,653	,694	-16,122	,022	10
Inter-Item Correlations	,291	-,054	,724	,779	-13,310	,035	10

The covariance matrix is calculated and used in the analysis.

Πίνακας 16.4: Περιγραφικά στατιστικά των μέσων τιμών, των διακυμάνσεων, των συν-διακυμάνσεων και των συντελεστών συσχέτισης των ερωτήσεων.

Στον *Πίνακα 16.5* παρατηρούμε το συντελεστή Cronbach's a. Παρατηρούμε ότι ο συντελεστής είναι αρκετά υψηλός που συνεπάγεται υψηλή εσωτερική συνοχή μεταξύ των ερωτήσεων από τις οποίες αποτελείται η κλίμακα. Επίσης, από τον *Πίνακα 16.3* και από την τελευταία στήλη «***Cronbach's alpha if item deleted***» διαπιστώνουμε ότι δεν υπάρχει καμία ερώτηση εκ των 10 που να βελτιώνει την εσωτερική συνοχή του ερωτηματολογίου όταν αφαιρεθεί.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,800	,804	10

Πίνακας 16.5: Συντελεστής εσωτερικής συνοχής Cronbach a.

16.3 Έλεγχος της αξιοπιστίας μεταξύ των κριτών και εντός των κριτών με τη χρήση του SPSS.

16.3.1 Intra-class correlation coefficient (ICC)

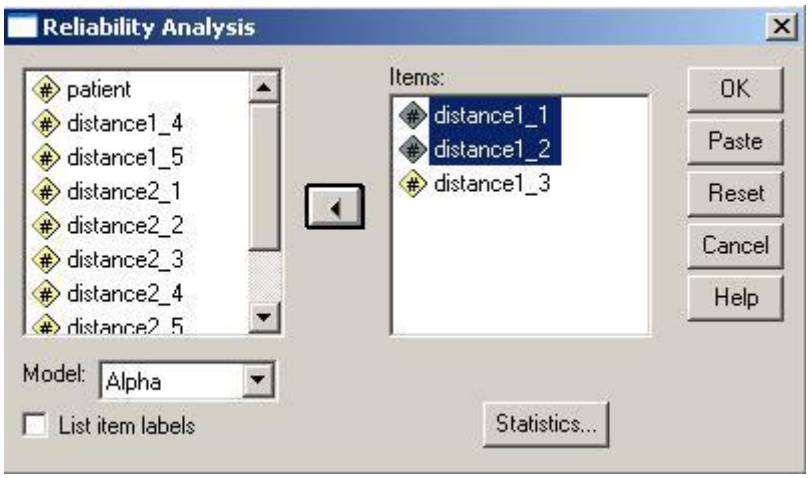
Ας υποθέσουμε ότι 2 ψυχολόγοι εφάρμοσαν ένα ερωτηματολόγιο αξιολόγησης της επιθετικότητας σε ένα δείγμα 10 παιδιών, υπολογίζοντας ένα συνολικό σκορ. Επιπλέον, ο κάθε ψυχολόγος συμπλήρωσε το ίδιο ερωτηματολόγιο για τα 10 αυτά παιδιά 3 φορές. Οι σκοποί της συγκεκριμένης μελέτης ήταν:

- να διερευνηθεί η μεταξύ αυτών των 2 ψυχολόγων αξιοπιστία, γιατί το μέγεθος δείγματος της μελέτης έχει υπολογιστεί να είναι αρκετά μεγάλο και συνεπώς ένας ψυχολόγος δεν μπορεί να αξιολογήσει την επιθετικότητα όλων των παιδιών.
- Να διερευνηθεί η αξιοπιστία του κάθε ψυχολόγου.

Δεδομένου ότι το σκορ που υπολογίζεται από την συμπλήρωση του ερωτηματολογίου είναι ένα ποσοτικό μέγεθος, το κατάλληλο στατιστικό κριτήριο για τον έλεγχο της «εντός των ψυχολόγων αξιοπιστίας» και της «μεταξύ των ψυχολόγων αξιοπιστίας» είναι το intra-class correlation coefficient (ICC), το οποίο υπολογίζεται ακολουθώντας τα εξής ***βήματα***:

Analyse ⟶ **Scale** ⟶ **Reliability analysis**

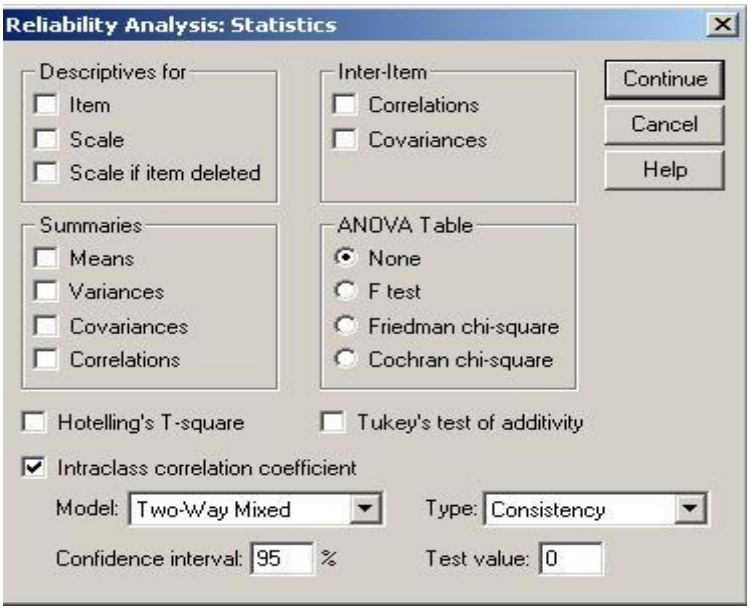
Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 16.3*.



Εικόνα 16.3: Πραγματοποίηση του ελέγχου «εντός των κριτών» και «μεταξύ των κριτών» αξιοπιστία**.**

Ας θεωρήσουμε ότι αρχικά θα υπολογίσουμε το ICC για την διερεύνηση της «εντός του κάθε ψυχολόγου αξιοπιστίας».

i. Ως «***Items***» δηλώνουμε τις 3 μεταβλητές του αρχείου στις οποίες έχουν καταχωρηθεί τα αποτελέσματα της συμπλήρωσης του ερωτηματολογίου για κάθε παιδί από τον πρώτο ψυχολόγο..
ii. Πατώντας το κουμπί επιλογών «***Statistics***» εμφανίζεται ένα νέο πλαίσιο διαλόγου που απεικονίζεται στην *Εικόνα 16.4*, από όπου ενεργοποιούμε την επιλογή «***intra-class correlation coefficient***».
iii. «***Continue***» & «***Ok***»



