
# ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΙΓΑΙΟΥ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ –ΕΙΣΑΓΩΓΗ ΚΑΤΕΥΘΥΝΣΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΝΑΛΟΓΙΣΤΙΚΩΝ-ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΩΝ ΜΑΘΗΜΑΤΙΚΩΝ

ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ


**Μεταπτυχιακή Διπλωματική εργασία με θέμα:**

# «Δομικά μοντέλα εξισώσεων: Θεωρία και εφαρμογές»


Της

Ρούση Αλεξάνδρας

Επιβλέπων: Ζήμερας Στυλιανός

Τριμελής Εξεταστική επιτροπή:

Ζήμερας Στυλιανός

Χατζησπύρος Σπύρος

Καραγρηγορίου Αλέξανδρος


Καρλόβασι, 2016

**ΠΕΡΙΕΧΟΜΕΝΑ**

Κεφάλαιο 3: Γενικευμένα γραμμικά μοντέλα



Κεφάλαιο 4: Πολυεπίπεδα μοντέλα

Κεφάλαιο 6: Στατιστική συμπερασματολογία και εκτίμηση παραμέτρων



Κεφάλαιο 7: Εφαρμογή των δομικών μοντέλων εξίσωσης



*Κεφάλαιο 8: Βιβλιογραφία*

**ΠΕΡΙΛΗΨΗ**

Τα δομικά μοντέλα εξισώσεων(SEM) θεωρείται μία επέκταση της παλινδρόμησης και της παραγοντικής ανάλυσης, η οποία ταυτόχρονα εξετάζει τις σχέσεις μιας ή περισσοτέρων εξαρτημένων μεταβλητών μεταξύ δύο ή περισσοτέρων ανεξάρτητων μεταβλητών.Ως παραγοντική ανάλυση εννοούμε τη μέθοδο για την διερεύνηση των σχέσεων ανάμεσα σε σύνολα από παρατηρήσιμες και λανθάνουσες μεταβλητές. Η παραγοντική ανάλυση εξετάζει τη συνδιακύμανση ανάμεσα στις παρατηρήσιμες μεταβλητές με σκοπό να συλλέξει πληροφορίες για τις υποκείμενες λανθάνουσες μεταβλητές.

Ουσιαστικά, τα δομικά μοντέλα εξισώσεων είναι μία στατιστική μεθοδολογία που χρησιμοποιείται ευρέως. Στο θεωρητικό μέρος συμπεριλήφθηκαν όλοι οι σκοποί αλλά και οι βασικοί όροι των δομικών μοντέλων εξισώσεων. Καθώς τα συστήματα δομικών εξισώσεων (SEM) είναι μία στατιστική μέθοδος που υιοθετεί μία επικυρωτική μέθοδο στηνπολυμεταβλητή ανάλυση ενός μοντέλου,που αφορά κάποιες παρατηρήσεις ή μετρήσεις, γίνεται μία αναφορά και στις πολυμεταβλητές μεθόδους. Επίσης γίνεται μία αναφορά στα γενικευμένα γραμμικά αλλά και στα ιεραρχικά μοντέλα. Τέλος, γίνεται αναφορά στα στατιστικά εργαλεία που χρησιμοποιούνται για την ανάλυση τους.

Στο πρακτικό μέρος τώρα της εργασίας χρησιμοποιούνται πραγματικά δεδομένα από μετρήσεις συστατικών για κάποιες ποικιλίες κρασιών. Για την εφαρμογή χρησιμοποιηθήκε το στατιστικό πρόγραμμα IBMSPSSStatistics 20 και το πρόγραμμα ανάλυσης των SEM, LISREL.

# *Κεφάλαιο 1 : Βασικές έννοιες γραμμικής άλγεβρας και θεωρίας πινάκων*

## *1.1:ΠΙΝΑΚΕΣ*

### *1.1.1:Γενικές έννοιες*

Ένας πίνακας $A$ με στοιχεία από το $F$ είναι μία διάταξη των $mn$στοιχείων(αριθμών) $a_{ij}$ του συνόλου $F$ σε σχήμα ορθογώνιο παραλληλόγραμμο της μορφής:

$$A = \begin{bmatrix} a_{11}\ a_{12}\ \dots a_{1n} \\ a_{21}\ a_{21}\ \dots a_{2n} \\ \dots\ \dots\ \dots\ \dots \\ a_{m1}a_{m2}\cdots a_{mn} \end{bmatrix}$$

Οι αριθμοί $a_{ij}$ ονομάζονται στοιχεία του πίνακα και είναι τοποθετημένοι σε $m$ γραμμές και $n$ στήλες. Οι δύο δείκτες $i$ και $j$ δείχνουν την θέση , που κατέχει το κάθε στοιχείο στον πίνακα. Ο πρώτος δείκτης $i$ ονομάζεται ***δείκτης γραμμής*** και ο δεύτερος $j$***δείκτης στήλης***.Ο συμβολισμός $a_{ij}$ βρίσκεται συγχρόνως στην $i-$γραμμή και στην $j-$στήλη. Λέμε ότι ο πίνακας $A$ είναι ένας $m \times n$ πίνακας.

Δύο ή περισσότεροι πίνακες με τον ίδιο αριθμό γραμμών και στηλών χαρακτηρίζονται ως πίνακες του ***ίδιου τύπου***. Αν διαγραφούν κάποιες γραμμές ή στήλες ή γραμμές και στήλες από τον πίνακα $A$, ο πίνακας που απομένει ονομάζεται ***υποπίνακας*** του $A$. Ένας πίνακας $A$ ονομάζεται ***σύνθετος***πίνακας, αν τα στοιχεία του είναι πίνακες μικρότερου μεγέθους από αυτό του $A$, η δε διαμέριση γίνεται κατά τέτοιον τρόπο ώστε τα στοιχεία – ***υποπίνακες***, που βρίσκονται στην ίδια γραμμή, έχουν όλα τον ίδιο αριθμό γραμμών και τα στοιχεία - υποπίνακες που βρίσκονται στην ίδια στήλη, έχουν όλα τον ίδιο αριθμό στηλών. Ο σύνθετος πίνακας συμβολίζεται με $A = (A_{ij})$, όπου $A_{ij}$ είναι το στοιχείο – υποπίνακας, που προκύπτει από την διαμέριση του αρχικού πίνακα $A$ και βρίσκεται στην $i-γραμμή$και $j-στήλη$ του πίνακα $A$.

Ένας $1 \times n$ πίνακας λέγεται ***πίνακας-γραμμή*** και ένας $m \times 1$ πίνακας λέγεται ***πίνακαςστήλη ή διάνυσμα***.

Ένας $m \times n$ πίνακας λέγεται ***μηδενικός*** αν όλα τα στοιχεία του είναι ίσα με μηδέν και συμβολίζονται με $\mathbb{O}_{m\times n}$ ή απλά $\mathbb{O}$.

Ένας $m \times n$ πίνακας λέγεται***τετραγωνικός***, αν $m = n$ δηλαδή τετραγωνικός είναι ο πίνακας που έχει ίσο αριθμό γραμμών και στηλών.

$$A = \begin{bmatrix} a_{11}\,a_{12}\dots a_{1n} \\ a_{21}\,a_{22}\dots a_{2n} \\ \vdots\ \ \vdots\ \ \ddots\ \ \vdots \\ a_{n1}a_{n2}\cdots a_{nn} \end{bmatrix}$$

Τα ***διαγώνια στοιχεία*** ενός $m \times n$ πίνακα $(a_{ij})$ είναι τα στοιχεία $a_{11}, a_{22}, \dots, a_{kk}$, όπου $k = \min\{\mathrm{m}, \mathrm{n}\}$, τα οποία βρίσκονται πάνω στην ***κύρια διαγώνιο*** του πίνακα. Ένας τετραγωνικός πίνακας $A = (a_{ij})$ λέγεται ***διαγώνιος***, αν για κάθε $i \neq j$ ισχύει $a_{ij} = 0$. Δηλαδή, ένας τετραγωνικός πίνακας είναι διαγώνιος, αν κάθε στοιχείο, που δε βρίσκεται στη διαγώνιο είναι ίση με το 0. Τους πίνακες αυτούς θα τους συμβολίζουμε με $A = diag(a_{11}, a_{22}, \dots, a_{nn})$. Ειδικότερα, ο διαγώνιος πίνακας που έχει όλα τα διαγώνια στοιχεία ίσα με τη μονάδα ονομάζεται ***μοναδιαίος*** και συμβολίζεται:$\boldsymbol{I_n = diag(1, 1, \dots, 1)}$.

$$A = \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

Το άθροισμα των στοιχείων της κύριας διαγωνίου του $m \times n$ πίνακα $A$, συμβολίζεται ιχν$A$ ή $\boldsymbol{trA}$ και ονομάζεται ***ίχνος***του $A$.

$$tr(A) = a_{11} + a_{22} + \cdots + a_{nn} = \sum_{i=1}^{n} a_{ii}$$

<u>*Ιδιότητες ίχνους*</u>

Αν $A$ και $B$ τετραγωνικοί πίνακες τότε το ίχνος έχει τις εξής ιδιότητες:

1. $tr(A + B) = tr(B + A) = tr(A) + tr(B)$
2. $tr(AB) = tr(BA)$

Ένας τετραγωνικός πίνακας $A = (a_{ij})$ λέγεται ***άνω τριγωνικός***, αν για κάθε $i > j$, έχουμε $a_{ij} = 0$. Δηλαδή ένας τετραγωνικός πίνακας λέγεται άνω τριγωνικός, αν τα στοιχεία, που βρίσκονται κάτω από την κύρια διαγώνιο, είναι ίσα με μηδέν. Ένας τετραγωνικός πίνακας $A = (a_{ij})$ λέγεται ***κάτω τριγωνικός***, αν για κάθε $i < j$, έχουμε $a_{ij} = 0$. Δηλαδή ένας τετραγωνικός πίνακας λέγεται κάτω τριγωνικός αν τα στοιχεία που βρίσκονται πάνω από την κύρια διαγώνιο, είναι ίσα με μηδέν.

Έστω $A = (a_{ij})$ ένας $m \times n$ πίνακας. Ο $n \times m$ πίνακας $(a_{ij})$ ονομάζεται ***ανάστροφος του $A$ και συμβολίζεται με $A^t$***. Η αναστροφή ενός πίνακα γίνεται μετατρέποντας τις γραμμές ενός πίνακα σε στήλες και τις στήλες σε γραμμές.

Ο ανάστροφος του πίνακα $A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$

είναι ο $A^t = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{12} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}$

Ένας τετραγωνικός πίνακας $(a_{ij})$λέγεται ***συμμετρικός***, αν για κάθε $i, j$, ισχύει $a_{ij} = a_{ji}$, ή ισοδύναμα αν και μόνο αν ισχύει ότι $A = A^t$. Τώρα, ένας τετραγωνικός πίνακας λέγεται ***αντισσυμετρικός*** αν για κάθε $i, j$, ισχύει $a_{ij} = -a_{ji}$, ή ισοδύναμα ένας πίνακας $A$ λέγεται αντισσυμετρικός αν και μόνο αν $A = -A^t$.

Έστω $A = (a_{ij})$ ένας $m \times n$ πίνακας. Ο $n \times m$ πίνακας, ο οποίος έχει ως στοιχεία του τα συζυγή στοιχεία του πίνακα $A$, δηλαδή $\bar{a}_{ij}$, ονομάζεται***συζυγής*** του $A$ και συμβολίζεται $\bar{A}$. Ο $n \times m$ πίνακας $(\bar{a}_{ji})$ ονομάζεται αναστροφοσυζυγής του πίνακα $A = (a_{ij})$ και συμβολίζεται $A^* = \bar{A}^t$.

### *1.1.2: Πράξεις πινάκων*

Έστω πίνακες $A = (a_{ij})$, $B = (b_{ij}) \in M_{m\times n}(\mathbb{F})$. Ως άθροισμα, $A + B$, των πινάκων $A$ και $B$ ορίζεται ο πίνακας $A + B = (\gamma_{ij}) \in M_{m\times n}(\mathbb{F})$, ο οποίος είναι επίσης του ίδιου τύπου με τους αρχικούς πίνακες και έχει ως στοιχεία το άθροισμα των ομολόγων στοιχείων των $A$ και $B$,

δηλαδή

$$A + B = (\gamma_{ij}) = (a_{ij} + \beta_{ij})$$

**Η πρόσθεση πινάκων ορίζεται μόνο για πίνακες του ίδιου τύπου**.

Δηλαδή, έστω $A = \begin{pmatrix} 3 & 7 & 1 \\ -2 & 4 & 5 \end{pmatrix}$ και $B = \begin{pmatrix} -3 & 2 & x \\ 0 & y & 1 \end{pmatrix}$ τότε:

$$A + B = \begin{pmatrix} 3 & 7 & 1 \\ -2 & 4 & 5 \end{pmatrix} + \begin{pmatrix} -3 & 2 & x \\ 0 & y & 1 \end{pmatrix} = \begin{pmatrix} 3 + (-3) & 7 + 2 & 1 + x \\ -2 + 0 & 4 + y & 5 + 1 \end{pmatrix}$$
$$= \begin{pmatrix} 0 & 9 & 1 + x \\ -2 & 4 + y & 6 \end{pmatrix}$$

Έστω, τώρα $A = (a_{ij}) \in M_{m\times n}(\mathbb{F})$ και $k \in \mathbb{F}$. ***Το γινόμενο kA*** του $k$ επί τον $A$ είναι πίνακας ίδιου τύπου με τον αρχικό πίνακα $A$, του οποίου τα στοιχεία προκύπτουν από τα αντίστοιχα του $A$ με πολλαπλασιασμό τους επί $k$. Δηλαδή, προκύπτει ο πίνακας $kA = (ka_{ij})$. ***Ο πολλαπλασιασμός αυτός ονομάζεται πολλαπλασιασμός πίνακα επί αριθμό ή βαθμωτός πολλαπλασιασμός***.

Για παράδειγμα, αν $k = 2$ και $A = \begin{pmatrix} 6 & 8 & 5 \\ 1 & 3 + i & 4 \end{pmatrix}$ τότε,

$$2A = \begin{pmatrix} 2 * 6 & 2 * 8 & 2 * 5 \\ 2 * (-1) & 2 * (3 + i) & 2 * 4 \end{pmatrix} = \begin{pmatrix} 12 & 16 & 10 \\ -2 & 6 + 2i & 8 \end{pmatrix}$$

Ειδικά, αν $k = -1$, ο πίνακας $(-1)A$ συμβολίζεται – $A$ και ονομάζεται αντίθετος του $A$. Ως διαφορά δύο πινάκων θεωρείται ο πίνακας $A - B$ και είναι αποτέλεσμα των πράξεων $A + (-1)B$.

Για παράδειγμα,

$$\begin{pmatrix} -4 & 5 \\ 1 & -7 \end{pmatrix} - \begin{pmatrix} 3 & 6 \\ -4 & 1 \end{pmatrix} = \begin{pmatrix} -4-3 & 5-6 \\ 1-(-4) & -7-1 \end{pmatrix} = \begin{pmatrix} -7 & -1 \\ 5 & -8 \end{pmatrix}.$$

### *1.1.3: Αντίστροφος πίνακας*

Ένας τετραγωνικός πίνακας $A \in M_n(\mathbb{F})$ λέγεται **αντιστρέψιμος** , αν υπάρχει πίνακας $B \in M_n(\mathbb{F})$ τέτοιος ώστε $\boldsymbol{AB = BA = I_n = I(1.1.3).}$

Ο πίνακας $B$ αν υπάρχει λέγεται αντίστροφος του $A$ και συμβολίζεται με $A^{-1}$. Πρέπει να τονίσουμε ότι αν υπάρχει ο πίνακας $B$ είναι μοναδικός, αφού σύμφωνα και με την σχέση $(\mathbf{1.1.3})$ ,μπορούμε να δείξουμε ότι:

$$\boldsymbol{\hat{B} = \hat{B}I = \hat{B}(AB) = (\hat{B}A)B = IB = B.}$$

Συνεπώς, αν υπάρχει ο $A^{-1}$ είναι μοναδικός και ισχύει ότι:

$$\boldsymbol{AA^{-1} = A^{-1}A = I_n}.$$

<u>*Ιδιότητες αντίστροφου πίνακα* (Αν $A, B$ αντιστρέψιμοι, τότε)</u>

- $A$ είναι αντιστρέψιμος και $(A^{-1})^{-1} = A$.
- $A^k$ αντιστρέψιμος και $(A^k)^{-1} = (A^{-1})^k$.
- Για κάθε $\lambda \neq 0, (\lambda A)^{-1} = \frac{1}{\lambda}A^{-1}$
- $detA^{-1} = (detA)^{-1}$
- $(A^*)^{-1} = (A^{-1})^*$
- Ο πίνακας $AB$ είναι αντιστρέψιμος και $(AB)^{-1} = B^{-1}A^{-1}$

### *1.1.4: Ορθογώνιος Πίνακας*

Ένας τετραγωνικός πίνακας $A$ ονομάζεται ορθογώνιος αν ισχύει ότι $AA^t = I$ ή ισοδύναμα $A^{-1} = A^t$, δηλαδή ο αντίστροφος του είναι ίσος με τον ανάστροφο του.

### *1.1.5: Ορίζουσα πίνακα*

Η ορίζουσα είναι μία ειδική συνάρτηση στο σύνολο των τετραγωνικών πινάκων και έχει πεδίο τιμών στο σύνολο **R** ή **C**. Για κάθε $\nu \times \nu$ πίνακα $A$ η ορίζουσα συμβολίζεται $|A|$ ή $detA$.

Για τον πίνακα

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} => \det(A) = a_{11}a_{22} - a_{21}a_{12}$$

Και για τον πίνακα

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$\det(A) = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} =$$
$$= a_{11}a_{22}a_{33} - a_{11}a_{32}a_{23} + a_{12}a_{31}a_{23} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}.$$

Για $\nu \geq 2$, η ορίζουσα του $\nu \times \nu$ πίνακα $A = \left[a_{ij}\right]_{i,j=1}^{\nu}$ είναι το άθροισμα

$$detA = a_{11}detA_{11} - a_{12}detA_{12} + \cdots + (-1)^{1+\nu}a_{1\nu}detA_{1\nu} =$$
$$= \sum_{j=1}^{\nu}(-1)^{1+j}a_{1j}detA_{1j}.$$

Ο αριθμός $M_{ij} = (-1)^{i+j}a_{1j}detA_{ij}$ ονομάζεται αλγεβρικό συμπλήρωμα του στοιχείου $a_{ij}$ και προφανώς έχουμε:

$$detA = a_{11}M_{11} + a_{12}M_{12} + \cdots + a_{1\nu}M_{1\nu}.$$

Το παραπάνω άθροισμα ονομάζεται **ανάπτυγμα ορίζουσας** ως προς την πρώτη γραμμή του πίνακα $A$.

Για τετραγωνικό πίνακα $A = \left[a_{ij}\right]_{i,j=1}^{\nu}$

$$detA = a_{i1}M_{i1} + a_{i2}M_{i2} + \cdots + a_{in}M_{in} =$$
$$= a_{1j}M_{1j} + a_{2j}M_{2j} + \cdots + a_{nj}M_{nj}.$$

*Ιδιότητες οριζουσών*(Για κάθε $\nu \times \nu$ πίνακα)

- Αν εναλλάξουμε δύο γραμμές (ή στήλες) του $A$, η ορίζουσα του νέου πίνακα ισούται με $-detA$.
- Αν τα στοιχεία μιας γραμμής (ή στήλης) του $A$ πολλαπλασιασθούν επί τον αριθμό $k$, η ορίζουσα του νέου πίνακα ισούται με $k(detA)$.
- Αν το πολλαπλάσιο των στοιχείων μιας γραμμής (στήλης) του $A$ προστεθεί σε μία άλλη γραμμή(στήλη) του πίνακα, η ορίζουσα του νέου πίνακα ισούται με $detA$.

### *1.1.6: Επαυξημένος πίνακας*

Έστω $m \times n$ πίνακας $A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$ θεωρείται ο πίνακας των συντελεστών, ο πίνακας – στήλη $x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ και ο πίνακας $b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$ ο πίνακας των σταθερών όρων στου συστήματος.

Ο $m \times (n+1)$ πίνακας $(A|b) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right)$ ονομάζεται επαυξημένος πίνακας του συστήματος.

### *1.1.7: Κλιμακωτός πίνακας*

Ένας πίνακας $A$ ονομάζεται κλιμακωτός αν:

- Το πρώτο μη μηδενικό στοιχείο κάθε μη μηδενικής γραμμής βρίσκεται σε θέση δεξιότερα από το αντίστοιχο μη μηδενικό στοιχείο της προηγούμενης γραμμής, και
- Οι μη μηδενικές γραμμές βρίσκονται πάνω από τις μηδενικές γραμμές.

- ***<u>Ανηγμένος κλιμακωτός</u>***

Ένας κλιμακωτός πίνακας ονομάζεται ανηγμένος κλιμακωτός αν επιπλέον ισχύουν:

- Το ηγετικό στοιχείο κάθε μη μηδενικής γραμμής είναι το 1 και,
- Το ηγετικό στοιχείο 1 είναι το μοναδικό μη μηδενικό στοιχείο της στήλης που το περιέχει.

### *1.1.8: Μέθοδος απαλοιφής του Gauss*

Κάθε γραμμικό σύστημα είναι ισοδύναμο με ένα γραμμικό σύστημα του οποίου ο επαυξημένος πίνακας είναι κλιμακωτός ή μπορεί να είναι ανηγμένος κλιμακωτός. Η μέθοδος αυτή ονομάζεται μέθοδος απαλοιφής Gauss.

Έστω γραμμικό σύστημα και ο αντίστοιχος επαυξημένος πίνακας:

$$\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots a_{2n}x_n = b_2 \\ \quad \vdots \qquad \vdots \quad \vdots \quad \vdots \quad \vdots \\ a_{m1}x_1 + a_{m2}x_{2+} \cdots a_{mn}x_n = b_m \end{array}, (A|b) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right)$$

<u>*Αλγόριθμος απαλοιφής*</u>

Έστω ότι, $a_{11} \neq 0$ (αν όχι κάνουμε ανάλογους μετασχηματισμούς). Χρησιμοποιώντας την πρώτη γραμμή και κάνοντας πράξεις μεταξύ των γραμμών μηδενίζοντας όλα τα στοιχεία κάτω από το στοιχείο $a_{11}$. Συνεχίζοντας την διαδικασία με την δεύτερη γραμμή μηδενίζουμε όλα τα στοιχεία της στήλης κάτω από το στοιχείο $a_{22}$. Επαναλαμβάνουμε την ίδια διαδικασία για όλες τις γραμμές των γραμμοισοδύναμων επαυξημένων πινάκων που προκύπτουν οπότε καταλήγουμε σε ένα γραμμοισοδύναμο με τον αρχικό επαυξημένο πίνακα της μορφής,

$$\left(\begin{array}{cccccc|c} \tilde{a}_{11} & \tilde{a}_{12} & \ldots & \tilde{a}_{1r} & \ldots & \tilde{a}_{1n} & \tilde{b}_1 \\ 0 & \tilde{a}_{22} & \ldots & \tilde{a}_{2r} & \ldots & \tilde{a}_{2n} & \tilde{b}_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & \tilde{a}_{rr} & \ldots & \tilde{a}_{rn} & \tilde{b}_r \\ 0 & 0 & \vdots & 0 & \vdots & 0 & \tilde{b}_{r+1} \\ \vdots & \vdots & \ldots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & & 0 & & 0 & \tilde{b}_m \end{array}\right)$$

Συνεπώς το αντίστοιχο σύστημα θα είναι:

$$\begin{aligned} &\tilde{a}_{11}x_{i1} + \tilde{a}_{12}x_{i2} + \cdots + \tilde{a}_{1r}x_{ir} + \cdots + \tilde{a}_{1n}x_{in} = \tilde{b}_1 \\ &\tilde{a}_{22}x_{i2} + \cdots + \tilde{a}_{2r}x_{ir} + \cdots + \tilde{a}_{2n}x_{in} = \tilde{b}_2 \\ &\qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots = \vdots \\ &\tilde{a}_{rr}x_{ir} + \cdots + \tilde{a}_{rn}x_{in} = \tilde{b}_r \\ &\qquad\qquad\qquad\qquad\qquad\qquad 0 = \tilde{b}_{r+1} \\ &\vdots = \vdots \\ &\qquad\qquad\qquad\qquad\qquad\qquad 0 = \tilde{b}_m \end{aligned}$$

Όπου, $\tilde{a}_{ii} \neq 0, i = 1,2, \ldots, r$

- Αν r<m και κάποιο από τα $\tilde{b}_{r+1}, \ldots, \tilde{b}_m \neq 0$, το σύστημα είναι αδύνατο.
- Αν $r \leq m$ και $\tilde{b}_{r+1} = \cdots = \tilde{b}_m = 0$ το σύστημα είναι συμβιβαστό.
  α)αν $r = n$ το σύστημα έχει ακριβώς μία λύση
  β)αν $r < n$ το σύστημα έχεις άπειρες λύσεις.

### *1.1.9: Ιδιοτιμές και ιδιοδιανύσματα*

Έστω $A$ ένας τετραγωνικός πίνακας $m \times m$

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix}$$

Και διάνυσμα $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$ θα ονομάζεται ιδιοδιάνυσμα του πίνακα $A$ αν υπάρχει λάθος λ, τέτοιος ώστε να ισχύει $\boldsymbol{Ax} = \boldsymbol{\lambda x}$.

Τότε ο αιθμός λ, ονομάζεται **ιδιότιμη** του πίνακα $\boldsymbol{A}$.

Από τον ορισμό του ιδιοδιανύσματος έχουμε ότι:

$$Ax = \lambda x => Ax = \lambda Ix => (A - \lambda I)x = 0$$

Η παραπάνω εξίσωση είναι ένα γραμμικό ομογενές σύστημα εξισώσεων με m αγνώστους. Για να έχει λύση διάφορη της μηδενικής πρέπει:

$$\det(A - \lambda I) = |A - \lambda I) = 0$$

Αλλά, $(A - \lambda I) = \begin{pmatrix} a_{11}\ a_{12} \dots a_{1m} \\ a_{21}\ a_{22} \dots a_{2m} \\ \vdots\ \ \vdots\ \ \vdots\ \ \vdots \\ a_{m1} a_{m2} \cdots a_{mm} \end{pmatrix} - \begin{pmatrix} \lambda 0 \dots 0 \\ 0 \lambda \dots 0 \\ \vdots\vdots\vdots\vdots \\ 0 0 \cdots \lambda \end{pmatrix}$

Και συνεπώς,

$$\det(A - \lambda I) = det \begin{pmatrix} a_{11} - \lambda & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} - \lambda \dots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots a_{mm} - \lambda \end{pmatrix} = 0$$

Η ορίζουσα αυτή είναι ένα πολυώνυμο βαθμού m ως προς λ. Άρα, η εξίσωση $\det(A - \lambda I) = 0$ έχει m λύσεις ως προς λ και επομένως ένας τετραγωνικός πίνακας $A$ τάξης $m \times m$ έχει $m$ ιδιοτιμές λ. Σε κάθε ιδιοτιμή λ, στο σύστημα $(A - \lambda I)x = 0$ προσδιορίζουμε το ιδιοδιάνυσμα $x$, αποκλείοντας το μηδενικό διάνυσμα που είναι και η προφανής λύση.

*Μεθοδολογία εύρεσης ιδιοτιμών και ιδιοδιανυσμάτων*

*Βήμα 1*. Υπολογισμός του πίνακα $[\mathrm{A} - \lambda \mathrm{I}]$,ο οποίος προκύπτει από τον πίνακα Α αφαιρώντας την παράμετρο λ από τα στοιχεία της κύρια διαγωνίου.
*Βήμα2. Υπολογισμός του* $\det(A - \lambda I)$
*Βήμα 3. Επίλυση της εξίσωσης* $\det(A - \lambda I) = 0$.Οι ρίζες αυτού του πολυωνύμου είναι οι ιδιοτιμές του Α.

*Ιδιότητες ιδιοτιμών και ιδιοδιανυσμάτων*

Έστω ένας τετραγωνικός πίνακας $A \in R^{n \times n}$ με ιδιοτιμές $\lambda_1, \lambda_2, \dots, \lambda_n$

1. Το άθροισμα των ιδιοτιμών είναι ίσο με το άθροισμα των στοιχείων της κύριας διαγωνίου, δηλαδή $\lambda_1 + \lambda_2 + \cdots + \lambda_n = tr(A)$.
2. Το γινόμενο των ιδιοτιμών είναι ίσο με την ορίζουσα του πίνακα $A$, $\lambda_1 . \lambda_2 . \dots \lambda_n = detA$.
3. Η ορίζουσα του $A$ είναι 0 αν και μόνο αν τουλάχιστον μία ιδιοτιμή είναι ίση με το 0.
4. Οι ιδιοτιμές ενός διαγώνιου ή ενός τριγωνικού είναι τα στοιχεία της κύριας διαγωνίου.
5. Οι $\lambda_1, \lambda_2, \dots, \lambda_n$ είναι ιδιοτιμές και του $A^T$.

6. Αν ο $A$ είναι αντιστρέψιμος τότε $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \dots, \frac{1}{\lambda_n}$ είναι ιδιοτιμές του $A^{-1}$ και $V_{1/\lambda_i}(A^{-1}) \equiv V_{\lambda_i}(A), \forall \lambda_i$.

### *1.1.10: Διαγωνοποίηση Πινάκων*

Ένας τετραγωνικός πίνακας $A \in M_n(\mathbb{F})$***ονομάζεται διαγωνοποιήσιμος***στο $\mathbb{F}$ αν είναι όμοιος με διαγώνιο πίνακα, δηλαδή αν υπάρχει αντιστρέψιμος πίνακας $P \in M_n(\mathbb{F})$ τέτοιος ώστε ο πίνακας $P^{-1}AP$ να είναι διαγώνιος.

Τότε εφαρμόζεται μία διαγωνοποίηση στον πίνακα $A$ και έχουμε:

$D = P^{-1}AP$ ή $A = PDP^{-1}$, όπου$D \in M_n(\mathbb{F})$.

Έτσι λέμε ότι ο $P$ διαγωνοποιεί τον $A$.

- Ένας πίνακας $A \in M_n(\mathbb{F})$ είναι διαγωνοποιήσιμος αν και μόνο αν ο πίνακας $A$ έχει n ανεξάρτητα ιδιοδιανύσματα.
  Δηλαδή:
  $P = (x_1 x_2 \dots x_n), D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \vdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$
- Αν ο πίνακας $A \in M_n(\mathbb{F})$ έχει n διακεκριμένες ιδιοτιμές, τότε ο πίνακας $A$ είναι διαγωνοποιήσιμος.

### *1.1.11: Γραμμικές απεικονίσεις*

Έστω, $U, V$ δύο $\mathbb{F}$ –διανυσματικοί χώροι. Μία απεικόνιση $f: U \to V$ ονομάζεται γραμμική απεικόνιση αν για κάθε $u_1, u_2 \in U, a \in \mathbb{F}$ ισχύουν τα εξής:

1. $f(u_1 + u_2) = f(u_1) + f(u_2)$
2. $f(au_1) = af(u_1)$

Όπου, $\mathbb{F}$ συμβολίζουμε το σύνολο των πραγματικών αριθμών R.

Το σύμβολο της πρόσθεσης στην πρώτη ισότητα αναφέρεται στην πράξη της πρόσθεσης στο διανυσματικό χώρο $U$, ενώ το σύμβολο της πρόσθεσης στην δεύτερη ισότητα αναφέρεται στην πράξη της πρόσθεσης στο διανυσματικό χώρο $V$.

Έστω, $A \in M_n(\mathbb{F})$, η απεικόνιση $c_A: M_{n\times 1}(\mathbb{F}) \to M_{m\times 1}(\mathbb{F})$ με $c_A(x) = Ax$ είναι γραμμική, διότι για κάθε $x, y \in M_{n\times 1}(\mathbb{F}), a \in (\mathbb{F})$, σύμφωνα με τις ιδιότητες των πράξεων πινάκων ισχύουν οι σχέσεις:

- $c_A(x + y) = A(x + y) = Ax + Ay = c_A(x) + c_A(y)$
- $c_A(ax) = A(ax) = a(Ax)$

Οι παραπάνω ισότητες επαληθεύουν τις ιδιότητες του ορισμού άρα η $c_A$ είναι γραμμική απεικόνιση.

# *Κεφάλαιο 2ο: Πολυμεταβλητή ανάλυση*

Τα συστήματα δομικών εξισώσεων (SEM) είναι μία στατιστική μέθοδος που υιοθετεί μία επικυρωτική μέθοδο στηνπολυμεταβλητή ανάλυση ενός μοντέλου,που αφορά κάποιες παρατηρήσεις ή μετρήσεις.Συνήθως χρησιμοποιούνται σαν μία επικυρωτική διαδικασία διαφόρων θεωρητικών υποθέσεων,καθώς δεν υπολογίζουν μόνο τις εκτιμήσεις για τους παράγοντες του μοντέλου αλλά εξετάζουν και τον βαθμό προσαρμογής τους με τα δεδομένα.

## *2.1:Εισαγωγή*

Η πολυμεταβλητή ανάλυση ασχολείται με στατιστικές μεθόδους συλλογής,περιγραφής και ανάλυσης δεδομένων που αποτελούνται από μετρήσεις πολλών μεταβλητών σε ένα πλήθος ατόμων ή γενικότερα πειραματικών μονάδων.Γιαπαράδειγμα,τα δεδομένα μπορεί να είναι μετρήσεις του ύψους και του βάρους κάποιων ατόμων.Στην περίπτωση αυτή,οι μεταβλητές είναι δύο: το ύψος και το βάρος.

Οι μέθοδοι της πολυμεταβλητής στατιστικής ανάλυσης αναφέρονται σε διαδικασίες ή μεθοδολογίες όπου προσπαθούμε να καταλήξουμε στη στατιστική συμπερασματολογία με χρήση πολλών μεταβλητών. Οι στατιστικές μέθοδοι είναι εκ'φύσεωςπολυμεταβλητές ή τουλάχιστον τα δεδομένα του ερευνητή είναι σχεδόν πάντα πολυμεταβλητά και έπειτα εξαρτάται από εκείνον, κατά το πόσο θα χρησιμοποιήσει όλα τα δεδομένα του για να αποκομίσει την μεγαλύτερη δυνατή πληροφορία.Όπως αναφέρεται οι πολυμεταβλητές δεν αναπτύχθηκαν ξεχωριστά από τις μονομεταβλητές,μόνο που εκείνες δεν είναι τόσο διαδεδομένες λόγω της πολυπλοκότητά τους.

Οι πολυμεταβλητές τεχνικές είναι ιδιαίτερα χρήσιμες για τους παρακάτω λόγους:

- Έχουμε περισσότερη πληροφορία(περισσότερες μεταβλητές ερμηνεύουν καλύτερα το φαινόμενο που μελετάμε).Όση περισσότερη πληροφορία έχει κανείς τόσο περισσότερο μπορεί να περιορίσει την αβεβαιότητα του και επομένως να εξαγάγει εγκυρότερα συμπεράσματα.
- Μελετάμε συσχετισμούς(μεταξύ μεταβλητών και μεταξύ υποκειμένων).Η ανακάλυψη συσχετίσεων (μεταξύ διαφορετικών πραγμάτων) ανάμεσα σε διαφορετικές μεταβλητές μπορεί από μόνη της να οδηγήσει σε καινούργιες ερμηνείες για τα υπό μελέτη φαινόμενα.Επομένως,είναι καλή ιδέα να μελετήσουμε συγχρόνως ένα σύνολο μεταβλητών με σκοπό να αντλήσουμε όσο γίνεται περισσότερα από τα δεδομένα μας.

Επίσης ,οι πολυμεταβλητές τεχνικές χρησιμοποιούνται για:

- *Την εύρεση και ερμηνεία συσχετίσεων μεταξύ των μεταβλητών*

  Πολλά φαινόμενα συνήθως παρουσιάζουν πολυπλοκότητα και χρειάζονται παραπάνω από μία μεταβλητή για την καλή ερμηνεία των αποτελεσμάτων. Επομένως χρειάζεται να εξετάσουμε και να συμπεριλάβουμε το πώς συσχετίζονται οι διάφορες μεταβλητές προκειμένου να αποκτήσουμε καλύτερη και εγκυρότερη γνώση για το φαινόμενο που εξετάζουμε.
- *Την δημιουργία ομάδων είτε από παρατηρήσεις είτε από μεταβλητές σύμφωνα με τα χαρακτηριστικά.*

  Για παράδειγμα, ένας γιατρός ενδιαφέρεται να ομαδοποιήσει διάφορες μεταβλητές, έστω ότι οι μεταβλητές αυτές είναι τα διάφορα χαρακτηριστικά του αρρώστου (τα οποία μπορεί να είναι ποσοτικά όπως ο αριθμός αιμοπεταλίων είτε ποιοτικά όπως η παρουσία ή απουσία ενός συμπτώματος), ο γιατρός ενδιαφέρεται να δει ποια από τα χαρακτηριστικά εμφανίζονται μαζί και επομένως κάποιες από τις εξετάσεις είναι πλεονάζουσες και θα μπορούσαν να αποφευχθούν.
- *Τη μείωση των διαστάσεων του προβλήματος*

  Στα φαινόμενα που μελετώνται συχνά υπάρχουν μεταβλητές οι οποίες είτε είναι άχρηστες για το σκοπό της έρευνας μας είτε η πληροφορία που μας προσφέρουν παρέχεται και σε κάποια άλλη μεταβλητή και επομένως εκείνες οι μεταβλητές θεωρούνται πλεονάζουσες. Οι πλεονάζουσες μεταβλητές από στατιστικής απόψεως θα μπορούσαν να δημιουργήσουν σημαντικά προβλήματα στην στατιστική ανάλυση. Άρα θα ήταν προτιμότερο να μπορούσε κανείς να μειώσει τις υποψήφιες μεταβλητές. Επομένως για να προχωρήσουμε στην στατιστική ανάλυση χρειάζεται να δημιουργήσουμε καινούργιες μεταβλητές οι οποίες κατά κάποιον τρόπο θα περιλαμβάνουν μεγάλο μέρος της πληροφορίας που είχε ένα μεγάλο μέρος των αρχικών μεταβλητών,ώστε να προκύψουν λίγες μεταβλητές και να είναι δυνατή η διεξαγωγή της στατιστικής ανάλυσης.
- *Την πρόβλεψη νέων τιμών*

  Αυτό έχει να κάνει με δύο διαφορετικές περιπτώσεις. Η πρώτη περίπτωση αφορά τις χαμένες παρατηρήσεις, αυτό σημαίνει ότι σε πολυμεταβλητά προβλήματα συχνό φαινόμενο είναι για κάποια παρατήρηση να λείπει η τιμή της. Η δεύτερη περίπτωση αναφέρεται στο ότι η τιμή που μας λείπει είναι η τιμή που μας έδειχνε σε ποια ομάδα ανήκει η παρατήρηση. Σε αυτή την περίπτωση θέλουμε, από παρατηρήσεις για τις οποίες η κατάταξη είναι σε ομάδες μασκαι είναι ήδη γνωστή, να κατασκευάσουμε κανόνες ώστε να μπορούμε να δημιουργήσουμε νέες παρατηρήσεις. Ένα παράδειγμα είναι το αν μπορεί να δοθεί ή όχι δάνειο σε κάποιον υποψήφιο πελάτη της τράπεζας. Από τα δεδομένα που έχει κάποιος εργαζόμενος της τράπεζας στα χέρια του γνωρίζει αν ένα δάνειο αποπληρώθηκε κανονικά ή όχι καθώς και όλα τα στοιχεία του δανειολήπτη. Έτσι λοιπόν, μπορεί να δημιουργήσει έναν ‘’κανόνα’’ σχετικά με το ποιά χαρακτηριστικά είναι που επηρεάζουν τον δανειολήπτη να αποπληρώσει το δάνειο του κανονικά ή όχι. Όταν λοιπόν ένας

καινούργιος πελάτης ζητήσει δάνειο θα μπορέσει ο ερευνητής χρησιμοποιώντας τον κανόνα που δημιούργησε να τον κατατάξει στην κατηγορία των ‘’κακών’’ ή ‘’καλών’’ πελατών ανάλογα με το αν θα αποπληρώσει το δάνειο ή όχι.

- *Μοντελοποίηση σε πολλές διαστάσεις*
  Σε πολλά πολυμεταβλητά μοντέλα υπάρχουν περισσότερες από μία εξαρτημένες μεταβλητές. Τέτοια μοντέλα μας επιτρέπουν να λάβουμε υπόψη τυχόν συσχετίσεις ανάμεσα στις μεταβλητές.
- *Ποσοτικοποίηση μη παρατηρήσιμων ποσοτήτων*
  Κάποιες από τις πολύμεταβλητέςτεχνικές επιτρέπουν να δημιουργηθούν συνδυασμοί άλλων μετρήσιμων μεταβλητών οι οποίοι στη συνέχεια θα θεωρηθούν ότι ποσοτικοποιούν την αφηρημένη και μη μετρήσιμη έννοια.

## *2.2: Πολυμεταβλητές μέθοδοι*

Υπάρχουν τα παρακάτω είδη πολυμεταβλητών μεθόδων:

- Παραγοντική Ανάλυση (Factor Analysis)
- Ανάλυση σε κύριες συνιστώσες (Principal Component Analysis)
- Ανάλυση κατά συστάδες (Cluster Analysis)
- Διαχωριστική Ανάλυση (Discriminant Analysis)
- Ανάλυση Αντιστοιχιών (Corresponde Analysis)
- Ανάλυση Κανονικών συσχετίσεων (Canonical Correlation Analysis)
- Πολυδιάστατη Κλιμακοποίηση (Multidimensional Scaling)
- Πολυμεταβλητό Γραμμικό μοντέλο (Multivariate Linear Model)

Εμάς, όμως μας απασχολεί να αναλύσουμε τις μεθόδους που χρησιμοποιούνταιαλλά και μας ενδιαφέρουν στα δομικά μοντέλα εξισώσεων. Άρα θα ασχοληθώ και θα αναλύσω τις παρακάτω πολυμεταβλητές μεθόδους:

### *2.2.1: Παραγοντική Ανάλυση(Factor Analysis)*

Πρόκειται για μία πολυμεταβλητή μέθοδο μείωσης της διάστασης,που σκοπός της είναι να περιγράψει τις σχέσεις συνδιασποράς ή εξάρτησης μεταξύ πολλών μεταβλητών με την βοήθεια κάποιων τυχαίων άγνωστων ποσοτήτων που λέγονται **παράγοντες(factors)**. Η παραγοντική ανάλυση προσπαθεί να απλοποιήσει πολύπλοκες και αντίθετες σχέσεις που υπάρχουν μεταξύ ενός συνόλου παρατηρήσιμων μεταβλητών ανακαλύπτοντας κάποιους νέους κοινούς παράγοντες που ομαδοποιούν με κάποιο τρόπο τις μεταβλητές αυτές.

Όταν πρωτοανακαλύφθηκε η μέθοδος αυτή από τον KarlPearson στην προσπάθεια του να ορίσει και να μετρήσει ένα μέγεθος, την‘’νοημοσύνη’’,υπήρχαν πολλές αμφιβολίες σχετικά με την χρησιμότητα της.Αρχικά χρησιμοποιήθηκε σε επιστήμες όπως η ψυχολογία και αργότερα με την εξέλιξη της τεχνολογίας και αντίστοιχα των ηλεκτρονικών υπολογιστών υπήρξε και η ανάπτυξη της παραγοντικής ανάλυσης.

Το παραγοντικό μοντέλο μπορεί να περιγράφει ως εξής:

Υποθέτουμε ότι οι μεταβλητές μπορούν να ομαδοποιηθούν με βάση τις συσχετίσεις τους.Έτσι όλες οι μεταβλητές σε μία συγκεκριμένη ομάδα είναι υψηλά συσχετισμένες μεταξύ τους αλλά έχουν σχετικά μικρές συσχετίσεις με μεταβλητές των άλλων ομάδων.Για κάθε ομάδα μεταβλητών ένας παράγοντας είναι υπεύθυνος για τις παρατηρούμενες συσχετίσεις.

*2.2.1.1: Είδη παραγοντικής Ανάλυσης*

Μία μελέτη παραγοντικής ανάλυσης μπορεί να γίνει εφαρμόζοντας δύο διαφορετικά πλαίσια ανάλυσης δεδομένων.**Τη διερευνητική παραγοντική ανάλυση** και την **επιβεβαιωτική παραγοντική ανάλυση**.

- ***Η διερευνητική παραγοντική ανάλυση (exploratory factor analysis)*** βασίζεται στην υπόθεση ότι η διακύμανση μιας παρατηρήσιμης μεταβλητής είναι συνάρτηση ενός αριθμού από παράγοντες,οι οποίοι αντιστοιχούν στις ποικίλες διαστάσεις της επίδοσης που εκπροσωπεί αυτή η μεταβλητή.Η εφαρμογή προτείνεται όταν ο τρόπος με τον οποίο οι παρατηρήσιμες μεταβλητές ανάγονται σε λανθάνουσες δομές είναιάγνωστος. Το μοντέλο που προκύπτει μετά από αυτή την μέθοδο δείχνει τις συνδέσεις ανάμεσα στις παρατηρήσιμες μεταβλητές και τους παράγοντες,καθώς και το πρότυπο συσχετίσεων μεταξύ των παραγόντων,όμως δεν περιγράφονται οι σχέσεις μεταξύ αυτών.Το μοντέλο αυτό λέγεται **μοντέλο μέτρησης**.Καθώς το μοντέλο που προκύπτει μπορεί να έχει μεγάλο ή μικρό αριθμό παραγόντων γίνεται δύσκολη η απόδοση του νοήματος στο μοντέλο.
- ***Η επιβεβαιωτική παραγοντική ανάλυση (confirmatory factor analysis)*** εφαρμόζεται για να ελεγχθεί ότι υπάρχει ένα συγκεκριμένο πρότυπο σχέσεων μεταξύ των παρατηρήσιμων μεταβλητών και των παραγόντων.Το μοντέλο που περιγράφει το πρότυπο των σχέσεων που θα ελεγχθεί και διατυπώνεται με βάση την προυπάρχουσα σχετική ερευνητική εμπειρίαονομάζεται πλήρες καθώς, αποτελείται τόσο από ένα μοντέλο μέτρησης που απεικονίζει τις συνδέσεις ανάμεσα στις παρατηρήσιμες και τις λανθάνουσες μεταβλητές και τις συσχετίσεις ανάμεσα στις τελευταίες αλλά και από ένα μοντέλο δομικών σχέσεων το οποίο απεικονίζει τις συσχετίσεις και σχέσεις εξάρτησης ανάμεσα στις ίδιες λανθάνουσες μεταβλητές.Η επιβεβαιωτική παραγοντική ανάλυση επιτρέπει τον προσδιορισμό του μέρους της διακύμανσης της κάθε μεταβλητής που εξηγείται από τους παράγοντες που συνδέονται με αυτή. Η σχέση αυτή περιγράφεται με την μορφή δομικής εξίσωσης,μπορεί δηλαδή να υπολογίσει ποιοι παράγοντες και σε ποιό βαθμό σχετίζονται με τις συγκεκριμένες παρατηρήσιμες μεταβλητές. Στόχος της μεθόδου είναι, ο σχεδιασμός ενός από πριν καθορισμένου μοντέλου το οποίο περιγράφει όλες τις σχέσεις που αναμένεται να υπάρχουν ανάμεσα τους και στους παράγοντες.

Η μέθοδος ελέγχει την καταλληλότητα ενός από πριν διατυπωμένου μοντέλου.

*2.2.1.2:Υπόδειγμα μεθόδου*

Έστω ένα σύνολο μεταβλητών $Y = [Y_1, Y_2, \dots, Y_p]'$ με μέση τιμή $\mu$ και πίνακα διασποράς-συνδιασποράς $\Sigma$, σύφωνα με το γενικό μοντέλο της παραγοντικής ανάλυσης μπορούν να δημιουργηθούν $m$ −παράγοντες οι οποίοι απεικονίζονται ως $F = [F_1, F_2, \dots, F_m]'$. Οπότε, το υπόδειγμα θα έχει την μορφή:

$$\boldsymbol{X_i = \sum_{j=1}^{m} W_{ij} F_j + \varepsilon_i = W_{i1} F_1 + W_{i2} F_2 + \cdots + W_{im} F_m + \varepsilon_i} \textbf{ (2.2.1.2)}$$

Όπου,$W_{ij}$:πίνακας των συντελεστών των παραγόντων

$\varepsilon$: είναι το διάνυσμα των σφαλμάτων

Το σφάλμα $\varepsilon_i$ θεωρείται ο μοναδικός παράγοντας της $i$ παρατήρησης και είναι το μέρος της μεταβλητής που δεν μπορεί να εξηγηθεί από τον παράγοντα (Johnson 1998).

Η τιμή των παραγόντων $m$πρέπει να είναι μικρότερη από το πλήθος των μεταβλητών $p$, διαφορετικά δεν επιτυγχάνεται περικοπή του όγκου του προβλήματος αλλά απλά ένας διαχωρισμό τους .

*2.2.1.3: ΟΡΘΟΓΩΝΙΟ ΠΑΡΑΓΟΝΤΙΚΟ ΜΟΝΤΕΛΟ(orthogonalfactormodel)*

Έστω διάνυσμα $X$, με $p$μεταβλητές το οποίο έχει μέσο $\mu$ και πίνακα διασπορών-συνδιασπορών $\Sigma$.

Το μοντέλο παραγοντικής ανάλυσης μπορεί να γραφεί στην μορφή πίνακα:

$$\boldsymbol{X - \mu = LF + \varepsilon} \textbf{ (2.2.1.3)}$$

Όπου,

$\boldsymbol{X}$:είναι το διάνυσμα αρχικών μεταβλητών μεγέθους $p \times 1$.

$\boldsymbol{\mu}$:είναι το διάνυσμα μέσων μεγέθους $p \times 1$.

$\boldsymbol{L}$:είναι πίνακας $p \times k$ όπου το $L_{ij}$ είναι η επιβάρυνση του παράγοντα $F_j$ στην μεταβλητή $X_j$.

$\boldsymbol{F}$:$k \times 1$ διάνυσμα με τους παράγοντες

$\boldsymbol{\varepsilon}$:σφάλμα ή ειδικός παράγοντας

Ή σε μορφή εξισώσεων όπως φαίνεται παρακάτω:

$$\begin{aligned}
\boldsymbol{X_1 - \mu_1} &= \boldsymbol{l_{11} F_1 + l_{12} F_2 + \cdots + l_{1m} F_m + \varepsilon_1} \\
\boldsymbol{X_2 - \mu_2} &= \boldsymbol{l_{21} F_1 + l_{22} F_2 + \cdots + l_{2m} F_m + \varepsilon_2} \\
&\vdots \\
\boldsymbol{X_p - \mu_p} &= \boldsymbol{l_{p1} F_1 + l_{p2} F_2 + \cdots + l_{pm} F_m + \varepsilon_p}
\end{aligned}$$

**Με βάση τις παραπάνω εξισώσεις πρέπει να επισημανθούν τα εξής:**

1. Τα $X_i$ δεν είναι παρατηρήσεις αλλά μεταβλητές,όπως επίσης και ότι το δεξί μέλος της εξίσωσης δεν είναι παραητρήσιμο και έτσι πρέπει να εκτιμηθεί.
2. Οι παράγοντες $F_i$ μπορούν να γραφούν σαν γραμμικό συνδυασμό των μεταβλητών.Αυτό είναι χρήσιμο να συμβαίνει όταν θέλουμε να δημιουργήσουμε νέες μεταβλητές. Οι συντελεστές κάθε παράγοντα, όταν εκφράζουμε τις μεταβλητές ως γραμμικό συνδυασμό των παραγόντων καλούνται επιβαρύνσεις. Ενώ αντίστοιχα, οι συντελεστές κάθε μεταβλητής όταν εκφράζουμε κάθε παράγοντα ως γραμμικό συνδυασμό των μεταβλητών, καλούνται συντελεστές των σκορ.
3. Οι παράγοντες έχουν την ίδια διακύμανση. Μία βασική διαφορά με την ανάλυση σε κύριες συνιστώσες είναι ότι θέλουμε οι κύριες συνιστώσες να είναι σε φθίνουσα σειρά διακύμανσης.
4. Άλλη μία κύρια διαφορά με την ανάλυση σε κύριες συνιστώσες είναι πως εδώ το μοντέλο προσπαθεί να εκφράσει τις μεταβλητές ως γραμμικό συνδυασμό των παραγόντων, ενώ στην ανάλυση σε κύριες συνιστώσες μας ένοιαζε περισσότερο να εκφράσουμε τις κύριες συνιστώσες ως γραμμικό συνδυασμό των αρχικών μεταβλητών.

