

Παραγοντική ανάλυση και SPSS

Πρόχειρες σημειώσεις

ΣΤΟΧΟΣ ΠΑΡΑΓΟΝΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

Η εύρεση της ύπαρξης κοινών παραγόντων ανάμεσα σε μία ομάδα μεταβλητών.

Τι επιτυγχάνεται?

1. Μείωση της διάστασης του προβλήματος.
2. Εξήγηση των συσχετίσεων που υπάρχουν στα δεδομένα.
3. Δημιουργία νέων μεταβλητών που ίσως ερμηνεύουν μη μετρήσιμες έννοιες!!!
4. Δημιουργία ενός συνόλου παραγόντων για να χρησιμοποιηθούν ως ασυσχέτιστες μεταβλητές (διόρθωση προβλήματος πολυσυγγραμμικότητας)
5. Εξακρίβωση ή επιβεβαίωση μιας κλίμακας.

Μειονεκτήματα

1. Υποθέσεις μη ρεαλιστικές και ελέγξιμες.
2. Δεν δίνει μοναδική λύση.
3. Οι παράγοντες που προκύπτουν δεν έχουν μοναδική ερμηνεία.
4. Ο αριθμός των παραγόντων δεν είναι αυστηρά – μαθηματικά καθορισμένος.

Είδη παραγοντικής ανάλυσης

1. Exploratory: Προσπαθεί να ανακαλύψει την ύπαρξη παραγόντων σε ένα μεγάλο σύνολο μεταβλητών.
2. Confirmatory: Προσπαθεί να ανακαλύψει αν ο αριθμός των παραγόντων καθώς και η σύνθεση τους επιβεβαιώνει τα αναμενόμενα από τη θεωρία.

Βήματα παραγοντικής ανάλυσης

1. Καταλληλότητα δεδομένων και μεταβλητών

- Θα πρέπει να υπάρχουν ικανοποιητικές συσχετίσεις για τη διεξαγωγή παραγοντικής ανάλυσης. Αν τα δεδομένα είναι ασυσχέτιστα δεν έχει νόημα να συνεχίσουμε. Απόλυτες τιμές του συντελεστή συσχέτισης μεγαλύτερες του 0.4 είναι ικανοποιητικές.
- Η ορίζουσα του δειγματικού πίνακα συσχέτισης είναι μεγαλύτερη από 0.00001
- Έλεγχος της υπόθεσης της σφαιρικότητας (Bartlett's test of sphericity)

$$H_0 : R = I_p,$$

είτε με το στατιστικό

$$L = - \left[-n - \frac{1}{6(2p+5)} \right] \ln |R| \stackrel{H_0}{\sim} X_{p(p-1)/2}^2,$$

είτε με το

$$L = - \left[-n - \frac{(2p^2 + p + 2)}{6p} \right] \left[\ln |S| - \ln \prod_{i=1}^p s_i^2 \right] \stackrel{H_0}{\sim} X_{p(p-1)/2}^2.$$

- Εξέταση της πολυσυγγραμμικότητας.

Χρησιμοποιώντας το στατιστικό των Kaiser Meyer Olkin (KMO) εξετάζεται αν τα δεδομένα μας είναι κατάλληλα για παραγοντική ανάλυση. Το στατιστικό αυτό παίρνει τιμές στο διάστημα $[0,1]$. Αν $KMO > 0.6$ συνεχίζουμε την παραγοντική ανάλυση.

- Κοιτούμε τον δείκτη Measure of Sampling Adequacy για να αποφανθούμε αν μία μεταβλητή είναι κατάλληλη για να χρησιμοποιηθεί στην ανάλυση. Τιμές μεγαλύτερες του 0.5 μας υποδεικνύουν την καταλληλότητα.

2. Καθορισμός ή υπολογισμός του αριθμού των παραγόντων

Kaiser criterion: ο αριθμός των παραγόντων είναι ίσος με τον αριθμό των ιδιοτιμών του πίνακα συσχέτισης που είναι μεγαλύτερες από τη μονάδα.

Scree plot test: ο αριθμός των παραγόντων προσδιορίζεται από το γράφημα των ιδιοτιμών του πίνακα συσχέτισης σε φθίνουσα σειρά. Είναι ίσος με το πλήθος των ιδιοτιμών πριν την τελευταία σημαντική πτώση του μεγέθους της ιδιοτιμής.