Εικόνα 16.4: Επιλογή πραγματοποίησης του intra-class correlation coefficient

iv. Ο *Πίνακας 16.6* παρουσιάζει τα αποτελέσματα της ανάλυσης. Πιο συγκεκριμένα, παρουσιάζεται η εκτίμηση του ICC. Διαπιστώνουμε ότι υπολογίζονται και παρουσιάζονται 2 διαφορετικές τιμές του ICC (Single measure & Average measure). Αν και δεν διαφέρουν πολύ οι τιμές αυτών των δύο ICC (π.χ. 0,997 & 0,999, αντίστοιχα), ο κάθε ένας από αυτούς θα πρέπει να χρησιμοποιείται ανάλογα με το αν στην κυρίως μελέτη θα πραγματοποιηθούν και οι 3 αξιολογήσεις της επιθετικότητας του παιδιού από τον ψυχολόγο (χρησιμοποιείται το average measure) ή θα πραγματοποιηθεί μόνο μία (χρησιμοποιείται το single measure). Βέβαια, ο πιο ευρέως χρησιμοποιούμενος ICC είναι ο single measure. Στην συγκεκριμένη περίπτωση, διακρίνουμε ότι ο 1^{ος} ψυχολόγος είναι εξαιρετικά αξιόπιστος αφού το ICC>0,9. Από τον ίδιο Πίνακα διαπιστώνουμε ότι πραγματοποιείται και ένας στατιστικός έλεγχος προκειμένου να ελέγξουμε αν η τιμή του ICC που υπολογίστηκε από τα δεδομένα μας είναι στατιστικά σημαντικά διάφορη του μηδενός ή όχι (F test with true value 0). Από το sig.<0,001 διαπιστώνουμε ότι η μηδενική υπόθεση (δηλαδή ότι το ICC=0) απορρίπτεται με πιθανότητα λάθους <0,001 που είναι πολύ μικρότερη από το επιτρεπτό όριο του 0,05.
Παρόμοιο, είναι και το συμπέρασμά μας, όσον αφορά στην αξιοπιστία του 2^{ου} ψυχολόγου, αφού και γι' αυτόν το ICC βρέθηκε να είναι >0,9 (*Πίνακας 16.7*).

Intraclass Correlation Coefficient

	Intraclass Correlation [a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,997[b]	,991	,999	944,692	9,0	18	,000
Average Measures	,999[c]	,997	1,000	944,692	9,0	18	,000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Πίνακας 16.6: Αποτελέσματα ελέγχου της «εντός του ψυχολόγου» αξιοπιστίας; Intra-class correlation coefficient για τον 1^{ο} ψυχολόγο.

Intraclass Correlation Coefficient

	Intraclass Correlation [a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,950[b]	,862	,986	57,946	9,0	18	,000
Average Measures	,983[c]	,949	,995	57,946	9,0	18	,000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Πίνακας 16.7: Αποτελέσματα ελέγχου της «εντός του ψυχολόγου» αξιοπιστίας; Intra-class correlation coefficient για τον 2^{ο} ψυχολόγο.

Ας υποθέσουμε τώρα ότι επιθυμούμε να ελέγξουμε την αξιοπιστία μεταξύ των 2 ψυχολόγων. Επαναλαμβάνουμε τα βήματα όπως αναφέρονται παραπάνω για τον έλεγχο της «εντός των ψυχολόγων» αξιοπιστίας. Σε αυτή την περίπτωση, όμως, ως «***Items***» δηλώνουμε εκείνες τις μεταβλητές στις οποίες έχουν καταχωρηθεί σκορ της εκτίμησης των 10 παιδιών από την 1^{η} αξιολόγηση για κάθε ένα ψυχολόγο. Από τον *Πίνακα 16.8* διαπιστώνουμε ότι και η «μεταξύ των ψυχολόγων» αξιοπιστία είναι εξαιρετικά υψηλή (ICC>0,9). Και σε αυτή την περίπτωση υπολογίζονται 2 διαφορετικές τιμές για το ICC. Το single measure πρέπει να χρησιμοποιείται όταν ο κάθε ψυχολόγος θα αξιολογήσει διαφορετικά παιδιά στην κυρίως μελέτη, ενώ το average measure χρησιμοποιείται αν στην κυρίως μελέτη και οι 2 ψυχολόγοι αξιολογήσουν όλα τα παιδιά με σκοπό να υπολογιστεί ο μέσος όρος των 2 αποτελεσμάτων.

Intraclass Correlation Coefficient

	Intraclass Correlation [a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	,905[b]	,664	,975	19,961	9,0	9	,000
Average Measures	,950[c]	,798	,988	19,961	9,0	9	,000

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

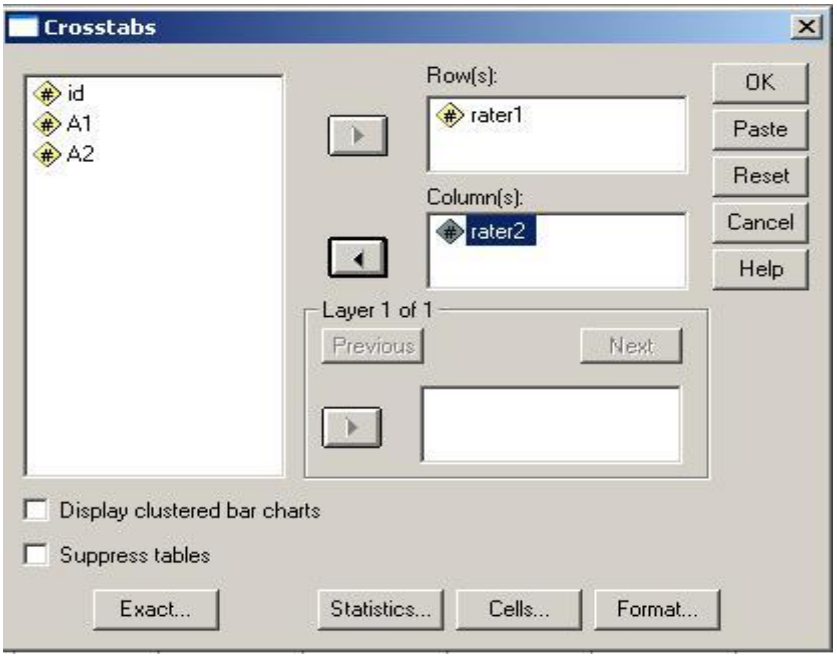
Πίνακας 16.8: Αποτελέσματα ελέγχου της «μεταξύ των ψυχολόγων» αξιοπιστίας; Intra-class correlation coefficient για την $1^{η}$ μέτρηση και των 2 ψυχολόγων.