*2.2.1.4: Βασικές προυποθέσεις*

Για να λειτουργήσει ένα υπόδειγμα πρέπει να ικανοποιούνται οι παρακάτω προυποθέσεις:

- *Οι μεταβλητές είναι ποσοτικές, συνεχείς*
- *Οι σχέσεις μεταξύ των μεταβλητών πρέπει να είναι γραμμικές*
- *Οι μεταβλητές θα πρέπει να συσχετίζονται επαρκώς μεταξύ τους, αλλά όχι υπερβολικά.*

Ακόμη,

**1.** $\mathrm{E}(\mathrm{F}) = 0$
**2.** $\mathrm{Cov}(\mathrm{F}) = \mathrm{I}$
**3.** $\mathrm{E}(\varepsilon) = 0$
**4.** $\mathrm{Cov}(\varepsilon) = \Psi$, όπου $\Psi$ είναι ο διαγώνιος πίνακας της μορφής:

$$\Psi = \begin{bmatrix} \psi_1 & 0 & \dots & 0 \\ \vdots & \psi_2 & \dots & \vdots \\ 0 & 0 & \dots & \psi_\mathrm{p} \end{bmatrix}$$

**5.** $\mathrm{Cov}(\varepsilon_i, F_i) = 0$ ,για κάθε$\mathrm{i} \neq \mathrm{j}$.

Με βάση την σχέση (2.2.1.2) και τις παραπάνω υποθέσεις μπορούμε να ορίσουμε συνοπτικά το ορθογώνιο παραγοντικό μοντέλο ως εξής:

Ορθογώνιοπαραγοντικόμοντέλομεmκοινούςπαράγοντες
(Orthogonalfactormodelwithcommonfactors)

**X = μ + LF + ε(2.2.1.4)**

(px1) (px1) (pxm)(mx1) (px1)

Όπου,

$$\mu_i = μέσοςτηςίμεταβλητής$$

$\varepsilon_i = iειδικόςπαράγοντας(ithspecificfactor)$
$F_j = jκοινόςπαράγοντας(jthcommonfactor) l_{ij} =$
$στάθμισητηςίμεταβλητήςστονjπαράγοντα$
*Ταμηπαρατηρήσιματυχαίαδιανύσματα*$F$*καιεικανοποιούνταεξής:*
$F$*καιεανεξάρτητα*

$$E(F) = 0$$
$$Cov(F) = I$$
$$E(\varepsilon) = 0$$

$Cov(\varepsilon) = \Psi$*,όπου*$\Psi$*είναιοδιαγώνιοςπίνακας*

Δηλαδή με βάση τις παραπάνω υποθέσεις, υποθέτουμε ότι: (Καρλής 2005)

- οι μοναδικοί και οι κοινοί παράγοντες είναι ασυσχέτιστοι(υπόθεση 5)
- οι παράγοντες αλλά και μοναδικοί παράγοντες είναι ασυσχέτιστοι μεταξύ τους(υποθέσεις2,4) και έχουν μηδενικές μέσες τιμές(υπόθεση 1)
- Η υπόθεση 2 σημαίνει ότι οι παράγοντες είναι ορθογώνιοι μεταξύ τους γι΄αυτό και το μοντέλο ονομάζεται ορθογώνιο.
- Τα δεδομένα προέρχονται από πολυμεταβλητούς κανονικούς πληθυσμούς, κάτι το οποίο χρησιμοποιείται ως βάση για ελέγχους καλής προσαρμογής του μοντέλου καθώς και για την εκτίμηση μέγιστης πιθανοφάνειας.

Όπως είχαμε δει από τις παραπάνω υποθέσεις η συνδιακύμανση μεταξύ F και ε είναι μηδέν. Συνεπώς, ο πίνακας συνδιακύμανσης μπορεί να διασπαστεί σε δύο μέρη, το πρώτο είναι το κομμάτι που ερμηνεύουν οι κοινοί παράγοντες και ονομάζεται ***εταιρικότητα (communality)***και το δεύτερο οφείλεται στους μοναδικούς παράγοντες και άρα το μοντέλο δεν μπορεί να ερμηνεύσει και ***ονομάζεται ιδιαιτερότητα (specificity).***

*2.2.1.5: Βήματα εφρμογής παραγοντικής ανάλυσης*

Τα βήματα που πρέπει να ακολουθήσει κανείς για την εφαρμογή της παραγοντικής ανάλυσης είναι τα εξής:

1. Έλεγχος ικανοποιητικών συσχετίσεων ώστε να εφαρμοστεί η παραγοντική ανάλυση.
2. Εύρεση του αριθμού των παραγόντων και εκτίμηση των παραγόντων του μοντέλου.

3. Περιστροφή του μοντέλου με στόχο να αυξήσουμε την ερμηνευτική του μοντέλου
4. Εκτίμηση των σκορ του μοντέλου για περαιτέρω στατιστική χρήση

*2.2.1.6:Αριθμός παραγόντων και εκτίμηση παραγόντων*

Για να βρεθεί ο αριθμός των παραγόντων μπορούν να χρησιμοποιηθούν οι τιμές των ιδιοτιμών του πίνακα διακύμανσης-συνδιακύμανσης, τιμές που εξηγούν κάποιο ποσοστό διακύμανσης ή scree-plot(διάγραμμα ιδιοτιμών ως προς τον άξονα των αριθμών τους).
Για να εκτιμήσει κανείς το μοντέλο χρειάζεται ο αριθμός των παραγόντων, έτσι ώστε να δουλέψει με διαδοχικά αυξανόμενο αριθμό παραγόντων και να κρατήσει το μοντέλο με βάση κάποιο κριτήριο καλής προσαρμοστικότητας.

- Από τον πίνακα επιβαρύνσεων μπορεί να εκτιμήσει τον πίνακα Σ.
- Έλεγχος λόγου πιθανοφανειών αν οι εκτιμήσεις έχουν γίνει με μέθοδο μέγιστηςπιθανοφάνειας.

Υπάρχουν δύο βασικοί μέθοδοι εκτίμησης: ***Η μέθοδος κύριων συνιστωσών*** και ***μέθοδος μέγιστης πιθανοφάνειας.***

- *Η μέθοδος κύριων συνιστωσών (principalcomponentmethod)*
  Ουσιαστικά είναι μία τροποποίηση της πολυμεταβλητής στατιστικής μεθόδου των κύριων συνιστωσώνπου χρησιμοποιείται από την παραγοντική ανάλυση. Με την μέθοδο αυτή δεν αλλάζουν οι επιβαρύνσεις των παραγόντων, δηλαδή οι τιμές των συντελεστών μένουν αμετάβλητες είτε προσθέσουμε ή αφαιρέσουμε κάποιον παράγοντα. Ακόμη, η μέθοδος αυτή μπορεί να εκτιμήσει όσους παράγοντες θέλει χωρίς όμως να ξέρουμε αν δουλεύει καλά ή όχι. Η μέθοδος κύριων συνιστωσών εξαρτάται από τις μονάδες μέτρησης κάτι που μας αναγκάζει να διαλέξουμε ανάμεσα σε ένα πίνακα διασποράς-συνδιασποράς $\Sigma$ είτε ενός πίνακας συσχέτισης $R$. Τέλος, με την εν λόγω μέθοδο μπορούν να υπολογιστούν ακριβώς τα σκορ.
- *Η μέθοδος μέγιστης πιθανοφάνειας (maximumlikelihoodmethod)*
  Θεωρείται κατάλληλη όταν τα δεδομένα μας ακολουθούν κανονική κατανομή και παρέχει ισχυρούς στατιστικούς ελέγχους για την επιλογή καταλληλότερου μοντέλου. Σε αντίθεση με την μέθοδο των κύριων συνιστωσών η εν λόγω μέθοδος μπορεί να περιλαμβάνει ασήμαντους παράγοντες στην ανάλυση.
  Όπως ακόμη, η μέθοδος μέγιστης πιθανοφάνειας αν έχει κάποιο πρόβλημα το μοντέλο υπάρχει περίπτωση να μην δουλεύει, σε αντίθεση με την μέθοδο των κύριων συσνιστωσών.
  Όπως και στην παραπάνω μέθοδο έτσι και σε αυτή, μπορούμε να χρησιμοποιήσουμε είτε έναν πίνακα διασποράς-συνδιασποράς $\Sigma$ είτε έναν πίνακα συσχέτισης $R$,με την μόνη διαφορά ότι η μέθοδος αυτή δεν μεταβάλλεται από τις μονάδες μέτρησης σε αντίθεση με την μέθοδο κύριων συνιστωσών *(Johnson&Wicher 1998)*.

Ο πίνακας συσχετίσεων περιέχει σαν στοιχεία τους συντελεστές του Pearson για κάθε ζευγάρι μεταβλητών. Ο συντελεστής συσχέτισης του Pearson μετράει την γραμμική συσχέτιση ανάμεσα στις μεταβλητές για ζεύγη ποσοτικών μεταβλητών.

*Πίνακας συσχετίσεων R*

$$R = \begin{bmatrix} 1 & r_{12} \cdots & r_{1p} \\ r_{21} & 1 & \cdots r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} r_{p2} \cdots & 1 \end{bmatrix}$$

Ο συντελεστής συσχέτισης του Pearson δύο μεταβλητών $X$ και $Y$ ορίζεται με βάση ένα δείγμα $p$ ζευγών παρατηρήσεων $(x_i, y_i), i = 1,2, \dots, p$συμβολίζεται με $r(X,Y)$ ή με $r$:

$$r = \frac{\sum_{i=1}^{p}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{p}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{p}(y_i - \bar{y})}}$$

*Έλεγχος συσχετίσεων*

*Για να εφαρμοστεί η παραγοντική ανάλυση όπως αναφέρθηκε και πιο πάνω πρέπει να υπάρχουν συσχετίσεις ανάμεσα στις μεταβλητές. Αν δεν υπάρχουν, δεν έχει νόημα η εφαρμογή τους καθώς δεν θα βρούμε κοινούς παράγοντες που να μας επιτρέψουν να εργαστούμε με αυτούς.*

Άρα, αυτό που μας ενδιαφέρει είναι:

- Να υπάρχουν αρκετά μεγάλες συσχετίσεις τουλάχιστον σε μεγάλο ποσοστό του πίνακα συσχετίσεων.
- Αν υπάρχουν μεταβλητές ασυσχέτιστες να τις αγνοήσουμε, καθώς αφού δεν θα συσχετίζονται με τις άλλες θα προκύψουν από μόνες τους ως ένας ξεχωριστός παράγοντας κάτι που θα δυσκολέψει την ερμηνεία των αποτελεσμάτων.

*Μερικός συντελεστής συσχέτισης*

Υπολογίζει τη συσχέτιση μεταξύ δύο μεταβλητών αγνοώντας τις υπόλοιπες. Ένας συντελεστής συσχέτισης ο οποίος υπολογίζει τη συσχέτιση αφού αφαιρέσει την επίδραση των υπολοίπων ονομάζεται, μερικός συντελεστής συσχέτισης. Για την εφαρμογή παραγοντικής ανάλυσης μας ενδιαφέρει οι μερικοί συντελεστές συσχέτισης να είναι μικροί.

Μέτρα

- ***Μέτρο Kaiser-Meyer-Olkin (KMO)****:* το οποίο χρησιμοποιείτε για να συγκρίνουμε το σχετικό μέγεθος των συντελεστών συσχέτισης με τους μερικούς συντελεστές συσχέτισης. Αν η τιμή του είναι μεγάλη τότε τα δεδομένα μας είναι κατάλληλα, τιμές $\leq 0{,}5$ είναι κακές τιμές, τιμές κοντά στο 0,8 είναι καλές και τιμές $\leq 0{,}8$ δεν είναι κατάλληλες καθώς δεν θα μας δώσουν ικανοποιητικά αποτελέσματα.
- ***Μέτρο δειγματικής καταλληλότητας*****:** το οποίο μας επιτρέπει την εξέταση των μεταβλητών μία-μία και κατά το πόσο είναι κατάλληλες να χρησιμοποιηθούν στην ανάλυση. Τιμές κοντά στο 1 είναι πολύ καλές.

*2.2.1.7: Περιστροφή*

Προσπαθούμε να κάνουμε τους παράγοντες πιο ερμηνεύσιμους, χωρίς να αλλάζουν κάποια από τα χαρακτηριστικά του μοντέλου παρά μόνο οι τιμές των επιβαρύνσεων.
***Μέθοδοι περιστροφής***

- *Varimax:*Ελαχιστοποιεί τον αριθμό των μεταβλητών με μεγάλες επιβαρύνσεις
- *Quartimax:*Ελαχιστοποιεί τον αριθμό των παραγόντων που εξηγούν μία μεταβλητή
- *Equimax*:Συνδυασμός των δύο παραπάνω
- *Oblique*:Μη ορθογώνια περιστροφή. Οι άξονες που προκύπτουν δεν είναι πια ορθογώνιοι και οι παράγοντες δεν είναι ανεξάρτητοι. Τον χρησιμοποιούμε όταν δεν θέλουμε οι παράγοντες που προκύπτουν να είναι ασυσχέτιστοι.

*2.2.1.8: Υπολογισμός σκορ*

Για να επιτευχθεί ο σκοπός της παραγοντικής ανάλυσης ο οποίος είναι η μείωση των μεταβλητών, αρκεί να δημιουργήσουμε καινούργιες μεταβλητές, τους παράγοντες ως γραμμικούς συνδυασμούς των αρχικών μεταβλητών έτσι ώστε ξεκινώντας από 10 μεταβλητές να μας μείνουν έστω 4 νέες, οι νέοι παράγοντες. Κάθε παράγοντας μπορεί να γραφτεί:

$$\boldsymbol{F_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1p}X_p}$$
$$\boldsymbol{F_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2p}X_p}$$
$$\vdots$$
$$\boldsymbol{F_k = \alpha_{k1}X_1 + \alpha_{k2}X_2 + \cdots + \alpha_{kp}X_p}$$

Όπου οι συντελεστές $\alpha_{ij}$ είναι τα σκορ της μεταβλητής $X_j$ στον παράγοντα $F_i$.

*2.2.1.9: Μορφές παραγοντικής Ανάλυσης*

Οι δύο μορφές παραγοντικής ανάλυσης είναι η R-mode παραγοντική ανάλυση και η Q-mode παραγοντική ανάλυση. Ο συμβολισμός R και Q αναπτύχθηκε στον κλάδο της ψυχολογίας από τον RayCattell το 1966. Ένας απλός πίνακας δεδομένων έχει στις

γραμμές τις παρατηρήσεις και στις στήλες τις μεταβλητές και συνήθως ο αριθμός των γραμμών είναι μεγαλύτερος από τον αριθμό των στηλών. Για να επιτευχθεί μείωση της πολυπλοκότητας των γραμμών χρησιμοποιείται R-mode, ενώ αντίστοιχα για την μείωση της πολυπλοκότητας των στηλών χρησιμοποιείται Q-mode.

- <u>R-mode παραγοντική ανάλυση</u>

  Στην περίπτωση αυτή το ζητούμενο είναι οι εσωτερικές σχέσεις μεταξύ των μεταβλητών υπολογίζοντας τους συντελεστές συσχέτισης (Παπαθεοδώρου 2009). Στο μοντέλο της παραγοντικής ανάλυσης υποθέτουμε ότι οι p μεταβλητές μπορούν να γραφούν ως γραμμικός συνδυασμός των k παραγόντων , δηλαδή

  $$\boldsymbol{X} = \boldsymbol{L}\mathrm{F} + \boldsymbol{\varepsilon} \quad \textbf{(2.2.1.9)}$$
  $$\mathrm{px1} \quad \mathrm{pxk}kx1 \quad px1$$

Όπου,

- $\boldsymbol{X}$**:** Διάνυσμα των αρχικών μεταβλητών px1.
- $\boldsymbol{L}$**:** είναι ο p × kπίνακας που περιέχει τους συντελεστές βαρύτητας $(L_{ij})$ του παράγοντα$F_i$ στην μεταβλητή $X_i$
- $\boldsymbol{F}$**:** είναι ο k × 1 πίνακας που περιέχει τους παράγοντες
- $\boldsymbol{\varepsilon}$: είναι το διάνυσμα των σφαλμάτων ή μοναδικός παράγοντας

Πρέπει να τονίσουμε ότι ο αριθμός k των παραγόντων πρέπει να είναι μικρότερος από τον αριθμό p των μεταβλητών αλλιώς δεν υφίσταται λόγος παραγοντικής ανάλυσης. Κάθε μεταβλητή μπορεί να γραφεί στην ακόλουθη μορφή:

$$\begin{aligned}
\boldsymbol{X_1} &= \boldsymbol{L_{11}F_1} + \boldsymbol{L_{12}F_2} + \cdots + \boldsymbol{L_{1k}F_k} + \boldsymbol{\varepsilon_1} \\
\boldsymbol{X_2} &= \boldsymbol{L_{21}F_1} + \boldsymbol{L_{22}F_2} + \cdots + \boldsymbol{L_{2k}F_k} + \boldsymbol{\varepsilon_2} \\
&\vdots \\
\boldsymbol{X_p} &= \boldsymbol{L_{p1}F_1} + \boldsymbol{L_{p2}F_2} + \cdots + \boldsymbol{L_{pk}F_k} + \boldsymbol{\varepsilon_p}
\end{aligned}$$

Βήματα στην R-mode παραγοντική ανάλυση:

1. Δημιουργία πίνακα δεδομένων
2. Υπολογισμός του πίνακα δεδομένων ή συντελεστών συσχέτισης
3. Υπολογισμός ιδιοτιμών και ιδιοδιανυσμάτων
4. Υπολογισμός παραγοντικών φορτίσεων (είναι οι συντελεστές συσχέτισης των μεταβλητών με τους παράγοντες)
5. Επιλογή διατηρούμενων παραγόντων
6. Περιστροφή των παραγόντων
7. Υπολογισμός των παραγοντικών τιμών

- <u>Q-mode παραγοντική ανάλυση</u>

Στην περίπτωση αυτή το ζητούμενο είναι η ανίχνευση εσωτερικών σχέσεων μεταξύ των δειγμάτων. Είναι μία δεύτερη μορφή παραγοντικής ανάλυσης κατά την οποία οι

ρόλοι των μεταβλητών και των θέσεων των παρατηρήσεων είναι αντεστραμμένοι. Η ανάλυση αυτή αναφέρεται σε αλληλοσυσχετίσεις μεταξύ των θέσεων των παρατηρήσεων. Οι στόχοι της είναι να μπορεί να διαμορφώσει ένα σύνολο παρατηρήσεων σε μία λογική σειρά έτσι ώστε να μπορεί να εξάγει σχέσεις μεταξύ των θέσεων των παρατηρήσεων.

Βήματα στην Q-mode παραγοντική ανάλυση:

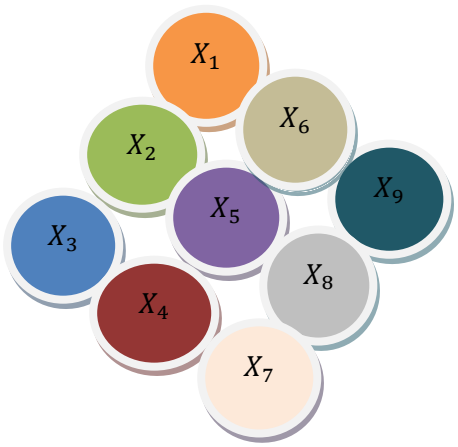
1. Δημιουργία πίνακα συντελεστή συσχέτισης μεταξύ των θέσεων των παρατηρήσεων
2. Εύρεση κύριων αξόνων με την βοήθεια των ιδιοτιμών και ιδιοδιανυσμάτων
3. Περιγραφή των παραγόντων σε θέση όπου η διακύμανση των φορτίων των παρατηρήσεων μεγιστοποιείται
4. Ερμηνεία ακραίων παρατηρήσεων για την εξαγωγή συμπερασμάτων

*2.2.1.10: Παράδειγματα*

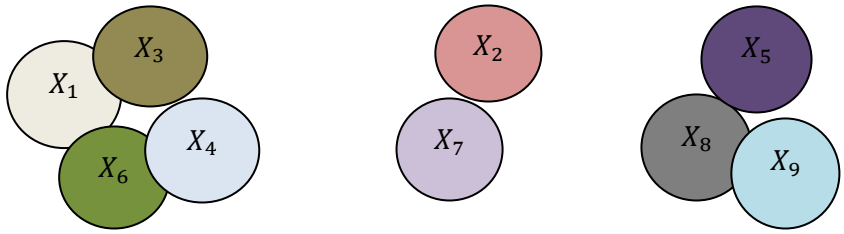
**<u>Παράδειγμα</u>**

Έστω οι μεταβλητές $X_1, X_2, \ldots, X_9$ τις οποίες θα τις μειώσουμε σε 3 κοινούς παράγοντες.

*Εικόνα 1*




Εννέα τυχαίες μεταβλητές



| Παράγοντας 1 | Παράγοντας 2 | Παράγοντας 3 |
|---|---|---|
| $X_1$<br>$X_3$<br>$X_4$<br>$X_6$ | $X_2$<br>$X_7$ | $X_5$<br>$X_8$<br>$X_9$ |

Πίνακας 1: Πίνακες παραγόντων

Όπως μπορούμε να παρατηρήσουμε από τα παραπάνω σχήματα ομαδοποιούνται οι μεταβλητές $X_1$,$X_3, X_4, X_6$ κάτι που σημαίνει ότι είναι υψηλά συσχετισμένες μεταξύ τους,δίνοντάς μας έτσι τον παράγοντα 1.Ομοίως οι μεταβλητές $X_2$και $X_7$ορίζουν τον παράγοντα 2 και οι μεταβλητές $X_5, X_8, X_9$ορίζουν τον παράγοντα 3.

### *2.2.2 : Ανάλυση σε κύριες συνιστώσες(PrincipalComponentsAnalysis)*

*Εισαγωγικά*

Μία μέθοδος η οποία είναι αρκετά γνωστή λόγω της ευκολίας και της απλής ερμηνείας των αποτελεσμάτων. *Ουσιαστικά ερμηνεύει τις συσχετίσεις μεταξύ των αρχικών μεταβλητών, μειώνει τις διαστάσεις του προβλήματος και επιτρέπει απλή ερμηνεία αλλά και δημιουργία απλών γραφημάτων. Σκοπός της μεθόδου αυτής είναι η ανεύρεση γραμμικών συνδυασμών των αρχικών δεδομένων ώστε να μην χάνεται πληροφορία καθώς και μείωση της διάστασης όταν όλες οι μεταβλητές, συνιστώσες είναι ασυσχέτιστες μεταξύ τους.*

Έστω$X_1, \dots, X_p$αρχικές μεταβλητές και $\sum_{i=1}^{p} X_i$= ολική μεταβλητή. Με τη μέθοδο αυτή επιδιώκεται η εύρεση λιγότερων μεταβλητών$Y_1, \dots, Y_k, k < p$οι οποίοι θεωρούνται γραμμικοί συνδυασμοί των $X_1, \dots X_p$και $\sum_{i=1}^{k} VarY_i \cong \sum_{i=1}^{k} X_i$ . Όπου $Y_1, \dots, Y_k$ είναι οι κύριες συνιστώσες. Οι κύριες συνιστώσες αντικαθιστούν τις αρχικές μεταβλητές και το αρχικό σύνολο δεδομένων ρμειώνεται σε k. Έτσι αντί να αναλύσουμε τα δεδομένα στον $R^p$τα αναλύουμε στον $R^k$.Άρα,μπορούμε να πούμε ότι η ανάλυση σε κύριες συνιστώσες μπορεί να ερμηνεύσει τη δομή διασποράς-συνδιασποράς χρησιμοποιώντας κάποιους γραμμικούς συνδυασμούς των αρχικών μεταβλητών με στόχο τη μείωση των δεδομένων και της ερμηνείας τους.(Καρλής 2005).

*2.2.2.1: Βασική ιδέα*

Η γραμμική άλγεβρα είναι η βασική ιδέα πάνω στην οποία αναπτύχθηκε η μέθοδος.
Έστω Α ένας τετραγωνικός πίνακαςδιαστάσεων $p \times p$.
Έχουμε ότι :Α = ΡΛΡ' όπου:
Λ:ένας $p \times p$διαγώνιος πίνακας όπου τα στοιχεία της διαγωνίου είναι οι ιδιοτιμές του πίνακα Α.

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ \vdots & \lambda_2 & \vdots \\ 0 & 0 & \lambda_p \end{bmatrix}$$

$P$:Ένας ορθογώνιος p × pπίνακας,δηλαδή ισχύει ότι: $P^{-1} = P$.

Ισχύουν οι παρακάτω ιδιότητες:

$\Lambda = P'AP$καθώς, $A = P\Lambda P' => P^{-1}A = P^{-1}\Lambda P' => P^{-1}AP = \Lambda P'P = \Lambda$.

Η παραπάνω αναπαράσταση του πίνακα $A$ονομάζεται φασματική ανάλυση.

Συμπεραίνουμε λοιπόν ότι ξεκινώντας από ένα τετραγωνικό πίνακα καταλήγουμε σε έναν διαγώνιο πίνακα $\Lambda$.Όπως βλέπουμε και από τις παραπάνω σχέσεις από έναν τετραγωνικό πίνακα μπορούμε να οδηγηθούμε σε έναν διαγώνιο πίνακα πολλαπλασιάζοντας με ένα κατάλληλο πίνακα $P$,και άρα αν ο τετραγωνικός πίνακας είναι ένας πίνακας διακύμανσης,καταλήγουμε σε ένα διαγώνιο πίνακα διακύμανσης.Δηλαδή το τυχαίο διάνυσμα που αντιστοιχεί στον πίνακα αυτόν είναι ασυσχέτιστο.

*2.2.2.2: Εύρεση κύριων συνιστωσών*

Έστω τώρα ένα σύνολο από ρμεταβλητές $X_1, \dots, X_p$ και θέλουμε την δημιουργία κύριων συνιστωσών $Y_1, \dots, Y_p$ οι οποίες να είναι γραμμικός συνδυασμός των αρχικών μεταβλητών,δηλαδή

$$\boldsymbol{Y_1 = \alpha_{11}X_1 + \alpha_{12}X_2 + \cdots + \alpha_{1p}X_p}$$
$$\boldsymbol{Y_2 = \alpha_{21}X_1 + \alpha_{22}X_2 + \cdots + \alpha_{2p}X_p}$$
$$\dots$$
$$\boldsymbol{Y_p = \alpha_{p1}X_1 + \alpha_{p2}X_2 + \cdots + \alpha_{pp}X_p}$$

Κάθε συνιστώσα είναι σταθμισμένο άθροισμα των $X_1, X_2, \dots, X_P$, όπου τα $\boldsymbol{\alpha_{ij}}$είναι οι συντελεστές (coefficients). Τα $\alpha_{ij}$ πρέπει να πληρούν κάποιες προυποθέσεις:

$$\sum_{i=1}^{p} {a_{ij}}^2 = 1, j = 1,2, \dots, p \;\&\; \sum_{i=1}^{p} a_{ij}a_{ik} = 0\,, j \neq k, j, k = 1,2, \dots, p$$

Οι παραπάνω περιορισμοί εξασφαλίζουν την ορθοκανονικότητα των συντελεστών.

Συγκεντρωτικά μπορούμε να πούμε τα εξής:

- Για την κατασκευή κύριων συνιστωσών χρειάζεται να βρούμε τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα $\Sigma$(όπου πίνακας Σ είναι ο πίνακας συν-διακυμάνσεων του τυχαίου διανύσματος $X$).
- Η μεγαλύτερη ιδιοτιμή και το ιδιοδιάνυσμα της αντιστοιχούν αντίστοιχα στην πρώτη τιμή κ.ο.κ.
- Η διακύμανση κάθε κύριας συνιστώσας είναι ίση με την αντίστοιχη ιδιοτιμή της.
- Οι κύριες συνιστώσες είναι ασυσχέτιστες μεταξύ τους, άρα και ο πίνακας διακύμανσης τους είναι ο διαγώνιος με διαγώνια στοιχεία τις ιδιοτιμές(π.χ.$\lambda_j$).
- Η συνολική διακύμανση των κύριων συνιστωσών θα είναι ίση με την συνολική διακύμανση των αρχικών μεταβλητών .
- Η γενικευμένη διακύμανση των κύριων συνιστωσών είναι ίδια με τη γενικευμένη διακύμανση των αρχικών μεταβλητών.

*2.2.2.3: Κύριες συνιστώσες σε πίνακα συν-διασποράς Σ*

Οι κύριες συνιστώσες είναι μερικοί γραμμικοί συνδυασμοί των $p$ τυχαίων μεταβλητών $X_1$,$X_2$,...,$X_p$ και εξαρτώνται από τον πίνακα συνδιασπορών $\Sigma$ των τυχαίων μεταβλητών.Η ανάπτυξη τους δεν απαιτεί πολυμεταβλητή κανονική κατανομή.
Έστω τυχαίο διάνυσμα $\mathrm{X}' = [X_1, X_2, \dots, X_p]$που έχει πίνακα συνδιασπορών με ιδιοτιμές$\lambda_1 \geq \lambda_2 \geq \geq \lambda_p \geq 0$.

Έστω γραμμικοί συνδυασμοί

$$\boldsymbol{Y_1 = I'_1 X = I_{11} X_1 + I_{21} X_2 + \cdots + I_{P1} X_P}$$
$$\boldsymbol{Y_2 = I'_2 X = I_{12} X_1 + I_{22} X_2 + \cdots + I_{P2} X_P}$$
$$\vdots$$
$$\boldsymbol{Y_p = I'_p X = I_{1p} X_1 + I_{2p} X_2 + \cdots + I_{Pp} X_P}$$

τότε,
$\mathrm{Var}(Y_i) = I'_i \Sigma I_i \mathrm{i} = 1,2, \dots, \mathrm{p}(1)$
$\mathrm{Cov}(Y_i, Y_k) = I'_i \Sigma I_k \mathrm{I}, \mathrm{k} = 1,2, \dots, \mathrm{p}(2)$
Από τις παραπάνω σχέσεις μπορούμε να παρατηρήσουμε ότι ο γραμμικός συνδυασμός με την μεγαλύτερη διασπορά είναι η πρώτη κύρια συνιστώσα, έτσι έχουμε μέγιστη διασπορά όπως φαίνεται και από την σχέση (1).
Έχουμε:
Πρώτη κύρια συνιστώσα=γραμμικός συνδυασμός $I'_1 X$που μεγιστοποιεί την $Var(I'_1 \mathrm{X})$ υπό τον περιορισμό $I'_1 I_1 = 1$.

Δεύτερη κύρια συνιστώσα=γραμμικός συνδυασμός $I'_2$Χπου μεγιστοποιεί την $Var(I''_2X)$, υπό τις συνθήκες ότι:$I'_2I_2 = 1$ και $cov(I'_1, X, I'_2X) = 0$ . Και γενικά κατά τον ίδιο τρόπο για την i-συνιστώσα.

*2.2.2.4: Παράδειγμα*

Έστω οι τυχαίες μεταβλητές $X_1$,$X_2$,$X_3$οι οποίες έχουν πίνακα συνδιασποράς $\Sigma$.

$$\Sigma = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Οι τιμές τόσο των ιδιοτιμών αλλά και τωνιδιοδιανυσμάτων του πίνακα $\Sigma$ είναι: $\lambda_1 = 5.83$ , $e'_1$ =[0.383,-0.924.0]

$\lambda_2 = 2.00$ , $e'_2 = [0,0,1]$

$\lambda_3 = 0.17$ , $e'_3 = [0.924,0.383,0]$

Επομένως, οι εξισώσεις των κύριων συνιστωσών είναι οι εξής :

$$Y_1 = e'_1X = 0.383X_1 - 0.924X_2$$
$$Y_2 = e'_2X = X_3$$
$$Y_3 = e'_3X = 0.924X_1 - 0.383X_2$$

Όπως μπορούμε να διακρίνουμε η μεταβλητή $X_3$ είναι η κύρια συσνιστώσα καθώς είναι ασυσχέτιστη με τις άλλες δύο μεταβλητές. Θέλουμε να δείξουμε ότι οι κύριες συνιστώσες έχουν διασπορές ίσες με τις ιδιοτιμές του πίνακα Σ ενώ παράλληλα είναι ασυσχέτιστες μεταξύ τους.

$$Var(Y_1) = Var(0.383X_1 - 0.924X_2)$$
$$= (0.383)\text{^}2\, Var(X_1) + (-0.924)\text{^}2\, Var(X_2) + 2 \times (0.383) \times (-0.924)Cov(X_1, X_2)$$
$$= 0.147 \times 1 + 0.854 \times 5 + 0.708 \times (-2) = 5.83 = \lambda_1$$

Επίσης,$Cov(Y_1, Y_2) = Cov(0.383X_1 - 0.924X_2, X_3)$
$$= 0.383Cov(X_1, X_3) - 0.924Cov(X_2, X_3) =$$
$$= 0.383 \times 0 - 0.924 \times 0 = 0$$

Έπειτα από***πρόταση***ξέρουμε ότι :

Αν τυχαίο διάνυσμα $X' = [X_1, X_2, \dots, X_p]$το οποίο έχει πίνακα συν-διασπορών Σ με ζεύγη ιδιοτιμών-ιδιοδιανυσμάτων $(\lambda_1, e_1), (\lambda_2, e_2), \dots, (\lambda_p, e_p)$όπου$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Έστω,επίσης οι κύριες συνιστώσες είναι : $Y_1 = e'_1X, Y_2 = e'_2X, \dots, Y_p = e'_pX$.Τότε ισχύει:

$$Var(Y) = \sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} Var(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$$

*(για απόδειξη βλέπεεργασία Παπαγεωργίου Ανδρέα, σελίδα15-16 – Παραγοντική Ανάλυση και Ανάλυση σε κύριες συνιστώσες (2009-2010), Πανεπιστήμιο Πατρών,)*

**Ένας τρόπος ερμηνείας των κύριων συνιστωσών είναι μέσω των συντελεστών των γραμμικών συνδυασμών**,έτσι μπορούμε να σχολιάσουμε ότι η πρώτη κύρια συνιστώσα $Y_1$ όπως φαίνεται και από τις παραπάνω σχέσεις λόγω των διαφορετικών προσήμων των συντελεστών,εκφράζει την αντίθεση μεταξύ των μεταβλητών $X_1, X_2$.Επίσης φαίνεται ότι η $X_2$επηρεάζει περισσότερο την $Y_1$ από ότι η $X_1$λόγω του μεγαλύτερου κατά απόλυτη τιμή συντελεστή.

**Ένας άλλος τρόπος ερμηνείας των κύριων συνιστωσών είναι μέσω των συσχετίσεων.**Για να δώσουμε την ερμηνεία μέσω συσχετίσεων θα χρησιμοποιήσουμε την παρακάτωπρόταση

**Πρόταση**

Αν $Y_1 = e'_1X, Y_2 = e'_2X, \dots, Y_p = e'_pX$, κύριες συνιστώσες που προέρχονται από πίνακα συνδιασπορώντότε:

$$\rho_{Y_iX_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \text{i, k} = 1,2, \dots, p\text{Q} = (e_{ij}) \text{ καισ}_{ij} = Var(X_i)$$

*(για απόδειξη βλέπε εργασία Παπαγεωργίου Ανδρέα, σελίδα 17-18 – Παραγοντική Ανάλυση και Ανάλυση σε κύριες συνιστώσες (2009-2010), Πανεπιστήμιο Πατρών)*

### *2.2.3: Ανάλυση Κανονικών Συσχετίσεων (CanonicalCorrelationAnalysis)*

Μέθοδος η οποία είναι παρόμοια με την ανάλυση σε κύριες συνιστώσες με τη διαφορά ότι οι συνιστώσες που προκύπτουν έχουν μεταξύ τους κάποια ελεγχόμενη συσχέτιση.Ουσιαστικά είναι μία άλλη τεχνική μείωσης της διάστασης,μας επιτρέπει να συνοψίσουμε τις σχέσεις σε μικρότερο αριθμό στατιστικών στοιχείων διατηρώντας παράλληλα τις βασικές πτυχές των σχέσεων. Χρησιμοποιήθηκε για την διερεύνηση των αλληλοσυσχετίσεων μεταξύ ομάδων μεταβλητών.

Έστω ένα παράδειγμα μεταβλητών που σχετίζονται με την άσκηση και την υγεία. Από την μία πλευρά έχουμε μεταβλητές που σχετίζονται με τη άσκηση, παρατηρήσεις όπως το ποσοστό σε ένα σκαλί, πόσο γρήγορα μπορείτε να τρέξετε, τον αριθμό των push-upsανά λεπτό κ.λ.π. Και από την άλλη έχουμε μεταβλητές υγείας όπως η αρτηριακή πίεση,τα επίπεδα χοληστερόλης,τα επίπεδα γλυκόζης, ο δείκτης μάζας του σώματος κ.λ.π. ΄Αρα γίνεται μελέτη μεταξύ των μεταβλητών της άσκησης και των μεταβλητών της υγείας.

Ως ένα δεύτερο παράδειγμα θα μπορούσε να θεωρηθεί: ότι μία ομάδα από εκπροσώπους πωλήσεων,στους οποίους έχουμε καταγράψει αρκετές μεταβλητές την απόδοση των πωλήσεων σε συνδυασμό μεδιάφορα μέτρα της πνευματικής και δημιουργικής ικανότητας. Έτσι, μπορεί να γίνει διερεύνηση των σχέσεων μεταξύ των μεταβλητών απόδοσης των πωλήσεων και των μεταβλητών επάρκειας.

*2.2.3.1: Διαδικασία ανάλυσης κανονικών συσχετίσεων*

Εάνοι τιμές δύο ομάδων μεταβλητών που έχουν ορισμένο θεωρητικό περιεχόμενο είναι διαθέσιμες για ένα δείγμα $N$ ατόμων ή αντικειμένων και υπολογιστούν οι τυποποημένες τιμές τους και παραστεί με $Z_1$:διάνυσμα με συντεταγμένες τις μεταβλητές της ομάδας $A$ και $Z_2$:διάνυσμα με συντεταγμένες τις μεταβλητές της ομάδας $B$.

Τότε, η ανάλυση κανονικών συσχετίσεων αναζητά καταρχήν ένα ζεύγος νέων τυποποιημένων μεταβλητών $x_1, y_1$όπου κάθε παράγοντας είναι ένας γραμμικός συνδυασμός των τυποποιημένων μεταβλητών των ομάδων$A, B$, έτσι ώστε οι $x_1, y_1$ να εμφανίζουν την μέγιστη δυνατή συσχέτιση. Ουσιαστικά, αναζητά δύο διανύσματα $c_1, d_1$τέτοια ώστε για τις τυποποιημένες μεταβλητές $x_1 = c'_1 z_1$και $y_1 = d'_1 z_2$, η ποσότητα $\frac{1}{N}\sum_{i=1}^{N} x_{1i} y_{1i}$ να γίνεται μέγιστη ,

όπου: $x_i = c' z_{1i}$και $y_i = d' z_{2i}$, i $= 1,2, \dots,$ Nκαι $z_{1i}, z_{2i}$:τιμές των $z_1, z_2$ για το $i-$άτομο ή αντικείμενο.

Έπειτα ένα δεύτερο ζεύγος αντικειμένων $x_2, y_2$ αναζητείται από τις ομάδες Α,Βέτσι ώστε κάθε παράγοντας να είναι ορθογώνιος(ασυσχέτιστος) προς τους $x_1, y_1$ και οι $x_2, y_2$ να εμφανίζουν μέγιστη δυνατή συσχέτιση. Αναζητούνται δηλαδή και πάλι διανύσματα $c_2, d_2$.Έτσι ώστε για τυποποιημένες μεταβλητές–παράγοντες $x_2 = c'_2 z_1$ και $y_2 = d'_2 z_2$, η ποσότητα $\frac{1}{N}\sum_{i=1}^{N} x_{2i} y_{2i}$ μέγιστη με την προυπόθεση ότι είναι ασυσχέτιστοι προς τους προηγούμενους εξαχθέντες παράγοντες. Η διαδικασία διακόπτεται όταν οι εξαχθέντες παράγοντες δεν είναι σημαντικοί. Οι εξαγόμενοι παράγοντες $x_i, y_i$ονομάζονται κανονικοί παράγοντες και οι αντίστοιχες συσχετίσεις τους συντελεστές κανονικής συσχέτισης.

Αν ισχύει ότι: $\frac{\sum_{j=1}^{\nu_1}[\alpha_{ij}^{(A)}]^2}{\nu_1}$ είναι το ποσοστό διακύμανσης που εξάγεται από την ομάδα $A$ και τον παράγοντα $x_1$και $\frac{\sum_{j=1}^{\nu_2}[\alpha_{ij}^{(B)}]^2}{\nu_2}$ είναι το ποσοστό διακύμανσης που εξάγεται από την ομάδα Β και τον παράγοντα $y_1$.

Όπου, $\nu_1, \nu_2$:πλήθος μεταβλητών

$[\alpha_{ij}^{(A)}]$:συντελεστής συσχέτισης του $x_i$με jσυντεταγμένη του διανύσματος $z_1$. $[\alpha_{ij}^{(B)}]$:συντελεστής συσχέτισης του $y_i$με jσυντεταγμένη του διανύσματος $z_2$.

Αν τώρα με $R_i$συμβολίζεται ο συντελεστής κανονικής συσχέτισης των κανονικών παραγόντων $x_i, y_i$ τότε η ποσότητα: $\boldsymbol{RD_{x_i}} = \frac{\sum_{j=1}^{\nu_1}[\alpha_{ij}^{(A)}]^2}{\nu_1}\boldsymbol{R_i^2}$η οποία ονομάζεται μέτρο επικάλυψης της ομάδας $A$ που εκφράζεται από την $x_i$. Το άθροισμα $\boldsymbol{RD_{x_i}}$για όλους τους παράγοντες που εξήχθησαν από την ομάδα $A$ εκφράζει το συνολικό μέτρο επικάλυψης, δηλαδή το πόσο συνολικά περισσεύει η ομάδα $A$όταν δίνεται η ομάδα $B$. Οι συντελεστές κανονικής συσχέτισης υπολογίζονται μέσω του προσδιορισμού των ιδιοτιμών της**:**

$$\boldsymbol{H = R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}}$$

όπου:

$\boldsymbol{R_{11}}$:η εκτίμηση των συντελεστών συσχέτισης των μεταβλητών(πλήθους $\nu_1$) της ομάδας Α

$\boldsymbol{R_{22}}$:η εκτίμηση των συντελεστών συσχέτισης των μεταβλητών(πλήθους $\nu_2$) της ομάδας Β

$\boldsymbol{R_{12}}$:η εκτίμηση των συντελεστών συσχέτισης των μεταβλητώντης ομάδας Α με τις μεταβλητές της ομάδας Β.

Η υπόθεση τώρα ότι η πρώτη ομάδα μεταβλητών δεν συσχετίζεται προς την δεύτερη,ελέγχετα με τη στατιστική συνάρτηση:$\chi^2 = -\left[(N-1) - \frac{1}{2}(\nu_1 + \nu_2 + 1)\right] log\Lambda$,

η οποία ακολουθεί κατά προσέγγιση την : $\chi^2$- κατανομή με $\nu_1\nu_2$ βαθμούς ελευθερίας και $\Lambda$ η στατιστική συνάρτηση $\Lambda$ του Wils, η οποία ορίζεται από τον τύπο $\Lambda = \prod_{i=1}^{\nu}(1-\lambda_i)$, όπου $\lambda_1, \lambda_1, \dots, \lambda_\nu$ είναι οι τιμές των ιδιοτιμών της **H**.

### *2.2.4: ΠολυδιάστατηΚλιμακοποίηση (MultidimensionalScaling, (MDS)*

Μαθηματική μέθοδος που σκοπό έχει να προβάλλει τις διαστάσεις του προβλήματος στο χώρο δύο ή περισσότερων διαστάσεων. Έτσι βελτιώνεται η ικανότητα ερμήνευσης των αποτελεσμάτων καθώς είναι πολύ πιο εύκολο να μελετηθεί ένα διάγραμμα λίγων διαστάσεων σε σχέση με δεδομένα πολλών διαστάσεων χωρίς ουσιαστικά κανένα εργαλείο απεικόνισής τους. Η μέθοδος αυτή με τον τρόπο αυτό καταφέρνει να δημιουργήσει δείκτες βασισμένους σε όλα τα δεδομένα οι οποίοι είναι πιο εύκολα κατανοητοί.

Ο όρος πολυδιάστατη κλιμακοποίηση μπορεί να χρησιμοποιηθεί με δυο διαφορετικούς τρόπους στην στατιστική:

- *Με την εύρεια έννοια*, δηλαδή αναφέρεται σε οποιαδήποτε τεχνική που παράγει μία πολυδιάστατη γεωμετρική αναπαράσταση του εικονικού χώρου των αντικειμένων των δεδομένων ,όπου ποιοτικές και ποσοτικές σχέσεις ανάμεσα στα αντικείμενα αυτά σχετίζονται με τη γεωμετρική τους αναπαράσταση.
- *Με τη στενή έννοια*,η πολυδιάστατη κλιμακοποίηση ξεκινά με κάποιες πληροφορίες που αφορούν την ομοιότητα ή ανομοιότητα των στοιχείων ενός συνόλου αντικειμένων με βάση κάποια μετρήσιμα από τις τυχαίες μεταβλητές χαρακτηριστικά τους.Έπειτα η μέθοδος προχωρά στην κατασκευή της γεωμετρικής αναπαράστασης ενός εικονικού χώρου με διακεκριμένα του σημεία τα αντικείμενα αυτά,αξιοποιώντας τις πληροφορίες.Τα δεδομένα που λειτουργούν ως εισροή στους αλγορίθμους της τεχνικής και κατ΄επέκταση στα λογισμικά που υποστηρίζουν την τεχνική αυτή είναι διάφορα μέτρα ομοιότητας ή ανομοιότητας των αντικειμένων από έρευνα.Έτσι τα δεδομένα μπορούν να θεωρηθούν ότι είναι ποσότητες οποιασδήποτε μορφής ομοιότητας ή ανομοιότητας .

Το πιο σημαντικό αποτέλεσμα της ανάλυσης που προσφέρει μία τέτοια μέθοδος είναι η γραφική-γεωμετρική αναπαράσταση του χώρου των αντικειμένων τα οποία αντιστοιχίζονται από μεμονωμένα και διακεκριμένα σημεία του γραφήματος αναπαράστασης του εικονικού τους χώρου.

Η μεθοδολογία όπως προαναφέρθηκε χρησιμοποιεί ως αρχικά δεδομένα τις εγγύτητες(ομοιότητες ή ανομοιότητες) μεταξύ όλων ανά δύο των αντικειμένων στα δεδομένα. Η εγγύτητα μεταξύ δύο αντικειμένων είναι μια αριθμητική ποσότητα που υποδεικνύει κατά το πόσο τα συγκεκριμένα δύο αντικείμενα είναι ή αντιλαμβάνονται όμοια ή συναφή. Με την χρήση λοιπόν όλων αυτών των ανά δύο εγγυτήτωντων αντικειμένων είναι δυνατό να κατασκευαστεί με επαναληπτικές αλγοριθμικές μεθόδους το ζητούμενο γράφημα αναπαράστασης. Η κατασκευή του γραφήματος θα πρέπει να υποστηρίζει μία ικανοποιητική αντιστοιχία μεταξύ των εγγυτήτων και των αντίστοιχων αποστάσεων στο γράφημα αυτό. Υποδεικνύεται έτσι ότι η μικρή απόσταση ανάμεσα σε δύο σημεία-αντικείμενα προυποθέτει μεγάλη ομοιότητα και αντίστροφα, δηλαδή όσο περισσότερα όμοια κρίνονται δύο αντικείμενα, τόσο πιο κοντά θα πρέπει να βρίσκονται τα αντίστοιχα σημεία τους γραφήματος που τα αντιπροσωπεύουν.

Γενικά, οι τεχνικές πολυδιάστατης κλιμακοποίησης μπορούν κάποιες φορές να εφαρμοστούν και σε δεδομένα που δεν αφορούν εγγύτητες μεταξύ αντικειμένων. Σε αυτές τις περιπτώσεις οι εγγύτητες λαμβάνονται έμμεσα από τη τεχνική σε κάποιο ενδιάμεσο πλαίσιο.

Είναι μία μέθοδος η οποία γίνεται ολοένα και δημοφιλέστερη και εφαρμόζεται σε αρκετούς επιστημονικούς κλάδους, όπως:

- Ψυχολογία, οι ψυχολόγοι την χρησιμοποιούν για να μελετήσουν αφενός τον τρόπο με τον οποίο οι άνθρωποι αντιλαμβάνονται και αξιολογούν διαφόρων ειδών ερεθίσματα καθώς και χαρακτηριστικά της προσωπικότητάς τους, αφετέρου διάφορες κοινωνικές καταστάσεις.
- Κοινωνιολογία, χρησιμοποιούν τέτοιες μεθόδους για να καθορίσουν την δομή των ομάδων και των οργανισμών, βασισμένοι στην αντίληψη των διαφορετικών πηγών και των μεταξύ τους αλληλεπιδράσεων.
- Ανθρωπολογία, χρησιμοποιούν την μέθοδο για να συγκρίνουν διαφορετικές πολιτιστικές ομάδες, βασισμένοι σε διάφορα χαρακτηριστικά τους όπως τη γλώσσα, τα ήθη και έθιμά τους.
- Οικονομολόγοι και ερευνητές μάρκετινγκ, οι οποίοι χρησιμοποιούν τέτοιες μεθόδους για να αξιολογήσουν αντιδράσεις καταναλωτών σε μία ευρεία ποικιλία ειδών προιόντων.
- Βιοχημικοί, οι οποίοι εφαρμόζουν τέτοιες μεθόδους για την τεχνική της δομής των πρωτεινών.

Η πολυδιάστατη κλιμακοποίηση είναι μία μέθοδος η οποία μας δίνει την λύση σε προβλήματα φύσεως όπως, το να μας ζητηθεί η κατασκευή ενός πίνακα αποστάσεων μεταξύ των σημείων των πόλεων σε έναν γεωγραφικό πίνακα ή και τοαντίστροφο.Η πολυδιάστατη κλιμακοποίηση είναι ένα μέσο για την απεικόνιση του επιπέδου της ομοιότητας των μεμονωμένων περιπτώσεων ενός συνόλου δεδομένων.Ένας αλγόριθμος MDSστοχεύει να τοποθετήσει κάθε αντικείμενο σε N-διάστατο χώρο, έτσι ώστε οι αποστάσεις μεταξύ αντικειμένων να είναι διανεμημένες όσο το δυνατόν καλύτερα.
Η εφαρμογή της είναι αρκετά περίπλοκη αφού συχνά τα δεδομένα περιέχουν σφάλματα, καθώς δεν είναι απαραίτητα απόλυτες οι αποστάσεις με μετρικές ιδιότητες αλλά είναι ποσότητες που λαμβάνονται με υποκειμενικό τρόπο και υποδεικνύουν διαφόρων ειδών ομοιότητας ή ανομοιότητας σε κάθε είδους αντικείμενα. Αποτελεί επίσης και μία τεχνική ελαχιστοποίησης του όγκου των δεδομένων, καθώς αναζητεί την βέλτιστη διαμόρφωση του εικονικού χώρου του αντικειμένου αυτών με μικρότερη δυνατή διάσταση.

*2.2.4.1: Κλασικό Πολυδιάστατο μοντέλο*

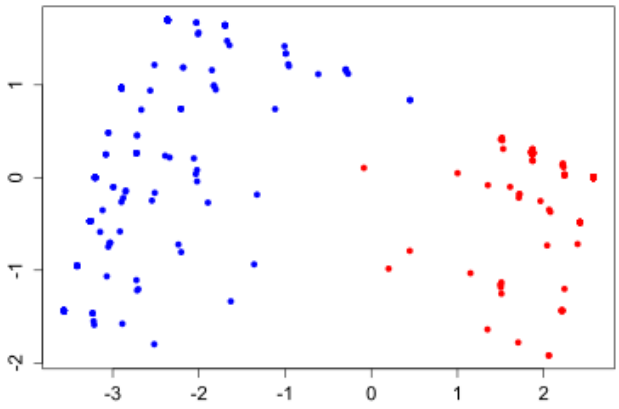
Το πολυδιάστατο μοντέλο είναι επίσης γνωστό ως ΄΄Ανάλυση Κύριων Συντεταγμένων΄΄. Απαιτεί έναν πίνακα είσόδου, δίνοντας τις διαφορές μεταξύ των ζευγών των στοιχείων και εξάγει έναν πίνακα συντεταγμένων του οποίου η διάταξη ελαχιστοποιεί την συνάρτηση απώλειας, αυτό ονομάζεται ΄΄*ένταση*.

Το κλασικό πολυδιάστατο μοντέλο υποθέτει ότι τα δεδομένα, ας πούμε ο πίνακας εγγύτητας μπορεί να είναι για παράδειγμα όπως μετριώνται οι αποστάσεις από έναν χάρτη. Το μοντέλο αυτό παρουσιάστηκε πρώτη φορά από τον *Torgerson(1952).*Έτσι, οι αποστάσεις σε ένα κλασικό πολυδιάστατο χώρο διατηρούν τα χρονικά διαστήματα και τις αναλογίες μεταξύ των εγγυτήτων όσο το δυνατόν καλύτερες.

Έστω ότι έχουμε το εξής πρόβλημα: κοιτάζουμε έναν χάρτη που δείχνει μία σειρά από πόλεις και ενδιαφερόμαστε για τις αποστάσεις μεταξύ τους. Αυτές οι αποστάσεις μπορούν να μετρηθούν εύκολα χρησιμοποιώντας έναν χάρακα, από μαθηματικής πλευράς όμως, γνωρίζοντας τις συντεταγμένες $x, y$η απόσταση μεταξύ δύο πόλεων $A$ και $B$θα δίνεται από:

$$\boldsymbol{d_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}} \textbf{(2.2.4.1)}$$

Σε πολλές εφαρμογές του κλασικού πολυδιάστατου μοντέλου, τα δεδομένα δεν είναι αποστάσεις όπως μετριώνται από έναν χάρτη αλλά δεδομένα εγγύτητας. Κατά την εφαρμογή του μοντέλου αυτού οι εγγύτητες υποτίθεται ότι είναι οι εγγύτητες οι οποίες παρουσιάζονται σαν τις μετρήσεις πραγματικών αποστάσεων. Ένα σημαντικό πλεονέκτημα της κλασικής πολυδιάστατης μοντελοποίησης είναι ότι παρέχει μία μοναδική λύση που δεν απαιτεί επαναληπτικές διαδικασίες.