(καλό για $n > 200$)

3. Εκτίμηση των παραγόντων

Μεθοδολογίες εκτίμησης των παραγόντων

- Principal component analysis (προτείνεται)

Η πιο ευρέως χρησιμοποιούμενη μέθοδος. Αναζητεί το γραμμικό συνδυασμό των μεταβλητών έτσι ώστε να επεξηγείται από τους παράγοντες η μέγιστη μεταβλητότητα των μεταβλητών.

- Maximum likelihood factoring

προϋποθέτει την ύπαρξη πολυδιάστατης κανονικής κατανομής. Ίσως οδηγεί σε πολλούς παράγοντες άλλα όχι περισσότερους από το ακέραιο μέρος του $p/2$, δηλαδή από $[p/2]$.

- Principal axis factoring παραλλαγή της PCA. Χρησιμοποιείται όταν ο πίνακας συσχέτισης είναι ιδιάζων.

- Image factoring χρησιμοποιείται κυρίως στη θεωρία εικόνας.

- Alpha factoring προσπαθεί να μεγιστοποιήσει την αξιοπιστία (δηλαδή το alpha του Cronbach, βλέπε Ανάλυση Αξιοπιστίας, Reliability Analysis)
- Unweighted least squares factoring ελαχιστοποιεί το άθροισμα τετραγώνων μεταξύ παρατηρούμενων και εκτιμώμενων πινάκων συσχέτισης.
- Generalizes least squares factoring: παραλλαγή της προηγούμενης.

4. Περιστροφή

Χρησιμοποιείται για να γίνουν τα αποτελέσματά μας πιο ερμηνεύσιμα. Ελπίζουμε ότι οι επιβαρύνσεις κάποιων παραγόντων θα είναι μεγάλες σε απόλυτη κλίμακα μόνο για κάποιες μεταβλητές και έτσι βλέποντας ποιες μεταβλητές εξαρτώνται με ποιους παράγοντες μπορούμε να τους ερμηνεύσουμε.

Μέθοδοι περιστροφής

- Varimax: ελαχιστοποιεί αριθμό μεταβλητών που έχουν μεγάλες επιβαρύνσεις για κάθε παράγοντα.
- Quartimax: ελαχιστοποιεί αριθμό παραγόντων που εξηγούν μία μεταβλητή
- Equimax: συνδυασμός των άνω
- Direct oblimin rotation και Promax οι οποίες δίνουν συσχετισμένους παράγοντες.