16.3.2 Kappa measure of agreement

Ας υποθέσουμε ότι 2 γιατροί αξιολόγησαν τις μαστογραφίες 50 γυναικών προκειμένου να τις κατατάξουν ως ασθενείς με καρκίνο μαστού ή ως υγιείς. Αυτό πραγματοποιήθηκε στα πλαίσια μιας πιλοτικής μελέτης προκειμένου να ελέγξουμε αν υπάρχει αυξημένη αξιοπιστία μεταξύ των ερευνητών έτσι ώστε και οι 2 γιατροί να χρησιμοποιηθούν στην κυρίως μελέτη (να αξιολογήσουν μισές μαστογραφίες ο κάθε ένας). Δεδομένου ότι το χαρακτηριστικό που μετράμε είναι ποιοτικό (κατάταξη των γυναικών στην κατηγορία των ασθενών ή την κατηγορία των υγιών), το κατάλληλο στατιστικό κριτήριο για τον έλεγχο της «μεταξύ των ερευνητών» αξιοπιστίας είναι το «kappa measure of agreement» και για να υπολογιστεί τα απαραίτητα ***βήματα*** είναι τα ακόλουθα:

Analyse → Descriptive statistics → Crosstabs

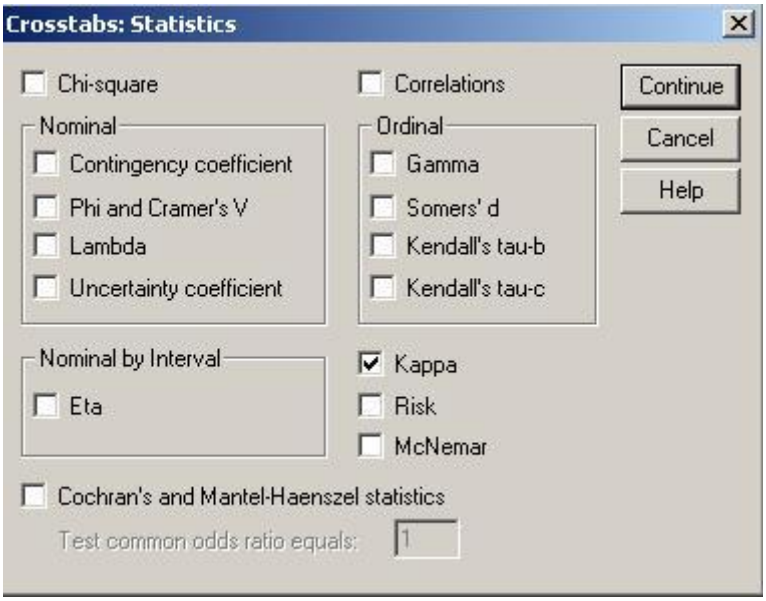
i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 16.5*.



Εικόνα 16.5: Πραγματοποίηση του «kappa measure of agreement»

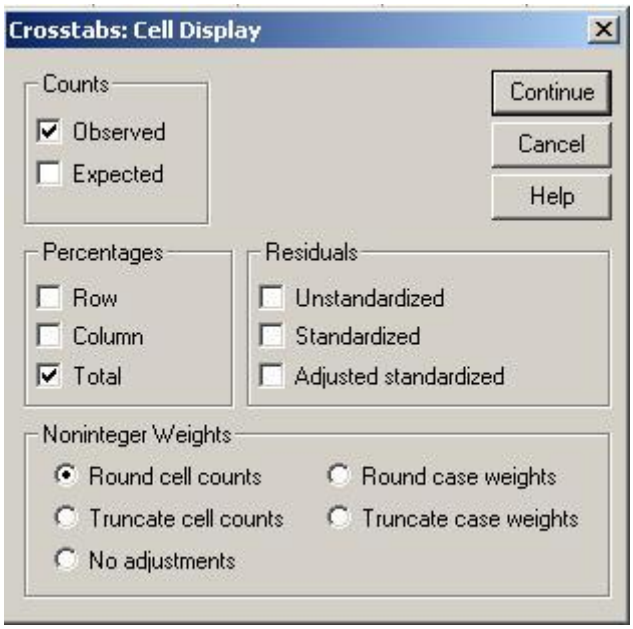
ii. Στο «***Rows***» και το «***Columns***» τοποθετούμε τις μεταβλητές στις οποίες είναι καταχωρημένα τα αποτελέσματα της αξιολόγησης της μαστογραφίας των 50 γυναικών από τους 2 γιατρούς.

iii. Πατώντας το κουμπί επιλογών «***Statistics***» ανοίγει το πλαίσιο διαλόγου της *Εικόνας 16.6*. Σε αυτό το πλαίσιο διαλόγου επιλέγουμε να υπολογιστεί το στατιστικό κριτήριο «kappa».



Εικόνα 16.6: Επιλογή του «kappa» από το πλαίσιο διαλόγου που ανοίγει πατώντας το κουμπί επιλογών «Statistics» της Εικόνας 16.5.

iv. Τέλος, πατώντας το κουμπί επιλογών «***Cells***» της *Εικόνας 16.5*, ανοίγει το πλαίσιο διαλόγου της *Εικόνας 16.7*, στο οποίο επιλέγουμε να εμφανιστεί στο output το ποσοστό συμφωνίας και διαφωνίας των 2 γιατρών, τσεκάροντας το «***Total***».



Εικόνα 16.7: Επιλογή εμφάνισης και των ποσοστών συμφωνία και διαφωνία των 2 κριτών, από το πλαίσιο διαλόγου που ανοίγει πατώντας το κουμπί επιλογών «Cells» της Εικόνας 16.5.

v. Στους *Πίνακες 16.9* & *16.10* παρουσιάζονται τα αποτελέσματα της παραπάνω ανάλυσης. Συγκεκριμένα, ο *Πίνακας 16.9* παρουσιάζει τα ποσοστά συμφωνίας και διαφωνίας των 2 γιατρών. Διαπιστώνουμε, λοιπόν, ότι οι 2 γιατροί

κατέταξαν από κοινού ως υγιείς το 38% των γυναικών και ως ασθενείς το 26% των γυναικών. Ενώ για τις υπόλοιπες γυναίκες (36%) οι 2 γιατροί διαφωνούσαν όσον αφορά στο συμπέρασμά τους αξιολογώντας τη μαστογραφία. Από τον *Πίνακα 16.10*, διαπιστώνουμε ότι το κ = 0,269, δηλαδή η συμφωνία μεταξύ των 2 κριτών δεν είναι ικανοποιητική. Στο ίδιο συμπέρασμα καταλήγουμε αξιολογώντας το p-value (Asymp sig. = 0,057), το οποίο να δείχνει ότι το κ δεν διαφέρει στατιστικά σημαντικά από το 0.

rater1 * rater2 Crosstabulation

			rater2		
			non-diseased	diseased	Total
rater1	non-diseased	Count	19	9	28
		% of Total	38,0%	18,0%	56,0%
	diseased	Count	9	13	22
		% of Total	18,0%	26,0%	44,0%
Total		Count	28	22	50
		% of Total	56,0%	44,0%	100,0%

Πίνακας 16.9: Ποσοστά συμφωνίας και διαφωνίας των 2 κριτών

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	,269	,137	1,906	,057
N of Valid Cases		50			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Πίνακας 16.10: Αποτελέσματα από την πραγματοποίηση του «kappa measure of agreement».