*Εικόνα 2:* Ένα παράδειγμα κλασικής πολυδιάστατης κλιμακοποίησης, όπου εφαρμόστηκε σε μοντέλα ψήφου στη Βουλή των αντιπροσώπων στις Ηνωμένες Πολιτείες. Κάθε κόκκινη κουκίδα αναπαριστά ένα Δημοκρατικό μέλος της Βουλής και κάθε μπλε κουκίδα έναν Δημοκράτη.

*Βήματα αλγορίθμου κλασικής πολυδιάστατης μοντελοποίησης:*

Ο κλασικός αλγόριθμος στηρίζεται στο γεγονός ότι, οι συντεταγμένες του πίνακα $X$μπορεί να προέρχονται από ιδιοτιμές αποσύνθεσης από τον βαθμωτό πίνακα $B = XX'$. Το πρόβλημα κατασκευής του πίνακα $B$από τον πίνακα εγγύτητας $P$λύνεται με πολλαπλασιασμό του τετραγώνου εγγυτήτων με τον πίνακα $J = I - n^{-1}11'$.Η διαδικασία αυτή ονομάζεται *διπλό κεντράρισμα.* Τα παρακάτω βήματα συνοψίζουν τον αλγόριθμο της κλασικής πολυδιάστατης μοντελοποίησης:

1. Ρύθμιση του πίνακα των τετραγώνων των εγγυτήτων: $P^{(2)} = [p^2]$.
2. Εφαρμόζεται ο πίνακας διπλών εγγυτήτων: $B = -\frac{1}{2}JP^{(2)}J$ , όπου ο πίνακας $J$ ορίζεται ως: $J = I - n^{-1}11'$, όπου $n$: είναι ο αριθμός των αντικειμένων.
3. Εξάγεται τις $m$μεγαλύτερες θετικές ιδιοτιμές $\lambda_1 \dots \lambda_m$των $B$ και των αντίστοιχων $m$ ιδιοδιανυσμάτων $e_1 \dots e_m$.

*2.2.4.2: Μέθοδος Συλλογής δεδομένων εγγυτήτων*

Εγγύτητες ονομάζονται οι αριθμητικές τιμές που υποδεικνύουν την ομοιότητα ή ανομοιότητα των αντικειμένων,τα οποία είναι και τα αρχικά δεδομένα για μία πολυδιάστατη κλιμακοποίηση. Μία ανάλυση της μεθόδου αυτής αναζητά την απεικόνιση των αντικειμένων αυτών στον χώρο με τρόπο ώστε οι αποστάσεις μεταξύ των αντικειμένων στο γράφημα να αντιστοιχίζονται όσο το δυνατόν καλύτερα στις εγγύτητες. Συχνά τα δεδομένα εγγύτητας είναι ταξινομημένα σε ένα τετραγωνικό πίνακα, το *λεγόμενο πίνακα εγγυτήτων.*

Οι δύο κύριες ομάδες τεχνικών για να συλλέξουμε έναν τέτοιο πίνακα δεδομένων, περιλαμβάνουν:

- ***Άμεσες μεθόδους συλλογής***
  Σε μια τέτοια προσέγγιση οι διαφορετικές πηγές μιας έρευνας, για παράδειγμα οι διαφορετικοί ερωτώμενοι μπορούν είτε: **να καθορίσουν οι ίδιοι μια αριθμητική τιμή ομοιότητας ή ανομοιότητας για κάθε ζεύγος αντικειμένων**, για αυτή την προσέγγιση για να παρθεί το σύνολο των δεδομένων πρέπει να το πάρουμε από κάποιες πηγές όπως ερωτώμενους-συμμετέχοντες στην έρευνα. Το μεγαλύτερο πλεονέκτημα αυτής της προσέγγισης συλλογής δεδομένων είναι ότι αυτά είναι έτοιμα για ανάλυση χωρίς περαιτέρω επεξεργασία. Έτσι είναι δυνατή και η ανάλυση κατά διαφορετική πηγή και η ενοποιημένη αθροιστική ανάλυση για όλες τις πηγές της έρευνας ταυτόχρονα, και ένα μειονέκτημά της είναι ο ραγδαία αυξανόμενος αριθμός από εγγύτητες μεταξύ των ζευγών αντικειμένων. **Είτε να καθορίσουν μία διάταξη των ζευγών αυτώνσε σειρά ομοιότητας ή ανομοιότητά τους.** Ένας τρόπος για να γίνει αυτό είναι να γράψουμε κάθε τέτοιο ζεύγος αντικειμένων σε κάρτες και αν ζητήσουμε από τις πηγές-συμμετέχοντες στην έρευνα να τις διατάξουν από αυτή των δύο αντικειμένων μικρότερης ομοιότητας έως αυτή των δύο αντικειμένων μεγαλύτερης ομοιότητας. Στη συνέχεια ορίζουμε μία αριθμητική πηγή εγγύτητας σε κάθε ομάδα, ανάλογα με την ομοιότητα των αντικειμένων κάθε ομάδας. Τέλος, μία διάταξη μεταξύ των ομοιοτήτων ή ανομοιοτήτων μπορεί να πραγματοποιηθεί με το να αναγράψουμε μόνο ένα αντικείμενο κάθε φορά σε μία κάρτα και να ζητήσουμε από τους συμμετέχοντες να ταξινομήσουν τα πιο όμοια συναφή κατά την γνώμη τους αντικείμενα. Πλεονεκτεί στο ότι κάνει χρήση της διαίσθησης των συμμετεχόντων, μειονεκτεί όμως στο ότι δεν μπορεί να γίνει ξεχωριστή ανάλυση των δεδομένων για κάθε συμμετέχοντα.
- ***Έμμεσες μεθόδους συλλογής***
  Οι μέθοδοι αυτές συλλογής εγγυτήτων μεταξύ των αντικειμένων δεν βασίζονται στις αριθμητικές τιμές ομοιότητας που θα απέδιδαν οι διαφορετικές πηγές. Η βάση τέτοιων μεθόδων είναι ότι οι εγγύτητες βασίζονται και συλλέγονται από διαφορετικές μετρήσεις, όπως**: δεδομένα σύγχυσης**όπου προκύπτουν όταν ο ερευνητής καταγράφει την συχνότητα με την οποία οι συμμετέχοντες αδυνατούν να διαχωρίσουν ένα αντικείμενο από

κάποιο άλλο. Το πλεονέκτημα είναι ότι η ομοιότητα των αντικειμένων κρίνεται κυρίως από την διαίσθηση των συμμετεχόντων χωρίς να απαιτείται από αυτούς κάποια ιδιαίτερη γνώση. **Ή πίνακες συσχετίσεων ή αποστάσεων ανάμεσα στα δεδομένα αυτά.**Όταν τα δεδομένα είναι μετρήσεις σε διαφορετικές κλίμακες και οι μετρήσεις αυτές είναι συσχετισμένες, τότε διαμορφώνεται ένας πίνακας συσχετίσεων. Οι συσχετίσεις μπορούν να λειτουργήσουν ως εγγύτητες, καθώς όσο περισσότερο συσχετισμένες είναι δύο μεταβλητές, τόσο πιο όμοιες μπορούν να θεωρηθούν και αντίστροφα. Ένα μειονέκτημα μιας τέτοιας εφαρμογής είναι ότι οι εγγύτητες πρέπει να κατασκευαστούν από επιπρόσθετες μετρήσεις. Και ένα πλεονέκτημα θεωρείται το γεγονός ότι οι εν λόγω μέθοδοι μπορούν να εφαρμοστούν σε δεδομένα συσχετίσεων ή αποστάσεων ακόμα και όταν η κλίμακα, οι διαστάσεις ή οι ιδιότητες των αντικειμένων υπό μελέτη είναι εκ των προτέρων γνωστά.

*2.2.4.3: Είδη πολυδιάστατης κλιμακοποίησης*

Ανάλογα με τον τύπο των αρχικών δεδομένων, διαχωρίζονται δύο είδη πολυδιάστατης κλιμακοποίησης. Σε κάποιες εφραμογές, οι εγγύτητες είναι γνωστές αριθμητικές τιμές και αναφέρονται σε πραγματικές αποστάσεις με κάποιο σφάλμα μέτρησης, ενώ σε κάποιες άλλες είναι τακτικά δεδομένα όλων των εγγυτήτων. Έτσι, τα είδη πολυδιάστατης κλιμακοποίησης είναι:

- η μετρική πολυδιάστατη κλιμακοποίηση
- η μη-μετρική πολυδιάστατη κλιμακοποίηση

***Η μετρική πολυδιάστατη κλιμακοποίηση*** προυποθέτει ότι τα αρχικά δεδομένα εφαρμόζονται με αριθμητικές τιμές και επιδεικνύουν εγγύτητες που είναι κλασματικά συνεχή δεδομένα και συνδέονται με τις αντίστοιχες αποστάσεις μέσω μιας γραμμικής σχέσης. Λειτουργεί ικανοποιητικά για πίνακα του οποίου τα στοιχεία συλλέγονται από μεθόδους άμεσου καθορισμού ομοιότητας ή ανομοιότητας.

Ουσιαστικά πρόκειται για ένα υπερσύνολο του κλασικού πολυδιάστατου μοντέλου που γενικεύει τη διαδικασία βελτιστοποίησης για μία ποικιλία συναρτήσεων απώλειας και πινάκων εισόδου των γνωστών αποστάσεων με βάρη κ.ο.κ. Μία χρήσιμη συνάρτηση απώλειας ονομάζεται ΄΄*στρες*΄΄. Μία μετρική πολυδιάστατη κλιμακοποίηση ελάχιστοποιεί την συνάρτηση κόστους που ονομάζεται ΄΄Stress΄΄ το οποίο είναι ένα υπολειπόμενο άθροισμα των τετραγώνων:

$$\boldsymbol{Stress_D(x_1, x_2, \dots, x_N) = \left(\sum_{i \neq j = 1,\cdots,N} (d_{ij} - \|x_i - x_j\|^2\right)^{1/2}} \quad \textbf{(2.2.4.3*)}$$

$$\boldsymbol{Stress_D(x_1, x_2, \dots, x_N) = \left(\frac{\sum_{i \neq j = 1,\cdots,N} (d_{ij} - \|x_i - x_j\|^2}{\sum_{i,j} d_{ij}^2}\right)^{1/2}} \quad \textbf{(2.2.4.3**)}$$

Η μετρική κλιμακοποίηση χρησιμοποιεί έναν μετασχηματισμό έντασης το οποίο ελέγχεται από την χρήση εκθέτη $p$:$d_{ij}^{p}$και $d_{ij}^{2p}$ για απόσταση.

***Η μη-μετρική πολυδιάστατη κλιμακοποίηση*** βασίζεται κυρίως στην αριθμητική διάταξη των εγγυτήτων. Τα δεδομένα συνήθως αποτελούνται από εγγύτητες οι οποίες είναι τακτικά αριθμητικά δεδομένα και συνδέονται με τις αντίστοιχες αποστάσεις μέσω μιας μονότονης συνάρτησης. Λειτουργεί αποτελεσματικά με δεδομένα που προκύπτουν από έμμεσες μεθόδους συλλογής.

Σε αντίθεση με την μετρική πολυδιάστατη κλιμακοποίηση, η μη μετρική πολυδιάστατη κλιμακοποίηση βρίσκει τόσο μια μη-παραμετρική μονότονη σχέση μεταξύ των διαφορών στο πίνακα στοιχείο-στοιχείο και την απόσταση μεταξύτων στοιχείων καθώς και την σχέση του κάθε στοιχείου στο χώρο. Έχουμε:

$$\boldsymbol{Stress} = \sqrt{\frac{\sum(\boldsymbol{f}(\boldsymbol{x})-\boldsymbol{d})^{\boldsymbol{2}}}{\sum \boldsymbol{d^2}}} \textbf{ (2.2.4.3***)}$$

Το $x$: δηλώνει τον φορέα εγγύτητας, $f(x)$: είναι ο μονότονος μετασχηματισμός αυτής, και $d$: είναι οι αποστάσεις των σημείων οι οποίες ελαχιστοποιούν το λεγόμενο *΄΄στρες΄΄*.

Βασικά βήματα ενόςαλγορίθμου μη μετρικής πολυδιάστατης κλιμακοποίησης

Ο βασικός αλγόριθμος μιας μη μετρικής πολυδιάστατης κλιμακοποίησης είναι μία διπλή διαδικασία βελτιστοποίησης. Πρώτον, πρέπει να βρεθεί ο μονότονος μετασχηματισμός των εγγυτήτων. Δεύτερον, τα σημεία της διαμόρφωσης πρέπει να κανονίζουν το βέλτιστο τρόπο έτσι ώστε οι αποστάσεις να ταιριάζουν με τις εγγύτητες όσο το δυνατόν περισσότερο. Τα βασικά στάδια ενός μη μετρικού πολυδιάστατου αλγορίθμου είναι:

1. Βρίσκεται μια τυχαία διαμόρφωση σημείων π.χ. με δειγματοληψία από μια κανονική κατανομή.
2. Υπολογισμός των $d$αποστάσεων των σημείων.
3. Βρίσκεται ο βέλτιστος μονότονος μετασχηματισμός των εγγυτήτων, προκειμένου να λάβει την βέλτιστη κλίμακα των δεδομένων $f(x)$.
4. Ελαχιστοποίηση του ΄΄στρες΄΄ μεταξύ της βέλτιστης κλίμακας των δεδομένων και των αποστάσεων με την εύρεση μιας νέας διαμόρφωσης των σημείων.
5. Συγκρίνουμε το ΄΄στρες΄΄ σε κάποιο κριτήριο. Αν το στρες είναι αρκετά μικρό τότε βγείτε από τον αλγόριθμο αλλιώς επιστρέφουμε στο 2.

### *2.2.5: Πολυμεταβλητό Γραμμικό Μοντέλο*

Το απλό γραμμικό μοντέλο θεωρείται ένα από τα πιο σημαντικά εργαλεία για την στατιστική συμπερασματολογία. Στο μοντέλο αυτό υπάρχει μία εξαρτημένη και πολλές συνήθως ανεξάρτητες μεταβλητές, οι οποίες στην περίπτωση της ανάλυσης διακύμανσης είναι κατηγορικές. Το μοντέλο αυτό μπορεί να γενικευτεί σε πολλές διαστάσεις, επιτρέποντας να υπάρχουν πολλές εξαρτημένες μεταβλητές οι οποίες να έχουν μεταξύ τους κάποια συσχέτιση. Γενικεύοντας το γραμμικό μοντέλο λοιπόν, προκύπτει η μέθοδος της πολυμεταβλητής παλινδρόμησης (Multivariateregression) και η μέθοδος της μεταβλητής ανάλυσης διακύμανσης(MANOVA).

Με το πολυμεταβλητό γραμμικό μοντέλο παλινδρόμησης εκτιμούμε μια γραμμική σχέση ανάμεσα σε δύο ή περισσότερες μεταβλητές $X_\kappa, \kappa = 1,2, \dots$ και μία ποσοτική μεταβλητή$Y$. Οι ανεξάρτητες μεταβλητές $X_i$ συνήθως είναι υπό τον έλγχο του ερευνητή. Στόχος είναι να μελετήσουμε πως οι μεταβολές των τιμών των $X_i$ επιδρούν γραμμικά στις τιμές που παίρνει η $Y$, δηλαδή πώς <<η $Y$ εξαρτάται γραμμικά από τις $X_\kappa$>>.

Επεκτείνουμε το απλό γραμμικό μοντέλο σε πολυμεταβλητό δεχόμενοι ότι η εξαρτημένη μεταβλητή $Y$είναι μία γραμμική συνάρτηση των $K$ ανεξάρτητων μεταβλητών $X_1, X_2, \dots, X_k$του σφάλματος $\varepsilon$. Αυτό αποτελεί φυσική επέκταση του απλού γραμμικού μοντέλου. Το πολυμεταβλητό γραμμικό μοντέλο έχει την μορφή:

$$\boldsymbol{Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k + \nu (2.2.5)}$$

όπου,
$Y$:εξαρτημένη μεταβλητή

$a$:ο $Y$-σταθερός όρος της επιφάνειας παλινδρόμησης

$b_i$:i $= 1,2, \dots,$ κείναι η κλίση της ευθείας της επιφάνειας παλινδρόμησης ως προς την αντίστοιχη μεταβλητή $X_i$, $x_1, \dots, x_k$τιμές από τις ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_k$αντίστοιχα.
ν:τυχαίο σφάλμα

Προυποθέσεις:

**1**.ν $\sim$ N$(0, S^2)$ ανεξάρτητο από άλλα σφάλματα

**2.** Οι μεταβλητές $X_i$ ανεξάρτητα με τα κατάλοιπα

Οι μεταβλητές καλούνται προβλεπτικές ή ερμηνευτικές.

Στόχος είναι η συνολική προβλεπτική-ερμηνευτική ικανότητά τους.

Οι παραδοχές για το σφάλμα είναι οι εξής:

- Το μοντέλο δίνεται από την προυπόθεση (1)

- Το τυχαίο σφάλμα ν έχει μέσο 0 και σταθερή διακύμανση$S^2$. Δηλαδή, Εν $=$ 0 καιVarν $= S^2$.
- Οι τυχαίες μεταβλητές $Y_1, Y_2, \dots, Y_n$ανεξάρτητες τυχαίες μεταβλητές.

*2.2.5.1: Μοντέλο 2 ανεξάρτητων μεταβλητών*

Τότε: $\mathbf{Y_i = \alpha + \beta_1 X_{\iota 1} + \beta_2 X_{i2} + \varepsilon_i}$**,i=1,2,…,n (2.2.5.1)**

όπου,
$Y_i$:τιμή εξαρτημένης μεταβλητής στην i-παράσταση

$X_{i1}, X_{i2}$:είναι οι τιμές των ανεξάρτητων μεταβλητών $X_1, X_2$ αντίστοιχα στην i-παράσταση
$\alpha, \beta_1, \beta_2$:παράμετροι του μοντέλου

$\varepsilon_i$:είναι ανεξάρτητες τυχαίες μεταβλητές οι οποίες αντιπροσωπεύουν τις αποκλίσεις από την ευθεία παλινδρόμησης και ακολουθούν την κανονική κατανομή με μέση τιμή 0 και διακύμανση $S^2$ ( $N(0, S^2)$.

Συνάρτηση παλινδρόμησης**:**

$$\boldsymbol{E(Y/X_1, X_2) = a + \beta_1 X_1 + \beta_2 X_2}$$

όπου,
$a$:σημείο τομής του άξονα Υκαι της συνάρτησης παλινδρόμησης

$\beta_1$:μεταβλητή του Ε(Υ) όταν το $X_1$αυξάνει κατά 1 μονάδα και το $X_2$ παραμένει σταθερό
$\beta_2$:μεταβλητή του Ε(Υ)όταν το $X_2$ αυξάνει κατά 1 μονάδα και το $X_1$ παραμένει σταθερό

*2.2.5.2: Μοντέλο κ-ανεξάρτητων μεταβλητών*

Τότε: $\mathbf{Y_i = \alpha + \beta_1 X_{\iota 1} + \beta_2 X_{i2} + \cdots + \beta_\kappa X_{\kappa 2} + \varepsilon_i, i = 1, 2, \dots, n}$ **(2.2.5.2)**

όπου,
$Y_i$:τιμή εξαρτημένης μεταβλητής στην i-παράσταση

$X_{i1}, X_{i2}, \dots, X_{i\kappa}$:είναι οι τιμές των ανεξάρτητων μεταβλητών στην i-παράσταση

α,$\beta_1$,$\beta_2$:παράμετροι του μοντέλου

$\varepsilon_i$:είναι ανεξάρτητες τυχαίες μεταβλητές οι οποίες αντιπροσωπεύουν τις αποκλίσεις από την ευθεία παλινδρόμησης και ακολουθούν την κανονική κατανομή με μέση τιμή 0 και διακύμανση $S^2$ ( N(0,$S^2$).

Και αντίστοιχα η συνάρτηση παλινδρόμησης:

$$\mathbf{E}(\boldsymbol{Y}|\boldsymbol{X_1},\dots,\boldsymbol{X_k}) = \boldsymbol{\alpha} + \boldsymbol{\beta_1}\mathbf{X_{\iota 1}} + \boldsymbol{\beta_2}\mathbf{X_{i2}} + \cdots + \boldsymbol{\beta_\kappa}\mathbf{X_{\kappa 2}}$$

$\alpha = \mathrm{E(Y)}$ όταν $X_1 = \cdots = X_k = 0$

$\beta_i$:για $\mathrm{i} = 1,2,\dots,$ κκαι δείχνει την μεταβλητή της E(Y) όταν η μεταβλητή $X_i$ αυξάνει κατά 1 μονάδα ενώ οι άλλες ανεξάρτητες μεταβλητές παραμένουν σταθερές.

*2.2.5.3: ΠολυμεταβλητήΑνάλυσηΔιακύμανσης* (Multivariate Analysis of Variance, MANOVA)

Θεωρείται μία γενίκευση της μονομεταβλητής ανάλυσης διακύμανσης. Είναι μία μέθοδος ελέγχου ύπαρξης διαφορών μεταξύ των μέσων δύο ή περισσοτέρων ομάδων και γενικεύοντας σε περιπτώσεις πολλών παραγόντων αν οι παράγοντες επιδρούν στην μέση τιμή.

- <u>*MANOVA για έναν παράγοντα*</u>

Έστω ότι ένα πείραμα μας δίνει τις μετρήσεις της μεταβλητής X σε J διαφορετικές ομάδες, $X_{ij}, i = 1,2,\dots,n_j$ και $j = 1,\dots,J$όπου με j συμβολίζουμε την ομάδα και I την παρατήρηση μέσα στην ομάδα n, όπου $X_{ij}$: είναι ένα διάνυσμα με στοιχεία. Έστω ακόμη, ότο ο πληθυσμός της j ομάδας ακολουθεί κανονική κατανομή με $N_p(\mu_i, \Sigma)$ με $\mu_i$: διάνυσμα των μέσων και $\Sigma$: πίνακας διασποράς-συνδιασποράς.

$$\boldsymbol{f(x)} = |\boldsymbol{2\pi\Sigma}|^{-\frac{1}{2}}\mathbf{exp}\{-\frac{\mathbf{1}}{\mathbf{2}}(\boldsymbol{x}-\boldsymbol{\mu})'^{\boldsymbol{\Sigma^{-1}}}(\boldsymbol{x}-\boldsymbol{\mu})\}$$

Ο έλεγχος είναι:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_J$$
$$H_1: \tau\upsilon\lambda\acute{\alpha}\chi\iota\sigma\tau\upsilon\nu\ \delta\acute{\upsilon}o\ \mu\acute{\varepsilon}\sigma o\iota\ \nu\alpha\ \delta\iota\alpha\varphi\acute{\varepsilon}\rho o\upsilon\nu$$

Χρησιμοποιώντας τη μέθοδο λόγου πιθανοφανειών όπως συνηθίζεται στους ελέγχους υποθέσεων υπολογίζουμε την πιθανοφάνεια κάτω από δύο υποθέσεις:

Στην μονομεταβλητή περίπτωση $\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$ και $\bar{x} = \frac{\sum_j \sum_i x_{ij}}{n}$.

Υπό την μηδενική υπόθεση η εκτίμηση για τον μέσο κάθε ομάδας είναι η ίδια και συγκεκριμένα είναι ο μέσος όλων των παρατηρήσεων.

Υποθέτοντας ότι $\sum_j n_j = n$ ο λογάριθμος πιθανοφάνειας είναι:

$$L_0 = -\frac{nlp}{2}[\log(2\pi) + \log(\Sigma)] - \frac{1}{2}\sum_j \sum_i (x_{ij} - \mu_{.j})' \Sigma^{-1}(x_{ij} - \mu_{.j})$$

Με αντίστοιχο τρόπο υπολογίζεται και η πιθανοφάνεια $L_1$, για την εναλλακτική υπόθεση.

- <u>Πολλαπλή παλινδρόμηση</u>

Στην περίπτωση αυτή μελετάται η σχέση μια εξαρτημένης μεταβλητής $Y$ με ένα πλήθος ανεξάρτητων επεξηγηματικών μεταβλητών $X_1, X_2, \cdots, X_p$. Το πλεονέκτημα είναι ότι μας επιτρέπει να λαμβάνουμε υπόψιν μας τον πίνακα συνδιασποράς. Το υπόδειγμα έχει την μορφή:

$$\underset{nxm}{\boldsymbol{Y}} = \underset{nxp}{\boldsymbol{X}} \ \underset{pxm}{\boldsymbol{B}} + \underset{nxm}{\boldsymbol{\varepsilon}}$$

όπου,

$Y$: πίνακας n × mδιαστάσεων με m:το πλήθος των μεταβλητών και n:το πλήθος των παρατηρήσεων

$X$: πίνακας n × pδιαστάσεων όπου p:το πλήθος των μεταβλητών και n:το πλήθος των παρατηρήσεων

$B$: πίνακας συντελεστών p × mδιαστάσεων

$\varepsilon$: πίνακας $n \times m$με τα τυχαία σφάλματα

# *Κεφάλαιο 3: Γενικευμένα γραμμικά μοντέλα (Generalized Linear Models – GLM)*

## *3.1: Εισαγωγή*

Τα γενικευμένα γραμμικά μοντέλα αποτελούν ένα καινούργιο σχετικά τομέα της στατιστικής όπου η θεματολογία τους στο μεγαλύτερο μέρος της ομαδοποιεί έννοιες και τεχνικές που προυπάρχουν δημιουργώντας ένα ενοποιημένο θεωρητικό και εννοιολογικό πλαίσιο.

Αναπτύχθηκαν από τους John Nelder και Robert Wed derburn το 1972, όπως σημαντική θεωρείται και η συμβολή του Peter Mc Cullagh ο οποίος είναι ο συγγραφέας μαζί με τον John Nelder του βιβλίου Generalized Linear Models του 1983. Μπορούμε επίσης να πούμε ότι αποτελούν τόσο μία σύνδεση αλλά και επέκταση γνωστών μοντέλων παλινδρόμησης τα οποία εμφανίζουν κοινές ιδιότητες και έχουν κοινή μέθοδο εκτίμησης παραμέτρων, ωστόσο τα κοινά χαρακτηριστικά των εννοιών που μελετώνται μας οδηγούν στην ομαδοποίηση των τεχνικών και δημιουργούν ένα σύνολο, αυτών των γενικευμένων γραμμικών μοντέλων. Αυτή η ομαδοποίηση λοιπόν, δημιούργησε προυποθέσεις για περαιτέρω μελέτη νέων τεχνικών για την αντιμετώπιση διαφόρων θεμάτων και σε συνδυασμό με την χρήση υπολογιστών μπορούμε να μελετήσουμε προβλήματα τα οποία δεν μπορούσαμε πριν την χρήση των γενικευμένων γραμμικών μοντέλων.

## *3.2:Γραμμικό μοντέλο*

Είναι της μορφής:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,, i = 1, 2, \dots, \nu (3.2)}$$

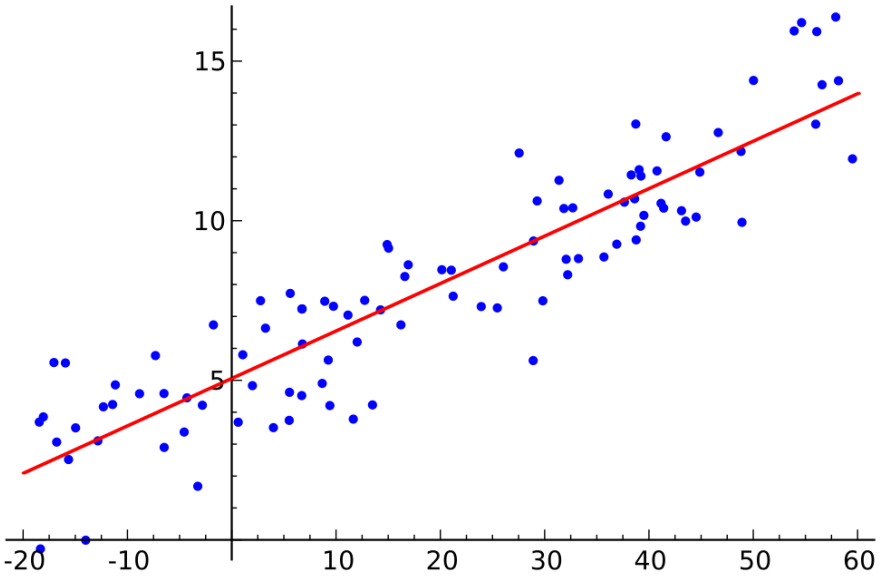
για το οποίο πρέπει να ισχύουν οι εξής υποθέσεις:

1. Οι ποσότητες $\boldsymbol{\beta_0}$και $\boldsymbol{\beta_1}$είναι άγνωστες παράμετροι.
2. Το $\boldsymbol{x_i}$είναι γνωστός αριθμός, πιο συγκεκριμέναείναι η τιμή της ανεξάρτητης μεταβλητής κατά την $i$-επανάληψη του πειράματος.
3. Το $\boldsymbol{Y_i}$είναι η τιμή της εξαρτημένης μεταβλητής κατά την $i$ -επανάληψη του πειράματος.
4. Τα $\boldsymbol{\varepsilon_i}$ είναι τυχαία σφάλματα με μέση τιμή 0 και διακύμανση $\sigma^2$, δηλαδή $E(\varepsilon_i) = 0$ και $Var(\varepsilon_i) = \sigma^2$.
5. Για ένα πείραμα με διαφορετικές επαναλήψεις τα σφάλματα $\varepsilon_i$ και $\varepsilon_j$ $(i \neq j)$θεωρούνται ασυσχέτιστα, δηλαδή $Cov(\varepsilon_i,\varepsilon_j) = 0\ γιαι \neq j$.
   Για να προκύψει η παραπάνω σχέση αξίζει να αναφερθεί ότι:
   $E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i) = \beta_0 + \beta_1 x_i + E(\varepsilon_i)$και $V(Y_i) = V(\beta_0 + \beta_1 x_i + \varepsilon_i) = V(\varepsilon_i) = \sigma^2$ και χρησιμοποιώντας την υπόθεση 4 προκύπτει ότι $E(Y_i) = \beta_0 + \beta_1 x_i$,$V(Y_i) = \sigma^2$. Έπειτα, η υπόθεση 5 εξασφαλίζει ότι τα σφάλματα δεν

συνδέονται μεταξύ τους με κάποιο συστηματικό τρόπο. Έτσι το σφάλμα $\varepsilon_i$ που αφορά την i-δοκιμή δε σχετίζεται με το σφάλμα $\varepsilon_j$ της j-δοκιμής ($i \neq j$). *Δοθέντος ότι ισχύει :*

$$Cov(Y_i, Y_j) = Cov(\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_0 + \beta_1 x_j + \varepsilon_j) = Cov(\varepsilon_i, \varepsilon_j) = 0,$$

οι μεταβλητές $Y_i, Y_j$ που αντιστοιχούν στις τιμές $x_i$ και $x_j$ είναι ασυσχέτιστες.




*Εικόνα 3: Απλή γραμμική παλινδρόμηση*

*Με μορφή διανύσματος μπορεί να γραφεί*:$Y = X\beta + \varepsilon$

Για μία παρατήρηση το μοντέλο μπορεί να γραφεί ως εξής:

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

$\varepsilon \sim N(0, \sigma^2)$ είναι η κανονικότητα, τότε και το Y είναι κανονική κατανομή.

Το παραπάνω μοντέλο με τις εξής υποθέσεις μπορεί να ονομαστεί είτε στατιστικό γραμμικό μοντέλο αλλά και απλό γραμμικό μοντέλο.Έχουμε και το κανονικό γραμμικό μοντέλο όπου το $\varepsilon_i \sim N(0, \sigma^2)$.

Για παράδειγμα, η σύγκριση δύο μέσων και η σχέση μιας συνεχούς εξαρτημένης μεταβλητής και μίας συνεχούς ανεξάρτητης σε δύο ομάδες χρησιμοποιεί τέτοιου είδους υπόδειγμα.

Ένα μοντέλο της μορφής όπως στην σχέση **(3.2)** το οποίο περιγράφει έναν τρόπο πρόβλεψης της μεταβλητής Y μέσω μιας γραμμικής συνάρτησης του x λέγεται μοντέλο παλινδρόμησης της Y πάνω στην X ή απλά μοντέλο παλινδρόμησης της Y στη X.

Σε κάθε πρόβλημα παλινδρόμησης διακρίνουμε συνήθως δύο είδη μεταβλητών:

- *Τις ανεξάρτητες ή ελεγχόμενες*: οι οποίες συμβολίζονται με Χ, είναι εκείνες στις οποίες μπορούμε να δίνουμε μία συγκεκριμένη τιμή ή παίρνουν τιμές που μπορούμε να παρατηρήσουμε αλλά όχι να ελέγξουμε.
- *Τις εξαρτημένες μεταβλητές ή μεταβλητές απόκρισης*:οι οποίες συμβολίζονται με Υκαι αντανακλούν το αποτέλεσμα των μεταβολών στις ελεγχόμενες μεταβλητές.
  (Μάρκος Κούτρας-Χ.Ευαγγελάρας,Εκδόσεις Αθ.Σταμούλης,Αθήνα 2010)

**Ένα γενικευμένο γραμμικό μοντέλο μπορεί να ορίζεται από ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $Y_1, \dots, Y_N$ όπου τα δεδομένα ακολουθούν κατανομές της εκθετικής οικογένειας κατανομών**.

## *3.3: Εκθετική Οικογένεια Κατανομών*

Μερικές από τις πιο συνηθισμένες οικογένειες κατανομών (π.χ. διωνυμική, Poisson, εκθετική, κανονική) έχουν πυκνότητες που μπορούν να γραφούν σε μία ειδική μορφή, αυτήν της εκθετικής οικογένειας κατανομών. Θα ξεχωρίσουμε τις περιπτώσεις όπου η παράμετρος είναι μονοδιάστατη δηλαδή, η οικογένεια κατανομών του Χ είναι μία εκθετική οικογένεια, θα λέμε ότι είναι μία Μονοπαραμετική Εκθετική Οικογένεια Κατανομών (ΜΕΟΚ), από την άλλη έχουμε και την Πολυπαραμετρική Εκθετική Οικογένεια Κατανομών(ΠΕΟΚ).

Έστω μία τυχαία μεταβλητή Υμπορεί να γραφεί ως:

$$\boldsymbol{f(y) = exp\left\{\frac{y\theta - b(\theta)}{\alpha(\varphi)} + c(y,\varphi)\right\},} \qquad \textbf{(3.3)}$$

(Τυχαία μεταβλητή ονομάζεται η δυνατότητα ορισμού μιας συνάρτησης Χ η οποία σε κάθε σημείο ω του δειγματικού χώρου Ω να αντιστοιχεί ένα πραγματικό αριθμό).

Όπου:$\boldsymbol{\alpha, b, c}$ : γνωστές συναρτήσεις που καθορίζουν την συγκεκριμένη κατανομή $\boldsymbol{\theta}$:παράμετρος θέσης, όπου με τις παραμέτρους αυτές προσπαθούμε να καθορίσουμε την κεντρική θέση της κατανομής, δηλαδή το σημείο που αντιστοιχεί στην τιμή της μεταβλητής από την οποία τείνουν να συγκεντρωθούν οι μεταβλητές του πληθυσμού.

$\boldsymbol{\varphi}$:παράμετρος διασποράς, είναι ένα μέτρο το οποίο μας δίνει με τρόπο περιληπτικό και αντικειμενικό τη μεταβλητότητα ή ανομοιογένεια, των παρατηρήσεων.

$\boldsymbol{y}$ :ανήκει στην εκθετική οικογένεια κατανομών

Για μερικές κατανομές, μέλη της εκθετικής οικογένειας ισχύει φ=1.0(όπως η διωνυμική , η Poison) εκτός από περιπτώσεις όπου πρέπει να λάβουμε υπόψη ακόμη έναν παράγοντα. Για σταθερό φ έχουμε την μονοπαραμετρική εκθετική οικογένεια κατανομών με την οποία θα ασχοληθούμε.

### *3.3.1 : Μονοπαραμετρική Εκθετική Οικογένεια Κατανομών*

Έστω μία απλή τυχαία μεταβλητή Y της οποίας η συνάρτηση πιθανότητας(αν είναι διακριτή) ή συνάρτηση πυκνότητας πιθανότητας(αν είναι συνεχής) εξαρτάται από μία παράμετρο θ.

Ορίζουμε ότι η κατανομή ανήκει στην μονοπαραμετρική εκθετική οικογένεια κατανομών αν μπορεί να γραφεί στην μορφή:

$$\boldsymbol{f(y;\theta) = \exp[A(\theta) + B(y) + C(\theta)D(y)]}, \boldsymbol{y \in Y, \theta \in \Theta} \quad \textbf{(3.3.1)}$$

Ο παραμετρικός χώρος Θ στην περίπτωση μας που μιλάμε για εκτίμηση μιας παραμέτρου είναι ένα διάστημα πραγματικών τιμώνή αλλιώς μπορούμε να πούμε ότι είναι το σύνολο των επιτρεπτών τιμών της παραμέτρου θ.

Η παραπάνω εξίσωση μπορεί να γραφεί στην κανονική μορφή ως:

$$\boldsymbol{f(y;\theta) = \exp[A(\theta) + B(y) + yC(\theta)]}$$

Επίσης, πρέπει να τονιστεί ότιτο στήριγμα της συνάρτησης πιθανότητας δηλαδή το σύνολο $S = \{y: f(y;\theta) > 0\}$ πρέπει να είναι ανεξάρτητο από την παράμετρο θ.

### *3.3.2 :Παράδειγμα*

**1.** Έστω Y=$(Y_1, \dots, Y_n)$ τυχαίο δείγμα από κανονική κατανομή(συνεχής κατανομή) $N(\theta, \sigma^2), \sigma > 0$ $γνωστήσταθερά.$
Η συνάρτηση πυκνότητας πιθανότητας είναι της μορφής:

$$f_1(y;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(y-\theta)^2}{2\sigma^2}\right\}$$

Αφού έχουμε τυχαίο δείγμα έχουμε:

$$f(y;\theta) = \prod_{i=1}^{n} f_1(y_i;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(y_i-\theta)^2}{2\sigma^2}\right\} =$$

$$= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} =$$

$$= exp\left\{-nlog\sigma - \frac{n}{2}log2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\} =$$

$$= exp\left\{-nlog\sigma - \frac{n}{2}log2\pi - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2} + \frac{\theta\sum_{i=1}^{n} y_i}{\sigma^2} - \frac{n\theta^2}{\sigma^2}\right\}$$

Άρα ,
με βάση την ΜΕΟΚ έχουμε ότι:

$$A(\theta) = nlog\sigma - \tfrac{n}{2}log2\pi - \tfrac{n\theta^2}{\sigma^2}, B(y) = -\tfrac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}, C(\theta) = \tfrac{\theta}{\sigma^2}, D(y) = \textstyle\sum_{i=1}^{n} y_i$$

## *3.4: Γενικό Γραμμικό Μοντέλο*

Ένα γενικό γραμμικό μοντέλο μπορούμε να ορίσουμε ότι είναι της μορφής:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i} \textbf{ (3.4)}$$

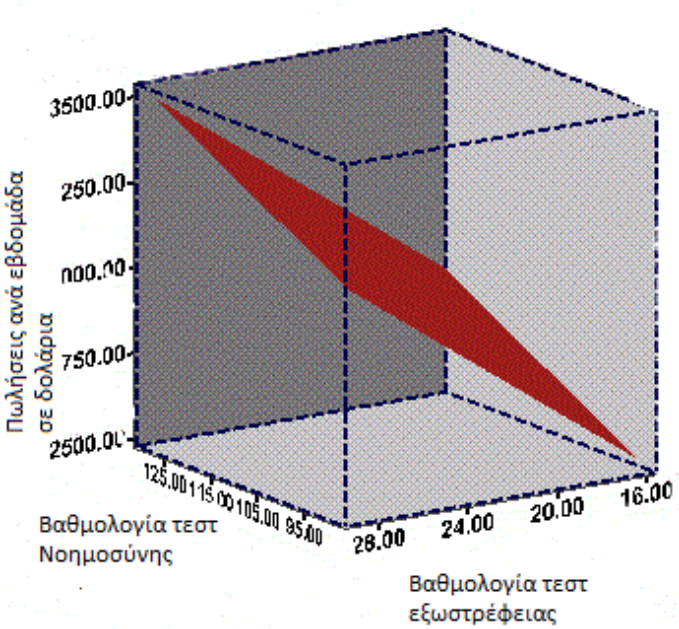
Όπου, $y_i$ είναι η απόκριση για $i = 1,2,\dots,n$ η οποία μοντελοποιείται από μία γραμμική συνάρτηση της επεξηγηματικής μεταβλητής $x_j, j = 1,2,\dots,p$ καθώς και έναν όρο σφάλματος όπως φαίνεται από την παραπάνω σχέση.

Ένα απλό γραμμικό μοντέλο είναι της μορφής: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Το μοντέλο είναι γραμμικό αν για παράδειγμα είναι της μορφής:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 {x_2}^2 + \varepsilon_i$$
$$y_i = \beta_0 + \gamma_1 \delta_1 x_1 + \exp(\beta_2)\, x_2 + \varepsilon_i$$

ενώ, αν δεν είναι γραμμικό είναι της μορφής:

$$y_i = \beta_0 + \beta_1 {x_1}^{x_2} + \varepsilon_i$$
$$y_i = \beta_0 \exp(\beta_1 x_i) + \varepsilon_i$$




*Εικόνα 4: Πολλαπλό γραμμικό μοντέλο σε τρισδιάστατη μορφή*

Το παραπάνω διάγραμμα αναφέρεται σε ένα παράδειγμα σχετικά με την μελέτη πωλήσεων της εταιρείας ABC σε κάποια καταστήματα για να καθορίσει αν η ευφυία και η εξωστρέφεια προβλέπουν την απόδοση των πωλήσεων των σημερινών εργαζομένων. Η λογική της εταιρείας αυτής είναι πως αν η νοημοσύνη και η εξωστρέφεια προβλέπει την απόδοση των πωλήσεων, τότε μια καλή στρατηγική για νέα καταστήματα είναι να προσλάβει έξυπνους εξωστρεφείς για τις θέσεις πωλήσεων. Για την διεξαγωγή της μελέτης όλες οι τρέχουσες λιανικές πωλήσεις στα υπάρχοντα καταστήματα έλαβαν ψυχολογικά τεστ, σχεδιασμένα να μετρούν την ευφυία και την εξωστρέφεια.Και όπως φαίνεται και από το παραπάνω διάγραμμα έχουμε 3 βαθμολογίες:

- Ένα τεστ νοημοσύνης(σε μία κλίμακα 50-150(χαμηλής-υψηλής νοημοσύνης))
- Ένα τεστ εξωστρέφειας(σε μία κλίμακα 15-30(χαμηλής-υψηλής εξωστρέφειας))
- Απόδοση των πωλήσεων που εκφράζεται ως μέσο ποσό σε δολλάρια που πωλούνται ανά εβδομάδα.

Έτσι το παραπάνω διάγραμμα μας δείχνει πως μία πολλαπλή παλινδρόμηση χρησιμοποιείται για να προβλέψει ως κριτήριο χρησιμοποιώντας δύο παράγοντες πρόβλεψης (οι δύο παράγοντες πρόβέψης είναι το τεστ νοημοσύνης και εξωστρέφειας).

*3.4.1:Όροι σφάλματος*

Υποθέτουμε ότι τα σφάλματα $\varepsilon_i$ είναι ανεξάρτητα και πανομοιότυπα κατανεμημένα έτσι ώστε:

$E(\varepsilon_i) = 0$ και $Var(\varepsilon_i) = \sigma^2$.

Συνήθως θεωρούμε $\varepsilon_i \sim N(0, \sigma^2)$

*3.4.2:Περιορισμοί στα γενικά γραμμικά μοντέλα*

Παρότι είναι πολύ χρήσιμα, υπάρχουν περιπτώσεις που τα γενικά γραμμικά μοντέλα δεν είναι κατάλληλα προς εφαρμογή, όπως:

- Όταν το εύρος y είναι περιορισμένο(π.χ. δυαδικό, μετρήσιμο)
- Όταν η διακύμανση της y εξαρτάται από το εύρος

Τα γενικευμένα γραμμικά μοντέλα τώρα επεκτείνουν το γενικό γραμμικό μοντέλο ώστε να καταφέρουν να αντιμετωπίσουν τα παραπάνω θέματα.

## *3.5: Γενικευμένο γραμμικό μοντέλο*

Στα γενικευμένα γραμμικά μοντέλα σημαντικό ρόλο παίζουν η κατανομή της απόκρισης και το μοντέλο που συνδέει τη μέση απόκριση με τις μεταβλητές παλινδρόμησης.Ένα γενικευμένο γραμμικό μοντέλο ορίζεται από ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $Y_1, Y_2, \dots, Y_N$ καθεμία από τις οποίες ακολουθεί μια κατανομή που ανήκει στην εκθετική οικογένεια κατανομών με τις ακόλουθες ιδιότητες:

1. Η κατανομή που ακολουθεί κάθε $Y_i$ έχει την κανονική μορφή και εξαρτάται από μία μόνο παράμετρο $\theta_i$.Τα $\theta_i$ δεν είναι απαραίτητο να είναι όλα ίδια,
$$f(y_i; \theta_i) = \exp[A(\theta_i) + B(y_i) + C(\theta_i)D(y_i)]$$
2. Οι κατανομές από όλα τα $Y_i$ είναι της ίδιας μορφής.
Έτσι η από κοινού συνάρτηση πυκνότητας πιθανότητας των $Y_1, Y_2, \dots, Y_N$ είναι:
$$f(y_1, y_2, \dots, y_N; \theta_1, \theta_2, \dots, \theta_N) = \prod_{i=1}^{N} \exp[A(\theta_i) + B(y_i) + y_i C(\theta_i)]$$
$$= \exp\left\{\sum_{i=1}^{N} A(\theta_i) + \sum_{i=1}^{N} B(y_i) + \sum_{i=1}^{N} y_i C(\theta_i)\right\}$$
Για τον προσδιορισμό του μοντέλου, συνήθως ενδιαφερόμαστε για ένα σύνολο παραμέτρων $b = (b_1, b_2, \dots, b_p)$,$\mathrm{p} < N$.
(Από κοινού συνάρτηση κατανομής π.χ. των τυχαίων μεταβλητών $X, Y$ονομάζεται η συνάρτηση $F$,που για κάθε $(x, y)$αντιστοιχεί την πιθανότητα:
$F(x, y) = P(X \le x, Y \le x)$. Ακόμη, από κοινού συνάρτης κατανομής των τυχαίων μεταβλητών $X_1, X_2, \dots, X_n$ ονομάζεται η συνάρτηση $F$, που για κάθε $(x_1, x_2, \dots, x_n)$ αντιστοιχεί η πιθανότητα:
$F(x_1, x_2, \dots, x_n) = P(X_1 \le x_1, \dots, X_n \le x_n)$).

### *3.5.1: Κατανομές στα Γενικευμένα Γραμμικά Μοντέλα*

Σημαντικές κατανομές στα γενικευμένα γραμμικά μοντέλα είναι η Εκθετική και η Γάμμα.Ειδικότερα η εκθετική είναι μία περίπτωση της Γάμμα αλλά κάθε μία έχει σημαντικές εφαρμογές.

- Η εκθετική με συνάρτηση πυκνότητας πιθανότητας:
$$f(y) = \frac{1}{\lambda} e^{-(y/\lambda)}, y > 0, \lambda > 0$$
- Η Γάμμα, η οποία εφαρμόζεται σε προβλήματα παλινδρόμησης όπου η απόκριση είναι συνεχής και η διακύμανση δεν είναι σταθερή αλλά ανάλογη του τετραγώνου του μέσου μ.Η συνάρτηση πυκνότητας πιθανότητας είναι:
$f(y) = \frac{1}{\Gamma(r)} \left(\frac{1}{\lambda}\right)' e^{-(y/\lambda)} y^{r-1}$,$\mathrm{r} > 0, \lambda > 0$.
όπου, r : παράμετρος κλίμακας.

### *3.5.2: Μεθοδολογία στα Γενικευμένα Γραμμικά Μοντέλα*

1. Οι ανεξάρτητες αποκρίσεις $y_1, y_2, \dots, y_n$ έχουν μέσους $\mu_1, \mu_2, \dots, \mu_n$.
2. Οι παρατηρήσεις $y_i$ ακολουθούν κατανομές που ανήκουν στην εκθετική οικογένεια κατανομών.
3. Το μοντέλο περιέχει μεταβλητές παλινδρόμησης της μορφής $x_1, x_2, \dots, x_k$.
4. Το μοντέλο κατασκευάζεται βάσει της γραμμικής παραμέτρου πρόβλεψης: $\eta = x'\beta = \beta_0 + \sum_{i=1}^{k} \beta_i x_i$.
   όπου: $\beta_i$:άγνωστοι παράμετροι οι οποίοι πρέπει να εκτιμηθούν από τα δεδομένα.
5. Το μοντέλο κατασκευάζεται μέσω της συνάρτησης σύνδεσης : $\eta_i = g(\mu_i), i = 1,2, \dots, n$
   η οποία δηλώνει συσχέτιση μεταξύ του μέσου και της γραμμικής παραμέτρου πρόβλεψης με μέση τιμή απόκρισης: $E(y_i) = g^{-1}(n_i) = g^{-1}(x_i'\beta)$.
6. Η συνάρτηση σύνδεσης είναι μονότονη και διαφοροποιήσιμη
7. Η διακύμανση $\sigma_i^2 (i = 1,2, \dots, n)$είναι συνάρτηση του μέσου μ. Αν όμως $\eta_i = \theta_i$τότε η $\eta_i$*:κανονική συνάρτηση σύνδεσης*.

Τα συνηθισμένο γραμμικό μοντέλο παλινδρόμησης χρησιμοποιεί γραμμικότητα για να περιγράψει τη σχέση μεταξύ της μέσης τιμής της μεταβλητότητας απόκρισης και μια σειρά επεξηγηματικών μεταβλητών. Τα γενικευμένα γραμμικά μοντέλα επεκτείνουν τα τυποποιημένα μοντέλα γραμμικής παλινδρόμησης συμπεριλαμβάνοντας μη-κανονικές κατανομές απόκρισης και πιθανόν μη γραμμικές λειτουργίες του μέσου.

***<u>Αποτελούνται από τρία βασικά εξαρτήματα:</u>***

- ***Τυχαία συνιστώσα:*** Καθορίζει την y μεταβλητή απόκρισης και την πιθανότητα κατανομής. Οι παρατηρήσεις $y = (y_1, \dots, y_n)^T$ αντιμετωπίζονται ως ανεξάρτητες. Δηλαδή, αποτελείται από μια μεταβλητή απόκρισης με $y$ ανεξάρτητες παρατηρήσεις και έχει πυκνότητα πιθανότητας από την εκθετική οικογένεια κατανομών. Περιορίζοντας έτσι το γενικευμένο γραμμικό μοντέλο στην εκθετική οικογένεια κατανομών παίρνουμε γενικές εκφράσεις για τις εξισώσεις μοντέλου πιθανότητας και έναν αλγόριθμο για την τοποθέτηση του μοντέλου.
- ***Γραμμικός συντελεστής πρόγνωσης:*** Για την παρατήρηση $i, i = 1,2, \dots, n$. Έστω $x_{ij}$ δηλώνουν την αξία της επεξηγηματικής μεταβλητής $x_j, j = 1,2, \dots, p$. Έστω, $x_i = (x_{i1}, \dots, x_{ip})$. Συνήθως θέτουμε $x_{ij} = 1$ή αφήνουμε τη πρώτη μεταβλητή να έχει 0 με $x_{i0} = 1$, έτσι ώστε να χρησιμεύει ως συντελεστής ενός όρου τομής στο μοντέλο. Ο γραμμικός προγνωστικός δείκτης στα γενικευμένα γραμμικά μοντέλα αφορά παραμέτρους $\{\eta\}$ σχετικά με την $E(y_i)$ με τις επεξηγηματικές μεταβλητές $x_1, \dots, x_p$ και χρησιμοποιούνται ως ένα γραμμικό συνδυασμό αυτών:

$$\eta_i = \underbrace{\sum_{j=1}^{n} \beta_j x_{ij}}, i = 1,2,\dots,n$$

Η παραπάνω σχέση αντικατοπρίζει το γεγονός ότι η έκφραση αυτή είναι γραμμικής στις παραμέτρους. Οι ίδιες οι επεξηγηματικές μεταβλητές μπορεί να είναι μη-γραμμικές συναρτήσεις όπως είναι ένας όρος αλληλεπίδρασης (π.χ. $x_{i3} = x_{i1}x_{i2}$)ή ένας τετραγωνικός όρος (π.χ.$x_{i2} = x_{i1}^2$). Σε μορφή πίνακα εκφράζεται ως:

$$\boldsymbol{\eta = X\beta}$$

όπου,$\eta = (\eta_1, \dots, \eta_n)^T$,

$\beta$: είναι ένας $p \times 1$διάνυσμα στήλης των παραμέτρων του μοντέλου,

$X$: ένας $n \times p$πίνακας των τιμών των επεξηγηματικών μεταβλητών.

Ο πίνακας $X$ καλείται μοντέλο πίνακα, έχει σειρές η μια για κάθε παρατήρηση και ρστήλες, μία για κάθε παράμετρο β.Συνήθως $p \leq N$, με στόχο του μοντέλου να είναι να συνοψίσει τα δεδομένα χρησιμοποιώντας ένα σημαντικά μικρότερο αριθμό παρατηρήσεων.