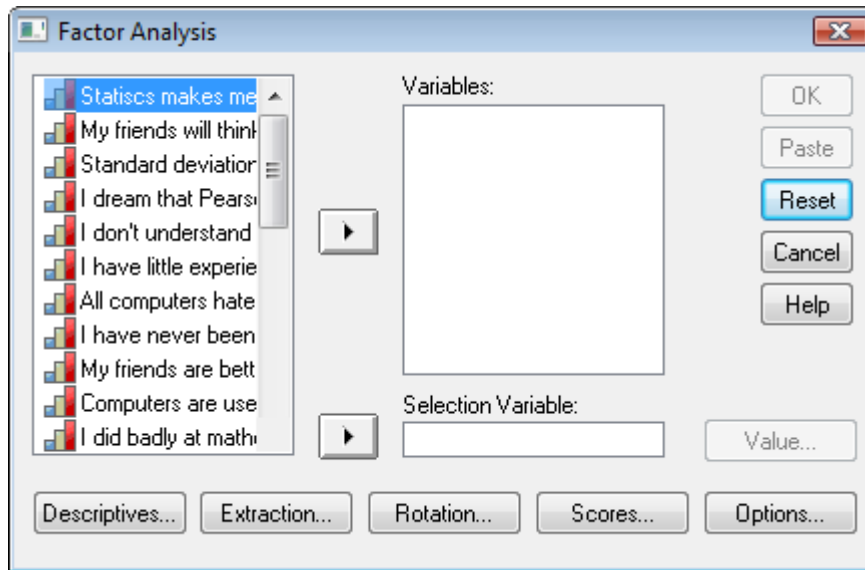
5. Ερμηνεία αποτελεσμάτων

Ας δούμε πως υλοποιούνται όλα τα παραπάνω μέσω του ακόλουθου παραδείγματος

Παράδειγμα

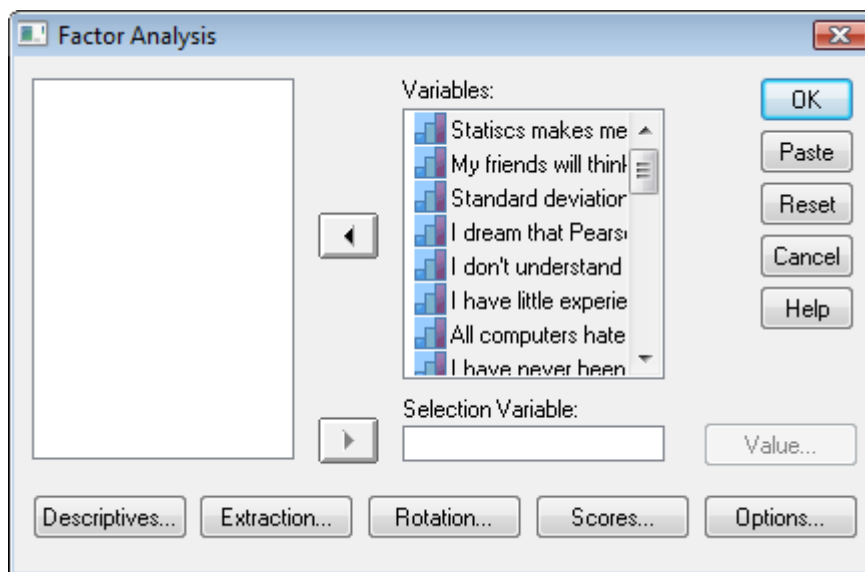
Στο αρχείο SAQ.sav καταγράφονται οι απαντήσεις σε 23 πιθανές ερωτήσεις για το άγχος των φοιτητών για το SPSS. Το ενδιαφέρον αρχικά επικεντρώνεται στην εύρεση πιθανών παραγόντων αιτιολόγησης του άγχους. (Παράδειγμα Andy Field (2000)).

Από το κεντρικό παράθυρο διαλόγου του Data View επιλέγουμε Analyze Data Reduction Factor. Στο νέο παράθυρο διαλόγου που προκύπτει:



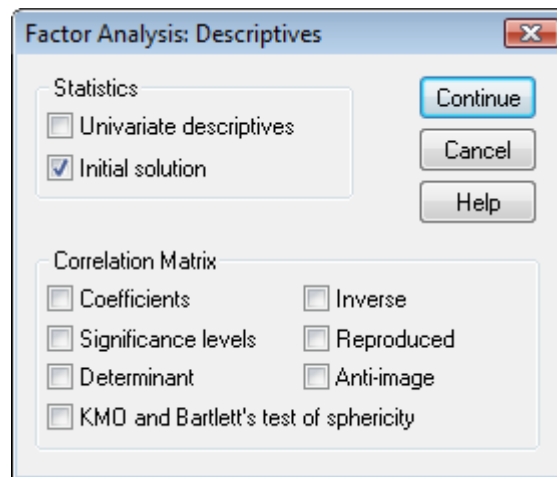
στο πεδίο Variables εισάγουμε τις μεταβλητές για τις οποίες θέλουμε να εξετάσουμε την ύπαρξη κοινών παραγόντων.

Επομένως έχουμε:



Στο πλαίσιο Selection Variable υπεισέρχεται η μεταβλητή εκείνη ως προς καθορισμένη τιμή της οποίας (υπεισέρχεται στο πεδίο Value) θέλουμε να διεξαχθεί η παραγοντική ανάλυση.

Από το πλαίσιο Descriptives έχουμε τις ακόλουθες επιλογές:



Univariate descriptives: μέση τιμή, τυπική απόκλιση και πλήθος πειραματικών μονάδων για κάθε μεταβλητή.

Initial solution μας δίνονται οι αρχικές ιδιοτιμές και το ποσοστό της μεταβλητότητας που εξηγείται.

Coefficients: ο πίνακας συσχέτισης (σχόλιο: επιθυμητό είναι οι απόλυτες τιμές των συντελεστών συσχέτισης να είναι μεγαλύτερες του 0.3).

Significance level: δίνονται οι p-τιμές για τον έλεγχο της υπόθεσης ότι ο πληθυσμιακός συντελεστής συσχέτισης κάθε ζεύγους μεταβλητών (σχόλιο: υπενθυμίζεται ότι θέλουμε οι μεταβλητές να είναι συσχετισμένες).

