**Μάθημα: ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ**

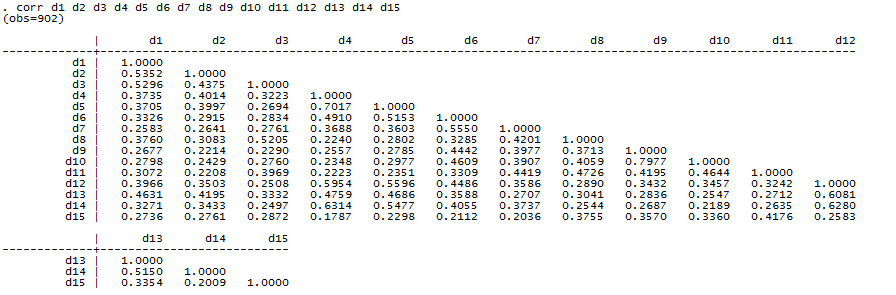
**Παραγοντική Ανάλυση ( Factor Analysis)**

**2ο Εργαστήριο- ΑΠΑΝΤΗΣΕΙΣ ME THN ΧΡΗΣΗ ΤΟΥ STATA**

1) α) Για τον έλεγχο της καταλληλότητας των δεδομένων ως προς την πραγματοποίηση παραγοντικής ανάλυσης αρχικά θα πρέπει να γίνει έλεγχος του πίνακα συσχετίσεων, ώστε να διαπιστωθεί ότι υπάρχουν αρκετές συσχετίσεις ( περίπου το 30-40% του συνόλου των στοιχείων του πίνακα συσχετίσεων). Έτσι θα εκτελέσουμε την εντολή

**corr d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15**

και στην συνέχεια προκύπτει το output του Stata



Από τα αποτελέσματα διαπιστώνουμε ότι έχουμε ικανό αριθμό συσχετίσεων μεταξύ των μεταβλητών με |r| > 0.3 ώστε να πραγματοποιήσουμε παραγοντική ανάλυση που θα καταλήξει σε ερμηνευτικά καλά αποτελέσματα.

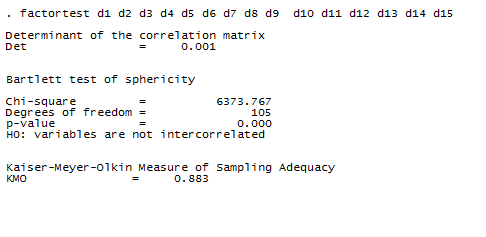
Στην συνέχεια θα προχωρήσουμε στον έλεγχο του κριτηρίου Kaiser-Meyer-Olkin και του ελέγχου σφαιρικότητας του Bartlett. To Kaiser-Meyer-Olkin Measure of Sampling Adequacy αποτελεί μέτρο συσχέτισης των δεδομένων του πίνακα συσχετίσεων ή διακύμανσης-συνδιακύμανσης της ανάλυσης και παίρνει τιμές από 0 έως 1, ενώ προτείνεται ως ελάχιστη τιμή που πρέπει να πάρει για να είναι αξιόπιστη η παραγοντική ανάλυση την τιμή 0,5, ενώ πολύ καλές θεωρούνται οι τιμές > 0,8. Επίσης ο έλεγχος σφαιρικότητας του Bartlett αποτελεί στατιστικό έλεγχο του οποίου μηδενική υπόθεση ( Ηο) είναι ότι πίνακας συσχετίσεων του πληθυσμού είναι ο μοναδιαίος πίνακας (δηλαδή οι τιμές της διαγωνίου είναι 1 και οι τιμές των υπόλοιπων στοιχείων είναι ίσες με 0 ή ο πίνακας διακύμανσης-συνδιακύμανσης είναι διαγώνιος ενώ η εναλλακτική υπόθεση (Ηα) ότι δεν είναι, δηλαδή ο πινακας συσχετίσεων δεν είναι ο μοναδιαίος και ο πίνακας διακύμανσης- συνδιακύμανσης δεν είναι διαγώνιος και άρα υπάρχουν συσχετίσεις μεταξύ των μεταβλητών της ανάλυσης.