- ***Συνάρτηση σύνδεσης (link function)***

  *Η συνάρτηση σύνδεσης συνδέει την τυχαία συνιστώσα με το γραμμικό προγνωστικό δείκτη*. Έστω $\mu = E(y_i)\,, i = 1,2,\dots,n$. Συνδέουν το $\eta$ με το $\mu$ με την σχέση $\eta = g(\mu)$, όπου: $g(\cdot)$:μονότονη και διαφοροποιήσιμη συνάρτηση.

  Έτσι η $g$ συνδέεται με την $\mu_i$ με επεξηγηματικές μεταβλητές μέσω του τύπου:

$$\boldsymbol{\eta = g(\mu_i) = \sum_{j=1}^{n} \beta_j x_{ij}}$$

  Όπου:

$$j = 1,2,\dots,N$$

  $\beta_1, \dots, \beta_p$: ένας μικρός αριθμός άγνωστων παραμέτρων, οπότε $p < N$, οι παράμετροι αυτοί βοηθάνε στον καθορισμό του μ.

$x_1, \dots, x_p$: μεταβλητές οι οποίες δημιουργούν την παραπάνω γραμμική πρόβλεψη.

Η συνάρτηση σύνδεσης $g$ που μετατρέπει το μ σε φυσική παράμετρο ονομάζεται κανονική συνάρτηση σύνδεσης.

Η συνάρτηση σύνδεσης, εξισώνει την φυσική παράμετρο με τον γραμμικό προγνωστικό δείκτη και δημιουργεί τις πιο συχνά χρησιμοποιούμενες συναρτήσεις των γενικευμένων.

Στα γενικευμένα γραμμικά μοντέλα τώρα υπάρχει συνάρτηση g και ένα σύνολο παραμέτρων $\beta = \beta_1, \dots, \beta_p, (p < N)$ τέτοια ώστε ένας γραμμικός συνδυασμός των $\beta$ να είναι ίσος με τη συνάρτηση αναμενόμενης τιμής $(\mu_i) = x_i^T \beta$.

***Συνεπώς ένα γενικευμένο γραμμικό μοντέλο έχει τρεις συνιστώσες:***

- Έστω μεταβλητές απόκρισης $Y_1, Y_2, \dots, Y_N$ οι οποίες μοιράζονται την ίδια κατανομή από τη εκθετική οικογένεια.
- Ένα σύνολο από παραμέτρους $\beta$ και επεξηγηματικές μεταβλητές $X$

$$\beta = \begin{pmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_\rho \end{pmatrix} \quad , \quad X = \begin{pmatrix} X_1^T \\ X_2^T \\ . \\ . \\ . \\ X_N^T \end{pmatrix} = \begin{bmatrix} X_{11} & \dots & X_{1p} \\ \vdots & \ddots & \vdots \\ X_{N1} & \dots & X_{Np} \end{bmatrix}$$

- Μία γνήσια και διαφορίσιμη συνάρτηση σύνδεσης g τέτοια ώστε: $g(\mu_i) = X_1^T \beta$ , όπου: $\boldsymbol{\mu_i = E(Y_i)}$**.**

Η συνάρτηση σύνδεσης συσχετίζει τη γραμμική παράμετρο με την αναμενόμενη τιμή $\mu$της μεταβλητής απόκρισης $y$. Στα κλασικά γραμμικά μοντέλα η μέση τιμή $\mu$ταυτίζεται με την γραμμική πρόβλεψη. Σε περιπτώσεις που έχουμε διακριτές τιμές και η κατανομή που ακολουθούν είναι η Poisson κατανομή, πρέπει να ισχύει $\mu > 0$. Μοντέλα με τέτοιου είδους μεταβλητές, εκφράζονται με τη λογαριθμική συνάρτηση σύνδεσης, $\eta = \log(\mu)$ ώστε να υπάρχει γραμμική σχέση.

Όταν $\eta = \theta$ , όπου $\theta$ είναι η κανονική παράμετρος, οι κανονικές συναρτήσεις σύνδεσης για τις παρακάτω κατανομές είναι:

- Κανονική: $\eta = \mu$
- Poisson: $\eta = \log(\mu)$
- Διωνυμική: $\eta = log(\frac{p}{1-\mu})$
- Γάμμα: $\eta = \mu^{-1}$

### *3.5.3: Καταλληλότητα του μοντέλου*

Ένα σημαντικό κομμάτι αφού δημιουργήσουμε το μοντέλο είναι να εξετάσουμε κατά το πόσο είναι ικανό να περιγράψει τα δεδομένα μας, δηλαδή να γίνει αξιολόγηση της σημαντικότητας των μεταβλητών στο μοντέλο. Ουσιαστικά, θέλουμε να ερμηνεύσουμε ποιες μεταβλητές $X_i$ είναι σημαντικές.

Μέσω στατιστικών ελέγχων γίνεται η επιλογή των στατιστικά σημαντικών μεταβλητών στο μοντέλο. Αυτοί οι έλεγχοι καταλληλότηταςθα μας βοηθήσουν να ερευνήσουμε ποιες μεταβλητές είναι σημαντικές στο μοντέλο και ποιες έχει. Ο έλεγχος αυτός θα γίνει μέσω ελέγχων που αφορούν τους συντελεστές $b$.

Τώρα, για την ερμηνεία **της επάρκειας ενός μοντέλου**, μπορεί να χρησιμοποιηθεί η δειγματική κατανομή της **στατιστικής συνάρτησης D(Deviance).** Κάτι το οποίο γίνεται εκτιμώντας τα D από τα δεδομένα και συγκρίνοντας την τιμή με την κατάλληλη $X^2_{N-p}$ κατανομή.

- *3.5.3.1: Επάρκεια του μοντέλου*

Η επάρκεια προσαρμογής ενός μοντέλου σε ένα σύνολο δεδομένων μπορεί να γίνει συγκρίνοντας την συνάρτηση πιθανοφάνειας αυτού του μοντέλου με την συνάρτηση πιθανοφάνειας του άλλου μοντέλου , το οποίο ονομάζεται πλήρες μοντέλο και περιγράφεται ως εξής:

1. Το πλήρες μοντέλο είναι ένα γενικευμένο γραμμικό μοντέλο με την ίδια κατανομή όπως το μοντέλο που μας ενδιαφέρει
2. Το πλήρες μοντέλο έχει την ίδια συνάρτηση σύνδεσης με το μοντέλο που μας ενδιαφέρει
3. Ο αριθμός των παραμέτρων στο πλήρες μοντέλο ισούται με τον αριθμό των παρατηρήσεων.

<u>*Έστω δύο υπόθέσεις:*</u>

$H_0$:Θεωρητικό μοντέλο

$H_1$:Πλήρες μοντέλο

Πρίν αναφέρουμε το οτιδήποτε, πρέπει να ορίσουμε την συνάρτηση πιθανοφάνειας: Έστω τυχαίο δείγμα $X = (X_1, \dots, X_n)$ από τον πληθυσμό $(X, Xf(x;\theta), \theta = (\theta_1, \dots, \theta_m) \in \Theta)$ η οποία είναι: $\boldsymbol{L(\theta) = \prod_{i=1}^{n} f(x_i;\theta), \theta \in \Theta,}$ καθότι εκφράζει πόσο πιθανοφανείς ή διαφορεικά πόσο σύμφωνες είναι με το συγκεκριμένο δείγμα $X = x$, οι διάφορες τιμές της παραμέτρου $\theta$.

Μπορούμε να πούμε ότι το πλήρες μοντέλο περιγράφει πλήρως τα δεδομένα και επομένως θεωρείται ένα επαρκές μοντέλο. Οι συναρτήσεις πιθανοφάνειαςτου πλήρους μοντέλου με το μοντέλο που μελετάμε μπορεί να πάρουν αντίστοιχες τιμές εκτίμησης μέγιστης πιθανοφάνειας$l_{H_0}$ και $l_{H_1}$και να προκύψουν αντίστοιχα οι τιμές $L(l_{H_0}; y)$και $L(l_{H_1}; y)$.

Αν το μοντέλο περιγράφει τα δεδομένα ικανοποιητικά τότε το $L(l_{H_1}; y)$ πρέπει να είναι κοντά στο $L(l_{H_0}; y)$ ενώ σε αντίθετη περίπτωση το $L(l_{H_1}; y)$ πρέπει να είναι μικρότερο από το $L(l_{H_0}; y)$.

Γι'αυτό γίνεται χρήση του γενικευμένου λόγου πιθανοφανειών:

$\boldsymbol{\lambda = \frac{L(l_{H_0};y)}{L(l_{H_1};y)}}$ **ή ισοδύναμα** $\boldsymbol{Log\lambda = log\big(L(l_{H_0}; y)\big) - log\Big(L\big(l_{H_1}; y\big)\Big) = l\big(l_{H_0}; y\big) - l\big(l_{H_1}; y\big).}$

Μεγάλες τιμές του $log\lambda$ είναι μία ένδειξη μη καλής προσαρμογής του μοντέλου.

- *3.5.3.2: Στατιστική συνάρτηση Deviance*

  Είναι συνάρτηση η οποία ερμηνεύει την επάρκεια ενός μοντέλου και ονομάστηκε έτσι από τους Nelder και Wedderburn (1972). Ονομάζεται αλλιώς και στατιστική συνάρτηση αναλογίας λογαριθμικής πιθανότητας και ορίζεται ως εξής: (Διπλωματική Νταιλανάς)

$$\boldsymbol{D = 2log\lambda = 2[l(b_{H_0}; y) - l(b_{H_1}; y)]}\textbf{(3.5.3.2)}$$

Και το αποτέλεσμα είναι:

$$\boldsymbol{D = 2\{[l(l_{H_0}; y)] - l(\beta_{H_0}; y)] - [l(b_{H_1}; y) - l(\beta_{H_1}; y)] + [l(\beta_{H_0}; y) - l(\beta_{H_1}; y)]}$$

**Ο 1ος όρος ακολουθεί $\boldsymbol{X^2_m}$**, όπου$m$ :ο αριθμός των παραμέτρων του πλήρες μοντέλου
**Ο 2ος όρος ακολουθεί $\boldsymbol{X^2_p}$**, όπου $p$:ο αριθμός των παραμέτρων στο μοντέλο που μας ενδιαφέρει
**Ο 3ος όρος είναι μία θετική σταθερά η οποία θα είναι κοντά το μηδέν, αν το μοντέλο μας περιγράφει τα δεδομένα πολύ καλά**.

*[Πλήρες θεωρείται το γενικευμένο γραμμικό μοντέλο το οποίο χρησιμοποιεί την ίδια κατανομή και έχει την ίδια συνάρτηση με το μοντέλο που μας ενδιαφέρει.]*

Επειδή στην συνέχεια για να ορίσουμε πλήρως την συνάρτηση deviance θα χρησιμοποιήσουμε ελέγχους υποθέσεων θα πρέπει από στατιστικής πλευράς και να τους ορίσουμε κιόλας.Στατιστική υπόθεση είναι μία εικασία για την κατανομή μιας τυχαίας μεταβλητής και στατιστικός έλεγχος είναι μια διαδικάσία στην οποία χρησιμοποιείται ένα δείγμα για να αποφασισθεί αν πρέπει να απορρίψουμε την υπόθεση ή να τη δεχθούμε σαν αληθινή. Οι στατιστικοί έλεγχοι χρησιμοποιούνται αρκετά συχνά αφού σε πολλές περιπτώσεις χρειάζεται να πάρουμε αποφάσεις βασιζόμενοι στο αποτέλεσμα ενός πειράματος τύχης.

<u>Η έννοια του ελέγχου υποθέσεων</u>

Έστω ένα απλό παράδειγμα, η γέννηση ενός παιδιού μπορεί να θεωρηθεί σαν ένα τυχαίο γεγονός με δύο δυνατά αποτελέσματα:

$A_1$: $\gamma\acute{\varepsilon}\nu\nu\eta\sigma\eta$ $\alpha\gamma o\rho\iota o\acute{\upsilon}$ $\kappa\alpha\iota$ $A_2$: $\gamma\acute{\varepsilon}\nu\nu\eta\sigma\eta$ $\kappa o\rho\iota\tau\sigma\iota o\acute{\upsilon}$.

Συνήθως θεωρούμε τα δύο ενδεχόμενα $A_1, A_2$ ισοπίθανα με $P(A_1) = P(A_2) = \frac{1}{2}$. Έστω, p η πιθανότητα του ενδεχομένου $A_1$, οπότε 1-p είναι η πιθανότητα του $A_2$. Τότε ζητάμε να ελέγξουμε την υπόθεση:

$H_0: p = 1 - p$ έναντι $H_1: p > 1 - p$

***Συνεχίζοντας***, μπορούμε να πούμε ότι, όταν υπάρχει είτε καλή είτε κακή προσαρμογή στο μοντέλο ισχύει: $\boldsymbol{D \sim X^2_{m-p}}$.

Αν το μοντέλο είναι κατάλληλο τότε η τιμή D πρέπει να είναι κοντά στο μέσο της κατανομής.

Αν ένα μοντέλο με ρπαραμέτρους περιγράφει καλά ένα σύνολο Ν παρατηρήσεων τότε $D \approx N - p$.

Για να γίνει έλεγχος υποθέσεων χρησιμοποιούμε την στατιστική συνάρτηση deviance, θα ορίσουμε ένα μοντέλο για κάθε υπόθεση και θα συγκρίνουμε την στατιστική συνάρτηση της καλής προσαρμογής για τα συγκεκριμένα μοντέλα. Τα μοντέλα που θα συγκριθούν θα πρέπει να έχουν την ίδια κατανομή και συνάρτηση σύνδεσης.

Έτσι έχουμε: $H_0: b_0 = \begin{bmatrix} b_0 \\ \vdots \\ b_q \end{bmatrix}$ και $H_1: b_1 = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}$ οι δύο υποθέσεις.

Γίνεται έλεγχος $H_0$ έναντι της $H_1$ χρησιμοποιώντας τη διαφορά των συναρτήσεων deviance: $\boldsymbol{\Delta D = D_0 - D_1.}$

Η συνάρτηση deviance ορίζεται ως εξής:$\boldsymbol{D = 2\left[l\left(b_{H_0}; y\right) - l\left(b_{H_1}; y\right)\right]}$οπότε έχουμε:

$$\boldsymbol{\Delta D = D_0 - D_1 = 2[l(b_{max}; y) - l(b_0; y)] - 2[l(b_{max}; y) - l(b_1; y)] =}$$
$$\boldsymbol{= 2[l(b_1; y) - l(b_0; y)]}$$

Αν τα μοντέλα περιγράφουν καλά τα δεδομένα τότε: $D_0 \sim X^2_{N-q}$και $D_1 \sim X^2_{N-p}$, τότε $\Delta D \sim X^2_{p-q}$.
Αν $D \sim X^2_{p-q}$τότε επιλέγεται το μοντέλο $H_0$ ως πιο απλό.

## *3.6: Κατάλοιπα*

Είχαμε ορίσει ως συνάρτηση του γραμμικού μοντέλου την συνάρτηση **(3.2)** η οποία είναι ουσιαστικά η πρόβλεψη της εξαρτημένης μεταβλητής $Y$ για δεδομένη τιμή $x$ της ανεξάρτητης μεταβλητής $X$.

Χρησιμοποιτώντας τις εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1$,των παραμέτρων $\beta_0$,η λογική πρόβλεψη για την τιμή του $Y$, όταν $X = x_i$, θα είναι:

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1,2, \dots, \nu$ και λέγεται προσαρμοσμένη τιμή της μεταβλητής απόκρισης $Y$ στη θέση $x_i$, οι $\hat{Y}_i$ πρέπει να τονιστεί ότι είναι τυχαίες μεταβλητές αφού εκφράζονται μέσω των τυχαίων μεταβλητών(εκτιμητριών) $\hat{\beta}_0, \hat{\beta}_1$.

Αν ορίσουμε την διαφορά $Y_i - \hat{Y}_i, i = 1,2, \dots, \nu$ και λόγω της σχέσης **(3.2)** έχουμε ότι:

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$$

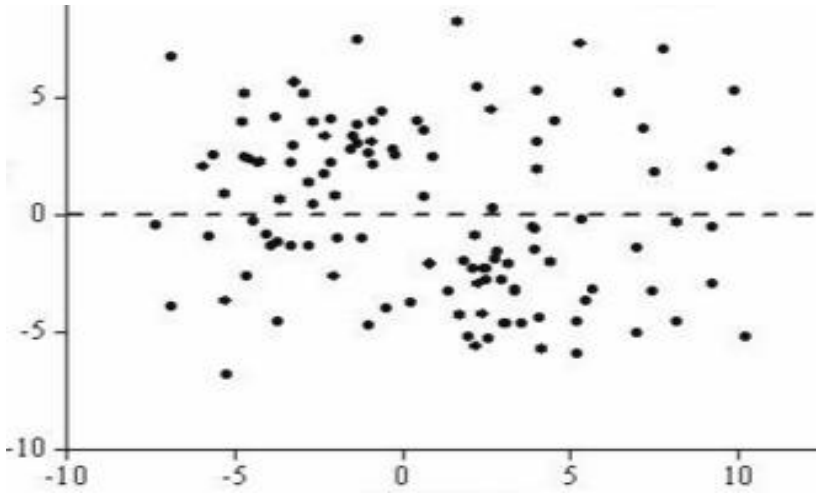
**Οπότε η διαφορά θα είναι:$\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}} = \boldsymbol{Y_i} - \left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}} \boldsymbol{x_i}\right) = \boldsymbol{Y_i} - \hat{Y}_i, \boldsymbol{i = 1, 2, \dots, \nu,}$ όπου τα $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}}$:ονομάζονται κατάλοιπα**.

Όπου $\boldsymbol{Y_i}$:παρατηρήσεις και$\hat{Y}_i$:εκτιμήσεις των παρατηρήσεων

Χρησιμοποιούνται για έλεγχο των χαρακτηριστικών των μοντέλων, καθώς επίσης μας δίνουν και πληροφορία για την καταλληλότητα του μοντέλου.

Τα κατάλοιπα $\hat{\varepsilon}_i, i = 1,2,\dots,\nu$, θα μπορούσε κανείς να τα δει ως τις παρατηρήσιμες τιμές των τυχαίων μεταβλητών $\varepsilon_i, i = 1,2,\dots,\nu$ υπό την προυπόθεση ότι το γραμμικό μοντέλο που ορίσαμε και στην αρχή του κεφαλαίου είναι σωστό. Επομένως θα ανέμενε κανείς ότι οι πορυποθέσεις που θέσαμε για τα σφάλματα $\varepsilon_i$ στο γραμμικό μοντέλο να μεταφέρονται κατάλληλα και στα κατάλοιπα$\hat{\varepsilon}_i$,οπότε πιθανή <<περίεργη>> συμπεριφορά των $\hat{\varepsilon}_i$να μπορεί να χρησιμοποιηθεί ως ένδειξη απόκλισης από τις υποθέσεις που τέθηκαν στο γραμμικό μοντέλο.
Η συνθήκη $E(\varepsilon_i) = 0, i = 1,2,\dots,\nu$ που ίσχυε στα σφάλματα εξακολουθεί να ισχύει και εδώ, δηλαδή, $E(\hat{\varepsilon}_i) = 0, i = 1,2,\dots,\nu$.

Η συνθήκη της σταθερότητας της διακύμανσης των σφαλμάτων $V(\varepsilon_i) = \sigma^2, i = 1,2,\dots,\nu$ που ίσχυε στα σφάλματα, δεν διατηρείται στα κατάλοιπα$\hat{\varepsilon}_i$, όπως επίσης και η συνθήκη που ισχύει στα σφάλματα ότι ανά δύο είναι ασυσχέτιστα δηλαδή $cov(\varepsilon_i, \varepsilon_j) = 0, για\ i \neq j$ δεν διατηρείται στα κατάλοιπα.




*Εικόνα 5: Διάγραμμα καταλοίπων για το κανονικό γραμμικό μοντέλο*

*3.6.1: Είδη καταλοίπων*

Ορίζουμε τα εξής κατάλοιπα:

1. *Κατάλοιπα deviance:*$r_D = sign(y - \mu)\sqrt{d_i}$ , $\sum_{i=1}^{N} r_D^2 = D$με $D = \sum_{i=1}^{N} d_i$
2. *Κατάλοιπα Pearson:*$r_p = \frac{y_i - \mu_i}{\sqrt{var(y_i)}}$ , i=1,2,…,N. Η κατανομή $r_p$ για μη-κανονικές κατανομές έχει δύσχρηστα υπόλοιπα και δεν έχει παρόμοιες ιδιότητες με αυτές των υπολοίπων των κανονικών κατανομών.
3. *Κατάλοιπα Anscode:*Ο Anscode προτείνει τον ορισμό της συνάρτησης$A(y)$, όπου $A(.)$, θα έκανε την κατανομή του $A(Y)$όσο το δυνατόν πιο κοντά στην κανονική.
   Η συνάρτηση$A(.)$ δίνεται από τον τύπο:

$A(\cdot) = \int \frac{d\mu}{var^{1/3}(\mu)}$. Έτσι τα Anscodeυπόλοιπα είναι:$r_A = \frac{\frac{3}{2}(y^{2/3} - \mu^{2/3})}{\mu^{1/6}}$.

4. *Κατάλοιπα Poisson*: $r_i = \frac{y_i - \hat{\lambda_i}}{\sqrt{\hat{\lambda_i}}}$ και $E(Y_i) = var(Y_i) = \lambda_i$.Τα κατάλοιπα Poisson μπορεί να θεωρηθούν σαν προσημασμένες τετραγωνικές ρίζες των όρων της στατιστικής συνάρτησης καλής προσαρμογής του Pearson: $\sum \frac{(o_i - e_i)^2}{e_i}$.
   Όπου,
   $o_i$: *είναι οι παρατηρηθείσες τιμές*
   $e_i$: *είναι οι προσαρμοσμένες τιμές* $\hat{\lambda}_\gamma$ *που αναμένονται απο το μοντέλο*.

Στα γενικευμένα γραμμικά μοντέλα ορίζονται τυποποιημένα κατάλοιπα με πολλούς διαφορετικούς τρόπους. Μία πρόταση είναι να γενικεύσουμε τον ορισμό των τυποποιημένων καταλοίπων για κανονικά μοντέλα και να προκύψει ο ορισμός των τυποποιημένων μοντέλων για γενικευμένα γραμμικά μοντέλα.

Έστω λοιπόν ότι $S_i = \sqrt{Var(\mu_i)}$η εκτιμηθείσα τιμή για την τυπική απόκλιση των προσαροσμένων τιμών $\hat{\mu}_i$.

Τα τυποπιημένα κατάλοιπα στα γενικευμένα γραμμικά μοντέλα ορίζονται ως:

$r_i = \frac{(y_i - \mu_i)}{s_i}$.

## *3.7: Εκτιμήσεις παραμέτρων*

**Για εκτιμήσεις παραμέτρων στην στατιστική χρησιμοποιούνται συνήθως δύο προσεγγίσεις οι οποίες είναι: << η μέθοδος της μέγιστης πιθανοφάνειας>> και <<η μέθοδος των ελαχίστων τετραγώνων>>. Υπάρχει βέβαια και <<η μέθοδος των σταθμισμένων ελαχίστων τετραγώνων>> η οποία θεωρείται μία παραλλαγή της μεθόδου ελαχίστων τετραγώνων.**

### *3.7.1: Είδη μεθόδων*

1. <u>*Μέθοδος Μέγιστης Πιθανοφάνειας*</u>
   Έστω $Y_1, Y_2, \dots, Y_N$τυχαίες μεταβλητές με από κοινού συνάρτηση πυκνότητας πιθανότητας $f(y_1, \dots, y_N; \theta_1, \dots, \theta_\rho)$.
   Εάν θεωρήσουμε ότι στην συνάρτηση $f$ τα $y_1, \dots, y_N$:σταθερά και $\theta_1, \dots, \theta_\rho$: μεταβλητές τότε: $\boldsymbol{L(\theta_1, \dots, \theta_\rho; y_1, \dots, y_N)}$ **είναι συνάρτηση πιθανοφάνειας.**

   Με μορφή πινάκων γράφεται ως:(θ; y), όπου: $y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_N \end{pmatrix}$και $\theta = \begin{pmatrix} \theta_1 \\ . \\ . \\ . \\ \theta_p \end{pmatrix}$

Η συνάρτηση πιθανοφάνειας$\mathrm{L}(\theta;\mathrm{y})$ είναι αλγεβρικά ίδια με την $f(\mathrm{y};\theta)$. Η αλλαγή στον συμβολισμό γίνεται για να δοθεί έμφαση στην μετατόπιση από την τυχαία μεταβλητή $y$ με σταθερά $\theta$, στην τυχαία μεταβλητή $\theta$ με σταθερά το $y$.

**Ισχύει ότι:** $\boldsymbol{L(\hat{\theta};y) \geq L(\theta;y) \forall \theta \in \Omega,}$

Όπου, θ:εκτιμητής μέγιστης πιθανοφάνειας

$\Omega$:σύνολο δυνατών τιμών των διανυσμάτων των παραμέτρων $\theta$.

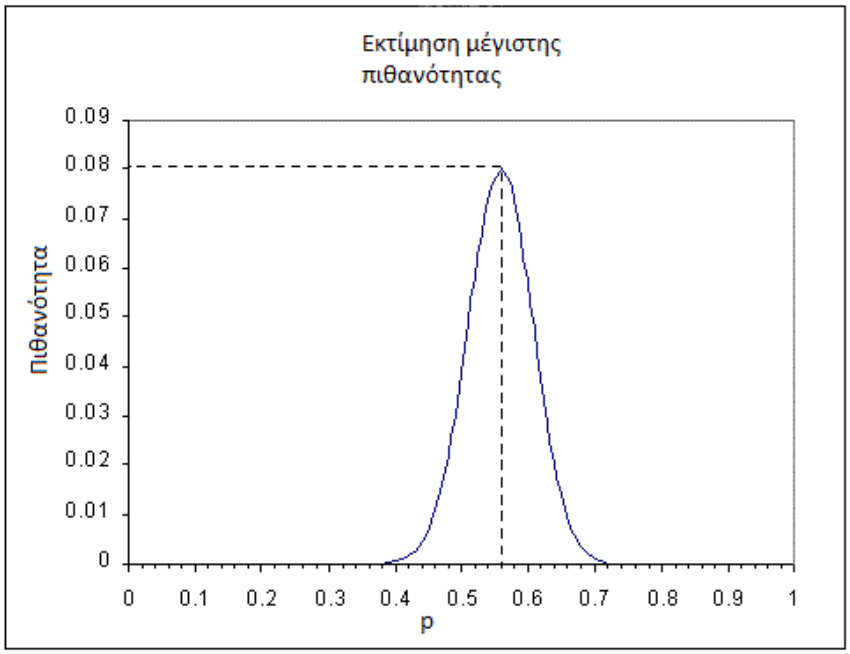
Επειδή, λογαριθμική συνάρτηση είναι μονότονη, ισοδύναμα ισχύει:

$l(\theta;y) = logL(\theta;y)$.

Στη συνέχεια λύνουμε το σύστημα εξισώσεων:

$\frac{\boldsymbol{dl(\theta;y)}}{\boldsymbol{d\theta_j}} = \mathbf{0}$για$\mathrm{j} = 1,2,\ldots,\mathrm{p}.$

Για να νατιστοιχούν οι λύσεις πράγματι σε μέγιστες τιμές του $l(\theta;y)$θα πρέπει ο πίνακας των δεύτερων παραγώγων $\frac{d^2 l(\theta;y)}{d\theta_j d\theta_k}$να είναι αρνητικά ορισμένος για την τιμή $\theta = \hat{\theta}$.




*Εικόνα 6: Διάγραμμα απεικόνισης εκτίμησης μέγιστης πιθανότητας*

Το παραπάνω διάγραμμα προκύπτει από το εξής πείραμα, αντί να υποθέσουμε ότι $p$είναι μια συγκεκριμένη τιμή (0,5) θέλουμε να βρούμε την μέγιστη εκτίμηση πιθανότητας (MLE) του $p$ δίνεται ένα συγκεκριμένο σύνολο δεδομένων. Κάνουμε μία δοκιμή να πετάξουμε ένα κέρμα 100 φορές και παρατηρούμε 56 κορώνα και 44 γράμματα. Έπειτα αντί να υποθέσω ότι $p = 0{,}5$ θέλουμε να βρούμε το MLEγια το $p$, και έπειτα αν η τιμή αυτή διαφέρει σημαντικά ή όχι από το 0,5. Βρίσκουμε την τιμή για το $p$ που καθιστά τα

δεδομένα πιο πιθανά. Τα παρατηρούμενα δεδομένα είναι πλέον σταθερά, άρα συνδέονται σε διωνυμικό μοντέλο πιθανοτήτων.
Έχουμε:
$n = 100$(συνολικός αριθμός που πετάμε το κέρμα)
$h = 56$(συνολικός αριθμός κορώνας)

Έστω ότι $p = 0{,}5$, συνδέουμε την τιμή αυτή με το μοντέλο ως εξής:

$$L(p = 0.5|data) = \frac{100!}{56!\,44!}0.5^{56}0.5^{44} = 0.0389$$

Αν η τιμή $p = 0{,}25$ τότε:

$$L(p = 0.52|data) = \frac{100!}{56!\,44!}0.52^{56}0.52^{44} = 0.0581$$

Αν πάρουμε και άλλες τιμές για το $p$ βλέπουμε πως αλλάζει η πιθανότητα. Το παραπάνω διάγραμμα δείχνει για τα δεδομένα όλο το εύρος των πιθανών τιμών για το $p$.Όπως φαίνεται η μέγιστη πιθανότητα εκτίμησης για το $p$ είναι περίπου 0,56. Σε μία τέτοια περίπτωση δεν χρησιμοποιείται η εκτίμηση μέγιστης πιθανότητας για να εκτιμηθεί το $p$. Σε πιο πολύπλοκα όμως μοντέλα και με πιο μεγάλο αριθμό παραμέτρων χρησιμοποιείται.

2. <u>*Μέθοδος Ελαχίστων τετραγώνων*</u>

Η μέθοδος αυτή συνιστάται στην εύρεση εκτιμητών $\hat{\beta}$ που ελαχιστοποιούν το άθροισμα των τετραγώνων των όρων των σφαλμάτων $e_i$.
Δηλαδή, προκύπτουν από την ελαχιστοποίηση της $S = \sum e_i^2$.
Υποθέτουμε ότι οι τυχαίες μεταβλητές $Y_1, \dots, Y_N$ με τιμές $(Y_i) = \mu_i\,, i = 1{,}2, \dots N$ και έστω $\mu_i$: συναρτήσεις των παραμέτρων $\beta_1, \dots, \beta_p, p < N$.
Έστω τώρα, β=$(\beta_1, \dots, \beta_p)^T$, θεωρούμε ότι: $Y_i = \mu_i + e_i, i = 1$ i=1,2,…,N.

Θέλουμε να ελαχιστοποιήσουμε το S
όπου,

$$\boldsymbol{S = \sum e_i^2 = \sum \left(Y_i - \mu_i(\beta)\right)^2}$$

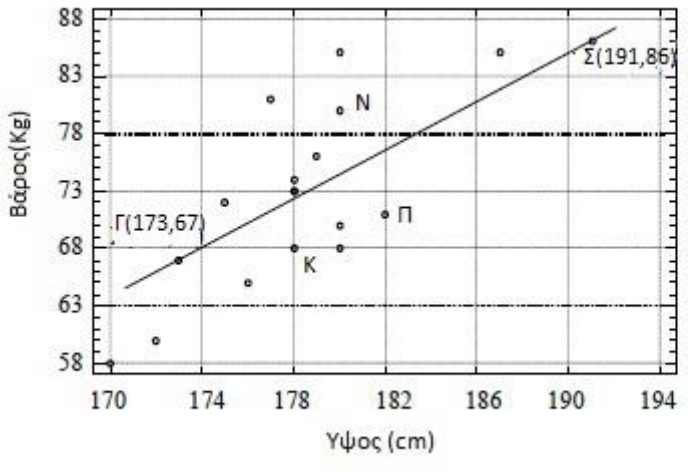
ή σε πίνακα ως εξής:

$$\boldsymbol{S = (y - \mu)^T (y - \mu)}$$

Όπου, $y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_N \end{pmatrix}$ ,$\mu = \begin{pmatrix} \mu_1 \\ . \\ . \\ . \\ \mu_N \end{pmatrix}$

Οι εκτιμητές του$\hat{\beta}$ βρίσκονται από τους παράγοντες του S ως προς κάθε στοιχείο του β, δηλαδή $\frac{dS}{d\beta_j} = 0$ γιαj $= 1,2,\ldots,$ p.

*Παράδειγμα*



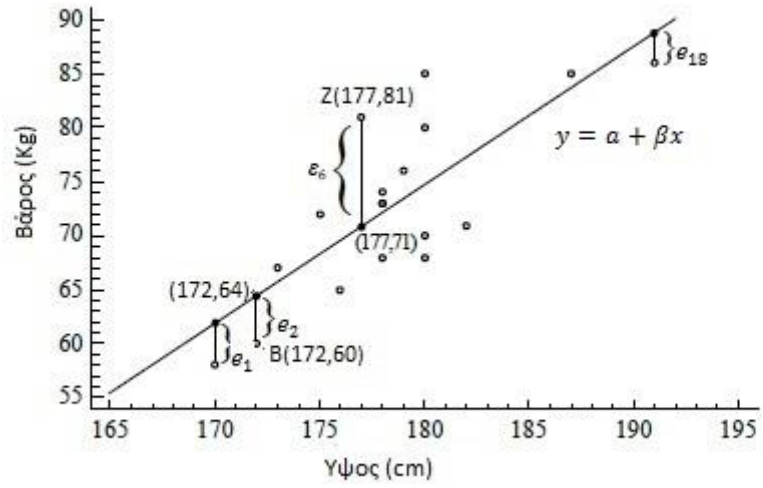

*Εικόνα 7: διασπορών για το ύψος και το βάρος 18 αγοριών Γ'Λυκείου*

Το παραπάνω δείγμα αποτελείται από το ζεύγος τιμών των συνεχών μεταβλητών $X$(ύψος(cm)) και $Y$(βάρος(kg)). Όπως μπορούμε να παρατηρήσουμε από το παραπάνω διάγραμμα τα σημεία $x, y$είναι συγκεντρωμένα γύρω από μία ευθεία.Θεωρούμε ως ανεξάρτητη την μεταβλητή $X$και ως εξαρτημένη την μεταβλητή $Y$, οπότε η ευθεία παλινδρόμησης που θα προσαρμόζει καλύτερα πάνω σε αυτά τα σημεία θα ονομάζεται ευθεία παλινδρόμησης της $Y$ πάνω στη $X$. Η εξίσωση της ευθεία θα είναι : $y = a + \beta x$, με όπου: $\alpha, \beta$:παράμετροι που θέλουμε να υπολογίσουμε.
Για τα σημεία $\Gamma(173,67)$ και $\Sigma(191,86)$ που επιλέξαμε πάνω στην ευθεία με το μάτι, αντικαθιστώντας στην παραπάνω εξίσωση θα έχουμε ότι:

$$\begin{cases} y_1 = a + \beta x_1 \\ y_2 = a + \beta x_2 \end{cases} <=> \begin{cases} 67 = a + 173\beta \\ 86 = \alpha + 191\beta \end{cases} <=> \alpha = -115{,}6 \; και \; \beta = 1{,}06$$

Άρα, $y = -115{,}6 + 1{,}06$, επόμενως η ευθεία που προσαρμόζεται καλύτερα στα σημεία του διαγράμματος διέρχεται από το σημείο $(0, -115{,}6)$και έχει συντελεστή διεύθυνσης 1,06.

Μία μέθοδος που χρησιμοποιείται για την εκτίμηση των παραμέτρων $\alpha, \beta$άρα και για την εύρεση της εξίσωσης της καλύτερης ευθείας που προσαρμόζεται στα δεδομένα, είναι η ΄΄μέθοδος ελαχίστων τετραγώνων΄΄.




*Εικόνα 8: προσαρμογής της ευθείας ελαχίστων τετραγώνων στο διάγραμμα διασποράς των παραπάνω δεδομένων*

Έτσι, για παράδειγμα όπως μπορεί να φανεί από το διάγραμμα, π.χ, για τον μαθητή $B(172{,}60)$ με ύψος $x_2 = 172cm$και κιλά $y_2 = 60kg$, ενώ σύμφωνα με την ευθεία που φέραμε, το βάρος του αναμένεται να είναι περίπου 60kg, δηλαδή έχουμε ένα σφάλμα $e_2 = 60 - 64 = -4$, δηλαδή βάρος 4kgλιγότερο από το αναμενόμενο, όμοια για τον μαθητή Z.

3. <u>*Σταθμισμένα Ελάχιστα Τετράγωνα*</u>

Χρησιμοποιείται στις περιπτώσεις όπου έχουμε επιπρόσθετη πληροφορία. Για την εισαγωγή της στους υπολογισμούς της απόδίδουμε βάρη στους όρους και ελαχιστοποιούμε το άθροισμα

$$S_w = \sum w_i \, [Y_i - \mu_i(\beta)]^2$$

Όπου, $w_i$:βάρη.

Γενικά, τα $Y_i$μπορεί να είναι συσχετισμένα.

Εάν V:πίνακας διακύμανσης-συνδιακύμανσης τότε οι εκτιμητές των σταθμισμένων ελαχίστων τετραγώνων προκύπτουν από την ελαχιστοποίηση του

$S_w = (y - \mu)^T V^{-1} (y - \mu)$.

Στην ειδική περίπτωση που οι όροι $\mu_i$:γραμμικοί συνδυασμοί των παραμέτρων $\beta_j (j = 1,2, \dots, p\,, p < N)$δηλαδή μ = Χβ για κάποιο N × pπίνακα Χ τότε: $S_w = (y - X\beta)^T V^{-1} (y - X\beta)$.

Οι παράγωγοι του $S_w$ ως προς τα στοιχεία $\beta_j$ του διανύσματος β είναι

$\frac{dS_w}{d\beta} = -2X^T V^{-1} (y - X\beta)$.

Θέλουμε:$\frac{dS_w}{d\beta}$=0 οπότε $2X^TV^{-1}(y-X\beta)$=0 συνεπώς $X^TV^{-1}X$b=$X^TV^{-1}y$.

Οπότε, οι εκτιμητές των ελαχίστων τετραγώνων b των διανυσμάτων παραμέτρων b είναι η λύση των κανονικών εξισώσεων $X^TV^{-1}X$b=$X^TV^{-1}y$.

*Παρατηρήσεις:*

1.Τα ελάχιστα τετράγωνα μπορούν να χρησιμοποιηθούν χωρίς να κάνουμε υποθέσεις για τις κατανομές των μεταβλητών $Y_i$

2.Στη μέθοδο μέγιστης πιθανοφάνειας χρειάζεται να καθορίσουμε την από κοινού πυκνότητα πιθανότητα των $Y_i$.

3.Για την εξήγηση της δειγματικής κατανομής των εκτιμητών ελαχίστων τετραγώνων, χρειάζονται επιπρόσθετες προυποθέσεις για τα $Y_i$.

### *3.7.2: Εκτιμήσεις για Γενικευμένα Γραμμικά Μοντέλα*

Οι υπολογισμοί των εκτιμητών γίνονται με δύο μεθόδους. Τη μέθοδο Newton – Raphson και την μέθοδο των score.

- ***Εκτίμηση με τη χρήση μεθόδου Newton – Raphson***

  Η μέθοδος αυτή χρησιμοποιείται για την εύρεση των λύσεων της εξίσωσης μίας μεταβλητής $f(x)=0$ όπου οι προσεγγίσεις των λύσεων για την περίπτωση αυτή δίνονται από:

$$x^{(m)} = x^{(m-1)} - \frac{f[x^{(m-1)}]}{f'[x^{(m-1)}]}$$

  Θέλουμε να υπολογίσουμε τους εκτιμητές των παραμέτρων β για τα γενικευμένα γραμμικά μοντέλα που περιγράφονται από:

$$f(y;\theta) = exp\left\{\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)\right\}$$

  Η λογαριθμική συνάρτηση πιθανοφάνειας για $Y_1, \dots, Y_N$ είναι:

$$l(\theta;y) = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i)$$

  για τις οποίες ισχύουν επίσης:

  $E(Y_i) = \mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)}$ και $g(\mu_i) = X_i^T\beta = \eta_i$

  όπου, g:μονότονη και παραγωγίσιμη συνάρτηση.

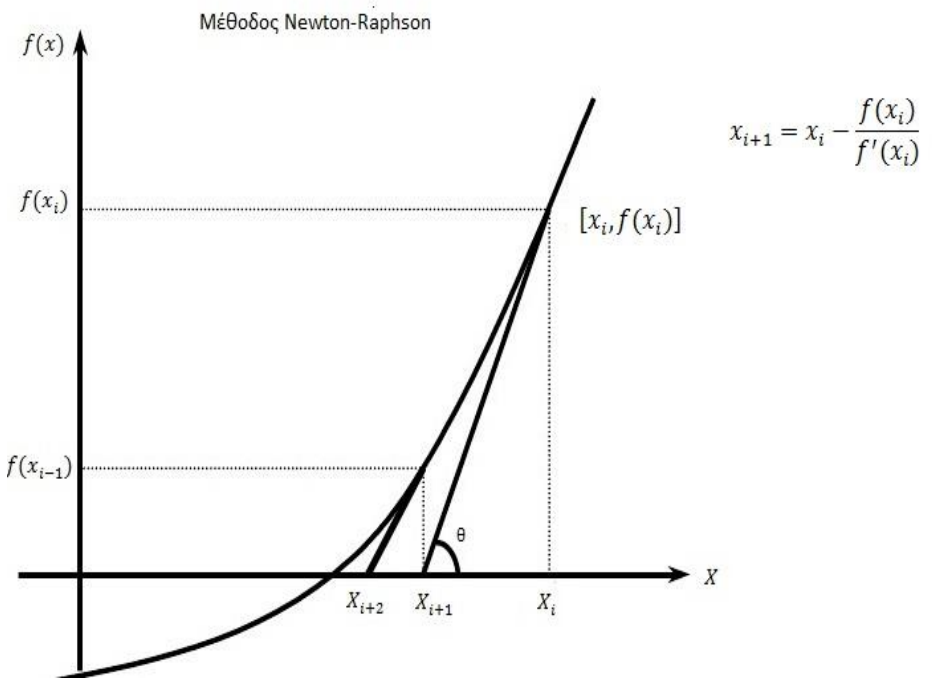
Επειδή οι κατανομές ανήκουν στην εκθετική οικογένεια κατανομών, το ολικό μέγιστο της λογαριθμικής συνάρτησης πιθανοφάνειας $l(\theta; y)$ δίνεται μοναδιαία από την λύση του συστήματος εξισώσεων $\frac{dl}{d\theta} = 0$ ή $\frac{dl}{d\beta} = 0$.

Ισχύει ότι:

$\frac{dl}{d\beta_j} = U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{d\mu_i}{d\eta_i}\right)$, όπου $x_{ij}$ είναι το j στοιχείο του $x^T i$. Στην γενική τους περίπτωση οι $U_j = 0, j = 1,2, \dots, p$ δεν είναι γραμμικές και οπότε πρέπει να λυθούν με μία επαναληπτική μέθοδο, η οποία είναι η μέθοδος Newton-Raphson, η οποία δίνει την m-οστή προσέγγιση από τη σχέση:

$$b^m = b^{(m-1)} - \left[\frac{d^2 l}{d\beta_j \beta_\kappa}\right]^{-1}_{\beta = b^{(m-1)}} U^{(m-1)}$$

όπου, $\left[\frac{d^2 l}{d\beta_j \beta_\kappa}\right]^{-1}_{\beta = b^{(m-1)}}$ είναι ο πίνακας των δεύτερων παραγώγων του $l$ για την τιμή του $\beta = b^{(m-1)}$ $και$ $U^{(m-1)}$ είναι το διάνυσμα των πρώτων παραγόντων $U_j = \frac{dl}{d\beta_j}$ για $\beta = b^{(m-1)}$.




*Εικόνα 9: Για την μέθοδο Newton-Raphson*

Ο τύπος$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$ : υπολογίζει την επόμενη εκτίμηση της ρίζας

$|e_a| = \left|\frac{x_{i+1} - x_i}{x_{i+1}}\right| \times 100$: απόλυτη σχετική προσέγγιση σφάλματος

- ***Εκτίμηση με χρήση της μεθόδου του score***

Είναι μία μέθοδος, κάποιες φορές πιο εύκολη από εκείνη του Newton-Raphson, η οποία προκύπτει αντικαθιστώντας τον πίνακα των δεύτερων παραγώγων με εκείνο των αναμενόμενων τιμών $E\left[\frac{d^2l}{d\beta_j\beta_\kappa}\right]$ το οποίο ισούται με τον πίνακα πληροφορίας $J$ ο οποίος είναι: $J = E[UU^T]$ και έχει:

$$J_{j_k} = E\left[U_j U_k\right] = E\left[\frac{dl}{d\beta_j}\frac{dl}{d\beta_k}\right] = -E\left[\frac{d^2l}{d\beta_j d\beta_k}\right].$$

Προκύπτει ότι:

$$b^{(m)} = b^{(m-1)} - \left[\frac{d^2l}{d\beta_j d\beta_k}\right]^{-1}_{\beta=b^{(m-1)}} U^{(m-1)} \text{ ή}$$

$$b^{(m)} = b^{(m-1)} + \left[j^{(m-1)}\right]^{-1} U^{(m-1)},$$

όπου ο όρος $J^{(m-1)}$ είναι ο πίνακας πληροφορίας που υπολογίζεται στο $b^{(m-I)}$.

Πολλαπλασιάζοντας με $f^{(m-I)}$ και έχουμε:

$$\boldsymbol{j^{(m-1)}b^{(m)} = j^{(m-1)}b^{(m-1)} + U^{(m-1)}(*)}$$

Για τα γενικευμένα γραμμικά μοντέλα μπορούμε να πούμε ότι το $(j,k)$ στοιχείο του $J$ είναι:$j_{jk} = \sum_{i=1}^{N} \frac{x_{ij}x_{ik}}{Var(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2$, $\acute{\alpha}\rho\alpha\ to\ J = X^T W X$,

όπου, $W$:ένας $NxN$ διαγώνιος πίνακας με στοιχεία:$w_{ii} = \frac{1}{Var(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2$

Στην σχέση **(*)** το δεξί μέλος είναι διάνυσμα με στοιχεία:

$$\sum_k \sum_i \frac{x_{ij}x_{ik}}{Var(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right)^2 b_k{}^{(m-1)} + \sum_i \frac{(y_i-\mu_i)x_{ij}}{var(Y_i)}\left(\frac{d\mu_i}{d\eta_i}\right).$$

Έτσι, προκύπτει ότι το δεξιό μέρος μπορεί αν γραφτεί ως: $X^T W_z$ όπου, το z παίρνει τις τιμές :$z_i = \sum_k x_{ik} b_k{}^{(m-1)} + (y_i - \mu_i)\left(\frac{d\eta_i}{d\mu_i}\right)$ όπου τα $\mu_i$ και $\frac{d\eta_i}{d\mu_i}$υπολογίζονται από το $b^{(m-1)}$.

Συνεπώς η επαναληπτική εξίσωση με τη μέθοδο του score είναι: $\boldsymbol{X^T W X b^{(m)} = X^T W_Z (**).}$

Μπορούμε να παρατηρήσουμε ότι η παραπάνω εξίσωση έχει την ίδια μορφή με εκείνη των σταθμισμένων ελαχίστων τετραγώνων με τη διαφορά το ότι πρέπει να λυθούν με μία επαναληπτική μέθοδο επειδή τα $\boldsymbol{z}$και $\boldsymbol{W}$ εξαρτώνται σε γενικές γραμμές από το $\boldsymbol{b}$.

**Επομένως, συμπεραίνουμε ότι οι εκτιμητές μέγιστης πιθανοφάνειας των γενικευμένων γραμμικών μοντέλων προκύπουν από μία επαναληπτική διαδικασία σταθμισμένων ελαχίστων τετραγώνων.**

## *3.8: Παραδείγμα*

1ο:Ξέρουμε ότι οι κτιμήσεις για γενικευμένα γραμμικάμοντέλα: ***Εκτίμηση Μεθόδου Newton-Raphson και εκτίμηση μεθόδου Fisher scoring***

Θεωρούνται αριθμητικές μέθοδοι οι οποίοι χρησιμοποιούνται για τον υπολογισμό της εκτιμήτριας μέγιστης πιθανοφάνειας. Με την βοήθεια ενός παραδείγματος θα προσπαθήσουμε να εξηγήσουμε την έννοια αυτών των δύο μεθόδων.

Έστω δεδομένα που εναφέρονται σε ώρες επιβίωσης που μετρούν την αντοχή συγκεκριμένων πλοίων σε συνθήκες πίεσης.

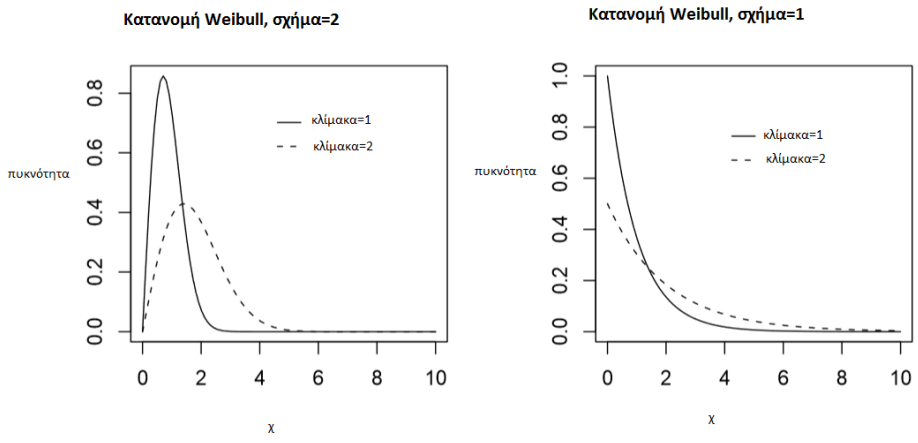
Ένα συγκεκριμένο μοντέλο που χρησιμοποιείται είναι το Weibull

$$\boldsymbol{f(y;\lambda;\theta;) = \frac{\lambda y_i^{\lambda-1}}{\theta^{\lambda}} e^{-\left(\frac{y_i}{\theta}\right)^{\lambda}}, y > 0}$$

Όπου, λ:παράμετρος που καθορίζει το σχήμα της κατανομής

θ:παράμετρος που καθορίζει την κλίμακα

Δίνονται τα γραφήματα της κατανομής Weibull για λ=1,2 και θ=1,2.



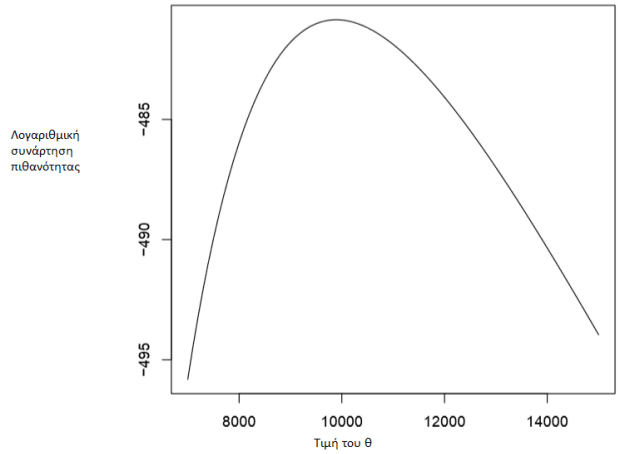

*Εικόνα 10: Κατανομή Weibull για λ=1,2 και θ=1,2*

Υποθέτουμε ότι τα δεδομένα μας ακολουθούν κατανομή Weibull με λ=2, καθώς έχουμε αποδεχτεί ότι η παράμετρος λ είναι γνωστή. Έστω, τώρα $y_1, \dots, y_n$ δεδομένα με λ γνωστό. Τότε η απο κοινού συνάρτηση πυκνότητας δίνεται από:

$$f(y_1, \dots, y_n; \theta) = \prod_{i=1}^{n} \frac{\lambda {y_i}^{\lambda-1}}{\theta^{\lambda}} e^{-\left(\frac{y_i}{\theta}\right)^{\lambda}}$$

Η συνάρτηση πιθανοφάνειας είναι:

$$L(\theta) = \log \mathrm{f}(y_1, \dots, y_n; \theta) = \sum_{i=1}^{n} \left\{ (\lambda - 1) log y_i + log\lambda - \lambda log\theta - \left(\frac{y_i}{\theta}\right)^{\lambda} \right\}$$




*Εικόνα 11: Γράφημα της συνάρτησης λογαριθμικής πιθανοφάνειας*

Παρατηρούμε ότι υπάρχει τιμή της θ η οποία μεγιστοποιεί την $L(\theta)$. ***Η συνάρτηση score δίνεται από:*** $\boldsymbol{\frac{dL}{d\theta} = U = \sum_{i=1}^{n} \{ -\frac{\lambda}{\theta} + \frac{\lambda {y_i}^{\lambda}}{\theta^{\lambda+1}} \} = -\frac{\lambda n}{\theta} + \frac{\lambda \sum_{i=1}^{n} y_i^{\lambda}}{\theta^{\lambda+1}}}$

Παρατηρούμε ότι για λ=2:

$$U(\theta) = 0 => \frac{-2n}{\theta} + \frac{2\sum_{i=1}^{n} {y_i}^2}{\theta^3} = 0 => \frac{n}{\theta} = \frac{\sum_{i=1}^{n} y_i^2}{\theta^3} =>$$
$$=> \theta^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} => \theta = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n}}.$$

Δηλαδή, η εκτίμηση μέγιστης πιθανότητας μπορεί να υπολογιστεί ακριβώς. Θα συγκρίνουμε το ακριβές αποτέλεσμα με εκείνο το οποίο μας δίνουν οι αναδρομικές μέθοδοι. Πρώτα όμως, τη μέθοδο Newton-Raphson.