Determinant: η ορίζουσα του πίνακα συσχέτισης. Πρέπει να είναι μεγαλύτερη από 0.00001.

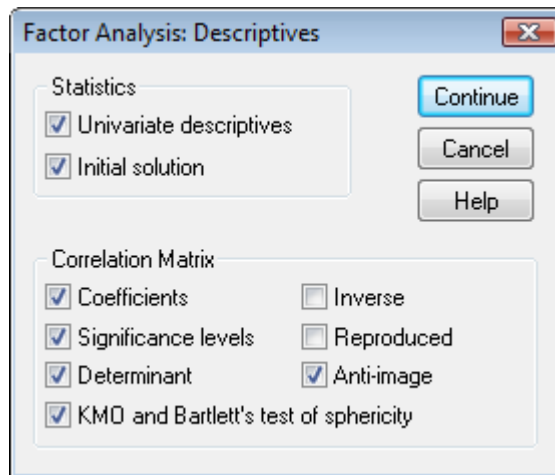
Inverse ο αντίστροφος του πίνακα συσχέτισης

Reproduced: ο εκτιμώμενος πίνακας συσχέτισης. Δίνονται και τα υπόλοιπα. Θα πρέπει να έχουν απόλυτες τιμές μικρότερες του 0.05

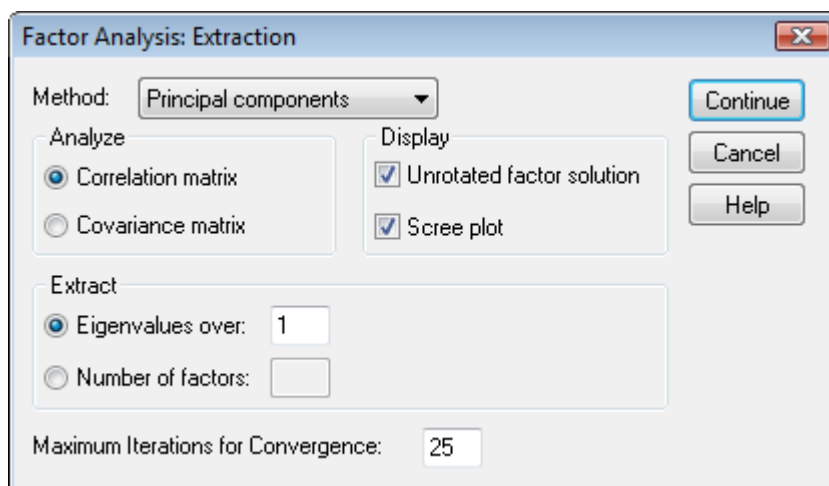
Anti-image: δίνονται οι συντελεστές μερικής συσχέτισης με αλλαγμένα πρόσημα στα μη διαγώνια στοιχεία, ενώ στα διαγώνια δίνεται ο δείκτης MSA (θέλουμε τιμές μεγαλύτερες του 0.5)

KMO and Bartlett test of sphericity. Η υπόθεση της σφαιρικότητας θα πρέπει να απορρίπτεται (p-τιμή του Bartlett test of sphericity<0.05) και ο δείκτης KMO θα πρέπει να είναι μεγαλύτερος από 0.6.

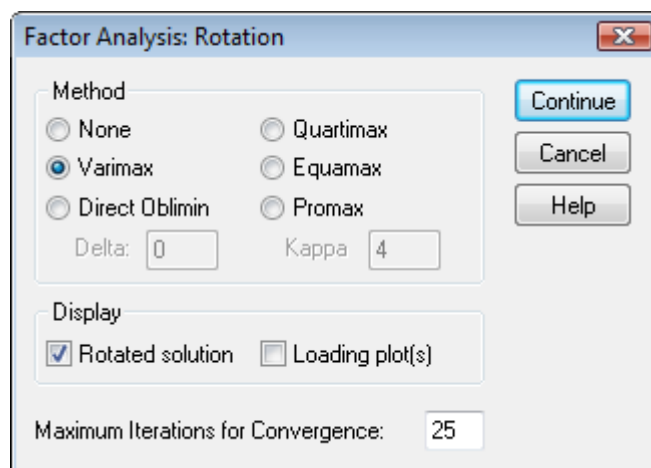
Ενδεικτικά επιλέγουμε τα ακόλουθα:



Από το πλαίσιο Extraction έχουμε τη δυνατότητα να καθορίσουμε τη μέθοδο εκτίμησης των παραγόντων (Principal components), το αν θα χρησιμοποιηθεί ο πίνακας συσχετίσεων ή ο συνδιακυμάνσεων (Correlation ή Covariance matrix), από το πεδίο Unrotated factor solution μας δίνεται η λύση της παραγοντικής ανάλυσης πριν την περιστροφή. Επιλέγοντας το πλαίσιο Scree plot προκύπτει το γράφημα των ιδιοτιμών του πίνακα συσχέτισης σε φθίνουσα σειρά. Τέλος από το πεδίο Extract είτε καθορίζουμε τον αριθμό των παραγόντων (πλαίσιο Number of factors) είτε επιλέγουμε τη χρησιμοποίηση του Kaiser Criterion (Eigenvalues over 1), δηλώνοντας και τον αριθμό των επαναλήψεων μέχρι να επιτευχθεί σύγκλιση.



Από το πλαίσιο Rotation επιλέγεται η μέθοδος περιστροφής (συνηθέστερα η Varimax) και δηλώνουμε το πλαίσιο Rotated Solution για να εμφανιστεί η λύση με περιστροφή στο παράθυρο των αποτελεσμάτων.



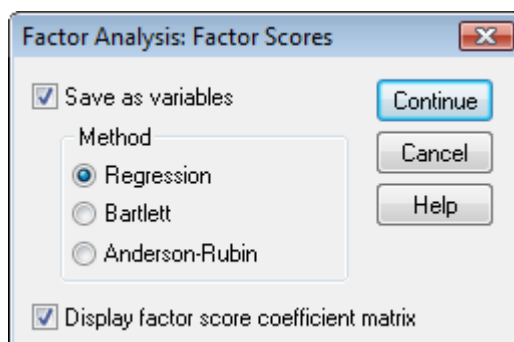
Από το πλαίσιο Scores επιλέγοντας το πεδίο Save as Variables δημιουργείται (με μία εκ των τριών διαθέσιμων μεθόδων) μία νέα μεταβλητή για κάθε παράγοντα. Οι μεταβλητές αυτές μπορούν να χρησιμοποιηθούν για την περαιτέρω έρευνα π.χ εξέταση της επίδρασης ποιοτικών μεταβλητών στα σκορ των παραγόντων. Αν είναι k το πλήθος οι παράγοντες οι νέες μεταβλητές προκύπτουν ως εξής:

$$F_1 = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p$$

$$F_2 = a_{21}X_1 + a_{22}X_2 + \dots + a_{2k}X_p$$

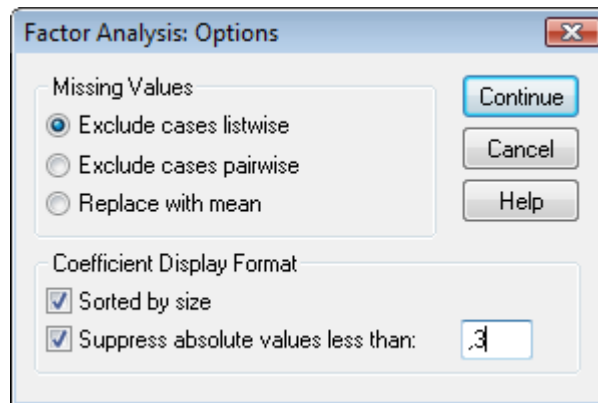
$$F_k = a_{p1}X_{11} + a_{p2}X_2 + \dots + a_{pk}X_p$$

Από το πεδίο Display factor score coefficient matrix: μας δίνονται οι συντελεστές με τους οποίους κάθε μεταβλητή πολλαπλασιάστηκε για τη δημιουργία των factor scores, δηλαδή οι τιμές των a_{ij} , $i = 1, \dots, p$, $j = 1, \dots, k$.