To Stata δεν έχει εξ ορισμού εντολή για να πραγματοποιεί τους δυο προαναφερόμενους ελέγχους. Για τον σκοπό αυτό υπάρχει εντολή γραμμένη από χρήστη την οποία θα την κατεβάσουμε χρησιμοποιώντας την εντολή findit. Η εντολή είναι

**findit factortest**

και στην συνέχεια εγκαθιστούμε το αντίστοιχο addon πρόγραμμα και εκτελούμε την εντολή για τις μεταβλητές της ανάλυσής μας

**factortest d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15**



Ελέγχοντας το κριτήριο ΚΜΟ διαπιστώνουμε ότι η τιμή του είναι 0,883 > 0,5 άρα τα δεδομένα μας είναι αρκετά συσχετισμένα για να προχωρήσουμε στην παραγοντική ανάλυση. Επίσης ο έλεγχος του p value ( p < 0,001 ) δείχνει ότι δεν αποδεχόμαστε την μηδενική υπόθεση σε επίπεδο σημαντικότητας α = 0,05, άρα ο πίνακας συσχετίσεων του πληθυσμού των ατόμων της μελέτης δεν είναι ο μοναδιαίος πίνακας και μπορούμε να συμπεράνουμε συνολικά ότι τα δεδομένα μας είναι επαρκώς συσχετισμένα μεταξύ τους για να προχωρήσουμε στην εφαρμογή της παραγοντικής ανάλυσης.

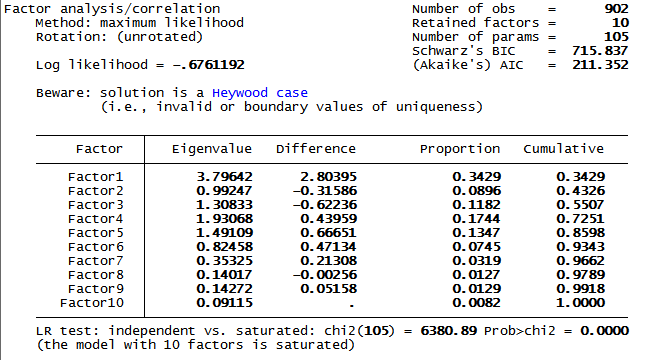
Η παρουσία πολυσυγγραμικότητας ( μεταβλητών με μεγάλη συσχέτιση μεταξύ τους) δημιουργεί προβλήματα στην παραγοντική ανάλυση καθώς δεν επιτρέπει να εκτιμηθεί η μοναδική συνεισφορά των υψηλά συσχετιζόμενων μεταβλητών με τους παράγοντες. Μία πρώτη αντιμετώπιση είναι η διαπίστωση μεταβλητών που έχουν μεγάλο συντελεστή συσχέτισης μεταξύ τους ( r >0.9) και η απομάκρυνσή τους από την ανάλυση. Ένας δεύτερος τρόπος να διαπιστωθεί η παρουσία πολυσυγγραμικότητας στα δεδομένα της ανάλυσης είναι ο έλεγχος της ορίζουσας του πίνακα συσχετίσεων της ανάλυσης. Τιμές ορίζουσας του πίνακα συσχετίσεων μεγαλύτερες του 0,000001 σημαίνουν ότι τα δεδομένα μας δεν παρουσιάζουν πολυσυγγραμικότητα και μπορεί να εφαρμοστεί η παραγοντική ανάλυση. Στην ανάλυσή η τιμή της ορίζουσας είναι 0,01 άρα τα δεδομένα μας δεν παρουσιάζουν πολυσυγγραμικότητα.