Γενικά θέλουμε να υπολογίσουμε την τιμή της x, για την οποία θα ισχύει ότι: $f(x) = 0$. Η εφαπτόμενη της $f(x)$ στο σημείο $x^{(m-1)}$ δίνεται από: $\left[\frac{df}{dx}\right]_{x=x^{(m-1)}} = f'\left(x^{(m-1)}\right) = \frac{f(x^{(m)})-f(x^{(m-1)})}{x^{(m)}-x^{(m-1)}}$.

Όπου η απόσταση $x^{(m)} - x^{(m-1)}$ είναι μικρή.

Αν το $x^{(m)}$ είναι η λύση της εξίσωσης $f(x) = 0$, δηλαδή $f\left(x^{(m)}\right) = 0$, έχουμε ότι:

$$\boldsymbol{x^{(m)} = x^{(m-1)} - \frac{f(x^{(m-1)})}{f'(x^{(m-1)})}}$$

Τώρα για $m = 1,2, \ldots$, και με αρχική τιμή $x^{(1)}$, βρίσκουμε διαδοχικές προσεγγίσεις έτσι ώστε $\left|x^{(m)} - x^{(m-1)}\right| < \varepsilon$. Ειδικά για την εκτιμήτρια μέγιστης πιθανοφάνειας έχουμε ότι:

$\theta^{(m)} = \theta^{(m-1)} - \frac{U(\theta^{(m-1)})}{U'(\theta^{(m-1)})}$.

Έχουμε ότι:$U(\theta) = -\frac{2n}{\theta} + \frac{2\sum_{i=1}^{n} {y_i}^2}{\theta^3} = 0$

Και $\frac{dU(\theta)}{d(\theta)} = U'(\theta) = \sum_{i=1}^{n} \frac{\lambda}{\theta^2} - \frac{\lambda(\lambda+1)y_i^{\lambda}}{\theta^{\lambda+2}}$ και για λ=2 γίνεται: $\frac{dU(\theta)}{d(\theta)} = \frac{2n}{\theta^2} - \frac{2*3*\sum y_i^2}{\theta^4}$.

Στη συνέχεια θέτουμε στην παραπάνω αναδρομική σχέση την

$E\left(U'(\theta)\right) = -J(\theta) => J(\theta) = -E\left(U'(\theta)\right) = \frac{\lambda^2 n}{\theta^2}$, η οποία ονομάζεται πληροφορία Fisher.

Έτσι καταλήγουμε σε μία τροποποίηση του αλγόριθμου Newton-Raphson ο οποίος ονομάζεται Fisher scoring ,δηλαδή :$\theta^{(m)} = \theta(m-1) + \frac{U(\theta^{(m-1)})}{J(\theta^{(m-1)})}$.

Από τα παραπάνω μπορούμε να πούμε ότι οι αναδρομικές σχέσεις που ορίζονται από τον αλγόριθμο Fisher-scoring συγκλίνουν στην ακριβή τιμή της $\hat{\theta}$.

## *3.9: Σχέση γεμικευμένων γραμμικών μοντέλων με δομικά μοντέλα εξισώσεων*

### *3.9.1: Εισαγωγή*

Στην μεγαλύτερη πλειοψηφία, οι στατιστικές διαδικασίες των παραμέτρων που χρησιμοποιούνται ευρέως είναι μέρος του γενικού γραμμικού μοντέλου(GLM) συμπεριλαμβανομένης της ανάλυση διακύμανσης (ANOVA), πολλαπλής παλινδρόμησης, περιγραφικής ανάλυσης διακρίνουσας, πολυπαραγοντική ανάλυση

διακύμανσης (MANOVA), ανάλυση κανονικής συσχέτισης (CCA) και δομικά μοντέλα εξισώσεων (SEM). Όλες αυτές οι διαδικασίες είναι ιεραρχικές, εκτός από ορισμένες που είναι ειδικές περιπτώσεις των άλλων. Όλες οι κλασικές μονοπαραγοντικές τεχνικές είναι ειδικές περιπτώσεις της πολλαπλής παινδρόμησης, οι μονοπαραγοντικές και πολυπαραγοντικές διαδικασίες των GLMείναι ειδικές περιπτώσεις της ανάλυσης κανονικών συσχετίσεων και όλες οι διαδικασίες των GLMείναι ειδικές περιπτώσεις των SEM.

Όλες οι στατιστικές διαδικασίες που βασίζονται στο GLMμοιράζονται μία σειρά από χαρακτηριστικά. Όλες οι διαδικασίες GLMείναι διαδικασίες ελαχίστων τετραγώνων που άμμεσα ή έμμεσα:

- Χρησιμοποιούνται τα βάρη ελαχίστων τετραγώνων, για την βελτιστοποίηση του σφάλματος του μοντέλου διακύμανσης
- Εστίαση στο συνθετικό των λανθάνουσων μεταβλητών που δημιουργούνται από την εφαρμογή των βαρών
- Απόδοση της διακύμανσης που αντιπροσωπεύει τα μεγέθη επίδρασης ανάλογα με $r^2$.

Παρά το γεγονός ότι αυτά τα βάρη, οι λανθάνουσες μεταβλητές και τα μεγέθη επίδρασης μπορεί να έχουν διαφορετικά ονόματα σε όλες τις διαδικασίες GLM, όλα είναι ανάλογα.

Πρέπει να τονιστεί ότι τα SEMείναι κάτι περισσότερο από ένα στατιστικό εργαλείο. Με την χρήση κατάλληλης συνάφειας και θεωρητικών παραδοχών, τα SEM μπορεί να χρησιμοποιηθούν ως μεθοδολογικό εργαλείο όχι μόνο για τον καθορισμό της πιθανότητας όπως γίνεται στις στατιστικές διαδικασίες.

**<u>Συγκεντρωτικά,</u>**
Μπορούμε να πούμε ότι, οι εκτιμητές μέγιστης πιθανοφάνειας των γενικευμένων γραμμικών μοντέλων προκύπτουν από μία επαναληπτική διαδικασία σταθμισμένων ελάχιστων τετραγώνων.

Γενικά,στα γενικευμένα γραμμικά μοντέλα παρατηρούμε τα εξής:

- Οι αναλυτές δεδομένων όταν χρησιμοποιούν μετασχηματισμούς επιλέγουν συχνότερα τη μέθοδο ελαχίστων τετραγώνωννα εφαρμόσουν στο μοντέλο.
- Η διακύμανση της απόκρισης υδεν είναι σταθερή για αυτό χρησιμοποιούμε τη μέθοδο σταθμισμένων ελαχίστων τετραγώνων για την εκτίμηση των παραμέτρων.
- Τα γενικευμένα γραμμικά μοντέλα πρέπει να υπερέχουν αναλύσεων που βασίζονταισε μετασχηματισμούς όταν ένα πρόβλημα εξακολουθεί να έχει σταθερή διακύμανση μετά το μετασχηματισμό.
- Η απόκλιση του μοντέλου μπορεί να χρησιμοποιεί τον έλγχο της καταλληλότητας του μοντέλου. Τα αποτελέσματα του ελέγχου του Waldμπορούν να χρησιμοποιούνται στον έλεγχο της υπόθεσης αλλά και στον κατασκευασμό διαστημάτων εμπιστοσύνης για κάθε παράμετρο του μοντέλου ξεχωριστά.

# *Κεφάλαιο 4: Πολυεπίπεδα μοντέλα*

## *4.1: Πολυεπίπεδα δεδομένα και πολυεπίπεδη ανάλυση*

***Η πολυεπίπεδη μοντελοποίηση είναι μία προσέγγιση που μπορεί να χρησιμοποιηθεί για να χειριστεί συγκεντρωμένα ή ομαδοποιημένα δεδομένα***. Έστω ότι προσπαθούμε να ανακαλύψουμε μερικούς παράγοντες που επηρεάζουν την απόδοση ενός παιδιού στα αγγλικά στην ηλικία των 16 ετών.Το δείγμα των μαθητών που συμμετέχουν στη μελέτη, θα πρέπει να διδάσκονται στις τάξεις μέσα στα σχολεία. Παίζει σημαντικό ρόλο η επίδραση ενός συνδυασμού παραγόντων σε κάθε μαθητή, για παράδειγμα, η κοινωνικοοικονομική κατάσταση των γονέων του παιδιού, όπως επίσης και παράγοντες σχετικά με το επίπεδο της τάξης.Ακόμη αν ενδιαφερόμαστε κυρίως σε παράγοντες που αφορούν το επιπέδο του μαθητή, θα πρέπει να ληφθεί υπόψη η ομαδοποίηση. Για παράδειγμα τα επίπεδα απόδοσης δύο παιδιών που βρίσκονται στην ίδια τάξη θα τείνει να είναι πιο όμοια συγκριτικά με τα επίπεδα δύο παιδιών που βρίσκονται σε διαφορετικές τάξεις.Αν χρησιμοποιηθούν στατιστικές τεχνικές που αγνοούν την ομαδοποίηση π.χ. πολλαπλή παλινδρόμηση, τότε τα τυπικά σφάλματα και τα διαστήματα εμπιστοσύνης θα είναι ρεαλιστικά και μπορούμε κάλλιστα να καταλήξουμε στο συμπέρασμα ότι υπάρχουν πραγματικά αποτελέσματα κατά τα οποία απλά εξετάστηκε η τυχαία διακύμανση.

Ακόμη η πολυεπίπεδη μοντελοποίηση μπορεί να χρησιμοποιηθεί για την ανάλυση δεδομένων επαναλαμβανόμενων μέτρων. Για παράδειγμα αν έχουμε τη μέτρηση της πίεσης του αίματος σε μία ομάδα ασθενών σε εβδομαδιαία διαστήματα, μπορούμε να σκεφτούμε τις διαδοχικές μετρήσεις όπως ομαδοποιούνται στο πλαίσιο των επιμέρους θεμάτων. Ένα πλεονέκτημα αυτής της προσέγγισης της πολυεπίπεδης μοντελοποίησης είναι ότι μπορεί να ασχοληθεί με δεδομένα στα οποία οι χρόνοι των μετρήσεων ποικίλουν από υποκείμενο σε υποκείμενο.

Η πολυεπίπεδη ανάλυση είναι μια κατάλληλη προσέγγιση για να λάβει υπόψην τις κοινωνικές συνθήκες, καθώς και τους επιμέρους ερωτηθέντες και θέματα.

Το ιεραρχικό μοντέλο είναι το είδος της ανάλυσης παλινδρόμησης για τα πολυεπίπεδα δεομένα, όπου η εξαρτημένη μεταβλητή είναι στο χαμηλότερο επίπεδο. Ενώ από την άλλη οι επεξηγηματικές μεταβλητές μπορούν να οριστούν σε οποιδήποτε επίπεδο. Ακόμη το ιεραρχικό γραμμικό μοντέλο είναι βολικό να εφαρμοστεί και σε διαμήκη δεδομένα.

## *4.2: Βασικά γραμικά πολυεπίπεδα μοντέλα και εκτιμήσεις τους*

### *4.2.1: Μοντέλο 2-επιπέδων και βασικοί συμβολισμοί*

Αρχικά, αναφερόμαστε σε κάποιους συμβολισμούς που θα χρησιμοποιηθούν παρακάτω

Ιεραρχικό γραμμικό μοντέλο*:*

$$i: υποδεικνύει\ το\ επίπεδο\ μία\ μονάδας(ατομικό)$$

$$j: υποδεικνύει\ το\ επίπεδο\ δύο\ ή\ περισσότερων\ μονάδων(ομάδα)$$

Μεταβλητές για μεμονωμένα i στην ομάδα j.

$$Y_{ij}: εξαρτημένη\ μεταβλητή$$

$$x_{ij}: επεξηγηματική\ μεταβλητή\ σε\ ένα\ επίπεδο$$

$$z_j: επεξηγηματική\ μεταβλητή\ σε\ επίπεδο\ δύο\ n_j, όπου\ n_j: μέγεθος\ ομάδας$$

Γίνεται εισαγωγή του μοντέλου 2-επιπέδων σε συνδυασμό με το βασικό συμβολισμό που χρησιμοποιείται. Εξετάζονται εναλλακτικοί τρόποι για την δημιουργία και την αξιολόγηση του μοντέλο καθώς και τον καθορισμό διαδικασιών για την εκτίμηση των παραμέτρων, που αποτελούν και την <<φόρμα>> για τον έλεγχο υποθέσεων και την κατασκευή διαστημάτων εμπιστοσύνης.

Έστω, ένα παράδειγμα από ένα σύνολο δεδομένων το οποίο αποτελείται από 728 μαθητές σε 50 δημοτικά σχολεία στο κέντρο του Λονδίνου.

Διακρίνουμε δύο περιπτώσεις:

Η 1η είναι, όταν οι μαθητές είναι στο 4ο έτος της εκπαίδευσής τους, δηλαδή όγδοο έτος της ηλικίας τους , και τρία χρόνια αργότερα στο τελευταίο έτος του δημοτικού σχολείου.

Έστω ένα απλό μοντέλο για ένα σχολείο που αφορά τις βαθμολογίες παιδιών έντεκα ετών με τις βαθμολογίες παιδιών οκτώ ετών.

$$\boldsymbol{y_i = a + \beta x_i + e_i} \quad \textbf{(4.2.1)}$$

όπου,

$\alpha$: $\sigma\eta\mu\epsilon\acute{\iota}o\ \tau o\mu\acute{\eta}\varsigma$
$\beta$: $\kappa\lambda\acute{\iota}\sigma\eta$
$e_i$: $\kappa\alpha\tau\acute{\alpha}\lambda o\iota\pi\alpha$

Αν θέλουμε τώρα, να περιγράψουμε ταυτόχρονα τη σχέση για πολλά σχολεία ταυτόχρονα θα έχουμε, για σχολείο $j$:

$$\boldsymbol{y_{ij} = a_j + \beta_j x_{ij} + e_{ij}}\textbf{(4.2.1*)}$$

όπου $j$: $\alpha\nu\alpha\phi\acute{\epsilon}\rho\epsilon\tau\alpha\iota\ \sigma\tau\eta\nu\ \mu o\nu\acute{\alpha}\delta\alpha\ \epsilon\pi\iota\pi\acute{\epsilon}\delta o\upsilon - 2$

και$i$: $\alpha\nu\alpha\phi\acute{\epsilon}\rho\epsilon\tau\alpha\iota\ \sigma\tau\eta\nu\ o\mu\acute{\alpha}\delta\alpha\ \epsilon\pi\iota\pi\acute{\epsilon}\delta o\upsilon - 1$

Πέρα από το παράδειγμα που αναφέραμε πιο πάνω, σε ένα πραγματικό μοντέλο 2-επιπέδων τα $a_j, \beta_j$γίνονται τυχαίες μεταβλητές. Στην συνέχεια αντικαθιστούμε το $a_j$με $\beta_{0j}$και το $\beta_j$ με $\beta_{1j}$και υποθέτουμε ότι:

$$\boldsymbol{\beta_{0j} = \beta_0 + U_{0j}}\textbf{(4.2.1**)}\text{και}\boldsymbol{\beta_{1j} = \beta_1 + U_{1j}}\ \textbf{(4.2.1***)}$$

Όπου,

$\beta_{0j}, \beta_{1j}$: είναι σταθεροί παράμετροι

$U_{0j}, U_{1j}$: είναι οι τυχαίες μεταβλητές με παραμέτρους

$$E(U_{oj}) = E(U_{1j}) = 0$$
$$Var(U_{0j}) = {\sigma_{U0}}^2\ , Var(U_{1j}) = {\sigma_{U1}}^2, cov(U_{0j}, U_{1j}) = \sigma_{U01}$$

Οπότε, με βάση τις εξισώσεις (4.2.1*),(4.2.1**) η (4.2.1) γίνεται:

$$\boldsymbol{y_{ij} = \beta_0 + \beta_1 x_{ij} + (U_{0j} + U_j x_{ij} + e_{0ij})}\textbf{(4.2.1****)}$$

Και $Var(e_{0ij}) = {S_{e0}}^2$.

Έτσι, σύμφωνα με τα παραπάνω η μεταβλητή απόκρισης ($\boldsymbol{y_{ij}}$), εκφράζεται ως το άθροισμα από ένα σταθερό μέρος και ένα τυχαίο τμήμα που υπάρχει μέσα στις παρενθέσεις.

Η σχέση **(4.2.1****)** σε μορφή πίνακα μπορεί να γραφεί ως εξής:

$$\boldsymbol{E(Y) = X\beta}\ \textbf{,με}\ \boldsymbol{Y = \{y_{ij}\}}$$

$$\boldsymbol{E(y_{ij}) = X_{ij}\beta = (X\beta)_{ij}}\ \ , \boldsymbol{X = \{X_{ij}\}}$$

Όπου,

{ }: $\delta\eta\lambda\acute{\omega}\nu o\upsilon\nu\ \pi\acute{\iota}\nu\alpha\kappa\alpha$

$X$: $\delta\eta\lambda\acute{\omega}\nu\epsilon\iota\ \tau o\nu\ \pi\acute{\iota}\nu\alpha\kappa\alpha\ \sigma\chi\epsilon\delta\iota\alpha\sigma\mu o\acute{\upsilon}\ \gamma\iota\alpha\ \tau\iota\varsigma\ \epsilon\pi\epsilon\xi\eta\gamma\eta\mu\alpha\tau\iota\kappa\acute{\epsilon}\varsigma\ \mu\epsilon\tau\alpha\beta\lambda\eta\tau\acute{\epsilon}\varsigma$

$$X_{ij}: είμαι\ η\ σειρά\ του\ ij-οστού\ X$$

Οι τυχαίες μεταβλητές αναφέρονται ως <<*κατάλοιπα*>> και στην περίπτωση ενός μεμονωμένου μοντέλου επιπέδου, το επίπεδο-1 του κατάλοιπου $e_{ij}$, γίνεται το συνηθισμένο γραμμικό μοντέλο της εναπομένουσας διάρκειας. Για να γίνει το μοντέλο συμμετρικό έτσι ώστε κάθε συντελεστής να έχει συνδεδεμένες επεξηγηματικές μεταβλητές, μπορούμε να ορίσουμε περαιτέρω επεξηγηματικές μεταβλητές για το σημείο τομής $\beta_0$ και τα συναφή κατάλοιπα, $U_{0j}$, δηλαδή $X_{0ij}$το οποίο περιλαμβάνει την τιμή 1, για να είναι όμως πιο απλό η μεταβλητή αυτή μπορεί να παραληφθεί.

### *4.2.2: Εκτίμηση παραμέτρων για τις συνιστώσες του μοντέλου διακύμανσης*

Η εξίσωση (4.2.1**) απαιτεί την εκτίμηση των δύο σταθερών συντελεστών $\beta_0, \beta_1$ και τεσσάρων άλλων παραμέτρων ${S_{U0}}^2, {S_{U1}}^2, S_{U01}, {S_{e0}}^2$.Αναφερόμαστε σε τέτοιες διακυμάνσεις και συσχετίσεις με τη μορφή τυχαίων παραμέτρων. Αρχικά λαμβάνεται υπόψη το πιο απλό μοντέλο 2- επιπέδων που περιλαμβάνει μόνο τις τυχαίες παραμέτρους ${S_{U0}}^2, {S_{e0}}^2$, το οποίο ονομάζεται *μοντέλο συνιστωσών διακύμανσης*διότι η διακύμανση της απόκρισης σχετίζεται με τη σταθερή συνιστώσα, ο σταθερός παράγοντας πρόβλεψης είναι:

$\boldsymbol{Var(y_{ij}/\beta_0, \beta_1, x_{ij}) = var(U_0 + e_{ij}) = {S_{U0}}^2 + {S_{e0}}^2}$, δηλαδή είναι το άθροισμα της διακύμανσης του επιπέδου-1 και της διακύμανσης του επιπέδου-2.Για το παράδειγμα που ανέφερα και παραπάνω με τους μαθητές και τα σχολεία, αν το μοντέλο συνεπάγεται ότι η συνολική διακύμανση για κάθε μαθητή είναι σταθερή τότε η συνδιακύμανση μεταξύ δύο μαθητών στο ίδιο σχολείο είναι:

$$\boldsymbol{cov(U_{0j} + e_{0i_1j}, U_{0j} + e_{0i_2j}) = cov(U_{0j}, U_{0j}) = {S_{U0}}^2}$$

δεδομένου ότι τα κατάλοιπα επιπέδου-1 υποτίθεται ότι είναι ανεξάρτητα.

Τότε η *συσχέτιση* μεταξύ δύο τέτοιων είναι:

$$\boldsymbol{r = \frac{{S_{U0}}^2}{({S_{U0}}^2 + {S_{e0}}^2)}}$$

η οποία αναφέρεται ως <<ενδό-επίπεδο-2 μονάδων συσχέτισης>>, και στην περίπτωση εδώ <<ενδοσχολική συσχέτιση>>. Αυτή η συσχέτιση μετρά το ποσοστό της συνδιακύμανσης που βρίσκεται μεταξύ των σχολείων.

Για παράδειγμα, σε ένα μοντέλο με 3 επίπεδα, τα οποία είναι σχολεία, τάξεις και μαθητές, έχουμε δύο τέτοιες συσχετίσεις. Η ενδοσχολική συσχέτιση μετρώντας το ποσοστό της διακύμανσης που είναι μεταξύ των σχολείων και τη συσχέτιση της ενδο-τάξης μετρώντας τη μεταξύ τους σχέση.

*Μία από τις μεθόδους εκτίμησης είναι εκείνη των συνηθών ελαχίστων τετραγώνων (OLS).* Στην πολλαπλή παλινδρόμηση είναι ανεφάρμοστα καθώς η εφαρμογή τους οδηγεί σε ανακριβή συμπεράσματα.

Έστω η δομή, ενός συνόλου δεδομένων σε 2–επίπεδα. Εστιάζοντας στην δομή συνδιασποράς έχουμε:

$$\begin{pmatrix} {\sigma_{U0}}^2 + {\sigma_{e0}}^2 & {\sigma_{U0}}^2 & {\sigma_{U0}}^2 \\ {\sigma_{U0}}^2 & {\sigma_{U0}}^2 + {\sigma_{e0}}^2 & {\sigma_{U0}}^2 \\ {\sigma_{U0}}^2 & {\sigma_{U0}}^2 & {\sigma_{U0}}^2 + {\sigma_{e0}}^2 \end{pmatrix}$$

Είναι ένας πίνακας συνδιασπορών για τις βαθμολογίες των φοιτητών σε ένα ενιαίο σχολείο. Για δύο σχολεία, ένα με 3 μαθητές και ένα με 2 μαθητές ο συνολικός πίνακας συνδιασποράς θα ήταν: $\begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}$

Όπου,

$$A = \begin{pmatrix} {\sigma_{U0}}^2 + {\sigma_{e0}}^2 & {\sigma_{U0}}^2 & {\sigma_{U0}}^2 \\ {\sigma_{U0}}^2 & {\sigma_{U0}}^2 + {\sigma_{e0}}^2 & {\sigma_{U0}}^2 \\ {\sigma_{U0}}^2 & {\sigma_{U0}}^2 & {\sigma_{U0}}^2 + {\sigma_{e0}}^2 \end{pmatrix}, B = \begin{pmatrix} {\sigma_{U0}}^2 + {\sigma_{e0}}^2 & {\sigma_{U0}}^2 \\ {\sigma_{U0}}^2 & {\sigma_{U0}}^2 + {\sigma_{e0}}^2 \end{pmatrix}$$

Το παραπάνω θεωρείται μία δομή <<μπλοκ-διαγώνια>>, η οποία αντανακλά το γεγονός ότι η συνδιακύμανση μεταξύ των μαθητών σε διάφορα σχολεία είναι μηδέν και σαφώς επεκτείνεται σε οποιδήποτε αριθμό του επιπέδου 2 μονάδων .

Ένας πιο μικρός τρόπος παρουσίασης του παραπάνω πίνακα είναι:

$$V_2 = \begin{pmatrix} {S_{U0}}^2 J_{(3)} + {S_{e0}}^2 I_{(3)} & 0 \\ 0 & {S_{U0}}^2 J_{(2)} + {S_{e0}}^2 I_{(2)} \end{pmatrix}$$

Όπου $I_{(n)}$:είναι ένας $(n \times n)$πίνακασταυτότητας και ο $J_{(n)}$:είναι ένας $(n \times n)$πίνακας αυτών. Ο δείκτης 2 στο $V$δείχνει ένα μοντέλο 2-επιπέδων. Σε OLSμοντέλα ενιαίου επιπέδου το ${S_{u0}}^2$είναι μηδέν και ο πίνακας συνδιακύμανσης στη συνέχεια μειώνει τη τυποποιημένη μορφή $S^2 I$, όπου $S^2$ είναι η υπολλειματική διακύμανση.

### *4.2.3: Γενικό 2-επίπεδο μοντέλο συμπεριλαμβανομένων των τυχαίων συντελεστών*

Η εξίσωση (5) συμπεριλαμβάνοντας περαιτέρω σταθερές επεξηγηματικές μεταβλητές είναι:

$$\boldsymbol{y_{ij} = \beta_0 + \beta_1 x_{ij} + \sum_{h=2}^{p} \beta_h x_{hij} + (u_{oj} + u_{1j} x_{1ij} + e_{0ij})} \quad \textbf{(4.2.3*)}$$

Και πιο συγκεντρωτικά,

$$\boldsymbol{y_{ij} = x_{ij}\beta + \sum_{h=0}^{1} U_{hj} Z_{nij} + e_{0ij} Z_{0ij}} \textbf{(4.2.3**)}$$

Όπου χρησιμοποιούνται νέες επεξηγηματικές μεταβλητές για τυχαίο μέρος του μοντέλου και γράφονται γενικότερα ως:

$$Z = \{Z_0 Z_1\}$$

$$Z_0 = \{1\}$$

$$Z_1 = \{X_{1ij}\}$$

Οι επεξηγηματικές μεταβλητές για το τυχαίο μέρος του μοντέλου είναι συχνά ένα υποσύνολο εκείνων στο σταθερό τμήμα. Επίσης κάποιες απο τις επεξηγηματικές μεταβλητές μπορεί να μετρηθούν σε οποιοδήποτε από τα επίπεδα , για παράδειγμα μπορεί να έχουμε:

Τα επίπεδα των μαθητών →επίπεδο 1

Τα επίπεδα του σχολείου→επίπεδο 2

Ο πίνακας$\Omega_2$ είναι ο πίνακας συνδιακύμανσης της τυχαίας τομής και κλίσης στο επίπεδο_2. Θα πρέπει να πραγματοποιηθεί προσεκτική διάκριση μεταξύ του πίνακα συνδιακύμανσης των απαντήσεων και του πίνακα συνδιακύμανσης των τυχαίων συντελεστών που δίνονται παρακάτω. Ο πίνακας $\Omega_1$ είναι ο πίνακας συνδιακύμανσης για το σύνολο του επιπέδου_1 τυχαίων συντελεστών, σε αυτή την περίπτωση υπάρχει μόνο ένας όρος διακύμανσης στο επίπεδο 1. Ακόμη μπορεί να τονιστεί ότι $\Omega = \{\Omega_i\}$ είναι το σύνολο των πινάκων συνδιακύμανσης.

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix}$$

$$A = ({\sigma_{u0}}^2 + 2\sigma_{u01}x_{1j} + {\sigma_{u1}}^2{x_{1j}}^2 + {\sigma_{e0}}^2)$$

$$B = ({\sigma_{u0}}^2 + \sigma_{u01}(x_{1j} + x_{2j}) + {\sigma_{u1}}^2 x_{1j}x_{2j}$$

$$C = ({\sigma_{u0}}^2 + 2\sigma_{u01}x_{2j} + {\sigma_{u1}}^2{x_{2j}}^2 + {\sigma_{e0}}^2)$$

Δίνοντας,

$$\begin{pmatrix} A & B \\ B & C \end{pmatrix} = X_j \Omega_2 {X_j}^2 + \begin{pmatrix} \Omega_1 & \\ & \Omega_1 \end{pmatrix}$$

$$X_j = \begin{pmatrix} 1 & x_{1j} \\ 1 & x_{2j} \end{pmatrix}, \Omega_2 = \begin{pmatrix} {\sigma_{u0}}^2 & \sigma_{u01} \\ \sigma_{u01} & {\sigma_{u1}}^2 \end{pmatrix}, \Omega_1 = {\sigma_{e0}}^2$$

### *4.2.4: Εκτίμηση στα πολυεπίπεδα μοντέλα*

*Η πολυεπίπεδη μοντελοποίηση χρησιμοποιεί επί τον πλείστον λειτουργίες ML εκτίμησης (μέγιστης πιθανοφάνειας). Έχουμε τα FML( τα οποία είναι τα σταθερά αποτελέσματα) και RML(τα οποία είναι τα τυχαία αποτελέσματα).*

Επιχειρήματα για την επιλογή μεταξύ σταθερού συντελεστή (F) και τυχαίου συντελεστή (R) μοντέλου:

1. Αν οι ομάδες είναι μοναδικές οντότητες και η εξαγωγή συμπερασμάτων θα πρέπει να επικεντρωθεί τότε σε αυτές τις ομάδες, οπότε επιλέγεται F(επιλέγεται συνήθως για μικρό αριθμό ομάδων).
2. Εάν οι ομάδες θεωρούνται ως δείγμα από ένα πληθυσμό και το συμπέρασμα θα πρέπει να επικεντρωθεί σε αυτό τον πληθυσμό, τότε επιλέγεται R (επιλέγεται συνήθως για μεγάλο αριθμό ομάδων).
3. Εάν το μέγεθος της ομάδας είναι μικρό και υπάρχουν πολλές ομάδες τότε το Rείναι η καλύτερη χρήση.
4. Εάν το επίπεδο των αποτελεσμάτων πρέπει να δοκιμάζεται τότε επιλέγεται R.
5. Εάν ο ερευνητής ενδιαφέρεται μόνο για αποτελέσματα εντός της ομάδας και οι διαφορές μεταξύ των ομάδων είναι πιθανές τότε η επιλογή είναι το F.
6. Εάν οι επιπτώσεις στην ομάδα $(U_{0j})$σχεδόν δεν ακολουθούν κανονική κατανομή τότε το Rείναι επικύνδινο.

Πέρα, όμως από την εκτίμηση μέγιστης πιθανοφάνειας για τους συντελεστές παλινδρόμησης χρησιμοποιούνται επίσης απλά ελάχιστα τετράγωνα (OLS) όπως και τα γενικευμένα ελάχιστα τετράγωνα (GLS), η εφαρμογή όμως των (OLS) είναι λιγότερο αποτελεσματική. Παρακάτω δίνεται μία γενική εικόνα της μεθόδου των επαναληπτικών γενικευμένων ελαχίστων τετραγώνων.

Έστω ένα απλό 2-επίπεδο συνιστωσών του μοντέλου διακύμανσης, το οποίο είναι:

$$\boldsymbol{y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{0j} + e_{0ij}}$$

Έστω, ότι γνωρίζουμε τις τιμές των διακυμάνσεων και έτσι θα μπορέσουμε να κατασκευάσουμε αμέσως το <<μπλοκ-διαγώνια>> του πίνακα $V_2$, ο οποίος αναφέρεται απλά ως $V$. Έπειτα, μπορεί να εφαρμοστεί η διαδικασία εκτιμήσεων των γενεικευμένων ελάχίστων τετραγώνων (GLS) για την απόκτηση του εκτιμητή για τους σταθερούς συντελεστές.

$$\boldsymbol{\widehat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y)}$$

όπου,

$$X = \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n_m m} \end{pmatrix}, Y = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_m m} \end{pmatrix}$$

Με $m$: $επίπεδο\ 2\ μονάδων$

$$n_j: επίπεδο\ 1\ μονάδας\ στην\ j - οστού\ επιπέδου\ 2\ μονάδων$$

Όταν τα κατάλοιπα έχουν κανονική κατανομή $Y$ δίνονται επίσης εκτιμήσεις μέγιστης πιθανοφάνειας.Η διαδικασία εκτίμησηςείναι επαναληπτική. Συνήθως, όμως ξεκινάμε από <<λογικές>> τις εκτιμήσεις των σταθερών παραμέτρων.

Υποθέτοντας ότι ${\sigma_{u0}}^2 = 0$ δίνουμε τις εκτιμήσεις OLSτων σταθερών συντελεστών $\hat{\beta}_0$.Έτσι σχηματίζονται τα πρώτα κατάλοιπα:

$$\tilde{\boldsymbol{y}}_{\boldsymbol{ij}} = \boldsymbol{y}_{\boldsymbol{ij}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x}_{\boldsymbol{ij}}$$

Ο φορέας των πρώτων καταλοίπων είναι: $\tilde{Y} = \{\tilde{y}_{ij}\}$.

Αν σχηματιστεί το προιόν του πίνακα $\tilde{Y}\tilde{Y}^T$, η αναμενόμενη τιμή του είναι απλά $V$. Αν γίνει αλλαγή του πίνακα ως προς τις στήλες, θα είναι γραμμένο ως $vec(\tilde{Y}\tilde{Y}^T)$και ομοίως μπορεί να γραφεί το διάνυσμα ως $vec(V)$.

*4.2.4.1: Άλλες εκτιμήσεις παραμέτρων*

Ο Longford(1987) ανέπτυξε μία διαδικασία που βασίζεται σε έναν αλγόριθμο <<Fisherscoring>>και ο Raudenbush (1994) έδειξε ότι επίσημα ισοδυναμεί με την διαδικασία IGLS.

Μια παραλλαγή των IGLSαναμένεται να είναι η πρόβλεψη γενικευμένων ελαχίστων τετραγώνων (EGLS), κάτι το οποίο επικεντρώνει το ενδιαφέρον για τις σταθερές παραμέτρους και χρησιμοποιεί την εκτίμηση του $V$που λαμβάνεται μετά την πρώτη επανάληψη, απλά πρέπει να ληφθεί μία συνεπής εκτιμήτρια των σταθερών συντελεστών χωρίς περαιτέρω επαναλήψεις.

*Μία διαφορετική προσέγγιση* είναι γενικά οι επεκτάσεις της Bayesianεκτίμησης, ως Μπευζιανό γραμμικό μοντέλο όπου $\beta_j$υποτίθεται ότι είναι ανταλλάξιμα και έχουν πορηγοούμενη κατανομή με διακύμανση ${S_{u0}}^2$.Η πλήρης εκτίμηση Bayesαπαιτεί την προηγούμενη κατανομή των τυχαίων συντελεστών καθώς και τις διακυμάνσεις επιπέδου 1,2. Μία άλλη εναλλακτική για την πλήρη εκτίμηση του Bayesείναι η <<Εμπειρική Bayes>>, η οποία αγνοεί τις προηγούμενες κατανομές των τυχαίων παραμέτρων και χρησιμοποιείται συνήθως για σκοπούς εξαγωγής συμπερασμάτων. Όταν υποτίθεται η κανονικότητα οι εκτιμήσεις αυτές είναι ίδιες όπως, IGLSή RIGLIS.

*Μία άλλη προσέγγιση*η οποία είναι παράλληλη με όλα αυτά είναι εκείνη των γενικευμένων εξισώσεων εκτίμησης (GEE) που εισήγαγε ο Liangκαι Zeger (1986). Η κύρια διαφορά είναι ότι οι εκτιμήσεις GEE αποκτάνε την εκτίμηση του $V$χρησιμοποιπώντας απλή παλινδρόμηση, κάτι το οποίο αφορά κατά κύριο λόγο την μοντελοποίηση των σταθερών συντελεστών και όχι την εξερεύνηση της δομής της τυχαίας συνιστώσας του μοντέλου. Όμως και εκείνη δεν είναι πλήρως αποτελεσματική παρόλο που οι προκύπτουσες εκτιμήσεις συντελεστών είναι συνεπείς. Κάποιες φορές όμως η GEEείναι προτιμότερη επειδή είναι πιο γρήγορη .

Επίσης, η GEEμπορεί να επεκταθεί και να χειριστεί αρκετά μοντέλα, που θα αναφερθούν και παρακάτω.

Πρόσφατα, η πλήρης Bayesianεκτίμηση έχει αναπτυχθεί από τους 'MarkovChainMonteCarlo' μεθόδους (MCMC) ιδιαίτερα με Gibbsδειγματοληψία. Το πλεονέκτημα αυτού είναι σε μικρά δείγματα, ώστε να λαμβάνει υπόψη την αβεβαιότητα που σχετίζεται με τις εκτιμήσεις των τυχαίων παραμέτρων και μπορεί να παρέχει ακριβή μέτρα αβεβαιότητας.

### *4.2.5: Κατάλοιπα*

Σε ένα ενιαίο επίπεδο μοντέλο όπως εκείνο της σχέσης (4.2.1), η συνθήκη εκτίμησης του κατάλοιπου $e_i$είναι απλά $\tilde{y}_i$, το πρώτο κατάλοιπο. Σε ένα πολυεπίπεδο μοντέλο όμως, θα έχουμε αρκετά κατάλοιπα σε διαφορετικά επίπεδα. Στη συνέχεια γίνεται εκτίμηση των επιμέρους καταλοίπων.

Λαμβάνοντας, όμως υπόψην τις εκτιμήσεις των παραμέτρων, εξετάζεται η πρόβλεψη ενός συγκεκριμένου καταλοίπου, δηλαδή, $u_{0j}$σε 2-επίπεδα συνιστωσών του μοντέλου διακύμανσης. Συγκεκριμένα, για κάθε μονάδα επιπέδου 2 χρειαζόμαστε:

$$\widehat{\boldsymbol{u}}_{\boldsymbol{0j}} = \boldsymbol{E}(\boldsymbol{u}_{\boldsymbol{0j}}|\boldsymbol{Y}, \widehat{\boldsymbol{b}}, \widehat{\boldsymbol{\Omega}})$$

Αγνοώντας τη διακύμανση της δειγματοληψίας που συνδέεται με τις εκτιμήσεις των παραμέτρων στην παραπάνω εξίσωση, έχουμε:

$$\left.\begin{array}{c} \boldsymbol{cov}(\tilde{\boldsymbol{y}}_{\boldsymbol{ij}}, \boldsymbol{u}_{\boldsymbol{0j}}) = \boldsymbol{var}(\boldsymbol{u}_{\boldsymbol{0j}}) = \boldsymbol{S_{u0}}^{\boldsymbol{2}} \\ \boldsymbol{cov}(\tilde{\boldsymbol{y}}_{\boldsymbol{ij}}, \boldsymbol{e}_{\boldsymbol{0ij}}) = \boldsymbol{S_{e0}}^{\boldsymbol{2}} \\ \boldsymbol{var}(\tilde{\boldsymbol{y}}_{\boldsymbol{ij}}) = \boldsymbol{S_{u0}}^{\boldsymbol{2}} + \boldsymbol{S_{e0}}^{\boldsymbol{2}} \end{array}\right\}$$

Θεωρούμε την εξίσωση **(4.2.3*)** ως γραμμική παλινδρόμηση του $u_{0j}$στο σύνολο της $\{\tilde{y}_{ij}\}$για την j-οστή μονάδα επιπέδου 2 και οι εξισώσεις **(4.2.3**)** ορίζουν τα απαιτούμενα για τις ποσότητες των συντελεστών παλινδρόμησης και ως εκ τούτου την $\hat{u}_{0j}$.

$$\left.\begin{array}{c} \widehat{\boldsymbol{u}}_{\boldsymbol{0j}} = \dfrac{\boldsymbol{n_j S_u}^{\boldsymbol{2}}}{(\boldsymbol{n_j S_u}^{\boldsymbol{2}} + \boldsymbol{S_{e0}}^{\boldsymbol{2}})} \tilde{\boldsymbol{y}}_{\boldsymbol{j}} \\ \tilde{\boldsymbol{e}}_{\boldsymbol{0ij}} = \tilde{\boldsymbol{y}}_{\boldsymbol{ij}} - \widehat{\boldsymbol{u}}_{\boldsymbol{0j}} \\ \tilde{\boldsymbol{y}}_{\boldsymbol{j}} = (\sum_{\boldsymbol{i}} \tilde{\boldsymbol{y}}_{\boldsymbol{ij}})/\boldsymbol{n_j} \end{array}\right\}$$

Όπου, $n_j$: $είναι οαριθμός των μονάδων στο επίπεδο\_1$.

Οι εκτιμήσεις των καταλοίπων δεν είναι αμερόληπτες αλλά συνεπείς. Ο συντελεστής πολλαπλασιασμού του μέσου των πρώτων καταλοίπων ($\tilde{y}_j$) αναφέρεται ως <<παράγοντας συρρίκνωσης>> δεδομένου ότι πάντοτε είναι μικρότερος ή ίσος της μονάδας. Καθώς αυξάνεται ο $n_j$, ο παράγοντας αυτός τείνει προς την μονάδα και καθώς ο αριθμός των μονάδων στο επίπεδο_1 σε μία μονάδα του επιπέδου_2 μειώνει τον <<εκτιμητή συρρίκνωσης>>, το $u_{0j}$ τείνει πιο κοντά στο μηδέν. Σε πολλές εφαρμογές τα κατάλοιπα υψηλότερου επιπέδου παρουσιάζουν ενδιαφέρον από μόνα τους και η αυξημένη συρρίκωνση για μια μικρή μονάδα επιπέδου_2 μπορεί να θεωρηθεί ως εκφράζουσα τη σχετική έλλειψη πληροφοριών στη μονάδα, έτσι ώστε η βέλτιστη εκτίμηση τοποθετεί τα προβλεπόμενα κατάλοιπα κοντά στην συνολική αξία πληθυσμού.

Αυτά τα κατάλοιπα λοιπόν μπορεί να έχουν δύο ερμηνείες:

- Μία ερμηνεία τους είναι ως τυχαίες μταβλητές, με κατανομή των οποίων οι τιμές των παραμέτρων μας λένε σχετικά με την διακύμανση 2 μονάδων, η οποία παρέχει επαρκής εκτίμηση για σταθερούς συντελεστές.
- Μία δεύτερη ερμηνεία είναι ως μεμονωμένες εκτιμήσεις για κάθε ομάδα επιπέδου_2, όπου υποθέτουμε ότι ανήκουν σε έναν πληθυσμό μονάδων για να προβλέψουμε τις τιμές τους.

Όπως και σε μοντέλα ενιαίου επιπέδου, έτσι και εδώ μπορούμε να χρησιμοποιήσουμε τα εκτιμώμενα κατάλοιπα για να τον έλεγχο των παραδοχών του μοντέλου. *Οι δύο συγκεκριμένες υποθέσεις που μπορούν να μελετηθούν άμεσα είναι η υπόθεση της ομαλότητας και ότι οι διακυμάνσεις είναι σταθερές*. Επειδή, οι διακυμάνσεις των εκτιμήσεων των καταλοίπων εξαρτώνται σε γενικές γραμμές από τις τιμές των σταθερών συντελεστών είναι σύνηθες να τυποποιήσουμε τα κατάλοιπα διαιρώντας με τα κατάλληλα τυπικά σφάλματα.

Όταν τα κατάλοιπα, σε υψηλότερα επίπεδα παρουσιάζουν ενδιαφέρον από μόνα τους, πρέπει να είναι σε θέση να παρέχουν το διάστημα εμπιστοσύνης και το τεστ σημαντικότητας καθώς και τις σημειακές εκτιμήσεις για αυτά ή τις λειτουργίες τους. Έτσι, τα εκτιμώμενα κατάλοιπα θα απαιτούν τις εκτιμήσεις των τυπικών σφαλμάτων. Τα κατάλοιπα επιπέδου_1 δεν παρουσιάζουν γενικά ενδιαφέρον από μόνα τους αλλά χρησιμοποιούνται για το μοντέλο ελέγχου, αφού πρώτα έχουν τυποποιηθεί με την χρήση των διαγνωστικών τυπικών σφαλμάτων.

### *4.2.6: Έλεγχοι υποθέσεων και διαστήματα εμπιστοσύνης*

Οι έλεγχοι υποθέσεων χρησιμοποιούνται δεδομένου ότι έχουν τη συνηθισμένη μορφή της μηδενικής υπόθεσης, ότι η τιμή της παραμέτρου ή μια συνάρτηση των τιμών των παραμέτρων είναι μηδέν. Ακομη, για αρκετά μεγάλα δείγματα μία μηδενική υπόθεση θα είναι σχεδόν βέβαιο ότι θα απορριφθεί. Αυτό που μας ενδιαφέρει όμως είναι αν η διαφορά είναι θετική ή αρνητική. Τα διαστήματα εμπιστοσύνης από την άλλη

τονίζουν την αβεβαιότητα που χαρακτηρίζει τις εκτιμήσεις των παραμέτρων και τη σημασία της ουσιαστικής σημασίας τους.

*4.2.6.1: Σταθεροί παράμετροι*

Οι εκτιμήσεις σταθερών παραμέτρων μαζί με τα τυπικά σφάλματα τους είναι κατάλληλα για τον έλεγχο υποθέσεων ή για διαστήματα εμπιστοτύνης που κατασκευάζονται ξεχωριστά για κάθε παράμετρο. Για τον έλεγχο υποθέσεων, που οφείλεται τις περισσότερες φορές σε ομαδοποιημένες ή κατηγοριοποιημένες επεξηγηματικές μεταβλητές, όπου οι επιδράσεις $n$ομάδων ορίζονται με βάση $n-1$ψευδομεταβλητές αντιθέσεις, και θέλουμε να ελέγξουμε αν αυτές οι αντιθέσεις είναι μηδέν. Μπορεί επίσης να μας ενδιαφέρει ένα ζεύγος από διαστήματα εμπιστοτύνης για τις εκτιμήσεις των παραμέτρων.

Καθορισμός ενός $(r \times p)$πίνακα αντίθεσης $C$.

Αυτό χρησιμοποιείται για το συνδυασμό γραμμικών ανεξάρτητων συναρτήσεων που $p$ −σταθεροί παράμετροι στο μοντέλο της μορφής:

$$\boldsymbol{f = C\beta}$$

έτσι ώστε, κάθε σειρά $C$ ορίζει μία ειδική γραμμική συνάρτηση. Παράμετροι οι οποίοι δεν εμπλέκονται έχουν τα αντίστοιχα στοιχεία μηδέν.

Έστω, ότι υποθέτουμε σαν συντελεστές την κοινωνική τάξη και το φύλο, οι οποίες είναι από κοινού μηδέν.

Ορίζουμε:

$$\boldsymbol{C} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{10} \\ \mathbf{0} & \mathbf{0} & \mathbf{01} \end{pmatrix} \ , \boldsymbol{f} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix}$$

Και η γενική μηδενική υπόθεση είναι : $H_0: f = k\ , k = \{0\}$.

Ορίζουμε:

$$\boldsymbol{R} = \left(\hat{\boldsymbol{f}} - \boldsymbol{k}\right)^{\boldsymbol{T}} \left[\boldsymbol{C}\left(\boldsymbol{X^T \hat{V}^{-1} X}\right)^{-1} \boldsymbol{C^T}\right]^{-1} \left(\hat{\boldsymbol{f}} - \boldsymbol{k}\right)$$

$$\hat{\boldsymbol{f}} = \boldsymbol{C}\hat{\boldsymbol{\beta}}$$

Αν η μηδενική υπόθεση $(H_0)$ αληθεύει τότε αυτό κατανέμεται προσεγγιστικά $C^2$με $r$ −βαθμούς ελευθερίας. Ο όρος $\left(X^T \hat{V}^{-1} X\right)^{-1}$ στην παραπάνω σχέση είναι ο εκτιμώμενος πίνακας συνδιακύμανσης των σταθερών συντελεστών. Αν βρούμε ένα στατιστικά σημαντικό αποτέλεσμα που μπορεί να επιθυμείτε να διερευνηθεί, ουσιαστικά ποιοι συγκεκριμένοι γραμμικοί συνδυασμοί των συντελεστών που συμμετέχουν διαφέρουν σημαντικά από το μηδέν. Το πιο απλό παράδειγμα θα μπορούσε να είναι όταν διαπιστώνουμε ότι $n - ομάδες$ διαφέρουν και θέλουμε να πραγματοποιηθούν όλες οι πιθανές συγκρίσεις ανά ζεύγη. Μία διαδικασία

ταυτόχρονης σύγκρισης που διατηρεί το συνολικό σφάλμα τύπου I, στο προβλεπόμενο επίπεδο περιλαμβάνει τη διεξαγωγή της ανωτέρω διαδικασίας είτε με ένα υποσύνολο των γραμμών της $C$ή ένα σύνολο γραμμικών ανεξάρτητων αντιθέσεων. Η τιμή του $R$που λαμβάνεται στη συνέχεια κρίνεται κατά τις κρίσιμες τιμές της $x^2$κατανομής με $r - \beta\alpha\theta\mu o\acute{\upsilon}\varsigma\ \varepsilon\lambda\varepsilon\upsilon\theta\varepsilon\rho\acute{\iota}\alpha\varsigma.$

$$\widehat{\boldsymbol{R}} = \left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)^{\boldsymbol{T}} \left[\boldsymbol{C}\left(\boldsymbol{X}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{C}^{\boldsymbol{T}}\right]^{-\mathbf{1}} \left(\boldsymbol{f} - \hat{\boldsymbol{f}}\right)$$

Αυτό δίνει μία τετραγωνική συνάρτηση των εκτιμώμενων συντελεστών, δίνοντας μια τρισδιάστατη r-ελλειψοειδή περιοχή.

Η $H_0$δίνει μία τιμή για την $x^2$ με 2 βαθμούς ελευθερίας και αντίστοιχη p-value=0.10.

Σε ορισμένες περιπτώσεις μπορεί να μας ενδιαφέρει σε ξεχωριστά διαστήματα εμπιστοσύνης για όλες τις πιθανές γραμμικές συναρτήσεις που αφορούν ένα υποσύνολο qπαραμέτρων ή qγραμμικά ανεξάρτητες συναρτήσεις των παραμέτρων, διατηρώντας παράλληλα μια σταθερή πιθανότητα ότι όλα τα διαστήματα περιλαμβάνουν τη πληθυσμιακή τιμή αυτών των συναρτήσεων των παραμέτρων.

Για ένα 100(1-α)% διάστημα εμπιστοσύνης $C_i$για την i-οστή σειρά του C, τότε ταυτόχρονα το (1-α)% διάστημα εμπιστοσύνης για $C_i\beta$ για όλα τα $C_i$δίνεται από:

$$\left(\boldsymbol{C_i}\widehat{\boldsymbol{\beta}} - \boldsymbol{d_i}, \boldsymbol{C_i}\widehat{\boldsymbol{\beta}} + \boldsymbol{d_i}\right)$$

Όπου, $\boldsymbol{d_i} = \left[\boldsymbol{C_i}\left(\boldsymbol{X}^{\boldsymbol{T}}\widehat{\boldsymbol{V}}^{-\mathbf{1}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{C_i}^{\boldsymbol{T}}\boldsymbol{X}_{\boldsymbol{q},(\boldsymbol{a})}{}^{\mathbf{2}}\right]^{\boldsymbol{a}}$

Όπου, ${X_{q,(a)}}^2$ είναι το σημείο α% της ${X_q}^2$κατανομής.

Ακόμη, μπορούμε να αναφέρουμε ότι για τον έλεγχο υποθέσεων μπορεί να χρησιμοποιηθεί το τεστ λόγου πιθανοφανειών για σταθερές παραμέτρους, αν και τα αποτελέσματα θα είναι παρόμοια. Η διαφορά προκύπτει στο ότι οι τυχαίες εκτιμήσεις που χρησιμοποιήθηκαν στην προηγούμενη σχέση είναι εκείνα που λαμβάνονται για το πλήρη μοντέλο αντί εκείνοι κάτων από την $H_0$, αν η τροποποίηση μπορεί να γίνει εύκολα.

*4.2.6.2: Τυχαίες παράμετροι*

Για πολύ μεγάλα δείγματα είναι δυνατόν να χρησιμοποιηθούν οι ίδιες διαδικασίες για ελέγχους υποθέσεων και διαστήματα εμπιστοσύνης ως προς τις σταθερές παραμέτρους. Για τον έλεγχο της $H_0$έναντι της $H_1$, που αφορούν την τοποθέτηση των πρόσθετων παραμέτρων που αποτελούν τον δείκτη πιθανότητας ή στατιστικής απόκλισης.

$$\boldsymbol{D_{01}} = -\mathbf{2}\boldsymbol{log_e}(\boldsymbol{\lambda_0}|\boldsymbol{\lambda_1}) \quad \textbf{(4.2.6.2)}$$

Όπου, $\lambda_0, \lambda_1$: πιθανότητες για μηδενικές και εναλλακτικές υποθέσεις και αυτό αναφέρεται στους πίνακες της $x^2$κατανομής μς βαθμούς ελευθερίας ίσους με την διαφορά qτου αριθμού των παραμέτρων.

Αν το $D_{01}$ έχει οριστεί στην τιμή του σημαίου α% της $x^2$ κατανομής με qβαθμούς ελευθερίας, τότε μία κατασκευή είναι σχεδιασμένη για το ικανοποιήει χρησιμοποιώντας μία κατάλληλη διαδικασία ανζήτησης.

*4.2.6.3: Κατάλοιπα*

Τα κατάλοιπα κατατάσσονται από το μικρότερο προς το μεγαλύτερο. Κατασκευάζουμε ένα διάστημα για κάθε κατάλοιπο, έτσι ώστε το κριτήριο για να κρίνουμε στατιστική σημαντικότητα σε (1-α)% επίπεδο για κάθε ζευγάρι καταλοίπων είναι κατά το πόσο τα διαστήματα εμπιστοσύνης τους συμπίπτουν.