Τέλος, από το πεδίο Options καθορίζουμε τον τρόπο χειρισμού των ελλিপών τιμών, τον τρόπο εμφάνισης των συντελεστών (κατά μέγεθος, Sorted by size), και δηλώνουμε αν δεν θέλουμε να εμφανίζονται συντελεστές με απόλυτη τιμή μικρότερη από καθορισμένο

αριθμός (συνήθως επιλέγεται το 0.3 έτσι ώστε ο προσδιορισμός των παραγόντων να είναι άμεσος)



Αποτελέσματα αυτών των επιλογών

Στο παράθυρο των αποτελεσμάτων λόγω των παραπάνω επιλογών προκύπτει ο πίνακας Descriptive Statistics όπου για κάθε μεταβλητή που δηλώθηκε στο πλαίσιο Variables δίνεται η μέση τιμή, η τυπική απόκλιση και το πλήθος των διαθέσιμων παρατηρήσεων. Ο πίνακας Correlation matrix μας δίνει τον πίνακα συσχέτισης, τις αντίστοιχες p-τιμές του μονόπλευρου ελέγχου ότι ο πληθυσμιακός συντελεστής συσχέτισης είναι ίσος με μηδέν κάθε δυνατού ζεύγους μεταβλητών (p-τιμές<0.05 για το συντριπτικό ποσοστών των ελέγχων) και την τιμή της ορίζουσας του δειγματικού πίνακα συσχέτισης (τιμή=0.01) . Από τον πίνακα KMO and Bartlett's test προκύπτει ότι τα δεδομένα είναι κατάλληλα για διεξαγωγή παραγοντικής ανάλυσης (ο δείκτης KMO=0.930>0.6, p-τιμή του Bartlett's test of sphericity<0.05). Επιπλέον όλες οι μεταβλητές μας είναι κατάλληλες για την παραγοντική ανάλυση καθώς από το πεδίο Anti-Image Matrices προκύπτει ότι ο δείκτης MSA για όλες τις υπό μελέτη μεταβλητές είναι μεγαλύτερος από 0.5 (αν υπήρχαν κάποιες με MSA<0.5 καλό θα ήταν να αποκλείονταν από τη μελέτη).

Στον πίνακα Communalities μπορούμε να «δούμε» το ποσοστό της μεταβλητότητας κάθε μεταβλητής που εξηγείται από τον αριθμό των παραγόντων που προσαρμόστηκε (άρα είναι ένας αριθμός μεταξύ 0 και 1). Η πρώτη στήλη (Initial) είναι πάντοτε 1 όταν χρησιμοποιείται η μέθοδος των κύριων συνιστωσών. Έτσι για παράδειγμα παρατηρούμε ότι το προσαρμοσμένο μοντέλο ερμηνεύει μόνο το 43,5% της πρώτης ερώτησης «Statistics makes me cry». Αυτό σημαίνει ότι το προσαρμοσμένο μοντέλο δεν είναι ιδιαίτερα καλό.

Communalities

	Initial	Extraction
Statistics makes me cry	1,000	,435
My friends will think I'm stupid for not being able to cope with SPSS	1,000	,414
Standard deviations excite me	1,000	,530
I dream that Pearson is attacking me with correlation coefficients	1,000	,469
I don't understand statistics	1,000	,343
I have little experience of computers	1,000	,654
All computers hate me	1,000	,545
I have never been good at mathematics	1,000	,739
My friends are better at statistics than me	1,000	,484
Computers are useful only for playing games	1,000	,335
I did badly at mathematics at school	1,000	,690
People try to tell you that SPSS makes statistics easier to understand but it doesn't	1,000	,513
I worry that I will cause irreparable damage because of my incompetence with computers	1,000	,536
Computers have minds of their own and deliberately go wrong whenever I use them	1,000	,488
Computers are out to get me	1,000	,378
I weep openly at the mention of central tendency	1,000	,487
I slip into a coma whenever I see an equation	1,000	,683
SPSS always crashes when I try to use it	1,000	,597
Everybody looks at me when I use SPSS	1,000	,343
I can't sleep for thoughts of eigen vectors	1,000	,484
I wake up under my duvet thinking that I am trapped under a normal distribution	1,000	,550
My friends are better at SPSS than I am	1,000	,464
If I'm good at statistics my friends will think I'm a nerd	1,000	,412

Extraction Method: Principal Component Analysis.