2) Στην συνέχεια θα προχωρήσουμε στην πραγματοποίηση της παραγοντικής ανάλυσης στις μεταβλητές d1-d15. Η εντολή που πραγματοποιεί παραγοντική ανάλυση στις μεταβλητές είναι η

**factor d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15, ml**

όπου η υποεπιλογή ml αποτελεί την επιλογή για να γίνει η εκτίμηση των παραμέτρων του παραγοντικού υποδείγματος της ανάλυσης με την μέθοδο της μέγιστης πιθανοφάνειας. Με την μέθοδο της μεγίστης πιθανοφάνειας θεωρείται ότι ισχύει η υπόθεση της πολυμεταβλητής κανονικότητας των δεδομένων μας (υπόθεση της κανονικότητας της κατανομής των σφαλμάτων και των παραγόντων της ανάλυσης). Με την εφαρμογή της μεθόδου της μεγίστης πιθανοφάνειας θεωρείται ότι οι μεταβλητές μας είναι συνεχείς ή δίτιμες. Στο παράδειγμα, λόγω της αποδοχής ότι σε ψυχομετρικά τέστ, οι διατάξιμες κατηγορικές μεταβλητές με άνω των 5 κατηγοριών μπορούν να θεωρηθούν ότι προέρχονται από την κανονική κατανομή, είναι δόκιμο να προχωρήσουμε στην ανάλυσή τους.

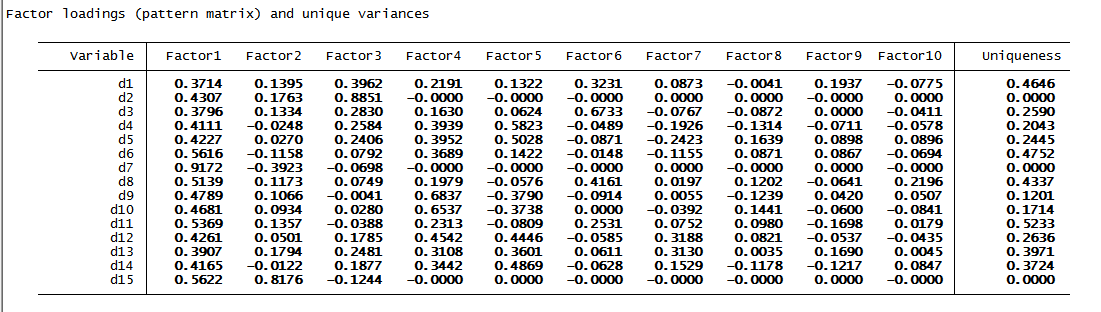
Στην συνέχεια προκύπτει το παρακάτω output



Αρχικά το STATA μας παρουσιάζει κάποια περιγραφικά χαρακτηριστικά για την ανάλυση, όπως την μέθοδο εξαγωγής των παραγόντων, το αν έχει γίνει περιστροφή ή όχι, καθώς και τις τιμές των στατιστικών κριτηρίων BIC και AIC, τα οποια χρησιμοποιούνται για την επιλογή του βέλτιστου υποδείγματος που εφαρμόζει στα δεδομένα. Όσο μικρότερη είναι η τιμή των κριτηρίων, τόσο καλύτερη είναι η εφαρμογή του υποδείγματος στα δεδομένα της μελέτης.