17. Αξιολόγηση διαγνωστικής ικανότητας δοκιμασιών.

17.1 Εισαγωγή

Ως διαγνωστικές δοκιμασίες χαρακτηρίζονται όλες οι κλινικές και εργαστηριακές διαγνωστικές εξετάσεις που έχουν ως σκοπό να διαπιστώσουν αν κάποιο άτομο πάσχει ή όχι από κάποια ασθένεια. Κατά κανόνα, οι διαγνωστικές δοκιμασίες **δεν είναι απόλυτα αξιόπιστες**, δεδομένου ότι βάσει των αποτελεσμάτων αυτών των δοκιμασιών είναι πιθανό κάποια υγιή άτομα να χαρακτηριστούν εσφαλμένα ως πάσχοντα, ενώ κάποια ασθενή άτομα να χαρακτηριστούν εσφαλμένα ως υγιή. Δεδομένου, ότι στόχος όλων είναι στην κλινική πράξη να χρησιμοποιούνται διαγνωστικές δοκιμασίες με ελαχιστοποιημένη πιθανότητα εμφάνισης των παραπάνω εσφαλμένων συμπερασμάτων, είναι απαραίτητο πριν αρχίσει να εφαρμόζεται κάποια δοκιμασία ευρέως να έχει ελεγχθεί η διαγνωστική ακρίβεια αυτής. Οι κατάλληλοι δείκτες για την αξιολόγηση της διαγνωστικής ακρίβειας μιας δοκιμασίας είναι: η **ευαισθησία (sensitivity)**, η **ειδικότητα (specificity)** και **η περιοχή κάτω από την καμπύλη λειτουργικών χαρακτηριστικών (ROC curve)**.

Ας πάρουμε την περίπτωση όπου ένας διαγνωστικός έλεγχος πραγματοποιείται (π.χ. μαστογραφία) και η απόφαση που ενδέχεται να ληφθεί αξιολογώντας τη μαστογραφία είναι: ασθενής ή υγιής. Προκειμένου να ελέγξουμε τη διαγνωστική ακρίβεια της μαστογραφίας, θα ακολουθήσουμε τα εξής βήματα:

1. Θα επιλέξουμε ένα δείγμα ασθενών και ένα δείγμα υγιών γυναικών.
2. Θα υποβληθούν όλες σε μαστογραφία, καθώς επίσης
3. Θα υποβληθούν και σε μία δοκιμασία, η οποία αποτελεί την πιο ακριβή δοκιμασία (gold standard) για την διάγνωση της νόσου (π.χ. βιοψία).
4. Χρησιμοποιώντας τα αποτελέσματα από τις 2 δοκιμασίες και θεωρώντας ως δεδομένο ότι η πραγματική κατάσταση της υγείας των γυναικών είναι αυτή που έχει προκύψει από την μαστογραφία, μπορεί να προκύψει η εξής κατηγοριοποίηση των γυναικών:
    - γυναίκες που έχουν θετικό αποτέλεσμα στην μαστογραφία και είναι όντως ασθενείς (βάσει βιοψίας),
    - γυναίκες που έχουν αρνητικό αποτέλεσμα στην μαστογραφία και είναι όντως υγιείς (βάσει βιοψίας),
    - γυναίκες που έχουν θετικό αποτέλεσμα στην μαστογραφία και δεν είναι ασθενείς (βάσει βιοψίας), και
    - γυναίκες που έχουν αρνητικό αποτέλεσμα στην μαστογραφία και δεν είναι τελικά υγιείς (βάσει βιοψίας).

Όπως γίνεται αντιληπτό οι δύο τελευταίες περιπτώσεις καλό θα ήταν να αποφεύγονται στη διαγνωστική πράξη διότι δημιουργούν πολλαπλά προβλήματα τόσο στα άτομα που έχουν εσφαλμένα κατηγοριοποιηθεί, όσο και στο ευρύτερο κοινωνικό τους περίγυρο. Λαμβάνοντας υπόψιν, λοιπόν, τα παραπάνω προκύπτει ο ακόλουθος τετράπτυχος πίνακας όπου περιέχει τα αποτελέσματα του διαγνωστικού ελέγχου/μαστογραφία (οριζόντια) και την πραγματική κατάσταση υγείας των ατόμων (κατακόρυφα), όπως αυτή έχει προκύψει από την βιοψία.

Όπως είναι φανερό ο αριθμός των ατόμων στα κελιά β και γ θα επιθυμούσαμε να είναι ελάχιστος.

		Κατάσταση υγείας (βάσει βιοψίας)		
Αποτελέσματα ελέγχου/μαστογραφίας		Νόσος		
		Παρούσα	*Απούσα*	Σύνολο
	Θετικός	α	β	α+β
	Αρνητικός	γ	δ	γ+δ
		α+γ	β+δ	N

17.1.1 Ευαισθησία και Ειδικότητα

Η **ευαισθησία** ορίζεται ως η ικανότητα του διαγνωστικού ελέγχου να εντοπίζει τα πραγματικά παθολογικά περιστατικά. Με βάση τον παραπάνω τετράπτυχο πίνακα η μαθηματική έκφραση της ευαισθησίας είναι:

Αριθμός ασθενών που έχουν θετικό έλεγχο/ Αριθμός παθολογικών ατόμων
$= \alpha / (\alpha + \gamma)$

Με λίγα λόγια εκφράζει την ικανότητα του ελέγχου να διακρίνει τα **ορθώς θετικά περιστατικά.** Συνήθως μια καλή ευαισθησία ορίζεται της τάξης του 80% και άνω. Μια τέτοια ευαισθησία υποδηλώνει ότι 80 από τους 100 παθολογικούς θα είχαν εντοπισθεί από το διαγνωστικό έλεγχο.

Ως **ψευδώς θετικό αποτέλεσμα** ορίζεται η λανθασμένη κατάταξη υγιών ατόμων ως ασθενών βάσει του διαγνωστικού ελέγχου. Η μαθηματική έκφραση είναι:

Αριθμός ασθενών που έχουν θετικό έλεγχο/ Αριθμός υγιών ατόμων
$= \beta / (\beta + \delta)$

Η **ειδικότητα** ορίζεται ως η ικανότητα της διαγνωστικής δοκιμασίας να ανιχνεύει τα πραγματικά φυσιολογικά περιστατικά. Η μαθηματική της έκφραση είναι:

Αριθμός υγιών που έχουν αρνητικό έλεγχο/ Αριθμός φυσιολογικών ατόμων
$= \delta / (\beta + \delta)$

Με λίγα λόγια εκφράζει τα **ορθώς αρνητικά περιστατικά.** Κατά αναλογία με την ευαισθησία καλή πιστότητα ορίζεται συνήθως της τάξης του 80% και άνω. Μια τέτοια ειδικότητα υποδηλώνει ότι το διαγνωστικό κριτήριο έχει καλή ικανότητα στη διάκριση των φυσιολογικών από τους παθολογικούς, μια και 80 από τους 100 φυσιολογικούς θα είχαν εντοπισθεί ορθά από τον προσυμπτωματικό διαγνωστικό έλεγχο.