Ο πίνακας Total Variance Explained περιέχει στη στήλη Initial Eigenvalues τις ιδιοτιμές και το ποσοστό της διακύμανσης που κάθε ιδιοτιμή ερμηνεύει, άρα κάθε κύρια συνιστώσα. Στη στήλη Extaction Sum of Squared Loadings μας δίνεται το ποσοστό της διακύμανσης που εξηγεί κάθε παράγοντας αν χρησιμοποιηθεί ο κριτήριο του προσδιορισμού του αριθμού των παραγόντων το κριτήριο του Kaiser. Τέλος στη στήλη Rotation Sum of Squared Loadings μας δίνεται το ποσοστό της διακύμανσης που εξηγείται από τους παράγοντες μετά την περιστροφή. Επομένως έχουμε 4 παράγοντες που εξηγούν το 50,3% της συνολικής διακύμανσης.

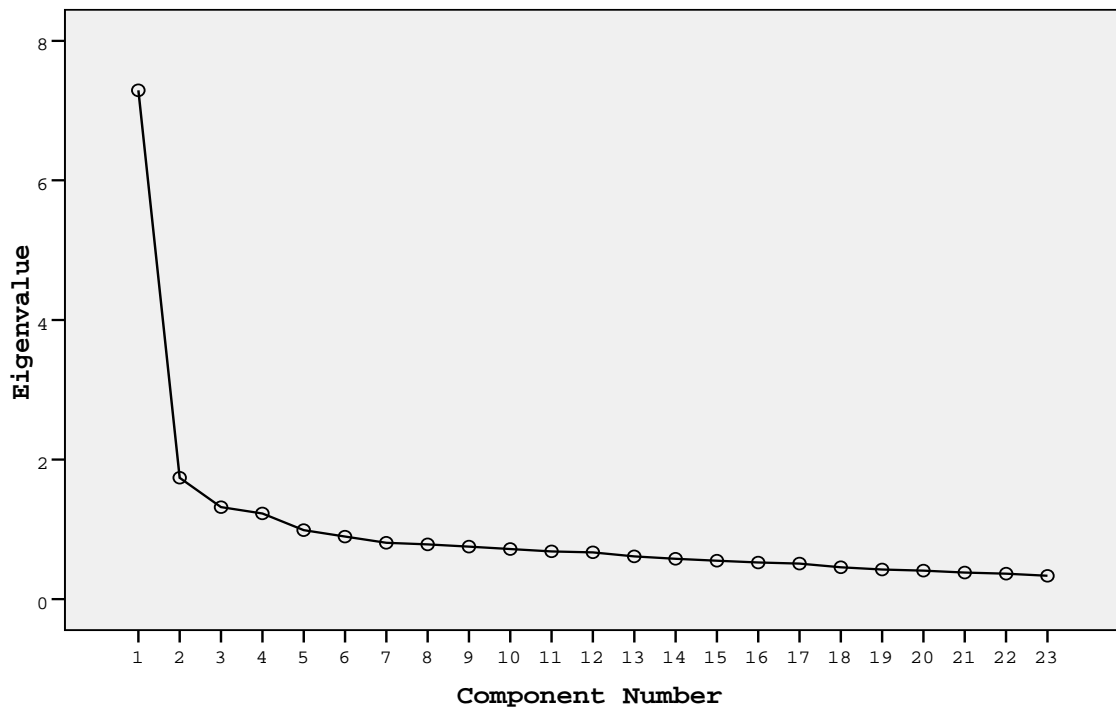
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7,290	31,696	31,696	7,290	31,696	31,696	3,730	16,219	16,219
2	1,739	7,560	39,256	1,739	7,560	39,256	3,340	14,523	30,742
3	1,317	5,725	44,981	1,317	5,725	44,981	2,553	11,099	41,841
4	1,227	5,336	50,317	1,227	5,336	50,317	1,950	8,476	50,317
5	,988	4,295	54,612						
6	,895	3,893	58,504						
7	,806	3,502	62,007						
8	,783	3,404	65,410						
9	,751	3,265	68,676						
10	,717	3,117	71,793						
11	,684	2,972	74,765						
12	,670	2,911	77,676						
13	,612	2,661	80,337						
14	,578	2,512	82,849						
15	,549	2,388	85,236						
16	,523	2,275	87,511						
17	,508	2,210	89,721						
18	,456	1,982	91,704						
19	,424	1,843	93,546						
20	,408	1,773	95,319						
21	,379	1,650	96,969						
22	,364	1,583	98,552						
23	,333	1,448	100,000						

Extraction Method: Principal Component Analysis.

Επιπλέον μας δίνεται το Scree plot. Δεν είναι ξεκάθαρο πόσους παράγοντες θα χρησιμοποιήσουμε (3 ή 4) .