Στον πρώτο πίνακα και στην πρώτη στήλη εμφανίζεται η ιδιοτιμή του κάθε παράγοντα, η οποία αντιστοιχεί στο ποσοστό της αρχικής μεταβλητότητας των δεδομένων που εκφράζει ο κάθε παράγοντας. Στην δεύτερη στήλη εμφανίζεται η διαφορά της ιδιοτιμής του κάθε παράγοντα από τον επόμενο, στην τρίτη στήλη εμφανίζεται το ποσοστό της αρχικής μεταβλητότητας που εκφράζει ο κάθε παράγοντας και στην τέταρτη στήλη το αθροιστικό ποσοστό. Θα πρέπει να τονιστεί η διαφορά που έχει η μέθοδος εξαγωγής μεγίστης πιθανοφάνειας από την ανάλυση σε κύριες συνιστώσες καθώς η σειρά που εξάγονται οι παράγοντες δεν είναι κατά φθίνουσα τιμή ιδιοτιμής. Οι παράγοντες με την μεγαλύτερη ερμηνευτική ικανότητα είναι ο παράγοντας 1,3, 4, 5. Ο LR test του μοντέλου χωρίς κάποια δομή έναντι του κορεσμένου μοντέλου, δηλαδή του μοντέλου που βάση τις υποθέσεις εφαρμογής έχει τους μέγιστους παράγοντες (δηλαδή 10). To STATA μας σημειώνει ότι η λύση που μας δίνει είναι Heywood, δηλαδή ο επαναληπτικός αλγόριθμος λόγω των στοχαστικών προσεγγίσεων που έχουν γίνει για τον υπολογισμό της τιμής του χι-τετράγωνο και των βαθμών ελευθερίας στην εκτιμηση της πιθανοφάνειας λειτουργεί με την υπόθεση ότι υπάρχει λύση. οδηγώντας έτσι στην 100% ερμηνεία της μεταβλητότητας μιας ή περισσοτέρων μεταβλητών από το υπόδειγμα. Παρόλα αυτά, επειδή η μέθοδος της μέγιστης πιθανοφάνειας έχει την τάση να δίνει Heywood λύσεις, τα αποτελέσματα αυτών ΠΑ θεωρούνται από την βιβλιογραφία ότι δίνουν πληροφορία στον αναλυτή ότι το μοντέλο δεν προσδιορίζεται σωστά από την ΠΑ.

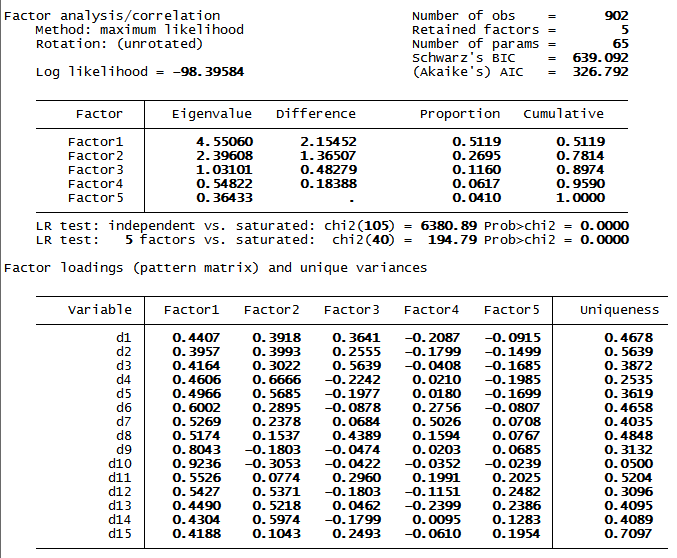
Στον δεύτερο πίνακα απεικονίζονται οι επιβαρύνσεις ( Loadings) των παραγόντων όπως εκτιμήθηκαν από την παραγοντική ανάλυση. Έτσι εκτιμήθηκαν 10 νέοι παράγοντες των οποίων οι επιβαρύνσεις ( οι οποίοι αποτελούν συντελεστές που εκφράζουν την συσχέτιση του κάθε παράγοντα με την κάθε μεταβλητή) απεικονίζονται στις δέκα επόμενες στήλες ενώ στην τελευταία στήλη απεικονίζεται η **ιδιαιτερότητα** της κάθε αρχικής μεταβλητής, η οποία είναι εκφράζει το ποσοστό της μεταβλητότητας της κάθε αρχικής μεταβλητής που δεν ερμηνεύεται από το παραγοντικό υπόδειγμα. Η **εταιρικότητα** της κάθε αρχικής μεταβλητής εκφράζει το ποσοστό της μεταβλητότητας της κάθε μίας αρχικής μεταβλητής που εκφράζεται από το παραγοντικό υπόδειγμα και υπολογίζεται αν αφαιρέσουμε από την μονάδα την ιδιαιτερότητα της κάθε μεταβλητής. Έτσι προκύπτει ο παρακάτω πίνακας