Ως **ψευδώς αρνητικό αποτέλεσμα** ορίζεται η λανθασμένη κατάταξη ασθενών ατόμων ως υγιών βάσει του διαγνωστικού ελέγχου. Η μαθηματική έκφραση είναι:

Αριθμός υγιών που έχουν αρνητικό έλεγχο/ Αριθμός παθολογικών ατόμων
$= \gamma / (\alpha + \gamma)$

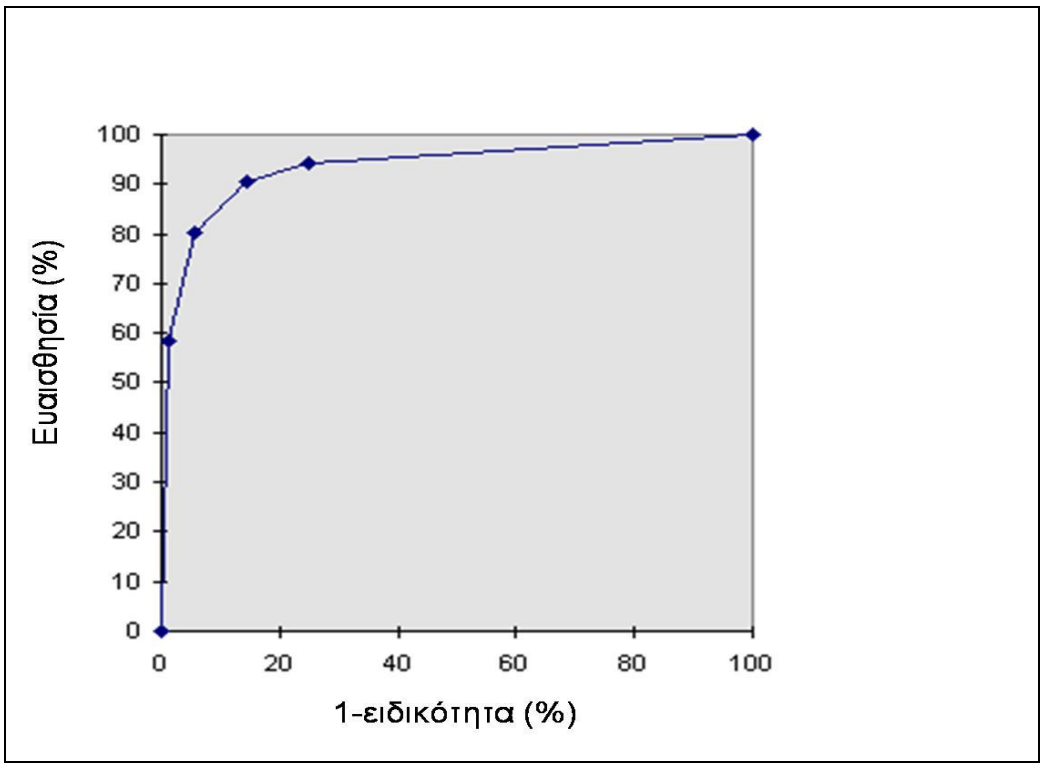
17.1.2 Καμπύλες λειτουργικών χαρακτηριστικών (ROC)

Οι ROC καμπύλες παρουσιάζουν ενδιαφέρον όταν η διαγνωστική δοκιμασία είναι ένας έλεγχος που μετράει κάποιο ποσοτικό χαρακτηριστικό (π.χ. κάποιος βιοχημικός δείκτης). Σε αυτήν την περίπτωση δεν μπορεί να υπολογιστεί μία μοναδική τιμή ευαισθησίας και ειδικότητας, παρά μόνο στην περίπτωση που έχει επιλεγεί ένα συγκεκριμένο διαχωριστικό όριο των τιμών του βιοχημικού δείκτη πάνω (ή κάτω) από το οποίο θεωρούνται παθολογικές τιμές (νόσος) και κάτω (ή πάνω) θεωρούνται φυσιολογικές τιμές (υγιείς).

Η ROC καμπύλη, λοιπόν, είναι ένας μαθηματικός τρόπος σύνδεσης της ευαισθησίας και της ειδικότητας. Πιο συγκεκριμένα, είναι ένα γράφημα που παρουσιάζει την ευαισθησία του διαγνωστικού ελέγχου σε σχέση με το (1- ειδικότητα) του ελέγχου για όλες τις πιθανές τιμές του βιοχημικού δείκτη (*Εικόνα 16.1*). Δηλαδή, για τον υπολογισμό της καμπύλης χρησιμοποιούνται όλα τα ζεύγη των τιμών:

{ορθώς θετικών = ευαισθησία} Χ {ψευδώς θετικών = 1 - ειδικότητα}

για όλες τις πιθανές οριακές τιμές διαχωρισμού του (ή των) διαγνωστικού κριτηρίου.



Εικόνα 17.1: Παράδειγμα ROC καμπύλης

Οι καμπύλες ROC χρησιμοποιούνται:

- για την αξιολόγηση και σύγκριση της διαγνωστικής ποιότητας των διαγνωστικών δοκιμασιών, καθώς επίσης και
- για την εύρεση του ιδανικού διαχωριστικού ορίου, δηλαδή της τιμής η οποία οριοθετεί το θετικό από το αρνητικό αποτελέσμα μιας δοκιμασίας.