Scree Plot



Στον πίνακα Component matrix δίνονται οι επιβαρύνσεις των παραγόντων που προκύπτουν με το μοντέλο για 4 παράγοντες. Ουσιαστικά ο πίνακας αυτός μας δίνει τις ακόλουθες πληροφορίες:

$$X_1 = L_{11}F_1 + L_{12}F_2 + \dots + L_{1k}F_k$$

$$X_2 = L_{21}F_1 + L_{22}F_2 + \dots + L_{2k}F_k$$

$$X_p = L_{p1}F_1 + L_{p2}F_2 + \dots + L_{pk}F_k$$

Σχόλιο: Από τις τιμές των L_{i1}, \dots, L_{ik} μπορούν να βρεθούν οι τιμές του πίνακα

Communalities στήλη Extraction με τη σχέση $\sum_{j=1}^k L_{ij}^2$, $i = 1, \dots, p$.

Ο πίνακας Rotated Component Matrix περιέχει τις επιβαρύνσεις των παραγόντων μετά την περιστροφή. Από τον πίνακα αυτό μπορούμε να «ερμηνεύσουμε» τους παράγοντες.

Rotated Component Matrix(a)

	Component			
	1	2	3	4
I have little experience of computers	,800			
SPSS always crashes when I try to use it	,684	,327		
I worry that I will cause irreparable damage because of my incompetence with computers	,647			
All computers hate me	,638	,327		
Computers have minds of their own and deliberately go wrong whenever I use them	,579	,360		
Computers are useful only for playing games	,550			
Computers are out to get me	,459			
I can't sleep for thoughts of eigen vectors		,677		
I wake up under my duvet thinking that I am trapped under a normal distribution		,661		
Standard deviations excite me		-,567		,368
People try to tell you that SPSS makes statistics easier to understand but it doesn't	,473	,523		
I dream that Pearson is attacking me with correlation coefficients	,320	,516	,314	
I weep openly at the mention of central tendency	,334	,514	,313	
Statistics makes me cry		,496	,356	
I don't understand statistics	,319	,429		
I have never been good at mathematics			,833	
I slip into a coma whenever I see an equation			,747	
I did badly at mathematics at school			,747	
My friends are better at statistics than me				,648
My friends are better at SPSS than I am				,645
If I'm good at statistics my friends will think I'm a nerd				,586
My friends will think I'm stupid for not being able to cope with SPSS		-,338		,543
Everybody looks at me when I use SPSS		-,372		,428

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 8 iterations.

Παρατηρούμε ότι υπάρχουν μεταβλητές που «φορτίζουν» σε περισσότερους από έναν παράγοντες. Συνηθέστερα στις κοινωνικές επιστήμες αυτές αποκλείονται και η παραγοντική ανάλυση διεξάγεται από την αρχή. Αν θέλαμε να δώσουμε μία ελεύθερη ερμηνεία στους 4 παράγοντες που προέκυψαν θα λέγαμε ότι ο πρώτος περιγράφει το άγχος για τη χρήση των υπολογιστών, ο δεύτερος για τη Στατιστική, ο τρίτος για τα μαθηματικά και ο τέταρτος για τις αντιδράσεις των υπολοίπων.

Τέλος στον πίνακα Component Transformation Matrix δίνεται ο πίνακας με τον οποίο πολλαπλασιάσαμε τον αρχικό πίνακα επιβαρύνσεων (Component Matrix) για να οδηγηθούμε στον τελικό πίνακα επιβαρύνσεων (Rotated Component matrix). Επομένως,

$$\text{Rotated Component matrix} = \text{Component Matrix} * \text{Component Transformation Matrix}$$

Component Transformation Matrix

Component	1	2	3	4
1	,635	,585	,443	-,242
2	,137	-,167	,488	,846
3	,758	-,513	-,403	,008
4	,067	,605	-,635	,476

Extraction Method: Principal Component Analysis.
 Rotation Method: Varimax with Kaiser Normalization.

Ενδεικτική Βιβλιογραφία

Field, A. P. (2000). *Factor Analysis Using SPSS*

<http://www.sussex.ac.uk/psychology/profile9846.html>

Coakes, S. and Steed, L (1999). *SPSS Analysis without Anguish*. Wiley.

Καρλής, Δ. (2005). Πολυμεταβλητή Στατιστική Ανάλυση. Εκδόσεις Σταμούλη.