Έτσι η ιδιαιτερότητα για την πρώτη μεταβλητή είναι 46,4% και η εταιρικότητα είναι (1-0,464=0,536 \*100=53,6%), δηλαδή το παραγοντικό υπόδειγμα εκφράζει το 53,6% της μεταβλητότητας της μεταβλητής d1 και αφήνει ανερμήνευτο το 46,4% της μεταβλητότητάς της. Για την μεταβλητή d2 το υπόδειγμα, όπως δόθηκε ως λύση από την ανάλυση, θεωρείται ότι ερμηνεύει το 100% της μεταβλητότητας του, κάτι που δεν ισχύει και είναι αριθμητικό αποτέλεσμα που προκύπτει για τον τερματισμό του αλγορίθμου. Αν διαπιστώναμε μικρή εταιρικότητα σε κάποια μεταβλητή που μας ενδιαφέρει ιδιαίτερα θα χρειαζόταν να προσθέσουμε έναν ή περισσότερους παράγοντες στην ανάλυσή μας ώστε να την αυξήσουμε.

3) Η εντολή που θα χρησιμοποιήσουμε είναι παρακάτω, με την προσθήκη της επιλογής factor ()

**factor d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15, ml factor(5)**

****

Παρατηρούμε ότι το νέο output δεν εμφανίζει το μήνυμα ότι έχουμε Heywood case, αρα το υπόδειγμα έχει εκτιμηθεί σωστά. Το κριτήριο BIC είναι μικρότερο τώρα από το κορεσμένο υπόδειγμα αλλά οι μοναδικότητες έχουν αυξηθεί γενικά σε όλες τις αρχικές μεταβλητές, γεγονός που τεκμηριώνει ότι το υπόδειγμα δεν ερμηνεύει αρκετή πληροφορία. Παρόλα αυτά επειδή έχουμε ορθή εκτίμηση του υποδείγματος και το BIC κριτήριο είναι μικρότερο, μπορούμε να θεωρήσουμε το παραπάνω υπόδειγμα ως το βέλτιστο για την ανάλυσή μας.

4) Από τον παραπάνω πίνακα μπορούμε να γράψουμε το στατιστικό υπόδειγμα των δύο πρώτων παραγόντων ( οι συντελεστές του υποδείγματος είναι οι επιβαρύνσεις του πίνακα 2)

d1= 0.4407F1+0.3918F2+0.3641F3 -0.2087F4-0.0915F5

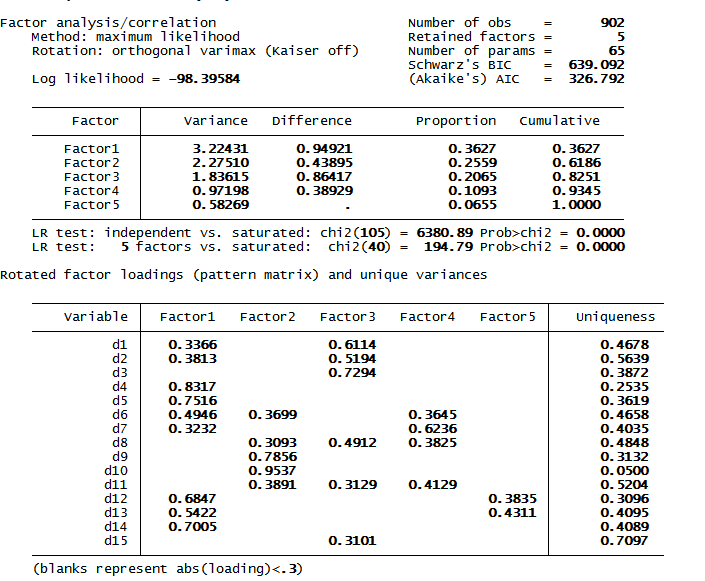
d2= 0.3957F1+0.3993F2+0.2555F3-0.1799F4-0.1499F5