Η γενική διαδικασία ορίζει ένα σύνολο διαστήματος εμπιστοσύνης για κάθε κατάλοιπο ως:

$$\widehat{\boldsymbol{u}}_{\boldsymbol{i}} \pm \boldsymbol{c}(\boldsymbol{se})_{\boldsymbol{i}}$$

Η τιμή $c$υπολογίζεται έτσι ώστε ο μέσος όρος για όλα τα πιθανά ζεύγη να είναι (1-α)%. Για κάθε δυνατό ζεύγος από διαστήματα υπάρχει ένα επίπεδο σημαντικότητας που σχετίζεται με το κριτήριο επικάλυψης και η τιμή $c$προσδιορίζεται έτσι ώστε ο μέσος όρος για όλα τα πιθανά ζεύγη είναι(1-α)%.

Για το $c$, όταν οι αναλογίες των τυπικών σφαλμάτων δεν διαφέρουν σημαντικά, δηλαδή όχι περισσότερο από 2:1, π.χ. η τιμή 1,4 μπορεί να χρησιμοποιηθεί για το $c$. Όσο η αναλογία αυξάνει το ίδιο κάνει και το $c$.

## *4.3: Επεκτάσεις των πολυεπίπεδων μοντέλων*

### *4.3.1: Πολύπλοκες δομές διακύμανσης*

Θα αναφερθούμε στις επεκτάσεις του βασικού μοντέλου ώστε να συμπεριλάβει τους περιορισμούς σχετικά με τις παραμέτρους, τον συντελεστή διόρθωσης της μονάδας, την τυπική εκτίμηση σφάλματος και αναλύσεις συνολικού επιπέδου.

Ως γενικό πρόβλημα θεωρείται η μοντελοποίηση της διακύμανσης του επιπέδου_1. Πιο πάνω αναφερθήκαμε στο 2-επίπεδο μοντέλο, για το 2-επίπεδο μοντέλο χρησιμοποιούμε το συμβολισμό $u_j$ , $e_{ij}$για την συνολική διακύμανση στα επιπέδα 2 και 1, και γράφουμε:

$$\boldsymbol{u_j = \sum_{h=0}^{r_2} u_{hj} Z_{hj}} \ \ , \boldsymbol{e_{ij} = \sum_{h=0}^{r_1} e_{hij} Z_{hij}} \textbf{(4.3.1)}$$

Όπου το $Z$είναι οι επεξηγηματικές μεταβλητές. Κανονικά $Z_{0j}$ , $Z_{0ij}$αναφέρεται στην σταθερά (=1) που ορίζει ένα βασικό όρο διακύμανσης σε κάθε επίπεδο.

Για μοντέλα 3_πιπέδων θα χρησιμοποιήσουμε τον συμβολισμό $v_\kappa$ , $u_{kj}$ , $e_{ijk}$ , όπου $i$: $αναφέρεται\ σε\ επίπεδο\ 1\ μονάδας$

$$j: αναφέρεται\ στο\ επίπεδο\ 2\ μονάδων$$

$$k: αναφέρεται\ στο\ επίπεδο\ 3\ μονάδων$$

$h$: $ευρύτερα\ οι\ επεξηγηματικές\ μεταβλητές\ και\ οι\ συντελεστές\ τους\ σε\ κάθε\ επίπεδο$

Ένα απλό μοντέλο για την διακύμανση επιπέδου_1 είναι να γίνει μία γραμμική συνάρτηση μιας απλής επεξηγηματικής μεταβλητής. Έχουμε:

$$\boldsymbol{y_{ij} = \beta_0 + \beta_1 x_{ij} + \left(u_j + e_{0ij} + e_{1ij} Z_{ij}\right),\ Z_{ij} = x_{ij}} \textbf{(4.3.1*)}$$

$$\boldsymbol{var\left(e_{0ij}\right) = S^2_{e_0}\ , var\left(e_{1ij}\right) = 0\ , cov\left(e_{0ij} e_{1ij}\right) = S_{e01}}$$

Έτσιώστε η συμβολή επιπέδου_1 στη συνολική διακύμανση να είναι γραμμική συνάρτηση του $Z_{ij}$, τότε έχουμε:

$$\boldsymbol{S^2_{e_0} + 2 S_{e01} Z_{ij}}$$

Αυτή η διάταξη του περιορισμού μιας παραμέτρου διακύμανσης να είναι μηδέν με την παρουσία ενός μια μη-μηδενικής συνδιακύμανσης χρησιμοποιείται για να ληφθεί η απαιτούμενη δομή της διακύμανσης.

Μία παράμετρος διακύμανσης μπορεί να είναι αρνητική εφ'όσον το συνολικό επίπεδο_1 διακύμανσης παραμένει θετικό εντός του εύρους των δεδομένων. Όταν ένας συντελεστής γίνεται τυχαίος σε ένα επίπεδο υψηλότερο από εκείνο στο οπίο ορίζεται η ίδια η επεξηγηματική μεταβλητή, τότε η προκύπτουσα διακύμανση μπορεί να ερμηνευθεί ως η μεταξύ-υψηλότερου επιπέδου μονάδα διακύμανσης της σχέσης εντός-μονάδας που περιγράφεται από το συντελεστή.

Το παραπάνω μοντέλο δεν περιορίζει την συνολική συμβολή επιπέδου_1, με οποιονδήποτε τρόπο. Ειδικότερα, είναι αρκετά πιθανό για διακύμανση επιπέδου_1 και ως εκ τούτου η συνολική διακύμανση να είναι αρνητική, κάτι το οποίο θα οδηγήσει σε προβλήματα της αριθμητικής εκτίμησης. Για την αντιμετώπιση αυτού αφαιρείται το μηδέν της διακύμανσης από το μοντέλο.

### *4.3.2: Μία 3-επίπεδη παραλλαγή του πολύπλοκου μοντέλου*

Η μεταβλητή απόκρισης που θα χρησιμοποιήσουμε είναι μία κλίμακα, στην κλίμακα 0-7, που ασχολείται με τις στάσεις απέναντι στην έκτρωση. Έστω ότι αυτό προέρχεται αθροίζοντας τις (0,1) απαντήσεις σε επτά ερωτήσεις και μπορεί να ερμηνευθεί ως ένδειξη εάν ο ερωτώμενος υποστηρίζει το δικαιώμα της γυναίκας ή όχι στην έκτρωση, με υψηλή βαθμολογία να δείχνει την ισχυρη υποστήριξη. Επεξηγηματικές μεταβλητές είναι η πολιτική υποταγή, αυτο-αξιολόγηση της κοινωνικής τάξης, το φύλο, η ηλικία, η θρησκεία και ο χρόνος. Μία σειρά από προκαταρκτικές αναλύσεις έχουν πραγματοποιηθεί και αποτελέσματα της πολιτικής

υποταγής, της κοινωνικής τάξης, του φύλου και της ηλικίς βρέθηκαν να είναι μικρά και όχι στατιστικά σημαντικά. Με βάση λοιπόν τα παρπάνω ένα βασικό 3-επίπεδο μοντέλο θα έχει την εξής μορφή:

$$\boldsymbol{y_{ijk} = \beta_0 + (\beta_1 x_{1ijk} + \beta_2 x_{2ijk} + \beta_3 x_{3ijk}) + (\beta_4 x_{4ijk} + \beta_5 x_{5ijk} + \beta_6 x_{6ijk}) + (u_k + u_{jk} + e_{ijk})}$$

με τις επεξηγηματικές μεταβλητές με δείκτες 1-3 να είναι ψευδομεταβλητές για της θρησκευτικές κατηγορίες 2-4 και εκείνες με δείκτες 4-6 είναι ψευδομεταβλητές για τα έτη 1984-1986(όπου έχουν παρθεί τα δεδομένα). Έχουμε τρεις τύπους, ένα σε κάθε επίπεδο στο τυχαίο μέρος του μοντέλου.

## *4.4: Περιορισμοί παραμέτρων*

Στο παράδειγμα που αναφέρθηκε παραπάνω σχετικά με τις στάσεις έκτρωσης, μερικές από τις σταθερές και τυχαίες παράμετροι για τα χρόνια και θρησκευτικές ομάδες ήταν παρόμοιες. Κάτι που δηλώνει ότι θα μπορούσε να χωρέσει ένα απλούστερο μοντέλο, όπως για παράδειγμα οι παρέμετροι να λαμβάνουν τις ίδιες αξίες και έτσι να μειώνονται τα τυπικά σφάλματα στο ίδιο μοντέλο. Έτσι απεικονίζεται η διαδικασία με την εκτίμηση του σταθερού μέρους για τα δεδομένα στάσεων σχετικά με τις εκτρώσεις.

Θεωρούμε τον γενικό γραμμικό περιορισμό για τις σταθερές παραμέτρους με μορφή : $Cb = k$, όπου $C$: είναι ένας $(n \times p)$πίνακας περιορισμού και $k$:είναι ένας φορέας που μπορεί να έχει αρκετά γενικές τιμές για τα στοιχεία τους.

Ο περιορισμένος εκτιμητής $\beta$ είναι:

$$\boldsymbol{\widehat{\beta}^c = \widehat{\beta} - LC(C^T LC)^{-1}(C^T \widehat{\beta} - k)} \textbf{(4.4)}$$

$$\boldsymbol{L = (X^T \widehat{V}^{-1} X)^{-1}} \textbf{ (23)}$$

Όπου, $\boldsymbol{\widehat{\beta}}$είναι η μη-περιορισμένη εκτίμηση. Ο πίνακας συνδιακύμανσης είναι:

$$\boldsymbol{M = I - LC(C^T LC)^{-1} C^T} \textbf{(4.4*)}$$

Ανάλογος τύπος υπάρχει για τυχαίες περιορισμένες παραμέτρους. Πέρα από τους γραμμικούς περιορισμούς μπορούμε να εφαρμόσουμε και μη γραμμικούς περιορισμούς. Για παράδειγμα, αν η εκτιμώμενη συσχέτισης μεταξύ της κλίσης και της τομής είναι -1,03, για να την περιορίσουμε ώστε να είναι ακριβώς -1,0 μετά από κάθε επανάληψη του αλγορίθμου υπολογίζουμε τη συνδιακύμανση ως συνάρτηση των διακυμάνσεων για να δώσει τη συσχέτιση αυτή. Έτσι μετά από την επανάληψη $t$υπολογίζουμε τη $\boldsymbol{S_{u01}^{t+1} = \widehat{S}^t{}_{u0} + \widehat{S}^t{}_{u1}}$και μετά περιορίζουμε την συνδιακύμανση ώστε να είναι ίση με την τιμή αυτή, ένα γραμμικό περιορισμό για την επανάληψη $t +$

1. Αυτή η διαδικασία επαναλαμβάνεται εως ότου ληφθεί η σύγκριση των μη-παραμετρικών αξιών.

## *4.5: Στάθμιση μονάδων*

Είναι συχνό σε δειγματολητπικές έρευνες να επιλέγουμε μονάδες επιπέδου_1 , για παράδειγμα για μέλη ενός νοικουριού, έτσι ώστε κάθε μονάδα πληθυσμού να έχει την ίδια πιθανότητα επιλογής. Τέτοια δείγματα αυτό-στάθμισης μπορεί να μοντελοποιηθούν χρησιμοποιώντας οποιαδήποτε από τα πολυεπίπεδα μοντέλα. Όμοια, αν το μοντέλο καθορίζει σωστά τη δομή του πληθυσμού, τα δείγματα μη στάθμισης αυτού μπορεί να μοντελοποιηθούν παρόμοια: οι πιθανότητες διαφορικής επιλογής δεν περιέχουν επιπλέον πληροφορίες για τις παραμέτρους του μοντέλου.

Διακρίνονται δύο είδη περιπτώσεων για τον καθορισμό των συντελεστών:

<u>*Στην πρώτη,*</u> τα βάρη είναι ανεξάρτητα από τις τυχαίες επιδράσεις στο επέπεδο. Στην περίπτωση αυτή έχουμε, έστω ένα μοντέλο 2-επιπέδου, δηλώνουμε όπου $w_j$: το βάρος που συνδέεται με την i-οστή μονάδα επιπέδου-2 και $w_{ij}$:το βάρος που συνδέεται με την i-οστή μονάδα επιπέδου-1 εντός της μονάδας j-οστού επιπέδου-2, έτσι ώστε: $\sum_i \boldsymbol{w_{i|j} = n_j}$ , $\sum_i \boldsymbol{w_j = J}$

Όπου το $J$:είναι ο συνολικός αριθμός των μονάδων επιπέδου-2 και $N = \sum_j n_j$:ο συνολικός αριθμός των μονάδων επιπέδου-1. Δηλαδή, τα χαμηλότερα βάρη του επιπέδου μέσα σε κάθε άμεση μονάδα κλίμακας υψηλότερου επιπέδου για να έχουν το ίδιο υψηλά επίπεδα. Για κάθε μονάδα επιπέδου-1, σήμερα αποτελούν το βάρος:

$$\boldsymbol{w_{ij} = \frac{N_{w_{i|j}w_j}}{\sum_{i,j} w_{i|j}w_j} = N_{w_{i|j}w_j} / \sum_j n_j w_j} \textbf{(4.5)}$$

Συμβολίζουμε με $Z_u, Z_e$ αντίστοιχα τα σύνολα των επεξηγηματικών μεταβλητών που καθορίζουν το επίπεδο-2 και το επίπεδο-1 τυχαίων συντελεστών στη μορφή:

$$\boldsymbol{Z_u^* = W_j Z_u, W_j = diag\{w_j^{-0.5}\}}$$

$$\boldsymbol{Z_e^* = W_{ij} Z_e, W_{ij} = diag\{w_{ij}^{-0.5}\}}$$

Εκτελώντας τώρα, ένα πρότυπο εκτίμησης, αλλά χρησιμοποιώντας $Z_u^*, Z_e^*$ όπως τις επεξηγηματικές μεταβλητές τυχαίων συντελεστών.

Για ένα 3-επίπεδο μοντέλο με προφανή επέκταση στο συμβολισμό, έχουμε τα εξής:

$$\boldsymbol{\sum_i w_{i|jk} = n_{jk}}, \boldsymbol{\sum_j w_{j|k} = J_k}, \boldsymbol{\sum_k w_k = K}, \boldsymbol{N = \sum_{jk} n_{jk}}, \boldsymbol{J = \sum_k J_k} \textbf{(4.5*)}$$

$$\boldsymbol{w_{ijk} = \frac{N w_{i|jk} w_{j|k} w_k}{\sum_{ijk} w_{i|jk} w_{j|k} w_k}}, \boldsymbol{w_{jk} = \frac{J w_{j|k} w_k}{\sum_{jk} w_{j|k} w_k}} \textbf{(4.5**)}$$

Συμβολίζεται με $V^*$ο πίνακας στάθμισης σ'αυτή την ανάλυση. Το σταθερό μέρος των συντελεστών εκτίμησης και του πίνακα συνδιακύμανσης δίνεται από:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T V^{*^{-1}} X})^{-1} \boldsymbol{X^T V^{*^{-1}} Y} \quad \textbf{(4.5***)}$$

$$\boldsymbol{cov}(\widehat{\boldsymbol{\beta}}) = \left(\boldsymbol{X^T V^{*^{-1}} X}\right)^{-1} \boldsymbol{X^T V^{*^{-1}} V V^{*^{-1}} X} (\boldsymbol{X^T V^{*^{-1}} X})^{-1}$$

Με ανάλογο τρόπο και για τις τυχαίες εκτιμήσεις των παραμέτρων.

Στις έρευνες συνήθως έχουν πρόσβαση μόνο στο τελικό επίπεδο_1 βαρών $w_{ij}$. Έστω, ένα 2-επίπεδο μοντέλο, παιρνουμε:

$$\boldsymbol{w'_j = \frac{W_j J}{\sum_j W_j}}, \qquad \boldsymbol{W_j = \frac{\sum_i w_{ij}}{n_j}}$$

Για ένα 3-επίπεδο μοντέλο, η διαδικασία πραγματοποιείται για κάθε μονάδα επιπέδου_3 και η προκύπτουσα $w_{jk}{}'$μεταμορφώνεται αναλόγως.

Χαρακτηριστικά:

1. Για ένα ενιαίο-επίπεδο μοντέλου η παραπάνω διαδικασία δίνει το σύνηθες σταθμισμένο εκτιμητή παλινδρόμησης.
2. Αν έχει δημιουργηθεί ένα ιδιαίτερο βάρος στο επίπεδο_1 στο μηδέν, τότε αυτό είναι ισοδύναμο με την άρση της μονάδας από την ανάλυση σε ένα μοντέλο 2-επιπέδων, από το επίπεδο_2 που παραμένει σταθμισμένη συνεισφορά. Παρόλα αυτά, αυτή η στάθμιση μπορεί να είναι κατάλληλη αν θέλουμε να αφαιρέσουμε το αποτέλεσμα της μονάδας μόνο σε επίπεδο_1, δηλαδή αν ήταν ένα ακραίο επίπεδο_1 ακραίων τιμών. Αν όμως θέσουμε ένα βάρος στο επίπεδο_2 στο μηδέν, τότε αυτό είναι ισοδύναμο με την αφαίρεση ολόκληρης της μονάδας επιπέδου_2. Αν θέλουμε να υπάρξουν εκτιμήσεις που ισοδυναμούν με την αφαίρεση της μονάδας επιπέδου_1 θα πρέπει να ορίζονται επεξηγηματικές μεταβλητές σε όλα τα επίπεδα_2 για την εν λόγω μονάδα επιπέδου στο μηδέν επίσης. Αυτό γίνεται ορίζοντας μία μεταβλητή δείκτη για την μονάδα με ένα μηδενικό που αντιστοιχεί στην εν λόγω μονάδα πολλαπλασιάζοντας όλες τις τυχαίες επεξηγηματικές μεταβλητές από αυτήν. Κατά τον υπολογισμό, μπορούμε επίσης να χρησιμοποιήσουμε τα βάρη. Αυτό οδηγεί στα ακόλουθα αποτελέσματα για το επίπεδο_2 καταλοίπων:

$$\boldsymbol{\widehat{\rho}_2 = \Omega_2 Z_u^{*T} V^{*-1} \widetilde{Y}}$$
$$\boldsymbol{cov(\widehat{\rho}_2) = \Omega_2 Z_u^{*T} V^{*-1}(V) V^{*-1} Z_u^* \Omega_2}$$
$$\boldsymbol{V = E(\widetilde{Y}\widetilde{Y}^T)}$$

Αυτό παρέχει έναν συνεπή εκτιμητή του πίνακα συνδιακύμανσης. Η μη-σταθμισμένη εκτιμήτρια για τα κατάλοιπα είναι επαρκής, οπότε ισχύουν οι συνήθεις τύποι.

## *4.6: Πολυμεταβλητό πολυεπίπεδο μοντέλο*

### *4.6.1: Πολυμεταβλητό πολυεπίπεδο μοντέλο*

Σε αυτό το σημείο γίνεται εξέταση μοντέλων όπου θέλουμε ταυτόχρονα να διαμορφώσουν διάφορες αντιδράσεις, όπως λειτουργίες των επεξηγηματικών μεταβλητών. Έστω ότι έχουμε στοιχεία από αποτελέσματα δύο συνιστωσών της εξέτασης επιστήμης στα έτη 1989-1905, μαθητές από 73 σχολείακαι κολέγια. Η εξέταση είναι το Γενικό Πιστοποιητικό Δευτεροβάθμιας εκπαίδευσης που λαμβάνονται στο τέλος της υποχρεωτικής σχολικής φοίτησης, συνήθως δηλαδή σε ηλικία 16 ετών. Το πρώτο στοιχείο είναι μία γραπτή ερώτηση( βαθμολογείται με σκορ 160) και το δεύτερο αποτελείται από μαθήματα( βαθμολογούνται με συνολικό σκορ 108) σύμφωνα με τα projectπου έχουν παραδωθεί σε όλη την διάρκεια, χαρακτηριζόμενα από τον καθηγητή για κάθε μαθητή.

### *4.6.2: Το βασικό 2 επίπεδο πολυμεταβλητό μοντέλο*

Για να οριστεί μια πολυμεταβλητή, στην περίπτωση του παραδείγματός μας, δύο περιγραφικές, για το μοντέλο αντιμετωπίζουμε τον κάθε μαθητή ως μονάδα επιπέδου_2 και των μετρήσεων <<υπό μαθητή>> ως μονάδα επιπέδου_1. Κάθε μέτρηση επιπέδου_1 έχει μία απάντηση η οποία είναι η γραπτή βαθμολογία σε χαρτί είτε η προφορική βαθμολογία των μαθημάτων. Οι βασικές επεξηγηματικές μεταβλητές είναι ένα σύνολο κινήσεων με μεταβλητές που αποτελούν ένδειξη και της μεταβλητής απόκρισης. Οι περαιτέρω επεξηγηματικές μεταβλητές ορίζονται από τον πολλαπλασιασμό αυτών των ψευδομεταβλητών με επεξηγηματικές μεταβλητές σε ατομικό επίπεδο, όπως για παράδειγμα το φύλο. Ο πίνακας δεδομένων για τρία άτομα, δύο από τα οποία έχουν και τα δύο μετρήσεις και το τρίτο ο οποίος έχει μόνο της γραπτή βαθμολογία δίνονται στο παρακάτω πίνακα. Ο πρώτος και ο τρίτος μαθητής είναι γυναίκες και ο δεύτερος άντρας.

| | | Παρακολουθήσεις | | Γένος | |
|---|---|---|---|---|---|
| Μαθητής | Απάντηση | Γραπτό | Μαθήματα | Γραπτό | Μαθήματα |
| 1(Γυναίκα) | $y_{11}$ | 1 | 0 | 1 | 0 |
| 1 | $y_{21}$ | 0 | 1 | 0 | 1 |
| 2(Άνδρας) | $y_{12}$ | 1 | 0 | 0 | 0 |
| 2 | $y_{22}$ | 0 | 1 | 0 | 0 |
| 3(Γυναίκα) | $y_{13}$ | 1 | 0 | 1 | 0 |

*Πίνακας 1:Πίνακας δεδομένων*

Το μοντέλο γράφεται ως:

$$\boldsymbol{y_{ij} = \beta_{01} z_{1ij} + \beta_{02} z_{2ij} + \beta_{11} z_{1ij} x_j + \beta_{12} z_{2ij} x_j + u_{1j} z_{1ij} + u_{2j} z_{2ij}}$$

$$\boldsymbol{z_{ij} = \begin{cases} 1 & \text{ανείναιγραπτή} \\ 0 & \text{αν είναι μαθήματα} \end{cases}, z_{2ij} = 1 - z_{ij}, x_j = \begin{cases} 1 & \text{ανείναιγυναίκα} \\ 0 & \text{αν είναι άντρας} \end{cases}}$$

$var(u_{1j}) = S_{u1}^2$, $var(u_{2j}) = S_{u2}^2$, $cov(u_{1j}, u_{2j}) = S_{u12}$

Χαρακτηριστικά αυτού του μοντέλου:

- Δεν υπάρχει επίπεδο_1 μεταβολής που ορίζεται επειδή το επίπεδο_1 υφίσταται μόνο και μόνο για να καθορίσει την πολυμεταβλητή δομή.
- Οι διακυμάνσεις και συνδιακυμάνσεις επιπέδου_2 είναι το κατάλοιπο μεταξύ μαθητή-διακυμάνσεων.
- Σε περίπτωση που έχει τοποθετηθεί μόνο το σημείο τομής των ψευδομεταβλητών και δεδομένου ότι κάθε μαθητής έχει δύο αποτελέσματα , το μοντέλο εκτίμησης των παραμέτρων αυτών γίνονται συνήθεις εκτιμήσεις διακυμάνσεων και συνδιακυμάνσεων μεταξύ των μαθητών.
- Οι πολυεπίπεδες εκτιμήσεις είναι στατιστικά αποτελεσματικές ακόμη και όταν λείπουν κάποιες απαντήσεις, και στην περίπτωση που οι μετρήσεις έχουν μία πολυμεταβλητή κανονική κατανομή τότε αποκτά μέγιστη πιθανότητα. Έτσι μπορούμε να πούμε ότι η τυποποίηση ενός 2-επίπεδου μοντέλου επιτρέπει την αποτελεσματική εκτίμηση ενός πίνακα συνδιασποράς με τις ελλιπούσες απαντήσεις.

### *4.6.3: Σχέδια περιστροφής*

Όπως αναφέραμε τα πλήρως ισορροπημένα πολυμεταβλητά σχέδια είναι περιττά και οι τυχαίες ελλειπούσες τιμές ελέγχονται αυτόματα. Η βασική διαμόρφωση ενός 2-επίπεδου μοντέλου δεν αναγνωρίζει επίσημα ότι μία απάντηση λείπει, δεδομένου ότι καταγράφονται μόνο οι παρευρυσκόμενες. Τώρα, εξετάζεται η περίπτωση όπου λείπουν απαντήσεις και ελέγχεται πως αυτό μπορεί να είναι χρήσιμο σε διάφορες περιστάσεις. Σε πολλά είδη ερευνών το πλήθος των πληροφοριών που απαιτούνται από τους ερωτηθέντες είναι τόσο μεγάλο, που είναι αναπόφευκτό ότι ο καθένας δεν θα απαντήσει σε όλες τις ερωτήσεις. Στην εκπαίδευση, σε έρευνες επιχειρήσεων αλλά ακόμα και σε έρευνες νοικοκυριών μπορεί να απαιτείται από ερωτηθέντες πλήρες απάντηση του ερωτηματολογίου, κάτι σημαντικό και για την σωστή διεξαγωγή αποτελεσμάτων.

Αν συμβολίζεται με $\{N\}$: $τοσύνολοτωναπαντήσεων$, και στη συνέχεια έστω $p$ υποσύνολα$\{N_i, i = 1, \dots, p\}$το καθένα από τα οποία είναι κατάλληλα για χορήγηση σε ένα υποκείμενο (επιπέδου_1). Κατά την επιλογή αυτών των υποσυνόλων μπορεί να υπολογιστεί μόνο η συνδιακύμανση μεταξύ αντικειμένου – επιπέδου μεταξύ των απαντήσεων που εμφανίζονται. Αν από την άλλη θέλουμε να εκτιμήσουμε την συνδιακύμανση για τις μονάδες υψηλότερου επιπέδου, όπως τα σχολεία, είναι αναγκαίο να εξασφαλιστεί το σχετικό ζεύγος των απαντήσεων που καταχωρείται σε μερικά σχολεία – ‘ένας αρκετά μεγάλο αριθμός’ θα παρέχει αποτελεσματικές εκτιμήσεις.

Κάθε υποσύνολο θεωρείται επίσημα ως πολυμεταβλητό διάνυσμα απάντησης με ελειπούσες τιμές, αν και οι ελειπούσες τιμές παράγονται από τον σχεδιασμό. Μπορούμε, όπως είδαμε και πιο πάνω να χωρέσουμε σε ένα πολυμεταβλητό μοντέλο

απάντηση γι αυτά τα δεδομένα και να λάβει και αποτελεσματικές εκτιμήσεις για τους σταθερούς συντελεστές και τις δομές συνδιακύμανσης σε οποιοδήποτε επίπεδο. Σε μία τέτοια διαμόρφωση, οι μεταβλητές που πρέπει να χρησιμοποιούνται ως επεξηγηματικές μεταβλητές πρέπει να μετριούνται για κάθε μονάδα επιπέδου_1.

### *4.6.4: Ανάλυση κύριων συνιστωσών*

Έχουμε ήδη αναφέρει ότι ο πίνακας συνδιακύμανσης για ένα πολυμεταβλητό διάνυσμα απαντήσεων, όπου υπάρχουν ελειπούντα δεδομένα μπορεί να υπολογιστεί από την οργάνωση για την πολυμεταβλητή δομή. Όταν οι μεταβλητές έχουν μία πολυμεταβλητή καντανομή των εκτιμήσεων που απορρέουν , τότε έχουνε μια μέγιστη πιθανότητα ή περιορισμένη μέγιστη πιθανότητα. Ο στόχος της ανάλυσης κύριων συνιστωσών είναι να βρεθεί μία γραμμική συνάρτηση ενός συνόλου διακυμάνσεων το οποίο έχει την μέγιστη διακύμανση, που υπόκειται σε κατάλληλο περιορισμό. Στην περίπτωση ενιαίου επιπέδου, για να μεγιστοποιηθεί η διακύμανση $\boldsymbol{w^T y}$,

όπου $w$: είναι το διάνυσμα των βαρών που ορίζει την γραμμική συνάρτηση των περιγραφικών $y$, $\Omega$: είναι ο πίνακας συνδιακύμανσης του $y$, έχουμε:

$$\boldsymbol{\Lambda = w^T \Omega w, \qquad w^T w = 1}$$

Η λύση δίνεται από το ιδιοδιάνυσμα που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή του $\Omega$, ότι είναι η λύση του:$|\Omega - \Lambda| = 0$.

Το $\Omega$ μπορεί να είναι πίνακας συνδιασποράς(ή συσχέτισης) μπορεί να έχει έναν υπολλειματικό πίνακα, μετά την παλινδρόμηση σε επεξηγηματικές μεταβλητές.

## *4.7: Λανθάνων κανονικά μοντέλα για πολυεπίπεδα δεδομένα*

### *4.7.1: Κανονικά πολυεπίπεδα πολυμεταβλητά μοντέλα*

Εξετάζονται μόνο απαντήσεις επιπέδου_1 και γενικεύεται η περίπτωση όπου οι απαντήσεις μπορεί να είναι μίξη διαφόρων τύπων. Ένα βασικό πολυμεταβλητό κανονικό μοντέλο περιγράφεται παρακάτω:

$$\boldsymbol{y_{ij} = X_{ij}\beta + z_{ij}u_j + e_{ij}}$$

Όπου, $y_{ij}$: είναι το διάνυσμα γραμμής που περιέχει τις απαντήσεις $p$.

Πιό πάνω είχαμε αναφερθεί στην εκτίμηση MCMC, η οποία έχει τα εξής βήματα:

***Βήμα 1***: Δείγμα επιπέδου_1 σταθερών συντελεστών από την παραπάνω σχέση με δεδομένες τις τρέχουσες εκτιμήσεις των παραμέτρων.

***Βήμα 2***: Δείγμα του επιπέδου_2 τυχαίων επιδράσεων $u_j$.

***Βήμα 3***: Υπολογίζετε το επίπεδο_1 καταλοίπων $e_{ij}$

***Βήμα 4***: Δείγμα του επιπέδου_1 του πίνακα συνδιασποράς

***Βήμα 5***: Δείγμα του επιπέδου_2 του πίνακα συνδιασποράς.

Σε περίπτωση που οποιαδήποτε ανταπόκριση λείπει, συμπληρώνονται οι τιμές της δειγματοληψίας υπό την προυπόθεση να μην λείπουν απαντήσεις. Σε όλες τις περιπτώσεις, όταν δοκιμάζονται οι υποκείμενες κανονικές μεταβλητές, δίνονται παρατηρούμενες μη-κανονικές μεταβλητές και ελέγχεται η συσχέτιση λανθάνουσων και παρατηρούμενων φυσιολογικών αντιδράσεων.

### *4.7.2: Δειγματοληψία δυαδικών απαντήσεων*

Για ένα δεδομένο δυαδικής απάντησης, θεωρείται η λανθάνουσα κανονική μεταβλητή $y_{ij,1}$το οποίο χωρίς κάποια απώλεια της γενίεκυσης μπορεί να ονομαστεί μεταβλητή_1, και δηλώνει εκεί τις απαντήσεις από $y_{ij}^{*}$, όπου αυτές υποτίθεται ότι έχουν μία πολυμεταβλητή κανονική κατανομή είτε επειδή έτσι έχουν παρατηρηθεί είτε επειδή προκύπτουν από τη δειγματοληψία των μη φυσιολογικών αντιδράσεων. Δοκιμάζεται η λανθάνουσα κανονική τιμή για $y_{ij,1}$. Λαμβάνοντας υπόψην τις τρέχουσες τιμές παραμέτρων, έχουμε την υπό όρους κατανομή:

$$y_{ij,1}|y_{ij}^{*} \sim N(X_{ij}\beta_1 + y_{ij}^{*}\beta_2 + z_{ij}u_j, 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$$

$$\beta_2 = \Sigma_{21}\Sigma_1^{-1}, cov\begin{pmatrix} y_{ij}^{*} \\ y_{ij,1} \end{pmatrix} = \Sigma = \begin{pmatrix} \Sigma_1 & \\ \Sigma_{12} & \Sigma_2 \end{pmatrix}, \Sigma_2 = 1$$

Το 1 παρουσιάζεται όταν $y_{ij,1} > 0$ τότε μία παρατηρηθείσα τιμή 1 έχουμε δείγμα από την κανονική κατανομή $N(0{,}1 - \Sigma_{12}\Sigma_1^{-1}\Sigma_{12})$ . Στο διάστημα $(-(X_{ij}\beta_1 + y_{ij}^{*}\beta_2 + z_{ij}u_j), \infty)$, διαφορετικά δείγματα από $\left(-\infty, -(X_{ij}\beta_1 + y_{ij}^{*}\beta_2 + z_{ij}u_j)\right)$. Το κατώτατο όριο παραμέτρων $\alpha$ καθορίζει την αθροιστική κατανομή πιθανότητας για την διατάξιμη απάντηση. Συνήθως, $a_1 = 0$έτσι ώστε το χαμηλότερο όριο απορροφάται στο σημείο τομής. Δεδομένου των τρέχουσων παραμέτρων, η συνιστώσα της πιθανότητας συνδέεται με μία συγκεκριμένη εντολή κατηγορικής απόκρισης που δίνεται από:

$$P_a = \prod_{j=1}^{J} \prod_{i=1}^{n_j} \prod_{k=1}^{p} \pi_{a,k}^{w_{ij,k}}$$

Για δεδομένο $a$όπου, $w_{ij,k} = 1$εάν η απάντηση για την μονάδα $ij$είναι στην κατηγορία $k$. Οι πιθανότητες$\pi_{a,k}$είναι αυτές που αντιστοιχούν στα κανονικά διαστήματα που ορίζονται παραπάνω.

Έστων ένα νέο σύνολο τιμών $a^{*}$χρησιμοποιώντας μία κατάλληλη κατανομή και ρυθμίζει νέες παραμέτρους με όριο = $a^{*}$ με πιθανότητα $min(1, P_{a^{*}}/P_a)$.

### *4.7.3: Δειγματοληψία διατάξιμων κατηγορικών απαντήσεων*

Έχουμε, μία διατάξιμη κατηογρική απάντηση, με διατάξιμη κατηγορία με αριθμό $1, \cdots, p$. Έστω το ίδιο υπό όρους μοντέλο με παραπάνω, αλλά προσθέτουμε κατώτατα όρια κατηγορίας $\{a_k, k = 1, \dots, p\}$. Έχουμε:

$$y_{ij,1}|y_{ij}^* \sim N(X_{ij}\beta_1 + y_{ij}^*\beta_2 + z_{ij}u_j, 1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$$

Όπου $\beta_2 = \Sigma_{21}\Sigma_1^{-1}$.

Αν παρατηρήσουμε το δείγμα $k(1 < k < p)$τότε έχουμε δείγμα από το κανονικό διάστημα $\left(a_{k-1} - \left(X_{ij}\beta_1 + y_{ij}^*\beta_2 + z_{ij}u_j\right), a_k - \left(X_{ij}\beta_1 + y_{ij}^*\beta_2 + z_{ij}u_j\right)\right)$με σχετική πιθανότητα $\pi_{a,k}$.

Αν $k = 0$έχουμε δείγμα από το $\left(-\infty, -\left(X_{ij}\beta_1 + y_{ij}^*\beta_2 + z_{ij}u_j\right)\right)$ και αν $k = p$έχουμε δείγμα από $\left(a_{p-1} - \left(X_{ij}\beta_1 + y_{ij}^*\beta_2 + z_{ij}u_j\right), \infty\right)$.

### *4.7.4: Δειγματοληψία δεδομένων μέτρησης*

Εξετάζουμε την κατανομή Poissonμε μέση τιμή $\theta$:

$$f(h;\theta) = \frac{e^{-\theta}\theta^h}{h!}, h = 0, \dots, p-1$$

Για τις πρώτες $p$κατηγορίες που παρατηρούνται με την αθροιστική κατανομή

$$F(h;\theta) = \sum_{g=0}^{h} f(g;\theta)$$

Επιλέγουμε την τιμή αναφοράς $h = 0$και δειγματοληψία μας παραλληλίζει για μία κατηγορηματική μεταβλητή. Για μία παρατηρούμενη καταμέτρηση στην πρώτη κατηγορία μπορούμε να δοκιμάσουμε μία τιμή από το ακόλουθο χρονικό διάστημα της κανονικής κατανομής με διακύμανση $(1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$

$$\left(-\infty, -\left(X_{ij}\beta_1 + y_{ij}^*\beta_2 + z_{ij}u_j\right)\right)$$

Και αν παρατηρούμε μία τιμή στην κατηγορία $h^*$έχουμε δείγμα από:

$$(\alpha, \beta), \alpha = \Phi^{-1}[F(h^*;\theta_i)], \qquad \beta = \Phi^{-1}[F(h^* - 1;\theta_i)]$$

Όπου $\Phi$: είναι η αθροιστική συνάρτηση κατανομής του $N(0.1 - \Sigma_{21}\Sigma_1^{-1}\Sigma_{12})$

### *4.7.5: Συνεχής δειγματοληψία μη κανονικών δεδομένων*

Για μία εύρεια κατηγορία κατανομών μπορούμε να εφαρμόσουμε μία ομαλοποίηση που είναι συνάρτηση μιας ή περισσοτέρων παραμέτρων και στην συνέχεια να

ενσωματώσει αυτό με παρόμοιο τρόπο με αυτόν που περιγράφεται για διακριτές αποκρίσεις. Για παράδειγμα, ο μετασχηματισμός Box-Coxγια το $y \geq 0$είναι:

$$y^{(\lambda)} = \begin{matrix} \left(y^{\lambda} - 1\right)\lambda^{-1}, \lambda \neq 0 \\ \log(y)\,,\ \ \lambda = 0 \end{matrix} \Bigg\}$$

Η σχετική συνιστώσα πιθανότητας για τα μη μετασχηματισμένα $y$είναι:

$$-\sum_{\iota} \frac{\left(y_i^{(\lambda)} - \mu_i\right)^2}{2\sigma^2} + (\lambda - 1)\sum_{i} log(y_i)$$

Όπου, $\mu_i$: σταθερός παράγοντας πρόβλεψης ,$\sigma^2$: είναι η υποσυνθήκη διακύμανση.

Ο δεύτερος όρος προέρχεται από Ιακωβιανούς μετασχηματισμούς.

#### *4.7.6: Δειγματοληψία από επέπεδο1 και επίπεδο2 πινάκων συνδιασποράς*

Στην περίπτωση που όλες οι απαντήσεις είναι φυσιολογικές, δοκιμάζουμε τους πίνακες συνδιακύμανσης επιπέδου1 και επιπέδου 2 χρησιμοποιώντας μία αντίστροφη κατανομή Wishart.

Για όλες τις κατηγορικές απαντήσεις οι διακυμάνσεις στο επίπεδο 1 είναι σταθερές και ίσες με 1,0, με μηδενικές συσχετίσεις μεταξύ των κατηγοριών της κάθε μη διατεταγμένης κατηγορικής μεταβλητής αλλά μη μηδενικές συσχετίσεις μεταξύ αυτών των κατηγορικών και άλλων κατηγορικών και συνεχών μεταβλητών.

### *4.8: Μη- γραμμικά πολυεπίπεδα μοντέλα*

#### *4.8.1: Μη γραμμικά μοντέλα*

Μέχρι τώρα έχουμε αναφερθεί σε γραμμικά μοντέλα, με την έννοια ότι η απόκριση είναι γραμμική συνάρτηση των παραμέτρων στο συαθερό τμήμα και τα στοιχεία $V$είναι γραμμικές συναρτήσεις των παραμέτρων. Όμως, σε πολλές εφαρμογές εξετάζονται μοντέλα όπου τα σταθερά ή τυχαία τμήματα του μοντέλου ή και τα δύο περιέχουν μη γραμμικές συναρτήσεις. Για παράδειγμα έχουμε την ακόλουθη συνάρτηση όπου περιγράφει την ανάπτυξη ως αναφορά το ύψος των μικρών παιδιών:

$$\boldsymbol{y_{ij} = \alpha_0 + \alpha_1 t_{ij} + u_{a0j} + u_{\alpha 1 j} t_{ij} + e_{aij} - \exp(\beta_0 + \beta_1 t_{ij} + u_{\beta 0 j} + u_{\beta 1 j} t_{ij} + e_{\beta ij})}$$

Όπου,

$t_{ij}$: είναι η ηλικία του i-οστού παιδιού

$\alpha_0, \alpha_1, \beta_0, \beta_1$: σταθεροί συντελεστές

$e_{aij}, e_{\beta ij}$: κατάλοιπα

$u$:αποκρίσεις

### *4.8.2: Μη γραμμικές συναρτήσεις των γραμμικών συνιστωσών*

Μπορούμε να ορίσουμε ένα γενικό μοντέλο με προσοχή σε μία 2-επίπεδη δομή, ως εξής:

$$\boldsymbol{y_{ij} = X_{1ij}\beta_1 + Z_{1ij}{}^{(2)}u_{1j} + Z_{1ij}{}^{(1)}e_{1ij} + f\left(X_{2ij}\beta_2 + Z_{2ij}{}^{(2)}u_{2j} + Z_{2ij}{}^{(1)}e_{2ij}\right) + \cdots}$$

Όπου,

Η συνάρτηση $f$: μία μη - γραμμική αυνάρτηση και η $(+\cdots)$:υποδηλώνει ότι μπορούν να συμπεριληφθούν πρόσθετες μη – γραμμικές συναρτήσεις,

$X$: σταθερό μέρος επεξηγηματικών μεταβλητών

$Z$:τυχαίο μέρος επεξηγηματικών μεταβλητών ($Z^{(1)}$ , $Z^{(2)}$

Το μοντέλο πρώτα γίνετια γραμμικό μέσα από μία κατάλληλη επέκταση σειράς Taylorκαι αυτό οδηγεί σε εξέταση ενός γραμμικού μοντέλου, όπου οι επεξηγηματικές μεταβλητές μετασχηματίζονται χρησιμοποιώντας, πρώτα και δεύτερα παράγωγα της μη γραμμικής συνάρτησης. Οι τυχαίες μεταβλητές σε ένα δεδομένο επίπεδο, για γραμμικά και μη γραμμικά συαστικά μπορεί να σχετίζονται.

### *4.8.3: Εκτίμηση των πληθυσμιακών μέσων*

Γίνεται εξέταση της αναμενόμενης τιμής της απόκρισης για ένα δεδομένο σύνολο συμμεταβλητών τιμών. Λόγω της μη-γραμμικότητας αυτό δεν είναι γενικά ίσο με την προβλεπόμενη τιμή όταν οι τυχαίες μεταβλητές στην μη-γραμμική συνάρτηση είναι μηδέν.

Για παράδειγμα, τα στοιχεία διακύμανσης του μοντέλου είναι:

$$\boldsymbol{p_{ij} = \exp(\beta_0 + \beta_1 x_{ij} + u_j)} \quad \textbf{(4.8.3)}$$

Και υποθέτοντας κανονικότητα για $u_j$ παίρνουμε:

$$\boldsymbol{E\left(p_{ij}\middle|x_{ij}\right) = \exp\left(\beta_0 + \beta_1 x_{ij}\right) \int_{-\infty}^{\infty} e^{u_j} f\left(u_j\right) du_j = \exp(\beta_0 + \beta_1 x_{ij} + S_u^2/2)} \quad \textbf{(4.8.3*)}$$

Όπου, $f$: είναι η συνάρτηση πυκνότητας της κανονικής καντανομής. Οι Zeger (1988) προτείνουν ένα μοντέλο <<μέσου πληθυσμού>> για την άμεση απόκτηση του πληθυσμού προβλεπόμενων τιμών με την εξάλειψη των τυχαίων μεταβλητών από τη μη γραμμική-συνιστώσα. Σε γενικές όμως γραμμές αυτή η προσέγγιση είναι λιγότερο αποδοτική όταν το πλήρες μοντέλο με τυχαίες μεταβλητές κατά την μη-γραμμική

συνάρτηση είναι το σωστό μοντέλο. Οι προβλεπόμενες τιμές του πληθυσμού, εξαρτώνται από συμπαράγοντες, μπορούν να ληφθούν αν απαιτείται όπως παραπάνω με την λήψη των προσδοκιών πάνω από τον πληθυσμό. Εναλλακτικά, μπορεί να δημιουργηθεί ένας μεγάλος αριθμός προσομοιωμένου συνόλου τιμών για τις τυχαίες μεταβλητές και για κάθε σετ αξιολογείται η συνάρτηση απόκρισης για να ληφθεί μία εκτίμηση της πλήρους κατανομής του πληθυσμού.

### *4.8.4: Μη γραμμικές συναρτήσεις για διακυμάνσεις και συνδιακυμάνσεις*

Υπάρχουν περιπτώσεις όπου μπορεί να θέλουμε να διαμορφώσουμε διακυμάνσεις ή σχυσχετίσεις ως μη-γραμμικές συναρτήσεις.

Έστω ότι το επίπεδο_1 διακύμανσης μειώνεται με την αύξηση των τιμών των επεξηγηματικών μεταβλητών έτσι ώστε να προσεγγίζεται ασυμπτωτικά μία σταθερή τιμή. Τότε για ένα 2-επίπεδο μοντέλο έχουμε:

$$\boldsymbol{var(e_{ij}) = \exp(\beta_0^* - \beta_1^* x_{ij})} \textbf{(4.8.4)}$$

Όπου, $\beta_0^*, \beta_1^*$: είναι παράμετροι που πρέπει να εκτιμηθούν. Ένα τέτοιο μοντέλο εγγυάται ότι το επίπεδο_1 διακύμανσης είναι θετκό, κάτι το οποίο συμβαίνει σε γραμμικά μοντέλα .

### *Μοντελοποίηση διακυμάνσεων και συσχετίσεων ως μη-γραμμικές συναρτήσεις*

Για τυχαίες παραμέτρους ($\beta^*$) ενός γραμμικού μοντέλου έχουμε:

$$\boldsymbol{Y^* = vec(\tilde{Y}\tilde{Y}^T) = X^*\beta^*} \ , \qquad \boldsymbol{E(Y^*) = vec(V)}$$

Μπορεί τώρα να εφαμροστεί η ίδια διαδικασία για τον προσδιορισμό και την εκτίμηση ενός μη γραμμικού μοντέλου.Αν το επίπεδο_1 διακύμανσης είναι μία εκθετική συνάρτηση της συμμεταβλητής $X_1^*$, δηλαδή το i-οστό στοιχείο του $X^*\beta^*$, η συμβολή της διακύμανσης στο επίπεδο_1 είναι:

$$\boldsymbol{S_{et}^2 = f(\beta^*) = \exp(\beta_0^* x_{0t}^* + \beta_1^* x_{1t}^*), X_1^* = \{x_{1t}^*\}, \beta^* = \begin{Bmatrix} \beta_0^* \\ \beta_1^* \end{Bmatrix}}$$

Επειδή, εκτιμήσαμε μόνο μη-γραμμικές συναρτήσεις και δεν προστέθηκαν προσεγγίσεις σε μία άλλη τυχαία συνιστώσα, οι εκτιμήσεις που προκύπτουν είναι εκτιμήσεις μέγιστης πιθανοφάνειας. Πρέπει να τονιστεί ότι οι παράμετροι $\beta_0^*, \beta_1^*$ δεν είναι απαραίτητα θετικοί όταν μοντελοποιούνται, μπορεί να θεωρηθούν απλά ως παράμετροι που πρέπει να εκτιμηθούν αν θεωρήσουμε κανονικότητα.

### *4.8.5: Εκτίμηση μη-γραμμικού μοντέλου*

### *Μοντελοποίηση γραμμικών συνιστωσών*

Θεωρούμε ένα μη-γραμμικό όρο της μορφής:

$$\boldsymbol{y_{ij} = f(X_{2ij}\beta_2 + Z_{2ij}^{(2)}u_{2j} + Z_{2ij}^{(1)}e_{2ij})} \quad \textbf{(4.8.5)}$$

Στη$(t+1)$ −οστή επανάληψη επεκτείνουμε και για δύο σταθερά και τυχαία μέρη, έχουμε ως εξής:

$$\boldsymbol{f_{ij}(H_t) + X_{ij}(\beta_{2,t+1} - \beta_{2,t})f_{ij}^{1}(H_t)}$$

$$\boldsymbol{+\left(Z_{2ij}^{(2)}u_{2j} + Z_{2ij}^{(1)}e_{2ij}\right)f_{ij}^{1}(H_t) + (Z_{2ij}^{(2)}u_{2j} + Z_{2ij}^{(1)}e_{2ij})^2 + f_{ij}^{2}(H_t)/2} \textbf{(4.8.5*)}$$

Στην παραπάνω σχέση η 1η γραμμή ενημερώνει το σταθερό μέρος του μοντέλου. Η ποσότητα $f_{ij}(H_t) - X_{ij}\beta_{2,t}f_{ij}^{1}(H_t)$ θεωρείται ως το αντιστάθμισμα που πρέπει να αφαιρείται από την μεταβλητή απόκρισης. Ο 1ος όρος στην 2η γραμμή είναι μία γραμμική τυχαία συνιστώσα με βάση τις επεξηγηματικές μεταβλητές που μετασχηματίζονται με τον πολλαπλασιασμό από την 1η απόκλιση. Καθορίζουμε το $H_t$και εξετάζεται η κατανομή του 2ου όρου της 2ης γραμμής.

Επιλέγουμε, $H_t = X_{2ij}\beta_{2,t}$ισούται με την διεξαγωγή της επέκτασης Taylorγύρω από το σταθερό μέρος της προβλεπόμενης τιμής. Επιλέγουμε $H_t = X_{2ij}\beta_{2,t} + Z_{2ij}^{(2)}\hat{u}_{2j} + Z_{2ij}^{(1)}\hat{e}_{2ij}$, επεκτείνεται γύρω από την τρέχουσα προβλεπόμενη τιμή για την ij-όστή μονάδα και αντικαθιστούμε την 2η γραμμή με

$$\boldsymbol{\left(Z_{2ij}^{(2)}(u_{2j} - \hat{u}_{2j}) + Z_{2ij}^{(1)}(e_{2ij} - \hat{e}_{2ij})\right)f_{ij}^{1}(H_t)}$$

$$\boldsymbol{+\left(Z_{2ij}^{(2)}(u_{2j} - \hat{u}_{2j}) + Z_{2ij}^{(1)}(e_{2ij} - \hat{e}_{2ij})\right)^2 f_{ij}^{1}(H_t)/2} \quad \textbf{(4.8.5**)}$$

Έτσι έχουμε μία επιπλεόν μετατόπιση από το γραμμικό όρο να προστεθεί με την απόκριση$\left(Z_{2ij}^{(2)}\hat{u}_{2j} + Z_{2ij}^{(1)}\hat{e}_{2ij}\right)f_{ij}^{1}(H_t)$.

Από την δεύτερη γραμμή της εξίσωσης (10) , όπου η επέκταση του Tayloreίναι περίπου μηδέν, έχουμε:

$$E\left(Z_{2ij}^{(2)}u_{2j} + Z_{2ij}^{(1)}e_{2ij}\right) = 0, \qquad E\left(Z_{2ij}^{(2)}u_{2j} + Z_{2ij}^{(1)}e_{2ij}\right)^2 = S_{zu}^2 + S_{ze}^2$$

$$S_{zu}^2 = Z_{2ij}^{(2)}\Omega{Z_{2ij}^{(2)}}^T, S_{ze}^2 = Z_{2ij}^{(1)}\Omega{Z_{2ij}^{(1)}}^T$$

Αν υποθέσουμε κανονικότητα , κάθε 3η ροπή που σχηματίζεται από το γινόμενο των δύο όρων στην 2η γραμμή είναι μηδέν και θα έχουμε:

$$var\left(Z_{2ij}^{(2)}u_{2j} + Z_{2ij}^{(1)}e_{2ij}\right)^2 = 2(S_{zu}^4 + S_{ze}^4)$$

Έτσι ώστε να καθορίζονται οι πρόσθετες τυχαίες μεταβλητές:

$$Z_u^* = \frac{S_{zu}^2 f^2(H_t)}{\sqrt{2}}, \qquad Z_e^* = \frac{S_{ze}^2 f^2(H_t)}{\sqrt{2}}$$

Τα οποία είναι ασυσχέτιστα και οι διακυμάνσεις περιορίζονται ώστε να είναι ίσες με 1,0.Αν η επέκταση του Taylorέχει ληφθεί σχετικά με τις τρέχουσες αξίες των καταλοίπων έχουμε:

$$E\left[Z_{2ij}^{(2)}\left(u_{2j} - \hat{u}_{2j}\right)\right]^2 + E\left[Z_{2ij}^{(1)}\left(e_{2ij} - \hat{e}_{2ij}\right)\right]^2$$

Μπορεί εδώ να εφαρμοστεί η συνηθής γραμμική επέκταση για τα κατάλοιπα, όπως έχει αναφερθεί και πιο πάνω.

Το παραπάνω μπορεί να επεκταθεί σε έναν απλό τρόπο, σε περισσότερα από δύο επίπεδα αλλά και σε πολυμεταβλητά μοντέλα.

Για γενικευμένα γραμμικά μοντέλα (Waclawinκαι Liang 1993) θεωρούν μία προσέγγιση εξισώσεων εκτίμησης χρησιμοποι΄πωντας μία μονάδα ειδικής πρόβλεψης. Μία πλήρης μέθοδος πιθανοφάνειας εκτίμησης βασίζεται σε ένα επαναλαμβανόμενο μοντέλο μέτρων με δυαδικές απαντήσεις που δίνονται από τον Garret(1993). Για μικρά δείγματα, πρέπει να χρησιμοποιείται μία αμερόληπτη διαδικασία για να ληφθούν οι αντίστοιχες αμερόληπτες εκτιμήσεις.