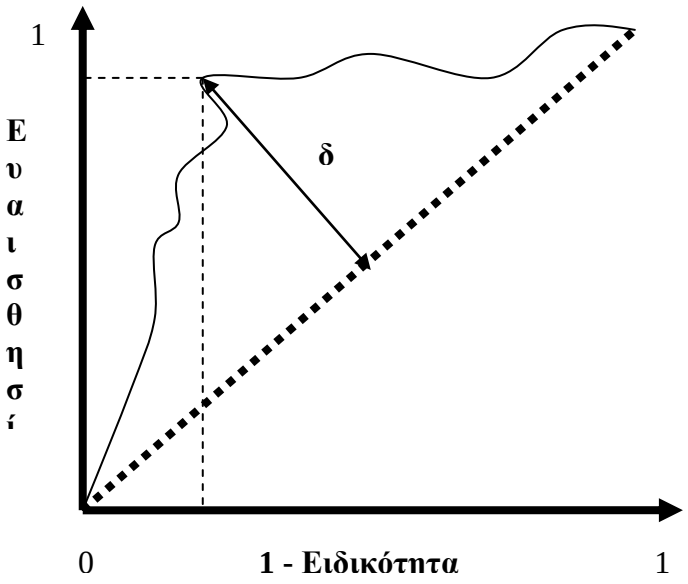
Για να αξιολογήσουμε μια διαγνωστική δοκιμασία ως άριστη είναι φανερό ότι επιθυμούμε η καμπύλη ROC να συγκλίνει στην πάνω αριστερή γωνία του τετραγώνου των 2 αξόνων. Στη θέση αυτή έχουμε ευαισθησία 100% και 1 - ειδικότητα = 0% ή ειδικότητα 100%. Αντιθέτως η κάτω δεξιά γωνία υποδηλώνει μια

πολύ κακή διαγνωστική δοκιμασία με ευαισθησία και ειδικότητα 0%. Άλλος ένας τρόπος για την αξιολόγηση μιας διαγνωστικής δοκιμασίας είναι με τον υπολογισμό του εμβαδού κάτω από την καμπύλη (Area Under the Curve, AUC). Μια τέλεια δοκιμασία έχει AUC = 1, ενώ μια κακή δοκιμασία έχει AUC<0.5.

Οι καμπύλες ROC χρησιμοποιούνται και για να απαντούν στο ερώτημα ποια από τις διαγνωστικές μεθόδους είναι καλύτερες. Όμως δεν πρέπει να αποτελούν μοναδική απάντηση στο ερώτημα περί διαγνωστικής ποιότητας των διαφόρων μεθόδων. Πολλές φορές η κλινική πράξη υποδεικνύει ποια διαγνωστική μέθοδος είναι η «καλύτερη» και «εφαρμόσιμη».

Η επιλογή του βέλτιστου σημείου του διαγνωστικού ελέγχου είναι κρίσιμης σημασίας για τις επιστήμες υγείας και αποτελεί πρόκληση για τη βιοστατιστική επιστήμη. Με τον όρο **διαχωριστικό σημείο** ή **κριτήριο θετικότητας** εννοούμε την τιμή στην κλίμακα του βιολογικού δείκτη (διαγνωστικού ελέγχου) πάνω από την οποία θεωρούνται τα παθολογικά περιστατικά και κάτω από αυτή θεωρούνται τα φυσιολογικά περιστατικά. Κατά καιρούς διάφορες μέθοδοι έχουν προταθεί για τον προσδιορισμό αυτού του σημείου.

Πρόσφατα προτάθηκε μία μέθοδος η οποία βασίζεται στην χρήση της ROC καμπύλης. Πιο συγκεκριμένα αφού υπολογισθεί η καμπύλη για κάθε δυνατό συνδυασμό ευαισθησίας και ειδικότητας, τότε το βέλτιστο σημείο είναι αυτό που έχει τη μεγαλύτερη απόσταση από την κύρια διαγώνιο. Με τον τρόπο αυτό επιτυγχάνεται η βέλτιστη ευαισθησία με την ελάχιστη απώλεια στην πιστότητα, και το αντίστροφο.



Η μέγιστη απόσταση ***δ*** επιτυγχάνεται από τη μεγιστοποίηση της συνάρτησης:

$$\delta = \frac{\|1 - ειδικότητα - ευαισθησία\|}{\sqrt{2}} \quad (1)$$

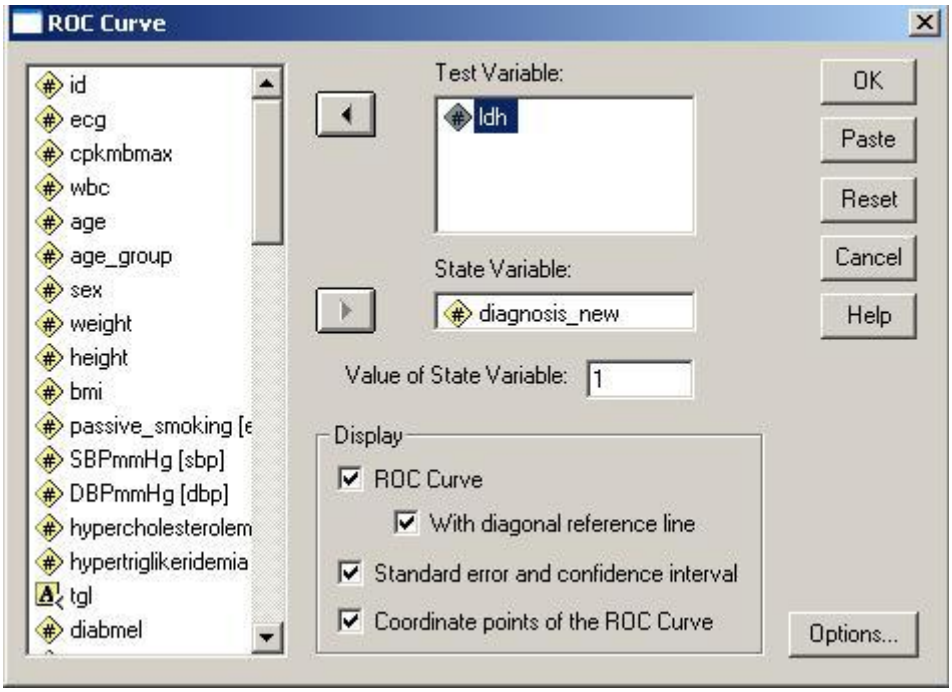
η οποία υπολογίζει την απόσταση των σημείων της καμπύλης ROC (1-ειδικότητα, ευαισθησία) από την κύρια διαγώνιο που είναι η ευθεία Y = χ.

17.2 Αξιολόγηση διαγνωστικής ικανότητας δοκιμασιών με τη χρήση του SPSS.

Ας υποθέσουμε ότι στόχος μας είναι να εξετάσουμε αν η LDH είναι καλός δείκτης διάγνωσης εμφράγματος του μυοκαρδίου έναντι ασταθούς στηθάγχης. Ο κατάλληλος τρόπος είναι να φτιάξουμε την κατάλληλη ROC καμπύλη και να ελέγξουμε αν η περιοχή κάτω από αυτήν διαφέρει σημαντικά από το 0,5 (κατώφλι που δείχνει πως δεν υπάρχει καμία διαγνωστική ικανότητα της LDH). Για να κάνουμε αυτό τον έλεγχο ακολουθούμε τα εξής βήματα:

Analyse → ROC curve

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 17.2*.