5) Οι παράμετροι του παραγοντικού υποδείγματος χωρίς περιστροφή εκτιμήθηκαν στην αρχική εντολή της ανάλυσης. Για να εφαρμόσουμε περιστροφή, θα χρησιμοποιήσουμε την εντολή **rotate** μετά την εκτέλεση της εντολής **factor** η οποία ακολουθείται από κόμμα και την επιλογή της περιστροφής του πίνακα των επιβαρύνσεων. Αντίστοιχα με την ανάλυση σε κύριες συνιστώσες, έχουμε περιστροφές ορθογώνιες ( οι οποίες κατασκευάζουν νέους παράγοντες ασυσχέτιστους με τους αρχικούς ) και μη ορθογώνιες ( οι οποίες κατασκευάζουν νέους παράγοντες που συσχετίζονται με τους αρχικούς ). Με την ορθογώνια περιστροφή **varimax** ελαχιστοποιούνται οι επιβαρύνσεις με μεγάλες τιμές σε κάθε παράγοντα ενώ με την ορθογώνια περιστορφή **quartimax** ελαχιστοποιείται ο αριθμός των παραγόντων με μεγάλες επιβαρύνσεις

Η εντολή για την εφαρμογή της ορθογώνιας περιστροφής v**arimax** έιναι

**rotate, varimax blanks(0.3)**

με την επιλογή blanks() επιλέγουμε από ποια τιμή και πάνω να εμφανίζονται επιβαρύνσεις στον πίνακα των αποτελεσμάτων και προκύπτει το αντίστοιχο output



ε) Η ερμηνεία των παραγόντων γίνεται με τον έλεγχο των στοιχείων του πίνακα μετά την περιστροφή και τον έλεγχο της αντιστοιχείας των μεταβλητών με τις μεγαλύτερες επιβαρύνσεις σε κάθε παράγοντα. Έτσι έχουμε

* Ο πρώτος παράγοντας συσχετίζεται περισσότερο με τις μεταβλητές d4, d5, d14 και d15 οι οποίες εκφράζουν την ικανοποίηση που αντλεί το άτομο από το κοινωνικό του περιβάλλον ( οικογένεια, σύντροφο, κοινωνικές σχέσεις) και μπορεί να χαρακτηριστεί ως « Θετικό κοινωνικό περιβάλλον»
* Ο δεύτερος παράγοντας συσχετίζεται περισσότερο με τις μεταβλητές d9, d10 και d11, οι οποίες έχουν σχέση με την εργασία και μπορούμε να θεωρήσουμε ότι εκφράζει την « Ικανοποίηση από την εργασία»
* Ο τρίτος παράγοντας συσχετίζεται περισσότερο με τις μεταβλητές d1, d2, d3, d13 και d15 που είναι σχετικές με συμπτώματα άγχους και κατάθλιψης και μπορούμε να θεωρήσουμε ότι αποτελεί τον παράγοντα του « Άγχους-Κατάθλιψης»
* O τέταρτος παράγοντας συσχετίζεται περισσότερο με την ικανοποίηση από την εργασία (d6,d7,d8) και θεωρούμε ότι αποτελεί τον παράγοντα «Εργασιακο περιβάλλον»
* Ο πέμπτος παράγοντας συσχετίζεται περισσότερο με την ικανοποίηση από την σχέση με τον σύντροφο (d12 και d13 ερωτήσεις) και θεωρούμε ότι αποτελεί τον παράγοντα «Σχέση με τον σύντροφο»

στ) Η αποθήκευση του score των παραγόντων ως νέων μεταβλητών γίνεται με την εντολή **predict** μετά την εκτέλεση της παραγοντικής ανάλυσης, ακολουθούμενη από την ονομασία των νέων μεταβλητών που θα αποθηκεύσει το Stata.

Έτσι μετά την ανάλυση εκτελούμε την εντολή

**predict x1 x2 x3 x4 x5**

και παρατηρούμε ότι έχουν προστεθεί πέντε νέες μεταβλητές στην λίστα των μεταβλητών, που αντιστοιχούν στα score των παραγόντων, όπως εκτιμήθηκαν από την ΠΑ.

**logit cvd\_event x1 x2 x3 x4 x5, or**

και προκύπτει το αντίστοιχο output