## *4.9:Πολυεπίπεδη παραγοντική ανάλυση, διαρθρωτική εξίσωση και μίξη μοντέλων*

### *4.9.1:2-επιπέδων παραγοντικό μοντέλο*

Η θεωρία των δομικών μοντέλων εξισώσεων συμπεριλαμβάνει δύο ειδικές περιπτώσεις, τα μοντέλα διαδρομής και τα μοντέλα παραγοντικής ανάλυσης.

Ο Raudenbush(1995) εφάρμοσε τον αλγόριθμο EMγια την εκτίμηση ενός 2-επίπεδου μοντέλου δομικών εξισώσεων και οι Row&Hill(997) έδειξαν πως ένα τυπικό πολυεπίπεδο λογισμικό μπορεί να χρησιμοποιηθεί για να παρέχει προσεγγίσεις σε εκτιμήσεις μέγιστης πιθανοφάνειας γενικά, πολυεπίπεδων μοντέλων διαρθρωτικών εξισώσεων.

Έστω ένας βασικός 2-επίπεδος παράγοντας μοντέλο, όπου έχουμε ένα σύνολο μετρήσεων για ένα μαθητή μέσα σε ένα δείγμα σχολείων, μαζί με μία σειρά από μετρήσεις σε επίπεδο σχολικής μονάδας, η οποία μπορεί να περιλαμβάνει συγκεντρωτικές μετρήσεις στάθμης των μαθητών. Οι μετρήσεις απόκρισης του ενδιαφέροντος των οποίων η δομή που επιθυμούμε να εξερευνήσουμε υποτίθεται ότι είναι κανονικά κατανεμημένη σε τυχαίες μεταβλητές. Το φύλο ή η κοινωνική τάξη μπορούν να θεωρηθούν για παράδειγμα, είναι περαιτέρω επεξηγηματικές μεταβλητές που μπορεί να επιθυμούν να ελεγχθούν.

Για το επίπεδο-1 $p$ απαντήσεων μπορεί μα γραφεί ένα πολυμεταβλητό μοντέλο με $p$ απαντήσεις όπου θα είναι:

$$\boldsymbol{y_{rij}} = (\boldsymbol{X\beta})_{\boldsymbol{rij}} + \boldsymbol{z_{rij}e_{rij}} + \boldsymbol{z_{rj}u_{rj}}$$

Όπου, $r$: οι δείκτες των απαντήσεων

Έχουμε μία σειρά από 1-επίπεδο τυχαίων μεταβλητών($\boldsymbol{e_{rij}}$)και ένα 2-επίπεδο τυχαίων μεταβλητών ($\boldsymbol{u_{rj}}$)**.** Μία γενική δομή παραγόντων για τις μεταβλητές επιπέδου-1 μπορεί να περιλαμβάνουν παράγοντες που ορίζονται τόσο σε επίπεδο-1 και επίπεδο-2 όπου έχουμε:

$$e_{rij} = \sum_{g} \lambda_{rg}^{(1)} f_{gij}^{(1)} + w_{rij}^{(1)}$$

$$u_{rj} = \sum_{g} \lambda_{rg}^{(2)} f_{gj}^{(2)} + w_{rj}^{(2)}$$

Όπου έχουμε $g$παράγοντες σε κάθε επίπεδο, αν και γενικά μπορούμε να έχουμε διαφορετικό αριθμό παραγόντων σε κάθε επίπεδο. $\lambda, f, \beta$: αντίστοιχα είναι οι φορτίσεις, παράγοντες και τα κατάλοιπα, και οι εκθέτες δείχνουν τα επίπεδα του παράγοντα.

Ένας τρόπος εκτίμησης των παραμέτρων του μοντέλου αυτού είναι να γίνει σε δύο στάδια. Το 1ο στάδιο περιλαμβάνει την εκτίμηση των ξεχωριστών επιπέδου-1 και επιπέδου-2 καταλοίπων των πινάκων συνδιακύμανσης για τις απαντήσεις. Το 2ο στάδιο περιλαμβάνει την ανάλυση των παραγόντων αυτών των ξεχωριστών πινάκων χρησιμοποιώντας οποιαδήποτε πρότυπη διαδικασία, όπως π.χ. μέθοδος μέγιστης πιθανοφάνειας, όπως δίνει ο McDonald(1993).

Η παραπάνω διαδικασία θα πρέπει να είναι αποτελεσματική, εκτός εάν τα δεδομένα είναι ισορροπημένα, με εξαιρετικά μεταβλητούς αριθμούς μονάδων επιπέδου-1 μέσα σε μονάδες επιπέδου-2.

Οι Row&Hill(1997) περιγράφουν μία διαδικασία η οποία εφαρμόζεται σε ορισμένα εκπαιδευτικά δεδομένα.

$$\boldsymbol{y_{1ij}} = \boldsymbol{\alpha_1} + \boldsymbol{\beta_1 x_{1ij}} + \boldsymbol{u_{1j}} + \boldsymbol{e_{1ij}}$$

$$\boldsymbol{y_{2ij}} = \boldsymbol{\alpha_2} + \boldsymbol{\beta_2 y_{1ij}} + \boldsymbol{u_{2j}} + \boldsymbol{e_{2ij}}$$

Όπου η $\boldsymbol{y_{1ij}}$εμφανίζεται και στις δύο εξισώσεις. Το κλασικό μοντέλο διαδρομής αντιμετωπίζει το $\boldsymbol{y_{1ij}}$στην δεύτερη εξίσωση υπό όρους έτσι ώστε να μπορεί να αντιμετωπιστεί ευθέως ως διμεταβλητό 2-επίπεδο μοντέλο. Μία δυνατότητα επιλογής ανάμεσα σε ένα τέτοιο μοντέλο και ένα ανευ-όρων μοντέλο θα εξαρτηθεί από ουσιαστικές εκτιμήσεις, ιδίως εκεί όπου υπάρχει μια χρονική διάταξη των

μεταβλητών, όταν το υπό όρους μοντέλο φαίνεται να είναι πιο κατάλληλο σε γενικές γραμμές.

### *4.9.2: Γενικό πολυεπίπεδο παραγοντικό μοντέλο:*

Ένα γενικό μοντέλο παράγοντα με φυσιολογικές αποκρίσεις ορίζεται ως εξής:

$$\boldsymbol{y_{rij} = \beta_r + \sum_k a_{rk} x_{kij} + \sum_{h=1}^{H} \lambda_{rh}^{(2)} f_{hj}^{(2)} + \sum_{g=1}^{G} \lambda_{rg}^{(1)} f_{gij}^{(1)} + u_{rj} + e_{rij}}$$

$$\boldsymbol{u_{rj} \sim N(0, \sigma_{ur}^2), e_{rij} \sim N(0, \sigma_{er}^2), f_j^{(2)} \sim MVN_H(0, \Omega_2), f_{ij}^{(1)} \sim MVN_G(0, \Omega_1)}$$

$$\boldsymbol{r = 1, \dots, R, \quad i = 1, \dots, n_j, \quad j = 1, \dots J, \quad \sum_{j=1}^{J} n_j = N}$$

Εδώ έχουμε $\boldsymbol{R}$απαντήσεις για $\boldsymbol{N}$άτομα, τα οποία μοιράζοντα μεταξύ $\boldsymbol{J}$επιπέδου 2 μονάδων. Έχουμε $\boldsymbol{H}$σύνολα παραγόντων, $\boldsymbol{f_{hj}^{(2)}}$: ορίζεται στο επίπεδο-2 και $\boldsymbol{G}$σύνολα παραγόντων , $\boldsymbol{f_{gij}^{(1)}}$: ορίζεται στο επίπεδο-2. Στο σταθερό τμήμα του μοντέλου, έχουμε σταθερούς ξεχωριστούς όρους τομής $\boldsymbol{\beta_r}$για κάθε απάντηση και επιτρέπουν συμμεταβλητές $\boldsymbol{x_{kij}}$. Οι φορτίσεις $\boldsymbol{\lambda^{(1)} \lambda^{(2)}}$ είναι προσδιορισμένες σε κάθε επίπεδο.

Αν γνωρίζουμε τις τιμές των φορτίσεων $\lambda$ τότε θα μπορούσαε να διαμορφώσουμε ένα μοντέλο 3-επιπέδων με τους φορείς φόρτωσης ως επεξηγηματικές μεταβλητές. Αντίθετα, αν γνωρίζουμε τις τιμές των τυχαίων επιδράσεων $f$θα μπορούσαμε να εκτιμήσουμε τις φορτίσεις ως σταθερούς συντελεστές σε ένα πολυμεταβλητό μοντέλο.Αυτές οι εκτιμήσεις δείχνουν ότι ένας αλγόριθμος ΕΜ μπορεί να χρησιμοποιηθεί όπου οι τυχαίες επιδράσεις θεωρούνται ελλιπούντα δεδομένα.

### *4.9.3: Δομικά μοντέλα εξισώσεων*

Στο βασικό παραγοντικό μοντέλο, οι ίδιοι οι παράγοντες δεν διαμορφώνονται περαιτέρω. Σε πολλές εφαρμογές όμως μπορεί να χρειαστεί να συμβεί αυτό, τα μοντέλα αυτά που μπορούν να το εκτελέσουν ονομάζονται γενικά, μοντέλα δομικών εξισώσεων. Αρχικά, έχουμε μία απλή επέκταση στο παραγοντικό μοντέλο 2-επιπέδνμ με απαντήσεις σε ένα μόνο επίπεδο και η παραγοντική δομή τοποθετείται στο επίπεδο-1 όπου ο παράγοντας είναι μία γραμμική συνάρτηση των επεξηγηματικών μεταβλητών.

$$\boldsymbol{y_{rij} = \beta_r + \lambda_r^{(1)} f_{ij}^{(1)} + e_{rij}, \qquad f_{ij}^{(1)} = a_1 x_{ij} + u_j + w_{ij}}$$

$$\boldsymbol{w_{ij} \sim N(0, \sigma_w^2), \quad e_{rij} \sim N(0, \sigma_e^2), \quad u_j \sim N(0, \sigma_u^2)}$$

Όπου, για αναγνωρισιμότητα θέτουμε: $\boldsymbol{\sigma_w^2 = 1}$

Σε αντικατάσταση αυτό δίνει:

$$\boldsymbol{y_{rij} = \beta_{0r} + \lambda_r^{(1)} a_1 x_{ij} + \lambda_r^{(1)} u_j + \lambda_r^{(1)} w_{ij} + e_{rij}}$$

Εμείς εξετάζουμε ένα πιο γενικό μοντέλο δομικών εξισώσεων που επιτρέπει τις σχέσεις μεταξύ των παραγόντων. Απεικονίζεται, ένα ενιαίο επιπέδου μοντέλο, το οποίο μπορεί να γραφεί με την μορφή πίνακα ως εξής:

$$\boldsymbol{A_1 f_1 = A_2 f_2 + w}$$

$$\boldsymbol{Y_1 = \Lambda_1 f_1 + U_1}$$

$$\boldsymbol{Y_2 = \Lambda_2 f_2 + U_2}$$

Όπου,

$\boldsymbol{Y_1, Y_2}$**:** πολυμεταβλητοί φορείς (διανύματα) των απαντήσεων

$\boldsymbol{A_1}$**:** ένας γνωστός πίνακας μετασχηματισμού,

$\boldsymbol{A_2}$**:** ένας πίνακας συντελεστών που καθορίζει ένα πολυμεταβλητό γραμμικό μοντέλο ανάμεσα στο σύνολο των μετασχηματισμένων παραγόντων

$\boldsymbol{\Lambda_1, \Lambda_2}$**:** είναι φορτίσεις

$\boldsymbol{w}$**:** είναι ένα τυχαίο διάνυσμα καταλοίπων

$\boldsymbol{U_1, U_2}$**:** αμοιβαία ανεξάρτητες με μηδενικά μέσα.

Το παρπάνω μοντέλο μπορεί να γενικευθεί περαιτέρω με την εξέταση $m$συνόλων μεταβλητών απόκρισης, $Y_1, Y_2, \dots, Y_m$και διάφορες συνδέσεις μεταξύ τους, πολλαπλές ομάδες δομικών σχέσεων με την k-οστή σχέση που έχει την γενική μορφή:

$$\boldsymbol{\sum_h V_h^{(k)} A_h^{(k)} = \sum_g V_g^{(k)} A_g^{(k)} + W^{(k)}}$$

# *Κεφάλαιο 5:ΔΟΜΙΚΑ ΜΟΝΤΕΛΑ ΕΞΙΣΩΣΕΩΝ (SEM)*

## *5.1: Ιστορική Αναδρομή*

Τα μοντέλα δομικών εξισώσεων εμφανίστηκαν για πρώτη φορά το 1920, όταν ο SewellWright, ένας γενετιστής προσπάθησε να λύσει παράλληλα κάποιες εξισώσεις με σκοπό να διευκολύνει τις γενετικές επιδράσεις που έχει η μία γενιά στην επόμενή της. Ο Wright ουσιαστικά αντιμετώπισε μία περίπτωση στην οποία τα γονίδια των γονέων (δηλαδή τα αίτια) ήταν γνωστά, τα χαρακτηριστικά των απογόνων ήταν γνωστά (δηλαδή οι συνέπειες) και η σύνδεση αιτίου και συνέπειας ήταν μίας κατεύθυνσης, χωρίς να υπάρχει ουσιαστικά ανάδραση. Κάτι το οποίο είναι γνωστό ως <<μιας κατεύθυνσης αιτιολογικό μοντέλο ροής>>. Αυτό είναι και το μοναδικό είδος μοντέλου το οποίο μπορεί να ονομαστεί ***ανάλυση διαδρομών(pathanalysis)***. (Θα γίνει αναφορά παρακάτω σχετικά με την ανάλυση διαδρομών). Σκοπός του Wright ήταν να εκτιμήσει το πλήθος των συνεπειών που μεταφέρονται από κάθε γονέα προς τον απόγονο. Η λύση αυτή μπορούσε να προσδιοριστεί γράφοντας το σύστημα των εξισώσεων, εκφράζοντας τις εξισώσεις χρησιμοποιώντας συσχετίσεις ανάμεσα στις διάφορες μεταβλητές και επιλύοντας ως προς τους αγνώστους.

### *5.1.1: Η ανάλυση διαδρομών (path analysis)* **στις κοινωνικές επιστήμες**

Ο Blalock(1964) και ο Duncan(1966) εισήγαγαν τις μελέτες τους στις κοινωνικές επιστήμες. Η διαδικασία επίλυσης της ανάλυσης διαδρομών ήταν οι εξής: οι παράμετροι εκτιμούνταν λύνοντας ένα σύστημα εξισώσεων με την χρήση της γραμμικής άλγεβρας ή χρησιμοποιώντας πολλαπλή παλινδρόμηση. Για παράδειγμα, ένα από τα χαρακτηριστικά πεδία στο οποίο βρήκαν εφαρμογή τα δομικά μοντέλα εξισώσεων ήταν οι μελέτες που προσπαθούσαν να εκτιμήσουν το επάγγελμα και την καριέρα που θα αποκτούσε ένα άτομο, λαμβάνοντας ως δεδομένο την κοινωνική θέση της οικογένειας, την απόκτηση μόρφωσης στο παρελθόν και το κύρος κάθε εργασίας. Αυτή η μελέτη η οποία δημιούργησε ο Duncan εξέταζε τις προυποθέσεις της

επιτυχίας απόκτησης μόρφωσης και εύρεσης απασχόλησης, όμως λόγω των μεγάλων χρονικών περιόδων των μοντέλων,σε γενικές γραμμές είχαν ***ροή μιας κατεύθυνσης.***

*5.1.1.1:Μοντέλα ροής μια κατεύθυνσης*

Στα μοντέλα τα οποία η υποθετική αιτιότητα είναι προς μία κατεύθυνση, η διαδικασία επίλυσης ήταν αρκετά άμμεση και οι μεθοδολογίες ήταν γνωστές από την εποχή του Wright. Για την εκτίμηση χρησιμοποιήθηκε αρχικά η άλγεβρα, επιλύοντας μία ή περισσότερες εξισώσεις για ένα πλήθος αγνώστων, ενώ αργότερα χρησιμοποιήθηκε και η ανάλυση παλινδρόμησης.

Σε μοντέλα τα οποία εκτιμούνταν το ίδιο πλήθος εξισώσεων και μεταβλητών, η παλινδρόμηση και η γραμμική άλγεβρα απέφεραν τα ίδια αποτελέσματα, κάτι το οποίο συνέβαινε επειδή η ίδια μοναδική λύση μπορεί να επιτευχθεί είτε επιλύοντας τις εξισώσεις με αλγεβρικό πίνακα είτε χρησιμοποιώντας παλινδρόμηση.

Σε μοντέλα με περισσότερους αγνώστους παρά εξισώσεις, δεν υπήρχε τόση πληροφόρηση ώστε να εκτιμηθούν μοναδικές τιμές για τις παραμέτρους. Το πρόβλημα που δημιουργούταν ήταν ότι υπήρχε άπειρος αριθμός εναλλακτικών λύσεων οι οποίες ήταν εξίσου αποδεκτές και δεν υπήρχε ασφαλής τρόπος επιλογής κάποιων από αυτές.

Τέλος, για τα μοντέλα στα οποία οι εξισώσεις ήταν περισσότερες από τις άγνωστες παραμέτρους οι εξισώσεις παρείχαν αρκετή πληροφόρηση ώστε να εκτιμηθούν οι παράμετροι. Έτσι προέκυπταν περισσότερες των μία λύσεων, οι οποίες δεν έδιναν το ίδιο αποτέλεσμα.

Για πρώτη φορά στην δεκαετία του 1970, η μεθοδολογία SEM΄΄πάντρεψε΄΄ την ψυχιομετρία και την οικονομετρία.

- Από την πλευρά της ψυχιομετρίας,τα μοντέλα SEM επιτρέπουν την ανακάλυψη λανθάνουσων δεικτών με πολλαπλούς δείκτες.
- Από την οικονομετρική πλευρά, τα SEM επιτρέπουν την επίλυση πολλαπλών εξισώσεων, που ενδεχομένως έχουν βρόχους ανατροφοδότησης.

Η προσέγγιση των SEM κατά την εποχή της ανάλυσης διαδρομών (δεκαετία του ‘60) χρησιμοποιούνταν για την επίλυση μοντέλων στα οποία υπήρχαν μίας κατεύθυνσης διαδρομές και χρησιμοποιούσαν τεχνικές πολλαπλής παλινδρόμησης. Τα μοντέλα αυτά ονομάζονταν <<συνηθής ανάλυση ελαχίστων τετραγώνων>>. Για την ανάλυση διαδρομών αυτές οι τεχνικές έδιναν τα ίδια αποτελέσματα με αυτά των υπάρχουσων τεχνικών, καθώς οι εκτιμήσεις με ελάχιστα τετράγωνα και μέγιστη πιθανοφάνεια ήταν ταυτόσημες.

Ένα μεγάλο και σημαντικό πλεονέκτημα των γενικότερων γραμμικών μοντέλων που χρησιμοποιούνται σε προγράμματα όπως, το **LISREL** και το **AMOS** είναι προγράμματα τα οποία χειρίζονται περισσότερους τύπους μοντέλων και έτσι δεν απαιτούν από τον χρήστη να μάθει ένα πλήθος διαφορετικών τεχνικών για

διαφορετικούς τύπους μοντέλων. Τα προγράμματα αυτά, αποτελούσαν και αποτελούν ακόμη και σήμερα εφαλτήριο για τα μοντέλα δομικών εξισώσεων.

## *5.2: Το πέρασμα από την ανάλυση διαδρομών στα μοντέλα δομικών εξισώσεων*

Παρατηρήθηκε ότι οι περισσότερες θεωρητικές μεταβλητές καθορίζονταν χωρίς ακρίβεια εξαιτίας της ανακρίβειας στο χειρισμό τους και στην ανακρίβεια μέτρησης των παρατηρούμενων μεταβλητών. Δηλαδή, η χρήση της ανάλυσης διαδρομών ήταν υπό αμφισβήτηση ως προς την έλλειψη λειτουργικότητάς της, το οποίο έκανε την παρουσιάση των SEM δυσκολότερη.

Το 1969 ο Joreskog και άλλοι ανέπτυξαν ένα γενικό γραμμικό μοντέλο που ξεπερνούσε τα περιοστιστικά εμπόδια των ελαχίστων τετραγώνων επιτρέποντας καλύτερη λειτουργικότητα στις θεωρητικές μεταβλητές. Ένα πλήθος υπολογιστικών προγραμμάτων που είναι ευρέως γνωστά τα τελευταία χρόνια είναι το LISREL, το EQS και το AMOS. Έτσι τα μοντέλο αυτό που ανέπτυξε ήταν τα δομικά μοντέλα εξισώσεων.

## *5.3: Τι είναι τα δομικά μοντέλα εξισώσεων*

Το δομικό μοντέλο εξισώσεων (SEM) είναι μία στατιστική μεθοδολογία που χρησιμοποιείται ευρέως από τους ερευνητές στις κοινωνικές κυρίως επιστήμες, καθώς και στις εκπαιδευτικές. Στο SEM λογισμικό σήμερα, τα μοντέλα είναι τόσο γενικά ώστε να περιλαμβάνουν τις περισσότερες από τις στατιστικές μεθόδους που χρησιμοποιούνται στις κοινωνικές επιστήμες.

Ο όρος SEM δεν υποδεικνύει μία ενιαία στατιστική τεχνική αλλά αναφέρεται σε μία οικογένεια σχετικών διαδικασιών και στατιστικών αναλύσεων. Χρησιμοποιούνται συνήθως σαν μία επικυρωτική διαδικασία θεωρητικών υποθέσεων επειδή δεν υπολογίζουν μόνο τις εκτιμήσεις για τους παράγοντες του μοντέλου αλλά εξετάζουν και τον βαθμό προσαρμογής τους με τα δεδομένα.

Κύριο χαρακτηριστικό τους είναι ότι ο ερευνητής πρέπει πρώτα να εκτιμήσει τη σχέση μεταξύ των μεταβλητών και να προτείνει το μοντέλο ανάλυσης και έπειτα να εξετάσει αν οι εκτιμήσεις αυτές επιβεβαιώνονται από τα δεδομένα. Οι σχέσεις αυτές εξετάζονται μέσω γραμμικών εξισώσεων και επιπλέον υπάρχει και η δυνατότητα της γραφικής αναπαράστασης του εκτιμώμενου μοντέλου για καλύτερη κατανόηση και επεξεργασία. Επίσης το μοντέλο αυτό έχει σαν κύριο σκοπό την δημιουργία απλών δομών από μετρήσιμα στοιχεία, δηλαδή προσπαθεί να ομαδοποιήσει τις ατομικές μετρήσιμες πληροφορίες όπως αυτές προκύπτουν σε μικρότερες σε αριθμό μονάδες.

***Τα SEM είναι μία επέκταση της παλινδρόμησης και της παραγοντικής ανάλυσης,η οποία όμως εξετάζει ταυτόχρονα τις σχέσεις μιας ή περισσοτέρων εξαρτημένων μεταβλητών μεταξύ δύο ή περισσοτέρων ανεξάρτητων μεταβλητών***.

Στην πολλαπλή παλινδρόμηση,ο αναλυτής θεωρεί ότι οι μεταβλητές που συμμετέχουν στον σχηματισμό του μοντέλου έχουν μετρηθεί απόλυτα ακόμα και αν αυτό δεν ισχύει. Το μοντέλο της παλινδόμησης εκτιμά και αναπαριστά μόνο άμεσες επιδράσεις. Στην περίπτωση που μία μεταβλητή δεν είναι απόλυτα παρατηρήσιμη δημιουργείται πρόβλημα μεροληψίας, όχι μόνο στην εκτίμηση του συντελεστή της συγκεκριμένης μεταβλητής αλλά και στις εκτιμήσεις των συντελεστών των υπολοίπων μεταβλητών του μοντέλου.

Στην ανάλυση παλινδρόμησης ο ερευνητής χρειάζεται να παρέχει το πρόγραμμα με πληροφορίες σχετικά με τις μεταβλητές που πρέπει να χρησιμοποιηθούν ως επεξηγηματικές και εξαρτημένες μεταβλητές. Και το πρόγραμμα στη συνέχεια καθορίζει αυτόματα τις παραμέτρους του μοντέλου. Αυτό όμως είναι κάτι το οποίο δεν λειτουργεί καλά στις εφαρμογές των δομικών μοντέλων εξισώσεων. Στα δομικά μοντέλα εξισώσεων είναι σημαντικό να δηλώνονται οι παράμετροι από την αρχή και να ρυθμίζουν σωστά το μοντέλο.

Αντίθετα ***το μοντέλο διαδρομής(pathmodel)*** που δημιουργείται στα SEM επιτρέπει την μελέτη επίδρασης των επιμέρους μεταβλητών στις οποίες μπορεί να αποσυντεθεί μία αρχική μεταβλητή.Με άλλα λόγια μία ανεξάρτητη μεταβλητή μπορεί να έχει άμμεση ή έμμεση αντίδραση σε μία εξαρτημένη μεταβλητή.Μια μεταβλητή μπορεί να φαίνεται μη σημαντική όταν αξιολογείται η άμεση επίδρασή της, αλλά μπορεί να γίνεται σημαντική όταν αξιολογείται η συνολική επίδραση, η οποία λαμβάνει υπόψη της τις διαδρομές σύνδεσης (pathways) που τη συνδέουν με την εξαρτημένη μεταβλητή.

Στην *παραγοντική ανάλυση (FactorAnalysis)* τόσο ο αριθμός των παραγόντων αλλά και τα βάρη των μεταβλητών δεν είναι γνωστά εκ των προτέρων. *Η παραγοντική ανάλυση χωρίζεται στην διερευνητική και επιβεβαιωτική ανάλυσ*η. Αντίθετα, στα SEM ο αναλυτής προσδιορίζει σε μεγαλύτερο βαθμό τη δομή του προβλήματος και πραγματοποιεί στατιστιστικούς έλεγχους σημαντικότητας.

Η γενική ιδέα των SEM περιγράφεται μέσω της υπόθεσης ότι:

$$\boldsymbol{H_0}: \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \quad \textbf{(5.3)}$$

Όπου, το $\boldsymbol{H_0}$**:** θεωρείται η αρχική υπόθεση του μοντέλου και δηλώνει ότι ο πίνακας Σ της συνδιακύμανσης του δείγματος είναι ο ίδιος με τον πίνακα που προκύπτει από το μοντέλο, το $\boldsymbol{\theta}$**:** είναι το διάνυσμα με τις παραμέτρους του μοντέλου. Η εναλλακτική υπόθεση δηλώνει ότι ο πίνακας Σ της συνδιακύμανσης του δείγματος είναι διαφορετικός από το μοντέλο.

Για να είναι επαρκές ένα μοντέλο θα πρέπει να ελαχιστοποιείται η διαφορά ανάμεσα στην συνδιακύμανση που προβλέπει το μοντέλο και την παρατηρούμενη συνδιακύμανση.

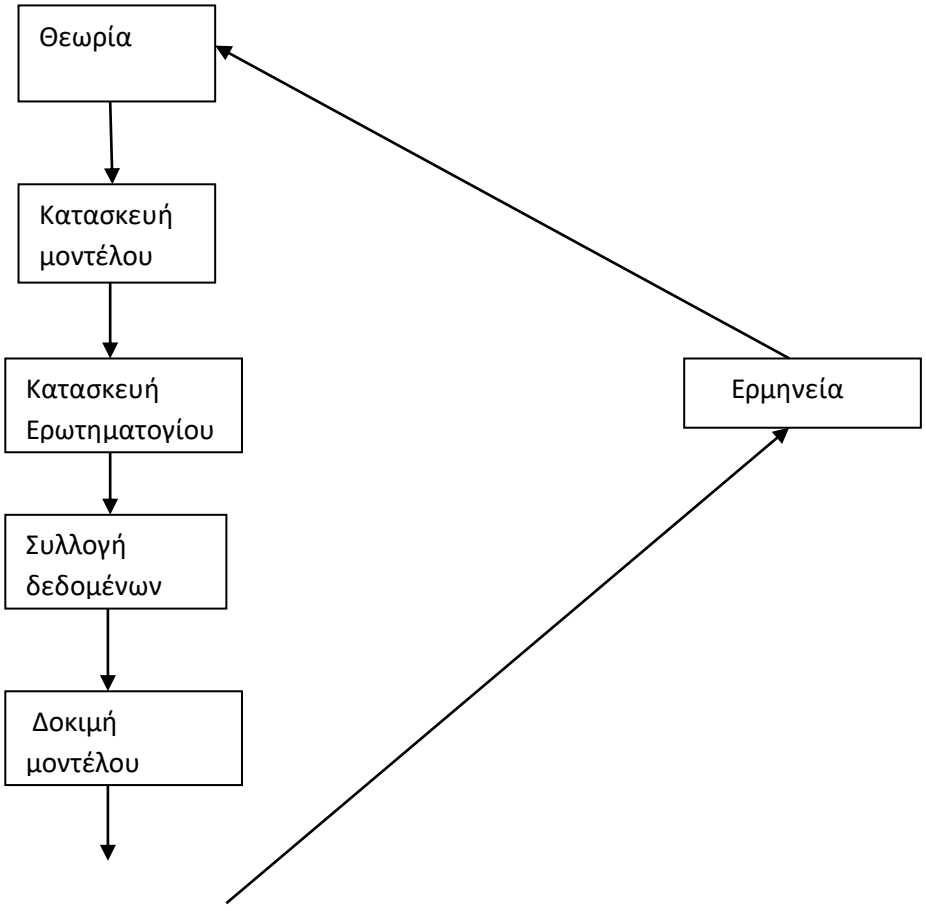
Για τον έλεγχο της προσαρμογής του εκτιμώμενου μοντέλου στα δεδομένα αλλά και τον προσδιορισμό σχέσεων ανάμεσα στις μεταβλητές χρησιμοποιούνται οι λεγόμενοι ***δείκτες προσαρμογής*** με τους οποίους θα ασχοληθούμε στην συνεχεια.

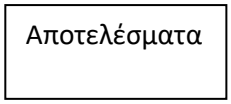
Επίσης, τα δομικά μοντέλα εξισώσεων έχουν την ικανότητα να προσδιορίζουν το σφάλμα μέτρησης αλλά και να το χρησιμοποιούν στη διαδικασία των λανθάνουσων μεταβλητών. Πρέπει να τονίσουμε ότι τα δομικά μοντέλα εξισώσεων περιλαμβάνουν δύο ειδών μεταβλητές: τις μετρήσιμες μεταβλητές και τις μη μετρήσιμες μεταβλητές ή λανθάνουσες.

Οι λανθάνουσες μεταβλητές είναι υποθετικές κατασκευές του ερευνητή για να περιγράψει τις έννοιες που θέλει να μετρήσει και δεν μετρούνται απευθείας αλλά αναπαριστώνται από πολλαπλές μετρήσιμες μεταβλητές οι οποίες χρησιμοποιούνται σαν δείκτες. Είναι μεταβλητές οι οποίες μετριώνται μέσω των υπολοίπων μεταβλητών αφού δεν μπορούν να μετρηθούν απευθείας. Χρησιμοποιούνται ευρέως σε κοινωνικές επιστήμες, στην ψυχολογία στην βιολογία και στις περιβαλλοντικές επιστήμες.

Στα δομικά μοντέλα εξισώσεων οι μεταβλητές είναι ενδογενείς ή εξωγενείς. Οι εξωγενείς μεταβλητές είναι οι ανεξάρτητες μεταβλητές οι οποίες είναι υπεύθυνες για τις διακυμάνσεις που παρατηρούνται σε άλλες μεταβλητές του μοντέλου όπως το φύλο, η ηλικία κ.τ.λ. Οι ενδογενείς μεταβλητές είναι οι εξαρτημένες μεταβλητές οι οποίες επηρεάζονται από τις υπόλοιπες μεταβλητές είτε άμεσα είτε έμμεσα.

Η κύρια προσέγγιση για να πραγματοποιηθεί μία ανάλυση SEMαπεικονίζεται στο παρακάτω σχήμα:






*Εικόνα 12: Προσέγγιση μοντέλου SEM*

Ο ερευνητής συνήθως, καθορίζει πρώτα το μοντέλο βασισμένος στην θεωρία, μετά προσδιορίζει πως θα το μετρήσει, συλλέγει τα δεδομένα και έπειτα εισάγει τα δεδομένα αυτά σε ένα λογισμικό πακέτο SEM. Το μοντέλο ταιριάζει τα δεδομένα αυτά με το μηδενικό μοντέλο και παράγει κάποια αποτελέσματα που περιλαμβάνουν την συνολική σύγκλιση του μοντέλου και τις εκτιμήσεις των παραμέτρων του. Στην ανάλυση εισάγεται ένας πίνακας συνδιακύμανσης των ετκιμώμενων μεταβλητών.

Άλλη μία διαφορά των δομικών μοντέλων εξισώσεων με άλλες στατιστικές μεθόδους είναι ότι τα SEM μπορούν να εκτιμούν και να αξιολογούν τις σχέσεις μεταξύ ποιοτικών μεταβλητών-εννοιών.Εξαιταίας αυτού τα δομικά μοντέλα εξισώσεων αποτελούνται από δύο κύρια μέρη:

### *5.3.1: Μετρικό – Δομικό μοντέλο*

1. ***Το μετρικό μοντέλο ή μοντέλο μέτρησης (measurement model)*** ερμηνεύει τις σχέσεις μεταξύ των παρατηρούμενων και μη παρατηρούμενων μεταβλητών.
   Οι παρατηρούμενες μεταβλητές μετριώνται χρησιμοποιώντας ερωτηματολόγια με προκαθορισμένες κλίμακες μέτρησης και επηρεάζουν μία ή και περισσότερες λανθάνουσες μεταβλητές. Για παράδειγμα, η εξυπηρέτηση και η ποιότητα ενός προιόντος. Είναι μία επιβεβαιωτική παραγοντική διαδικασία,η οποία διευκρινίζει το πόσο καλά μία ερώτηση ή μία μετρήσιμη μεταβλητή συνεισφέρει στο καθορισμό ενός παράγοντα.
   Σε μορφή πίνακα οι εξισώσεις της ανάλυσης σε κύριους άξονες από την μορφή $\boldsymbol{Y = PF + e}$ είναι:
   $\boldsymbol{Y = \Lambda_y \eta + \varepsilon}$για τις ενδογενείς μεταβλητές
   $\boldsymbol{X = \Lambda_x \xi + \delta}$ για τις εξωγενείς μεταβλητές
   Αν έχουμε:
   $\boldsymbol{q}$**:**αρχικές μεταβλητές $X$και $\boldsymbol{p}$**:**αρχικές μεταβλητές $Y$, από τις οποίες προκύπτουν $\boldsymbol{n}$: αφανείς εξωγενείς μεταβλητές $\xi$, και $\boldsymbol{m}$**:**αφανείς ενδογενείς μεταβλητές $n$.
   **Y**: είναι οι μετρήσιμες μεταβλητές οι οποίες συνδέονται με τις ενδογενείς μεταβλητές
   $\boldsymbol{X}$:είναι οι μετρήσιμες μεταβλητές που συνδέονται με τις εξωγενείς μεταβλητές
   $\boldsymbol{P}$: ο πίνακας των συντελεστών παλινδρόμησης
   $\boldsymbol{F}$: το διάνυσμα των μεταβλητών
   $\boldsymbol{e}$: το διάνυσμα των σφαλμάτων
   $\boldsymbol{\Lambda_y}$:Ο πίνακας $(p \times m)$των φορτίσεων των ενδογενών μεταβλητών των αρχικών μεταβλητών $Y$

$\boldsymbol{\Lambda_x}$:Ο πίνακας $(q \times n)$των φορτίσεων των εξωγενών μεταβλητών των αρχικών μεταβλητών $X$.

$\boldsymbol{\eta}$:είναι οι ενδογενείς μεταβλητές (ο πίνακας $m \times 1$ των ενδογενών μεταβλητών)

$\boldsymbol{\xi}$:είναι οι εξωγενείς μεταβλητές (ο πίνακας $n \times 1$των ενδογενών μεταβλητών)

$\boldsymbol{\delta}$:είναι τα σφάλματα μέτρησης (ο πίνακας$q \times 1$των σφαλμάτων των $q$μεταβλητών $X$)

$\boldsymbol{\varepsilon}$:είναι το σφάλμα μέτρησης (ο πίνακας $p \times 1$των σφαλμάτων των $p$ μεταβλητών $Y$.

Οι εξισώσεις **Y** και $\boldsymbol{X}$ πρέπει να εκφραστούν σε πίνακες διακύμανσης-συνδιακύμανσης των μετρήσιμων μεταβλητών. Πολλαπλασιάζοντας από δεξιά κάθε μέρος των εξισώσεων της ανάλυσης σε κύριους άξονες με τον αντιμεταθετικό τους προκύπτουν οι αναμενόμενες τιμές.

Έτσι, για την $\boldsymbol{Y = \Lambda_y \eta + \varepsilon,}$ η εξίσωση είναι:

$$\Sigma_{YY} = \Lambda_Y \eta\eta' \Lambda_Y{}' + \Theta_\varepsilon$$

Για την $\boldsymbol{X = \Lambda_x \xi + \delta,}$ η εξίσωση είναι:

$$\Sigma_{XX} = \Lambda_X \xi\xi' \Lambda_X{}' + \Theta_\delta$$

- ***Το δομικό μοντέλο*** από την άλλη, καθορίζει τις αιτιώδεις σχέσεις μεταξύ των λανθάνουσων μεταβλητών, καθώς και της προσαρμοστικότητας των μοντέλων με τα δεδομένα. Με τον όρο αιτιώδη σχέση ορίζεται η υπόθεση ότι θεωρώντας σταθερά όλα τα υπόλοιπα στοιχεία που καθορίζουν το μοντέλο, μία αλλαγή στην μεταβλητή που βρίσκεται στην ουρά του βέλους, τι είδους μεταβολή θα προκαλέσει στην μεταβλητή που βρίσκεται στο κεφάλι του βέλους και αν αυτή είναι στατιστικά σημαντική. Λανθάνουσες ή ενδογενείς είναι οι μη παρατηρούμενες μεταβλητές, δηλαδή μεταβλητές για τις οποίες στο ερωτηματολόγιο δεν υπάρχει ερώτηση για την απευθείας μέτρησή τους. Για παράδειγμα, η εξυπνάδα και η καταναλωτική αφοσίωση.

  Με άλλα λόγια, μπορούμε να πούμε ότι είναι το μέρος της παλινδρόμησης με λανθάνουσες μεταβλητές των δομικών μοντέλων εξισώσεων. Οι βασικές διορθώσεις ανάμεσα στα δομικά μοντέλα εξισώσεων και στα μοντέλα <<ανάλυσης διαδρομών>> είναι:
  - Οι μεταβλητές στα δομικά μοντέλα εξισώσεων τυπικά δεν μετρώνται και ότι
  - Όλα τα μοντέλα μπορεί κανείς να τα χρειαστεί από μία γενική εξίσωση παλινδρόμησης

Οι μεταβλητές στις εξισώσεις παλινδρόμησης είναι τα η και τα ξ από το μοντέλο μέτρησης. Αυτές οι μεταβλητές σχετίζονται μέσω της γενικής εξίσωσης παλινδρόμησης. Η εξίσωση για το δομικό μοντέλο είναι:

$$\boldsymbol{\eta = \beta\eta + \Gamma\xi + \zeta}$$

Όπου,

$\boldsymbol{\eta}$**:** Διάνυσμα των ενδογενών μεταβλητών

$\boldsymbol{\beta}$**:** Είναι ο πίνακας $(m \times m)$των σχέσεων μεταξύ των ενδογενών μεταβλητών
$\boldsymbol{\Gamma}$**:**Είναι ο πίνακας $(m \times n)$των σχέσεων μεταξύ ενδογενών και εξωγενών μεταβλητών
$\boldsymbol{\zeta}$**:**Είναι ο πίνακας $(m \times 1)$των σφαλμάτων του δομικού μοντέλου

## *5.4: Ανάλυση διαδρομών (Path analysis)*

### *5.4.1: Αναδρομή*

Ο S.Wright ανέπτυξε ένα μοντέλο διαδρομών την περίοδο 1918-1920, το οποίο χρησιμοποιήθηκε για να περιγράψει τις άμμεσες εξαρτήσεις μεταξύ ενός συνόλου μεταβλητών. Χρησιμοποιήθηκε ευρέως στην κοινωνιολογία και σε άλλες επιστήμες και αφορά στη ταυτόχρονη επίλυση ενός συνόλου εξισώσεων παλινδρόμησης το οποίο περιγράφει τις σχέσεις μεταξύ των καταγεγραμμένων μεθόδων που περιλαμβάνονται στο μοντέλο διαδρομών. Επόμένως μπορεί να περιγράψει οποιοδήποτε πλήθος εξαρτημένων αλλά και ανεξάρτητων μεταβλητών. Όπως η μέθοδος παλινδρόμησης έτσι και η μέθοδος ανάλυσης διαδρομών εξετάζει τις σχέσεις μεταξύ των μεταβλητών στοχεύοντας στην περιγραφή των άμμεσων και έμμεσων επιδράσεων μεταξύ των μεταβλητών του μοντέλου. Η ανάλυση διαδρομών αφορά τη συσχέτιση και όχι την αιτιότητα μεταξύ των μεταβλητών.

Ο Goodman το 1973 εξέτασε την ανάλυση διαδρομών των δυαδικών μεταβλητών με την χρησιμοποίηση λογισμικών προτύπων παλινδρόμησης και συζήτησε τα αποτελέσματα των παραμέτρων. Από την άλλη, ο Hagennars το 1998 έκανε μία γενική συζήτηση της ανάλυσης διαδρομής των επαναλαμβανόμενων περιστασιακών συστημάτων των κατηγορικών μεταβλητών με την χρησιμοποίηση της κατευθυνόμενης γραμμικής λογαριθμικής πρότυπης προσέγγισης.

Η ανάλυση διαδρομών είναι μία στατιστική τεχνική που χρησιμοποιείται για να εξετάσει αιτιώδεις σχέσεις μεταξύ δύο ή περισσοτέρων μεταβλητών, ουσιαστικά είναι μία τεχνική παλινδρόμησης και ένα παρακλάδι της οικογένειας δομικών μοντέλων εξισώσεων. Παρά το γεγονός ότι είναι το παλαιότερο μέλος της οικογένειας των δομικών μοντέλων εξισώσεων δεν είναι καθόλου ξεπερασμένο. Στόχος της είναι να παράγει εκτιμήσεις για το μέγεθος και τη σημασία της υποθετικής αιτιώδης σχέσης μεταξύ μεταβλητών.

Ο Ράιτ έκανε προσέγιση της ανάλυσης διαδρομής με τον δικό του τρόπο, ο οποίος έχει τα εξής βήματα:

1. Γράφονται εξισώσεις του μοντέλου που αφορούν μετρήσιμες μεταβλητές
2. Ασχούνται συσχετίσεις μεταξύ τους ως προς τις άγνωστες παραμέτρους του μοντέλου
3. Προσπαθούν να λύσουν το σύστημα των εξισώσεων (ένα κάθε φορά στο οποίο οι συσχετισμοί αντικαθίστανται από τις συσχετίσεις του δείγματος).

Για να δομηθεί μία ανάλυση διαδρομών γράφουμε αρχικά τις μεταβλητές και βάζουμε βέλη προς τις μεταβλητές που υποθέτουμε ότι σχετίζονται, έτσι δημιουργείται διάγραμμα εισροών και εκροών.

Το διάγραμμα εισροών δημιουργείται πρώτο, αντιπροσωπεύει τις αιτιώδεις σχέσεις και σκοπός του είναι να μας βοηθήσει στην περαιτέρω ανάλυση.

Το διάγραμμα εκροών απεικονίζει τα αποτελέσματα της στατιστικής ανάλυσης και δείχνει εάν επαληθεύονται οι υποθέσεις.

### *5.4.2: Περιορισμοί στην εφαρμογή της*

Χρησιμοποιείται σε πολλούς τομείς της τεχνολογίας λόγω της απλής μοντελοποίησης που προσφέρει, όμως έχει και κάποιους περιορισμούς.

Οι περιορισμοί είναι οι εξής:

- Μπορούμε να βρούμε τις αιτιώδεις σχέσεις αλλά όχι την κατεύθυνσης της αιτιότητας.
- Μπορεί να χρησιμοποιηθεί καλύτερα όταν έχουμε μικρό αριθμό υποθέσεων.
- Δεν χρησιμοποιείται όταν στο διάγραμμα υπάρχουν βρόγχοι οι οποίοι αλλάζουν συνεχώς, αλλά πρέπει να χρησιμοποιηθεί όταν έχουμε ξεκαθαρίσει τις υποθέσεις και έχουμε μία κατά το δυνατόν στατιστική κατάσταση.

Στην στατιστική, η ανάλυση διαδρομής χρησιμοποιείται για να περιγράψει τις κατευθυνόμενες εξαρτήσεις ανάμεσα σε ένα σύνολο μεταβλητών. Αυτό περιλαμβάνει τα μοντέλα με οποιαδήποτε μορφή της πολλαπλής ανάλυσης παλινδρόμησης, ανάλυση παραγόντων, κανονική ανάλυση συσχέτισης, διακριτική ανάλυση και γενικότερα τις οικογένειες των μοντέλων στην πολυμεταβλητή ανάλυση διακύμανσης και συνδιακύμανσης.

**Θεωρείται ειδική περίπτωση των δομικών μοντέλων εξισώσεων μόνο και μόνο λόγω των δεικτών που χρησιμοποιούνται για κάθε μια από τις μεταβλητές στο μοντέλο αιτιώδους συνάφειας. Δηλαδή, η ανάλυση διαδρομής από την μία είναι το SEM με ένα δομικό μοντέλο αλλά δεν υπάρχει μοντέλο μέτρησης. Άλλοι όροι που μπορούν να πουν ότι περιλαμβάνει η ανάλυση διαδρομής αλλά και χρησιμοποιούνται σε αυτή είναι τα αιτιώδη μοντέλα, η ανάλυση των δομών συνδιακύμανσης και τα μοντέλα λανθάνουσων μεταβλητών.**

Η ανάλυση διαδρομής είναι μία επέκταση του μοντέλου παλινδρόμησης και χρησιμοποιείται για να εξεταστεί η προσαρμογή της συσχέτισης εναντίον δύο ή

περισσοτέρων μοντέλων συνάφειας. Το μοντέλο συνήθως απεικονίζεται σε σχήμα κύκλου και μονών η διπλών βελών, εκ των οποίων τα μονά βέλη αντιπροσωπεύουν τους συντελεστές παλινδρόμησης και τα διπλά βέλη δείχνουν τις συνδιακυμάνσεις μεταξύ των παραγόντων. Τα βέλη επίσης συνδέουν τους όρους σφάλματος με τις αντίστοιχες ενδογενείς μεταβλητές. Μερικές φορές το πλάτος των βελών στο μοντέλο διαδρομής σύρεται σε ένα πλάτος το οποίο είναι ανάλογο προς το απόλυτο μέγεθος των αντίστοιχων συντελεστών διαδρομής.

Το μοντέλο διαδρομής που δημιουργείται στα δομικά μοντέλα εξισώσεων επιτρέπει την μελέτη επίδρασης των επιμέρους μεταβλητών στις οποίες μπορεί να αποσυντεθεί μία αρχική μεταβλητή. Με άλλα λόγια μία ανεξάρτητη μεταβλητή μπορεί να έχει άμμεση είτε έμμεση επίδραση σε μία εξαρτημένη μεταβλητή. Μπορούμε να αναφέρουμε ότι τα μοντέλα διαδρομής θεωρούνται ειδική περίπτωση των μοντέλων διαρθρωτικής εξίσωσης. Για να αναλύσει κανείς ένα μοντέλο διαδρομής μπορεί να χρησιμοποιήσει προγράμματα όπως: SEM, EQS, LISREL.

Το μοντέλο διαδρομής είναι ένα διάγραμμα που σχετίζεται με ανεξάρτητες, ενδιάμεσες και εξαρτημένες μεταβλητές. Ενιαία βέλη δείχνουν την αιτιώδη συνάφεια μεταξύ των εξωγενών ή ενδιάμεσων μεταβλητών και της εξαρτημένης μεταβλητής. Τα βέλη συνδέουν τους όρους σφάλματος με τις αντίστοιχες εξωγενείς μεταβλητές. Διπλά βέλη δείχνουν συσχέτιση μεταξύ των ζευγών των εξωγενών μεταβλητών . Μερικές φορές το πλάτος των βελών στο μοντέλο διαδρομής **μετατρέπεται** σε ένα πλάτος που είναι ανάλογο του απόλυτου μεγέθους των αντίστοιχων συντελεστών διαδρομής.

***5.4.3: Διαφορές μεταξύ ανάλυσης διαδρομής και δομικών μοντέλων εξίσωσης:***

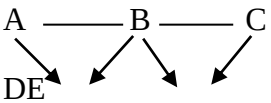
- Η ανάλυση διαδρομής είναι μία ειδική περίπτωση των SEM.
- Περιέχει μόνο παρατηρούμενες μεταβλητές και κάθε μεταβλητή έχει μόνο έναν δείκτη ανάλυσης.
- Η ανάλυση διαδρομής υποθέτει ότι όλες οι μεταβλητές μετριούνται χωρίς σφάλματα.
- Τα SEMχρησιμοποιούν λανθάνουσες μεταβλητές για να λογοδοτήσουν για το σφάλμα της μέτρησης.
- Η ανάλυση διαδρομής έχει ένα περιοριστικό σύνολο υποθέσεων από SEM.
- Τα περισσότερα από τα μοντέλα που υπάρχουν στις βιβλιογραφίες είναι SEMκαι όχι αναλύσεις διαδρομής

## *5.5:Αιτιώδες διαδρομές(casual paths)*

Αιτιώδες διαδρομές για μία δεδομένη μεταβλητή περιλαμβάνει:

1. Άμμεσες διαδρομές από τα βέλη που οδηγούν σε αυτό
2. Συσχετιζόμενες διαδρομές από ενδογενείς μεταβλητές με βέλη που οδηγούν στην δεδομένη μεταβλητή.

Για παράδειγμα




A,B,C:εξωγενείς μεταβλητές

D,E: ενδογενείς μεταβλητές, όχι όροι σφάλματος

Έχουμε τις διαδρομές από το A εως το D, από το B στο A και οι διαδρομές από το B στο A εώς D και από το C στο A εως D και από το C στο B με τον D. Οι διαδρομές που εμπλέκονται δύο συσχετίσεις είναι από το C στο B, A εως D δεν είναι σχετικές. Ομοίως για τις διαδρομές που πηγαίνουν προς τα πίσω (E στο B εως D ή από το B στο A εως D).

### *5.5.1:Εξωγενείς και ενδογενείς μεταβλητές*

Οι εξωγενείς μεταβλητές σε ένα μοντέλο διαδρομής είναι εκείνες με τις ρητές αιτίες(δεν υπάρχουν βέλη μεταξύ τους εκτός από την μέτρηση του λάθους). Εάν οι εξωγενείς μεταβλητές συσχετίζονται αυτό υποδεικνύεται από ένα διπλό βέλος που τις συνδέει.

Έπειτα οι ενδογενείς μεταβλητές είναι εκείνες που έχουν τα εισερχόμενα βέλη. Περιλαμβάνουν μεταβλητές που μεσολάβησαν σαν και τις εξαρτώμενες. Οι ενδογενείς μεταβλητές, είναι οι μεταβλητές που καθορίζονται στο πλαίσιο του μοντέλου.

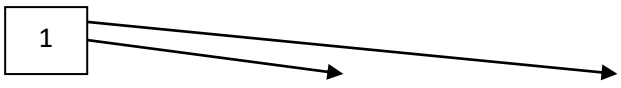
### *5.5.2:Συντελεστής διαδρομής(coefficient path)*

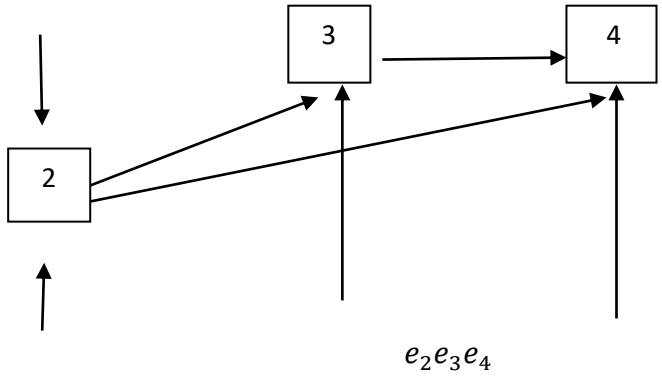
Ένας συντελεστής διαδρομής είναι ένας τυποποιημένος συντελεστής παλινδρόμησης που δείχνει το άμμεσο αποτέλεσμα μιας ανεξάρτητης σε μια εξαρτημένη μεταβλητή στο μοντέλο διαδρομής. Έτσι όταν το μοντέλο έχει δύο ή περισσότερες μεταβλητές συνάφειας, οι συντελεστές διαδρομής είναι μερικοί συντελεστές παλινδρόμησης που μετρούν την έκταση της επίδρασης μιας μεταβλητής σε ένα άλλο μοντέλο διαδρομής. Για διμεταβλητή παλινδρόμηση όπου ο συντελεστής β για τυποποιημένα δεδομένα είναι το ίδιο με τον συντελεστή συσχέτισης, έτσι για μεταβλητή εξαρτώμενη από μία μόνο εξωγενή μεταβλητλη ο συντελεστής διαδρομής σε αυτή την ειδική περίπτωση είναι ένας συντελεστής μηδενικής τάξης.