Εικόνα 17.2: Ανάλυση διαγνωστικής ικανότητας δοκιμασιών.

ii. Στο «***Test Variable***» τοποθετούμε τον δείκτη του οποίου τη διαγνωστική ικανότητα επιθυμούμε να ελέγξουμε (π.χ LDH)

iii. Στο «***State Variable***» τοποθετούμε την μεταβλητή στην οποία είναι καταχωρημένη η πραγματική κατάσταση της υγείας των ατόμων (diagnosis_new όπου 1:έμφραγμα μυοκαρδίου & 0: ασταθής στηθάγχη).

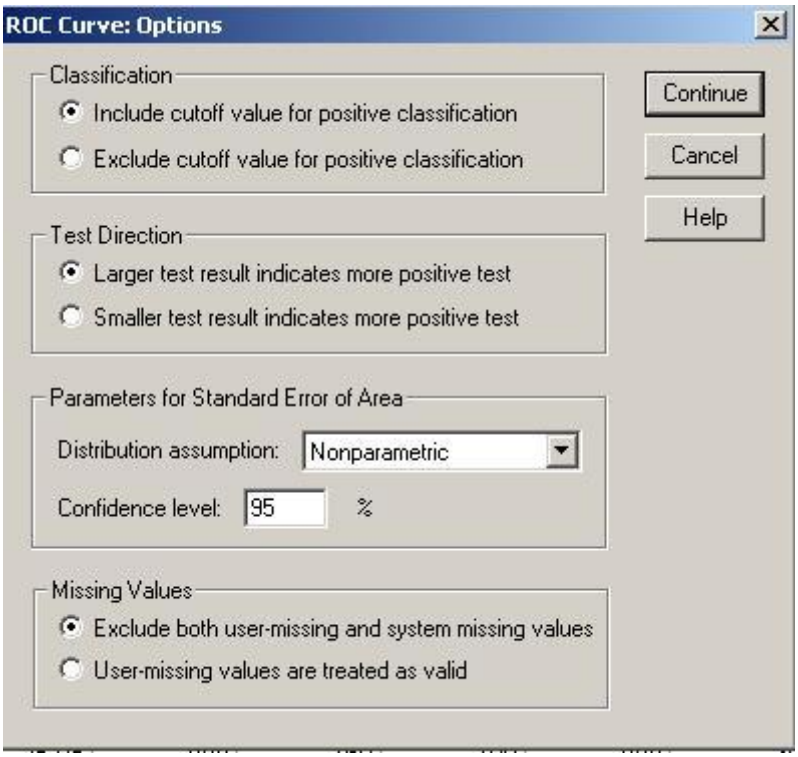
iv. Στο «***Value of State Variable***» πρέπει να δηλώσουμε την τιμή με την οποία είναι κωδικοποιημένη η νόσος για την οποία ελέγχουμε την διαγνωστική ικανότητα του δείκτη.

v. Στη συνέχεια, μπορούμε να επιλέξουμε στο output του SPSS να εμφανιστεί η διαγώνιος στο γράφημα (with diagonal reference line), να εμφανιστούν το τυπικό σφάλμα και το διάστημα εμπιστοσύνης της περιοχής κάτω από την ROC καμπύλη (standard error and confidence interval) και η ευαισθησία και

το (1-ειδικότητα) μαζί με το αντίστοιχο κατώφλι του δείκτη (coordinate points of the ROC curve).

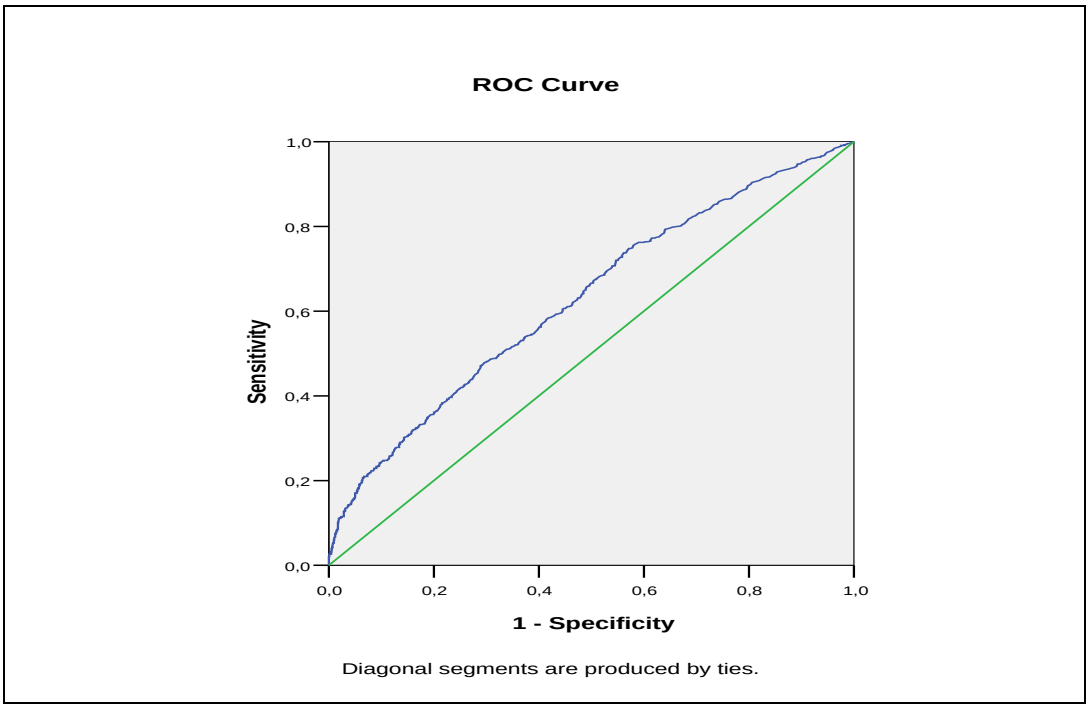
vi. Στο SPSS είναι προεπιλεγμένο ότι μεγαλύτερες τιμές του δείκτη του οποίου τη διαγνωστική ικανότητα εξετάζουμε υποδηλώνουν πιο πιθανή νόσο. Υπάρχουν, όμως και οι περιπτώσεις στις οποίες χαμηλότερες τιμές του δείκτη σχετίζονται με την εμφάνιση της νόσου. Σε αυτή την περίπτωση, οφείλουμε να το δηλώσουμε στο SPSS. Αυτό πραγματοποιείται πατώντας το κουμπί επιλογών «***Options***» όπου εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 17.3*. Εκεί επιλέγουμε, στο «Test Direction» το «Smaller test result indicates more positive test».

vii. «***Continue***» & «***Ok***»



Εικόνα 17.3: Επιλογές που εμφανίζονται πατώντας το κουμπί επιλογών «Options»

viii. Τα αποτελέσματα της παραπάνω ανάλυσης, παρουσιάζονται στον *Πίνακα 17.1* και στην *Εικόνα 17.4*. Συγκεκριμένα, από τον *Πίνακα 17.1* διαπιστώνουμε ότι η περιοχή κάτω από την ROC καμπύλη είναι 0,628, και ο στατιστικός έλεγχος δείχνει ότι διαφέρει στατιστικά σημαντικά από την τιμή 0,5 που υποδηλωνεί μηδενική ικανότητα της LDH να διαγιγνόσκει έμφραγμα του μυοκαρδίου, αφού sig.=0,00…1<0,05.



Εικόνα 17.4: ROC καμπύλη αναφορικά με την ικανότητα της LDH να διαγιγνόσκουμε έμφραγμα του μυοκαρδίου.

Area Under the Curve

Test Result Variable(s): ldh

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
,628	,013	,000	,601	,654

The test result variable(s): ldh has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Πίνακας 17.1: Περιοχή κάτω από την ROC καμπύλη αναφορικά με την ικανότητα της LDH να διαγιγνόσκουμε έμφραγμα του μυοκαρδίου.

Σημείωση*:** Αν επιθυμούμε να υπολογίσουμε το βέλτιστο διαχωριστικό όριο της LDH για την διάγνωση του εμφράγματος του μυοκαρδίου, οφείλουμε να μεταφέρουμε σε ένα αρχείο excel τα αποτελέσματα του Πίνακα «Coordinates of the Curve***» που εμφανίζεται πατώντας την κατάλληλη επιλογή της *Εικόνας 17.2* «***coordinate points of the ROC curve***» και να εφαρμόσουμε την συνάρτηση (1) της ενότητας 17.1. Η τιμή της LDH που θα αντιστοιχεί την ελάχιστη τιμή μεταξύ αυτών που θα προκείψουν εφαρμόζοντας την συνάρτηση θα είναι το βέλτιστο διαχωριστικό όριο

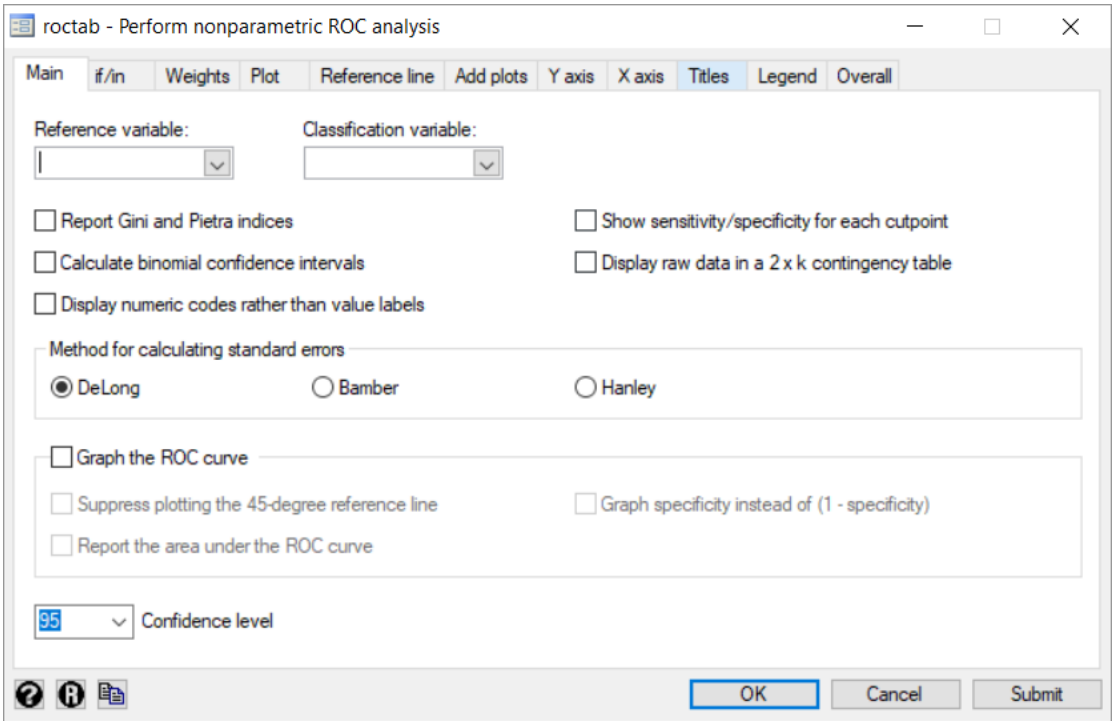
.

17.3 Αξιολόγηση διαγνωστικής ικανότητας δοκιμασιών με τη χρήση του STATA.

Για να κατασκευάσουμε την καπύλη ROC και να εφαρμόσουμε ταυτόχρονα τον κατάλληλο έλεγχο υπόθεσης ακολουθούμε τα εξής ***βήματα***:

Statistics → Epidemiology and related → ROC analysis → Nonparametric ROC analysis without covariates

i. Εμφανίζεται το πλαίσιο διαλόγου της *Εικόνας 17.5*.



Εικόνα 17.5: Ανάλυση διαγνωστικής ικανότητας δοκιμασιών.

ii. Στο «***Reference Variable***» τοποθετούμε την μεταβλητή στην οποία είναι καταχωρημένη η πραγματική κατάσταση της υγείας των ατόμων

iii. Στο «***Classification Variable***» τοποθετούμε τον δείκτη του οποίου τη διαγνωστική ικανότητα επιθυμούμε να ελέγξουμε

iv. Στο «***Graph the ROC curve***» έχουμε την δυνατότητα να κατασκευάσουμε και την καμπύλη ROC.

Στο STATA χρησιμοποιείται η εντολή ***roctab*** και η βασική της σύνταξη φαίνεται στην *Εικόνα 17.6*:

Syntax

`roctab` *refvar classvar* [*if*] [*in*] [*weight*] [, *options*]

roctab_options	Description
Main	
`lorenz`	report Gini and Pietra indices
`binomial`	calculate exact binomial confidence intervals
`nolabel`	display numeric codes rather than value labels
`detail`	show details on sensitivity/specificity for each cutpoint
`table`	display the raw data in a $2 \times k$ contingency table
`bamber`	calculate standard errors by using the Bamber method
`hanley`	calculate standard errors by using the Hanley method
`graph`	graph the ROC curve
`norefline`	suppress plotting the 45-degree reference line
`summary`	report the area under the ROC curve
`specificity`	graph sensitivity versus specificity
`level(#)`	set confidence level; default is `level(95)`
Plot	
`plotopts(`*plot_options*`)`	affect rendition of the ROC curve
Reference line	
`rlopts(`*cline_options*`)`	affect rendition of the reference line
Add plots	
`addplot(`*plot*`)`	add other plots to the generated graph
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***

`fweights` are allowed; see **[U] 11.1.6 weight**.

plot_options	Description
marker_options	change look of markers (color, size, etc.)
marker_label_options	add marker labels; change look or position
cline_options	change the look of the line

Εικόνα 17.6: Βασική σύνταξη της εντολής roctab