### *5.5.3: Παραδείγματα*

*Παράδειγμα 1:*

Έστω ότι έχουμε 3 ενδογενείς μεταβλητές (2,3,4) και μία εξωγενή μεταβλητή (1). Όπως φαίνεται από το παρακάτω διάγραμμα:






*Εικόνα 13: Απεικόνιση μεταβλητών*

Όπως φαίνεται από το παραπάνω διάγραμμα δεν υπάρχουν προς τα πίσω διαγράμματα. Και όπως αναφέρθηκε απεικονίζονται 3 ενδογενείς μεταβλητές και 1 εξωγενής. Κάθε ενδογενής μεταβλητή απεικονίζεται από μία ή περισσότερες μεταβλητές στο μοντέλο συν τους όρους σφάλματος ($e_2 - e_4$). Ακόμη, μία ενδογενής μεταβλητή μπορεί να είναι αιτία μιας άλλης ενδογενής μεταβλητής αλλά όχι μιας εξωγενούς μεταβλητής. Έχουμε τις εξής διαδρομές:

- 1 έως 2,3,4
- 2 έως 3,4
- 3 έως 4

Μερικές παραδοχές για το μοντέλο διαδρομής:

1. Όλες οι σχέσεις είναι γραμμικές
2. Οι όροι σφάλματος είναι ασυσχέτιστες με τις υπόλοιπες μεταβλητές στο μοντέλο.
3. Η αιτιώδης ροή είναι μονόδρομος
4. Η μέτρηση των μεταβλητών έγινε στο διάστημα της κλίμακας
5. Οι μεταβλητές μετριούνται χωρίς σφάλματα

Υπολογισμός συντελεστών διαδρομής:

Οι μεταβλητές μας είναι σε μία τυποποιημένη μορφή σκορ (βαθμολογία $z$). Για το παράδειγμα μας, οι εξισώσεις για τις τέσσερις μεταβλητές είναι:

$$z_1 = e_1$$
$$z_2 = p_{21}z_1 + e_2$$
$$z_3 = p_{31}z_1 + p_{32}z_2 + e_3$$
$$z_4 = p_{41}z_1 + p_{42}z_2 + p_{43}z_3 + e_4$$

Η πρώτη μεταβλητή δεν εξηγείται από οποιαδήποτε άλλη μεταβλητή στο μοντέλο. Στη γλώσσα διαδρομής, το $e$ σημαίνει ότι προκαλεί έξω από το μοντέλο. Η δεύτερη μεταβλητή οφείλεται εν μέρη στην πρώτη μεταβλητή και εν μέρει σε σφάλμα ή

ανεξήγητες αιτίες. Κάθε $z$καθορίζεται από τις διαδρομές που οδηγούν κατευθείαν σε αυτό, και όχι τις έμμεσες διαδρομές.Παρακάτω δίνονται και οι συσχετίσεις:

$$r_{12} = \frac{1}{N}\sum z_1 z_2$$

Η οποία είναι η φόρμουλα για$r$με βαθμολογία $z$.

*Για την εξίσωση διαδρομής $z_2$, έχουμε:*

$$r_{12} = \frac{1}{N}\sum z_1(p_{21}z_1 + e_2) \; <=> r_{12} = p_{21}\frac{\sum z_1 z_1}{N} + \frac{\sum z_1 e_2}{N}$$

Στην παραπάνω εξίσωση ο 1ος όρος είναι οι χρόνοι των συντελεστών διαδρομής, που είναι η διακύμανση των $z_1$. Ενώ ο 2ος όρος είναι η συσχέτιση μεταξύ $z_1$ και $e_2$, όμως γνωρίζουμε πως αυτή η συσχέτιση είναι μηδενική, διότι αυτό είναι μία από τις παραδοχές της ανάλυσης διαδρομής. Έτσι, με βάση τα παραπάνω αν έχουμε να κάνουμε με τα αποτελέσματα $z$ο συντελεστής διαδρομής θα είναι: $r_{12} = p_{21}1 + 0$

*Για την εξίσωση διαδρομής $z_3$, έχουμε:*

$$r_{13} = \frac{1}{N}\sum z_1 z_3 <=> r_{13} = \frac{1}{N}\sum z_1(p_{31}z_1 + p_{32}z_2)$$
$$<=> r_{13} = p_{31}\frac{\sum z_1^2}{N} + p_{32}\frac{\sum z_1 z_2}{N}$$
$$<=> r_{13} = p_{31} + p_{32}r_{12}$$

Καθώς δεν γνωρίζουμε τους συντελεστές $p_{31}$ και $p_{32}$, μπορούμε να ορίσουμε και άλλη μία εξίσωση ως προς την συσχέτιση. Έχουμε:

$$r_{23} = \frac{1}{N}\sum z_2 z_3 <=> r_{23} = \frac{1}{N}\sum z_2(p_{31}z_1 + p_{32}z_2)$$
$$<=> p_{31}\frac{\sum z_2 z_1}{N} + p_{31}\frac{z_2^2}{N} <=>$$
$$<=> r_{23} = p_{31}r_{12} + p_{32}$$

Έχουμε:

$\boldsymbol{r_{13} = p_{31} + p_{32}r_{12}}$ **(1)**

$\boldsymbol{r_{23} = p_{31}r_{12} + p_{32}}$ **(2)**

Η **(1)** γίνεται: $\boldsymbol{r_{13} - p_{32}r_{12} = p_{31}}$ **(3)**

**Λόγω των (1), (3) η (2) γίνεται:**

$\boldsymbol{r_{23} = (r_{13} - p_{32}r_{12})r_{12} + p_{32} <=> p_{32} = \frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}}$.

Ακόμη:

$$r_{14} = \frac{1}{N}\sum z_1 z_4 <=> r_{14} = \frac{1}{N}\sum z_1(p_{41}z_1 + p_{42}z_2 + p_{43}z_3)$$
$$<=> r_{14} = p_{41}\frac{\sum z_1^2}{N} + p_{42}\frac{\sum z_1 z_2}{N} + p_{43}\frac{\sum z_1 z_3}{N}$$

Οπότε έχουμε:

$$r_{14} = p_{41}1 + p_{42}r_{12} + p_{43}r_{13}$$

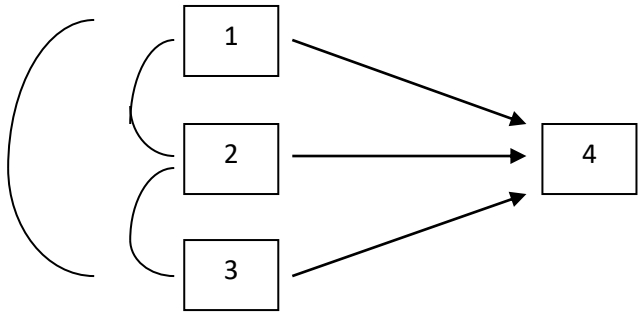
Οι άλλες δύο συσχετίσεις είναι:

$$r_{24} = p_{41}r_{12} + p_{42} + p_{43}r_{23}$$

$$r_{34} = p_{41}r_{13} + p_{42}r_{23} + p_{43}$$

Αν θεωρήσεουμε την 4η μεταβλητή ως κύρια, και η οποία να δέχεται μόνο επιδράσεις αλλά όχι να δίνει στις άλλες μεταβλητές.

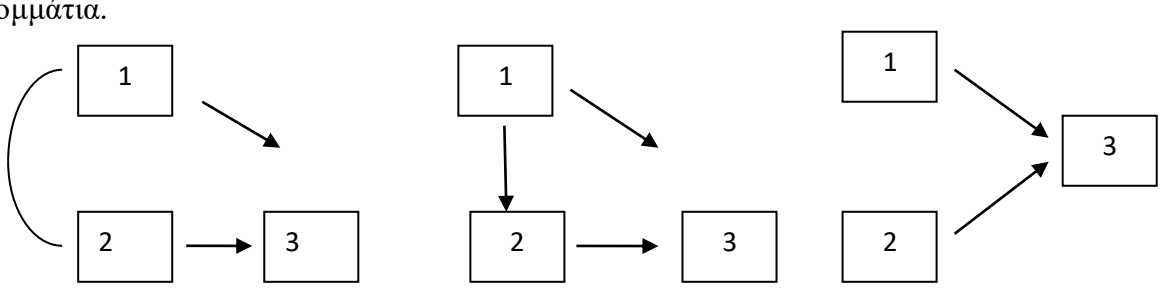
Έχουμε:




*Εικόνα 14: Διάγραμμα διαδρομής*

*Αποσύνθεση συσχετίσεων*

Ένα διάγραμμα διαδρομής συνεπάγεται ότι οι συσχετίσεις χτίζονται από πολλά κομμάτια.




*Α.ΣυσχέτισηΒ. ΜεσολάβησηΓ. Ανεξάρτητα*

*Εικόνα 15: Απεικόνιση των 3 μοντέλων*

Στο συσχετιζόμενο μοντέλο Α, το μέρος της συσχέτισης μεταξύ 1 και 3 οφείλεται στην άμμεση επίδραση της 1 στις 3 (μέσω $p_{31}$). Μέρος της συσχέτισης αυτής θα οφείλεται στη συσχέτιση του 1 με 2, επειδή η 2 επηρεάζει επίσης την 3, δηλαδή

$r_{12}p_{32}$. Όμως ο παραπάνω ο συσχετισμός του μοντέλου δεν μπορεί να αναλυθεί καθώς οι 1,2 είναι εξωγενείς.

Στο μοντέλο μεσολάβησης Β, μόνο η μεταβλητή 1 είναι εξωγενής. Σε αυτό μοντέλο η 1 επηρεάζει άμεσα την 3 ($p_{31}$)αλλά και έμμεσα μέσω της 2 ($p_{21}$ $και$ $p_{32}$), άρα η συσχέτιση μεταξύ των 2,3 αποτελείται από άμεσες και έμμεσες επιδράσεις. Ακόμη μεταξύ τους υπάρχει η συσχέτιση ($r_{23}$). Η συσχέτιση αυτή θα αντικατοπτρίζει το άμεσο αποτέλεσμα της 2 στη 3 ($p_{32}$). Εάν η διαδρομή από την 2 στην 3 ήταν μηδέν, ολόκληρη η συσχέτιση μεταξύ 2 και 3 θα λήταν ψευδή, επειδή αυτό οφείλεται στην μεταβλητή 1. Ωστόσο, στο συγκεκριμένο μοντέλο μόνο ένα μέρος συσχέτισης μεταξύ 2 και 3 είναι πλαστό. Το πλαστό μέρος είναι: $r_{23} - p_{32}$ή $p_{31}p_{21}$.

Στο μοντέλο Γ οι μεταβλητές 1,2 είναι ανεξάρτητες, σε μία τέτοια περίπτωση ο συντελεστής διαδρομής είναι ίσος με την παρατηρούμενη συσχέτιση. Η παρατηρούμενη συσχέτιση αποτελείται από 4 κομμάτια:

1. ΄Αμεσο αποτέλεσμα (DE) λόγω της διαδρομής
2. Έμμεση επίδραση (IE) λόγω διαδρομών μέσω ενδιάμεσων μεταβλητών
3. Υπερανάλυση (U) λόγω συσχετισμένων εξωγενών μεταβλητών
4. Οι παρασιτικές (S) λόγω τρίτης μεταβλητής αιτιών.

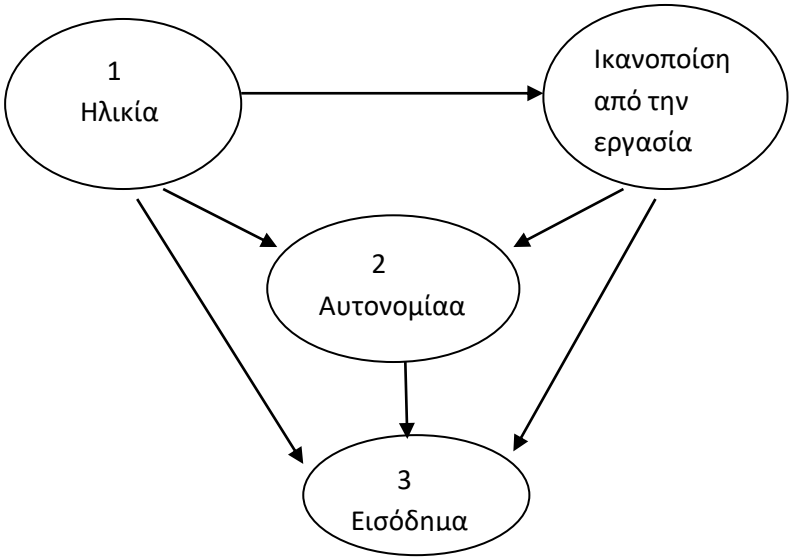
Μία σειρά από εξισώσεις με βάση το 1ο σχήμα είναι :

| | |
|---|---|
| $r_{12} = p_{21}$ | $r_{14} = p_{41} + p_{42}r_{12} + p_{43}r_{13}$ |
| $r_{13} = p_{31} + p_{32}r_{12}$ | $r_{24} = p_{41}r_{12} + p_{42} + p_{43}r_{23}$ |
| $r_{23} = p_{31}r_{12} + p_{32}$ | $r_{34} = p_{41}r_{13} + p_{42}r_{23} + p_{43}$ |

*Πίνακας 2:Πίνακας εξισώσεων*

*<u>Παράδειγμα 2:</u>*

Έστωμοντέλομετιςεξήςεξισώσειςπαλινδρόμησης:




*Εικόνα 16: Αεπικόνιση των μεταβλητών*

Εξίσωση 1:$Ικανοποίηση = \beta_{11}ηλικία + \beta_{12}αυτονομία + \beta_{13}εισόδημα + \varepsilon_1$

Εξίσωση 2:$Εισόδημα = \beta_{21}ηλικία + \beta_{22}αυτονομία + \varepsilon_2$

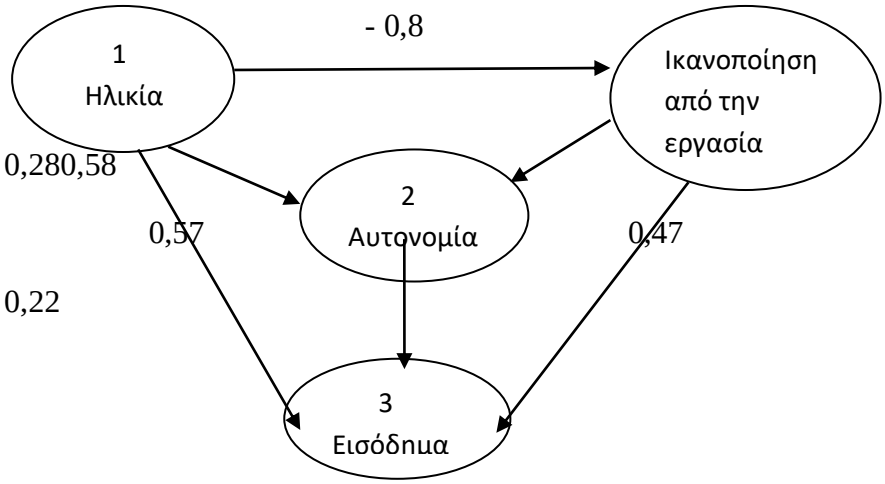
Εξίσωση 3:$Αυτονομία = \beta_{31}ηλικία + \varepsilon_3$

όπου, $\beta_{11}, \beta_{12}, \beta_{13}, \beta_{21}, \beta_{22}, \beta_{31} = συντελεστέςπαλινδρόμησης$

για παράδειγμα, $\beta_{21}$:είναι ο συντελεστής στην εξίσωση 2 για την μεταβλητή 1 η οποία είναι η ηλικία.

Οι συντελεστές διαδρομής που είναι οι βήτα συντελεστές σε αυτές τις εξισώσεις, είναι έτσι οι τυποποιημένοι μερικοί συντελεστές παλινδρόμησης της κάθε προηγούμενης ενδογενής μεταβλητής. Δηλαδή, ο συντελεστής διαδρομής είναι το μερικό βάρος λαμβάνοντας υπόψη και προηγούμενες μεταβλητές για την δεδομένη εξαρτημένη μεταβλητή.

Παλαιότερα ονομάζονταν p συντελεστές, τώρα οι συντελεστές διαδρομής ονομάζονται απλά βάρη, με βάση τη χρήση σε πολλαπλά μοντέλα παλινδρόμησης. Σύμφωνα με τον Bryman και Cramer οι συντελεστές διαδρομής=τυποποιημένους συντελεστές παινδρόμησης= βήτα βάρη να είναι:




*Εικόνα 17: Διάγραμμα με τα βάρη*

Συσχετιζόμενες εξωγενείς μεταβλητές: Εάν οι εξωγενείς μεταβλητές συσχετίζονται, είναι σύνηθες να ονομαστεί το αντίστοιχο διπλό βέλος μεταξύ τους συντελεστής συσχέτισης.

Όροι διαταραχής: Ονομάζονται αλλιώς υπολειμματικοί όροι σφάλματος , αντανακλούν ανεξήγητη διακύμανση στο σφάλμα μέτρησης. Η συσχέτιση μεταξύ δύο όρων διαταραχής είναι η μερική συσχέτιση των δύο ενδογενών μεταβλητών, χρησιμοποιώντας όλες τις κοινές τους αιτίες.Οι όροι διαταρραχής πρέπει να είναι θετικά ορισμένοι.Είναι ουσιαστικά οι μη παρατηρούμενες εξωγενείς μεταβλητές και αποτελούν παράγοντες που παραλείπονται από το μοντέλο, αλλά κρίνονται ότι είναι σημαντικές για την εξήγηση της συμπεριφοράς των μεταβλητών στο μοντέλο.

### *5.5.4: Κανόνας διαδρομής πολλαπλασιασμού*

Η αξία της κάθε ένωσης διαδρομής είναι το προιόν των συντελεστών διαδρομής του. Έστω μία απλή διαδικασία τριών μεταβλητών σύνθετης διαδρομής όπου η εκπαίδευση προκαλεί αιτίες,εισόδημα, συντηριτισμό.

Αν ο συντελεστής παλινδρόμησης του εισοδήματος για την εκπαίδευση είναι 1000: για κάθε έτος της εκπαίδευσης το εισόδημα ανεβαίνει 1000$.

Αν ο συντελεστής παλινδρόμησης του συντηρισμού στο εισόδημα είναι 0,0002: για κάθε εισόδημα το δολάριο ανεβαίνει, ανεβαίνει ο συντηριτισμός 0,0002 μονάδες σε μία κλίμακα 5 σημείων.

Έτσι αν η εκπαίδευση συνεχίζεται μέχρι 1 χρόνο, το εισόδημα ανεβαίνει 1000$ που σημαίνει ότι ο συντηριτισμός ανεβαίνει 0,2 πόντους. Είναι το ίδιο μετο αν ο πολλαπλασιαμός των συντελεστών είναι: 1000*0,0002=0,2. Το ίδιο θα ίσχυε αν υπήρχαν περισσότερες σχέσεις ή συνδέσεις στην διαδρομή.

Αν χρησιμοποιηθούν τυποποιημένοι συντελεστές διαδρομής, ο κανόνας διαδρομής πολλαπλασιασμού θα εξακολουθεί να ισχύει αλλά η ερμηνεία θα είναι σε τυποποημένους όρους.

### *5.5.5: Αποτέλεσμα αποσύνθεσης*

Οι συντελεστές διαδρομής μπορεί να χρησιμοποιηθούν για να αποσυντεθούν οι συσχετισμοί στο μοντέλο σε άμμεσες και έμμεσες επιδράσεις, που αντιστοιχούν φυσικά στις άμμεσες και έμμεσες διαδρομές που αντανακλώνται στα βέλη στο μοντέλο. Αυτό βασίζεται στο γεγονός ότι σε ένα γραμμικό σύστημα, η συνολική αιτιώδης επίδραση της μεταβλητής i σε μεταβλητή j είναι το άθροσιμα όλων των από το i στο j. Στο παραπάνω μοντέλο έχοντας υπόψη την <<ικανοποίηση>> ως εξαρτημένη μεταβλητή και την <<ηλικία>> ως ανεξάρτητη, οι έμμεσες επιπτώσεις υπολογίζονται με πολλαπλασιασμό των συντελεστών διαδρομής για κάθε διαδρομή από την ηλικία για την ικανοποίηση:

Ηλικία→εισόδημα→ικανοποίηση είναι $0{,}57 * 0{,}47 = 0{,}26$

ηλικία → αυτονομία→ ικανοποίηση είναι $0{,}28 * 0{,}58 = 0{,}16$

ηλικία→ αυτονομία →εισόδημα →ικανοποίηση είναι $0{,}28 * 0{,}22 \times 0{,}47 = 0{,}03$

συνολικό έμμεσο αποτέλεσμα$= 0{,}45$

Δηλαδή, η συνολική έμμεση επίδραση της ηλικίας στην ικανοποίηση είναι$+0{,}45$. Σε σύγκριση, το άμεσο αποτέλεσμα είναι $-0{,}08$.Η συνολική αιτιώδης επίδραση της ηλικίας στην ικανοποίηση είναι$(-0{,}8 + 0{,}45) = 0{,}37$. Η επίδραση αποσύνθεσης είναι ισοδύναμη με τα αποτελέσματα ανάλυσης παλινδρόμησης με μία εξαρτημένη μεταβλητή, όπως επίσης η ανάλυση διαδρομής μπορεί να χειριστεί επίδραση αποσύνθεσης για την περίπτωση δύο ή περισσοτέρων εξαρτημένων μεταβλητών.

Σε γενικές γραμμές, κάθε διμεταβλητή συσχέτιση μπορεί να αναλυθεί σε ψευδείς και συνολικής αιτιώδους συνάφειας και το συνολικό αποτέλεσμα αιτιώδης μπορούν να αναλυθούν σε άμμεση και έμμεση επίδραση. Η συνολική αιτιώδης επίδραση είναι ο συντελεστής με όλους εκ των προτέρων σε μία παλινδρόμηση αλλά δεν παρεμβαίνει στις μεταβλητές του μοντέλου για x και y. Το ψευδές αποτέλεσμα είναι το συνολικό αποτέλεσμα μείον το συνολικό αποτέλεσμα συνάφειας. Το άμμεσο αποτέλεσμα είναι ο μερικός συντελεστής για τον έλεγχο y επί του x για όλες τις προηγούμενες μεταβλητές και όλες τις παρεμβενόμενες μεταβλητές στο μοντέλο. Το έμμεσο αποτέλεσμα είναι το συνολικό αιτώδες αποτέλεσμα μείον το άμμεσο αποτέλεσμα και μετρά την επίδραση των παρεμβαινόντων μεταβλητών.

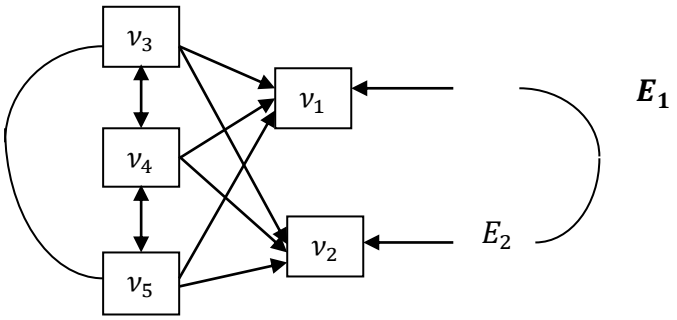
### *5.5.6: Παράδειγμα ανάλυσης διαδρομών*

Εξέταση των επιδράσεων πολών μεταβλητών στην ακαδημαική απόδοση πρωτοετών φοιτητών πανεπιστημίου.

5 εκπαιδευτικά μέτρα συλλέχθηκαν από δείγμα $\mathrm{N} = 150$ πρωτοετών φοιτητών. Παρατηρούμενες μεταβλητές:

1. Μέσος βαθμός που λαμβάνουν στα υποχρεωτικά μαθήματα
2. Μέσος βαθμός που λαμβάνουν στα κατ'επιλογήν υποχρεωτικά μαθήματα
3. Βαθμολογία γυμνασίου(γενικές γνώσεις)
4. Νοημοσύνη και βαθμολογία κατά το τελευταίο έτος λυκείου
5. Εκπαιδετυική βαθμολογία στο τελευταίο έτος του γυμνασίου

$\nu_1 - \nu_5$: παρατηρούμενες μεταβλητές, $E_1 - E_5$: όροι σφάλματος




*Εικόνα 18: Σχέση μεταβλητών*

Υπάρχουν βέλη που συνδέουν τις ανεξάρτητες μεταβλητές. Δεν υπάρχουν υπολείμματα ή σφάλαμτα μέτρησης που σχετίζονται με κάποιες από τις ανεξάρτητες μεταβλητές.
Οι εξαρτημένες μεταβλητές συνδέονται με τους εναπομένοντες όρους. Περιέχουν τόσο το σφάλμα μέτρησης και πρόβλεψης. Το καμπυλωτό βέλος που συνδέει τους όρους σφάλματος $E_1$ και $E_2$ αντιπροσωπεύει την πιθανή συσχέτισης τους, που έχει συμπεριληφθεί στο μοντέλο.

Ο σκοπός αυτής της ανάλυσης διαδρομής είναι για αν εξετάσει τη δύναμη των πολλών μεταβλητών στην ακαδημαική απόδοση πρωτοετών φοιτητών πανεπιστημίου.
$\nu_1, \nu_2$:εξαρτημένες μεταβλητές

$\nu_3, \nu_4, \nu_5$:ανεξάρτητες μεταβλητές

$E_1, E_2$:υπολειματικοί όροι

## 5.6: Ανάλυση παραγόντων (Factor Analysis)

Η ανάλυση παραγόντων ή αλλιώς παραγοντική ανάλυση είναι παλαιότερη και η πιο γνωστή μέθοδος για τη διερεύνηση των σχέσεων ανάμεσα σε σύνολα από παρατηρήσιμες και λανθάνουσες μεταβλητές. Εξετάξει την συνδιακύμανση ανάμεσα στις παρατηρήσιμες μεταβλητές με σκοπό να συλέξει πληροφορίες για τις υποκείμενες λανθάνουσες μεταβλητές τους (δηλαδή τους παράγοντες).Έχει αναλυθεί περαιτέρω η παραγοντική ανάλυση και στο 2ο κεφάλαιο.

### *5.6.1:Υπόδειγμα μεθόδου*

Έστω ένα σύνολο μεταβλητών $Y = [Y_1, Y_2, \dots, Y_\rho]'$ με μέση τιμή $\mu$ και πίνακα διασποράς-συνδιασποράς $\Sigma$, σύφωνα με το γενικό μοντέλο της παραγοντικής ανάλυσης μπορούν να δημιουργηθούν $m$ −παράγοντες οι οποίοι απεικονίζονται ως $F = [F_1, F_2, \dots, F_m]'$. Οπότε, το υπόδειγμα θα έχει την μορφή:

$$\boldsymbol{X_i = \sum_{j=1}^{m} W_{ij} F_j + \varepsilon_i = W_{i1} F_1 + W_{i2} F_2 + \cdots + W_{im} F_m + \varepsilon_i} \textbf{ (5.6.1)}$$

Όπου,$W_{ij}$:πίνακας των συντελεστών των παραγόντων
$\varepsilon$: είναι το διάνυσμα των σφαλμάτων
Το σφάλμα $\varepsilon_i$ θεωρείται ο μοναδικός παράγοντας της $i$ παρατήρησης και είναι το μέρος της μεταβλητής που δεν μπορεί να εξηγηθεί από τον παράγοντα (Johnson 1998).

Η τιμή των παραγόντων $m$πρέπει να είναι μικρότερη από το πλήθος των μεταβλητών $p$, διαφορετικά δεν επιτυγχάνεται περικοπή του όγκου του προβλήματος αλλά απλά έναν διαχωρισμό τους.

### *5.6.2: Διαχωρισμός ανάλυσης παραγόντων*

Η ανάλυση παραγόντων χωρίζεται σε δύο βασικούς τύπους:

- Η διερευνητική ανάλυση παραγόντων
- Η επιβεβαιωτική ανάλυση παραγόντων

*5.6.2.1: Η διερευνητική ανάλυση παραγόντων*

Η μέθοδος αυτή βασίζεται στην υπόθεση ότι η διακύμανση μιας παρατηρήσιμης μεταβλητής είναι συνάρτηση ενός αριθμού από παράγοντες(λανθάνουσες

μεταβλητές), οι οποίοι αντιστοιχούν στις ποικίλες διαστάσεις της επίδοσης που εκπροσωπεί αυτή η μεταβλητή.

Για παράδειγμα, ένας αριθμός από παρατηρήσιμες μεταβλητές, που αντιστοιχούν σε μετρήσεις της ικανότητας εκτέλεσης αριθμητικών πράξεων θα μπορούσε να αναχθεί σε δύο τουλάχιστον παράγοντες: ο ένας θα εκπροσωπούσε τη γενική μαθηματική ικανότητα και ο άλλος την ικανότητα συγκράτησης στην μνήμη των αριθμών και των διαδικασιών εκτέλεσης των πράξεων για όσο διάστημα απαιτείται ώσπου να γίνει η αριθμητική πράξη (Demetriou, Efklidew, &Platsidou, 1993). Κατά τον ίδιο τρόπο, οι παρατηρήσιμες μεταβλητές σε ένα δείγμα δεδομένων μπορεί να είναι οι αυτοαναφορές των εφήβων για το φόβο αποτυχίας, την ανάγκη για επιτυχία, το άγχος ως ανησυχία, το άγχος ως συναισθηματικότητα και το άγχος ως κατάσταση. Το σύνολο αυτών των μεταβλητών θα μπορούσε να αποτελεί δύο παράγοντες: ο ένας θα εκπροσωπούσε τον προσανατολισμό προς την επιτυχία και ο άλλος το άγχος .

Η εφαρμογή της μεθόδου αυτής συνιστάται όταν ο τρόπος με τον οποίο οι παρατηρήσιμες μεταβλητές ανάγονται σε λανθάνουσες δομές είναι άγνωστος. Η προσέγγιση θεωρείται διερευνητική καθώς δεν υπάρχει καμία προηγούμενη γνώση για το πως οι παρατηρήσιμες μεταβλητές μπορούν να χρησιμεύσουν ως μέτρηση των παραγόντων. Το μοντέλο που προκύπτει δείχνει τις συνδέσεις ανάμεσα στις παρατηρήσιμες μεταβλητές και τους παράγοντες, καθώς και το πρότυπο συσχετίσεων μεταξύ των παραγόντων, όμως δεν περιγράφονται οι σχέσεις μεταξύ αυτών. Το μοντέλο αυτό λέγεται μοντέλο μέτρησης.

Σύμφωνα με τα παραπάνω, όλα αυτά έχουν σαν αποτέλεσμα οι πληροφορίες που παίρνουμε από το μοντέλο αυτό να είναι συχνά ανεπαρκείς ή ασαφείς για την ψυχολογική ερμηνεία των δεδομένων. Συχνά, το μοντέλο μέτρησης που προκύπτει, μπορεί να έχει π.χ. μεγάλο αριθμό παραγόντων, να υπάρχουν μεταβλητές που μοιράζουν την διακύμανσή τους σε πολλούς δύσκολα ερμηνεύσιμους παράγοντες, να υπάρχουν συναθροίσεις μεταβλητών σε παράγοντες με τρόπο που να είναι στατιστικά έγκυρος αλλά ψυχολογικά χωρίς νόημα, έτσι γίνεται δυσχερής η απόδοση νοήματος στο μοντέλο μέτρησης. Η δυνατότητα παρέμβασης από τη μεριά του ερευνητή στη διεξαγωγή της ανάλυσης είναι περιορισμένη.

Ορθογώνιο παραγοντικό μοντέλο:

**Α) *Το μοντέλο***

Όπου $Y_1 = X_1 - \mu_1$,$X_1$: είναι το διάνυσμα των αρχικών μεταβλητών και $\mu_1$: είναι το διάνυσμα των μέσων.

$$Y_1 = l_{11}F_1 + l_{12}F_2 + \cdots + l_{1k}F_k + \varepsilon_1$$

$$Y_2 = l_{21}F_1 + l_{22}F_2 + \cdots + l_{2k}F_k + \varepsilon_2$$

$$\vdots$$

$$Y_p = l_{p1}F_1 + l_{p2}F_2 + \cdots + l_{pk}F_k + \varepsilon_p$$

$l_{ij} = Cov(Y_i, F_j)$και $Var(Y_i) = \sum_{j=1}^{k} l_{ij}^2 + y_i$

Όπου, $\varepsilon_p$: Μοναδικός παράγοντας (uniquefactor) της $Y_i$ή σφάλμα

$l_{pk}$: Επιβάρυνση (loading) της $Y_i$στον παράγοντα $F_j$

$$Var(Y_i) = Var\left(l_{i1}F_1 + l_{i2}F_2 + \cdots l_{ij}F_k + \varepsilon_i\right) = \underbrace{l_{i1}^2 + l_{i2}^2 + \cdots + l_{ij}^2} + \underbrace{y_i}$$

Εταιρικότητα (Commynality)Μοναδικότητα (Specifity)

Το μέρος της μεταβλητότητας της $Y_i$που οφείλεται στην ύπαρξη των κοινών παραγόντων $F$.

Το μέρος της μεταβλητότητας της $Y_i$ που δεν οφείλεται στην ύπαρξη των κοινών παραγόντων $F$.

$$\Sigma = Cov(Y) = Cov(LF + \varepsilon) = LL^T + Y$$

Στο δείγμα: $S = LL^T + \hat{Y}$. Στην πραγματικότητα παρατηρούμε τον πίνακα $S$. Όλη η ανάλυση συνίσταται στο να βρούμε κατάλληλους πίνακες $L$ και $Y$, τέτοιους ώστε να αναπαράγουν τον $S$.

Ο πίνακας $S$ μπορεί να αντιστοιχεί:

- Μοντέλα όπου η συνδιακύμανση οφείλεται στο γεγονός ότι η μία μεταβλητή συσχετίζεται με την άλλη
- Μοντέλα με διαφορετικό αριθμό παραγόντων
- Μοντέλα με ίδιο αριθμό παραγόντων αλλά διαφορετικές επιβαρύνσεις

## ***Β) Μεθοδολογία***

Υπάρχουν κάποια αξιώματα του παραγοντικού μοντέλου για την αντιμετώπιση των προβλημάτων:

- *Αξίωμα παραγοντικής αιτιότητας* (προυποθέτει ότι όλη η συνδιακύμανση μεταξύ των μεταβλητών οφείλεται στην ύπαρξη των κοινών παραγόντων)
- *Αξίωμα του απλούστερου μοντέλου* (από τα μοντέλα που προσαρμόζονται καλά στα δεδομένα επιλέγεται εκείνο με το μικρότερο αριθμό παραγόντων, το αξίωμα αυτό είναι σύνηθες και σε άλλες μεθόδους στατιστικής ανάλυσης.
- *Περιστροφή του μοντέλου* (υπάρχουν τεχνικές περιστροφές των αξόνων των παραμέτρων όπως θα δούμε και παρακάτω, προκειμένου η μορφή του πίνακα επιβαρύνσεων να οδηγεί σε απλούστερο μοντέλο).

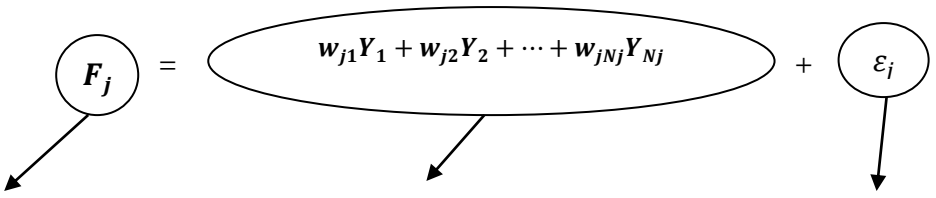
## ***Γ) Τα στάδια***

Τα διαδοχικά στάδια της εφαρμογής της διερευνητικής παραγοντικής ανάλυσης είναι:

1. *Έλεγχος ύπαρξης επαρκων συσχετίσεων μεταξύ των αρχικών μεταβλητών-καταλληλότητα μοντέλου:*
   →Έλεγχος ύπαρξης σημαντικών συσχετίσεων (θέλουμε συσχετίσεις ανά δύο να μην είναι πολύ μικρές).
   →Έλεγχος σφαιρικότητας του Bartlett($H_0: \Sigma = 1$)
   →Έλεγχος μετικών συσχτίσεων: Μέτρο δειγματικής καταλληλότητας ότν MKO=1
   →Έλεγχος μερικών συσχετίσεων, κατάλληλη Χ όταν MSA=1
2. *Επιλογή μεθόδου εκαγωγής των παραγόντων:*
   →Κύριων Συνιστωσών(PCA): Μέθοδος η οποία χρησιμοποιείται συνήθως για μείωση των δεδομένων, σήμερα δεν συνίσταται όμως ιδιαίτερα.
   →Μέγιστης πιθανοφάνειας(ML): Η μέθοδος αυτή προυποθέτει ότι τα δεδομένα ακολουθούν κανονική κατανομή. Όταν τα δεδομένα είναι κανονικά είναι η προτιμότερη μέθοδος.
   →Κύριωναξόνων(PrincipalAixsFactoring): Όταντα δεδομένα δεν είναι κανονικά προτιμάται σαν μέθοδος.
3. *Περιαστροφή αξόνων:*
   - Ορθογώνια περιστροφή: Οι παράγοντες παραμένουν ασυσχέτιστοι
     →Varimax: Ελαχιστοποιεί τον αριθμό των μεταβλητών (Y) που έχουν υψηλές επιβαρύνσεις ($l$) σε έναν παράγοντα $F$ −απλοποιούνται οι παράγοντες.
     →Quartimax: Ελαχιστοποιεί τον αριθμό των παραγόντων ($F$)στον οποίο έχει υψηλές επιβαρύνσεις ($l$)μία μεταβλητή ($Y$) - μειώνεται η παραγοντική πολυπλοκότητα.
     →Equamax: Συνσυάζει τα δύο παραπάνω προκειμένου να απλοποιηθεί το μοντέλο.
   - Μη ορθογώνια περιστροφή: Οι παράγοντες πλέον συσχετίζονται
     →DirectOblimin: Ο βαθμός συσχέτισης εξαρτάται από την παράμετρο δ. Όταν δ=0 έχουμε μεγαλύτερη συσχέτιση, μετά την περιστροφή ενώ μειώνεται όσο πιο αρνητικό είναι το δ.
     →Promax: Παρόμοια με το παραπάνω, αλλά γρηγορότερη στους υπολογισμούς οπότε και αποτελεσματικότερη για μεγαλύτερα σετ δεδομένων.

4. *Υπολογισμός των σκορ των παραγόντων*
   Τα σκορ των παραγόντων χρησιμεύουν στην μέτρηση π.χ. για την βαθμολογία/μέτρηση του κάθε ατόμου. Ουσιαστικά έχουμε:

$$\boldsymbol{F = wY}$$



Τα σκορ κάθε παράγοντα $F_j$

Ένας γραμμικός συνδυασμός των μεταβλητών $Y_{Nj}$,που αντιστοιχούν στον παράγοντα, πολλαπλασιασμένοι με κάποια βάρη $w_j$.

Σφάλμα που δεν οφείλεται στην μεταβλητότητα που δεν εξηγείται από το μοντέλο

5. *΄Ελεγχος καλής προσαρμογής του μοντέλου (goodnessoffit)*
   Η καλή προσαρμογή του μοντέλου συνδέεται και με την επιλογή του αριθμού των παραγόντων.
   Μόνο η μέθοδος της μέγιστης πιθανοφάνειας μας δίνει την δυνατότητα να εφαρμόσουμε το τεστ καλής προσαρμογής του μοντέλου.

*5.6.2.2: Η Επιβεβαιωτική ανάλυση παραγόντων*

Σε αντίθεση με τα παραπάνω, η μέθοδος αυτή εφαρμόζεται στον έλγχο της υπόθεσης ότι υπάρχει ένα συγκεκριμένο πρότυπο σχέσεων μεταξύ των παρατηρήσιμων μεταβλητών και των παραγόντων. Το μοντέλο αυτό ονομάζεται πλήρες, καθώς αποτελείται τόσο από ένα μοντέλο μέτρησης που απεικονίζει συνδέσεις ανάμεσα στις παρατηρήσιμες και λανθάνουσες μεταβλητές και τις συσχετίσεις ανάμεσα στις λανθάνουσες μεταβλητές, όσο και από ένα μοντέλο δομικών σχέσεων που απεικονίζει τις συνδέσεις (συσχετίσεις και σχέσεις εξάρτησης) ανάμεσα στις ίδιες λανθάνουσες μεταβλητές (Byrne,1994).

**Μπορούμε να αναφέρουμε ότι η ανάλυση παραγόντων τόσο ως επιβεβαιωτική ανάλυση αλλά και ως διερευνητική στηρίζεται στη υπόθεση ότι οι παρατηρούμενες μεταβλητές μπορούν να εκφραστούν ως ένας γραμμικός συνδυασμός κάποιων μη μετρήσιμων κοινών και ειδικών παραγόντων. Δηλαδή, η κάθε μεταβλητή εξηγείται από ένα αριθμό παραγόντων.**

Η επιβεβαιωτική ανάλυση επιτρέπει να προσδιοριστεί το μέρος της διακύμανσης της κάθε μεταβλητής που εξηγείται από τους παράγοντες που συνδέονται με αυτήν. Η σχέση αυτή περιγράφεται με την μορφή δομικής εξίσωσης. Μπορεί δηλαδή να υπολογιστεί ποιοί παράγοντες και σε ποιό βαθμό σχετίζονται με τις συγκεκριμένες παρατηρήσιμες μεταβλητές. Είναι δυνατό, να καθοριστούν οι συνδέσεις μεταξύ των παραγόντων δηλαδή πώς σχετίζονται, αλληλοεξαρτώνται οι παράγοντες και κατ'επέκταση οι διαστάσεις που εκπροσωπούν.

Για κάθε παρατηρούμενη μεταβλητή προσδιορίζονται στην δομική εξίσωση οι λανθάνουσες μεταβλητές. Στόχος της μεθόδου αυτής είναι ο έλεγχος ενός από πριν καθορισμένου μοντέλου το οποίο περιγράφει όλες τις σχέσεις που αναμένεται να υπάρχουν ανάμεσα στις μεταβλητές, τις μεταβλητές και τους παράγοντες και τέλος, τους παράγοντες μεταξύ τους. Η μέθοος αυτή ελέγχει την καταλληλότητα ενός από πριν διατυπωμένου μοντέλου για το ποιοι παράγοντες ερμηνεύουν τη διακύμανση των μεταβλητών και πως συνδέονται μεταξύ τους.

Στόχος της εφαρμογής της επιβεβαιωτικής ανάλυσης παραγόντων είναι να βρεθεί το μοντέλο που περιγράφει με τον πιο κατάλληλο αλλά και οικονομικό τρόπο τη δομή των πραγματικών δεδομένων. Άρα, ένα μοντέλο θα πρέπει όχι μόνο να έχει τους κατάλληλους στατιστικούς δείκτες αλλά και να ερμηνεύει τα δεδομένα με συνοπτικό τρόπο, χρησιμοποιώντας τον ελάχιστο αριθμό παραμέτρων.

Συνήθως, ο ερευνητής καταλήγει σε δύο μοντέλα που φαίνεται να προβλέπουν το ίδιο καλά τη δομή των πραγματικών δεδομένων, αλλά διαφέρουν στον αριθμό των παραμέτρων που εμπλέκουν στις δομικές εξισώσεις. Γενικά, ανάμεσα σε δύο μοντέλα που έχουν το ίδιο καλούς στατιστικούς δείκτες προτιμότερο είναι εκείνο που χρησιμοποιεί τις λιγότερες παραμέτρους για να ερμηνεύσει τη δομή τωνδεδομένων. Από την άλλη μπορούμε να συγκρίνουμε δύο μοντέλα για να αποφασίσουμε με έναστατιστικό κριτήριο αν το μοντέλο με τις περισσότερες παραμέτρους είναι στατιστικώς καλύτερο από εκείνο με τις λιγότερες παραμέτρους.

Το μοντέλο με το μεγαλύτερο πλήθος παραμέτρων έχει υψηλή τιμή $x^2$ και περισσότερους βαθμούς ελευθερίας από το άλλο μοντέλο. Ένας τρόπος για να συγκρίνουμε δύο μοντέλα είναι να υπολογίσουμε την διαφορά των $x^2$ και των βαθμών ελευθερίας που έχουν μεταξύ τους και να βρούμε αν η διαφορά αυτή, για τους αντίστοιχους βαθμούς ελευθερίας είναι στατιστικά σημαντική. Αν είναι σημαντική, δείχνει ότι το μοντέλο με τις περισσότερες παραμέτρους έχει μεγαλύτερη ερμηνευτική ισχύ. Αν η διαφορά των $x^2$ τους δεν είναι σημαντική, τότε η προσθήκη των παραμέτρων στο μοντέλο δεν συμβάλει στην καλύτερη ερμηνεία των δεδομένων, τότε το μοντέλο θεωρείται πλεοναστικό.

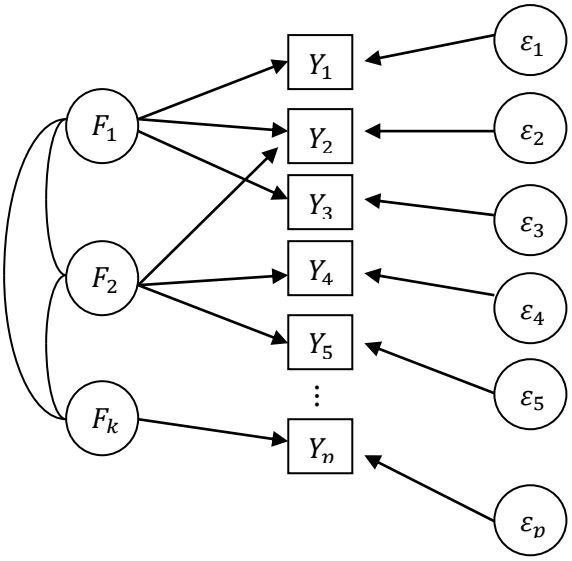
Το κατάλληλο μοντέλο μέσα από μία σύγκριση επιλέγεται όχι μόνο με βάση τα στατιστικά κριτήρια αλλά λαμβάνοντας υπόψη και το είδος της σχέσης που επιχειρείται να αποκατασταθεί μεταξύ των μεταβλητών και των παραγόντων. Αυτό το μοντέλο θεωρείται προτιμότερο.

Η στατιστική σημαντικότητα των παραγοντικών φορτίσεων για κάθε παράμετρο μπορεί να εκτιμηθεί με τον υπολογισμό των τυπικών τιμών(ή z-τιμών), με τις οποίες μπορεί κανείς να κρίνει ποιες από τις φορτίσεις των μεταβλητών είναι στατιστικά σημαντικές και ποιές όχι. Η z-τιμή για κάθε παράμετρο υπολογίζεται από την διαίρεση της φόρτισης της παραμέτρου με το τυπικό της σφάλμα.

- Αν πρόκειται για δίπλευρο τεστ, όταν η z-τιμή της παραμέτρου είναι >2.557 σημαίνει ότι το αποτέλεσμα είναι στατιστικά σημαντικό σε επίπεδο σημαντικότητας 0,01, ενώ αν είναι >1.960, τότε το αποτέλεσμα είναι στατιστικά σημαντικό σε επίπεδο σημαντικότητας 0.05.
- Αν πρόκετα για μονόπλευρο τεστ, όταν η z-τιμή της παραμέτρου είναι >2.326 σημαίνει ότι το αποτέλεσμα είναι στατιστικά σημαντικό σε επίπεδο σημαντικότητας 0,01, ενώ αν είναι >1.65 τότε το αποτέλεσμα είναι στατιστικά σημαντικό σε επίπεδο σημαντικότητας 0.05.

Οι z-τιμές θεωρούνται ενδεικτικές τιμές ως προς τα συμπεράσματα στα οποία οδηγούν για την καταλληλότητα του υπό-ελέγχου μοντέλου. Δηλαδή, αν κάποιες παράμετροι δεν έχουν στατιστικά σημαντικές φορτίσεις δεν θα πρέπει απαραίτητα να αφαιρεθούν εφόσον υποστηρίζονται από το θεωρητικό πλαίσιο και το μοντελό που τις περιλαμβάνει έχει καή προσαρμογή στα δεδομένα. Από την άλλη, για κάποιο <<φτωχό>> μοντέλο στο οποίο όλες οι φορτίσεις είναι σημαντικές αλλά οι δείκτες του μοντέλου είναι χαμηλοί, τότε: το μοντέλο είναι ελλιπές στο να περιγράφει τη δομή των πραγματικών δεδομένων.

Σε αντίθεση με την διερευνητική παραγοντική ανάλυση, η επιβεβαιωτική δεν μπορεί να εφαμροστεί χωρίς να οριστεί προηγουμένως πλήρως το μοντέλο, αυτό μπορεί να γίνει μέσω ενός μοντέλου διαδρομής (pathdiagram).




*Εικόνα 19: Διάγραμμα διαδρομής*

Τα ερωτήματα συμβολίζονται με τετράγωνα.

Οι παράγοντες συμβολίζονται με κύκλους.

Τα βέλη υποδεικνύουν τον παράγοντα από τον $F$ στο $Y$.

Με την βοήθεια των μέτρων καλής προσαρμογής (Goodnessoffitindices) ελέγχεται αν τα δεδομένα μας υποστηρίζουν το μοντέλο, δηλαδή αν είναι κατάλληλο για να εξηγηθεί.

*Goodness of fit indices(Μέτρακαλήςπροσαρμογής)*

- ✓ Chi-square έλεγχος: ανάλογα με το μέγεθος του δείγματος, χρησιμοποιείται το relativechi-square ($x^2/df$). Η επιθυμητή τιμή είναι κοντά στο 2.
- ✓ Absolutefitindices: Εξετάζεται από τον δείκτες RMR, RMSEA, GFI, που θα αναλυθούν παρακάτω.
- ✓ Relativefitindices: Εξετάζεται από τους δείκτες NNFI, CFI, που θα αναλυθούν παρακάτω.

Γενικά, μπορούμε να πούμε ότι ουσιαστικά τα δομικά μοντέλα εξισώσεων στην πραγματικότητα είναι η σύζευξη της <<ανάλυσης διαδρομών>>(pathanalysis) με την <<παραγοντική ανάλυση>>(factoranalysis).

## *5.7: Μεθοδολογία στα δομικά μοντέλα εξισώσεων*

Τα βήματα τα οποία πρέπει να ακολουθήσει κανείς για την επιτυχή εφαρμογή τους είναι τα εξής:

1. Model specification – Προσδιορισμός του μοντέλου: Το πρώτο πράγμα που πρέπει να κάνει κανείς είναι ο καθορισμός των σχέσεων που υπάρχουν είτε δεν υπάρχουν ανάμεσα στις παρατηρούμενες και αφανείς μεταβλητές. Κάθε ακαθόριστη σχέση ισούται με το μηδέν. Οι σχέσεις μεταξύ των μεταβλητών θεωρούνται ως παράμετροι. Θεωρείται το πιο σημαντικό στάδιο, πρέπει να πραγματοποιείται μετά από μελέτη του θεωρητικού μοντέλου και αν δεν γίνει σωστή εκτίμηση όλα τα υπόλοιπα στάδια θα οδηγήσουν σε άτοπο.
2. Model indentification – Προσδιορισμός της αναγνωρισιμότητας του μοντέλου: Γίνεται έλεγχος για την ύπαρξη τιμών για τις παραμέτρους του μοντέλου. Οι παράμετροι του μοντέλου μπορεί να προσεγγιστούν με τρείς τύπους παραμέτρων οι οποίες είναι:
    - Σταθερές: Λαμβάνουν συγκεκριμένη τιμή
    - Ελεύθερες: Μπορούν να λάβουν οποιαδήποτε τιμή
    - Περιορισμένες: Εξ'ορισμού είναι ίσες με άλλες παραμέτρους αλλά ουσιαστικά είναι άγνωστες.

    Για να υπολογίσουμε τους βαθμούς ελευθερίας του προβλήματος, αφαιρούμε τον αριθμό των παραμέτρων που έχουν εκτιμηθεί από τον αριθμό των γνωστών συσχετίσεων:

    a) BE>0 τότε το μοντέλο έχει υπερεκτιμηθεί
    b) BE=0 τότε το μοντέλο είναι απόλυτα σωστό
    c) BE<0 τότε δεν μπορεί να γίνει εκτίμηση
3. Data(δεδομένα): Γενικά το μέγεθος του δείγματος πρέπει να είναι της τάξης του 200 κάτι το οποίο όμως εξαρτάται και από την επιθυμητή ισχύ του μοντέλου, την πολυπλοκότητα και την μηδενική υπόθεση. Προβλήματα που μπορεί να παρουσιαστούν είναι να υπάρχουν πλεονάζουσες μεταβλητές (δηλαδή μεγάλες συσχετίσεις μεταξύ των μεταβλητών), ασυνήθιστα σκορ των εξεταζόμενων, να μην ακολουθούν ομαλή κατανομή τα σταστικά στοιχεία και να παραλείπεται απο τους συμμετέχοντες η απάντηση σε μερικές ερωτήσεις.
4. Estimation(εκτίμηση): Μέσω κατάλληλων υπολιγιστικών προγραμμάτων γίνεται εκτίμηση των τιμών των άγνωστων παραμέτρων και των σφαλμάτων τους. Αν δεν υπάρχει τρόπος να επιλυθούν τα τυχόν σφάλματα που έχουν βρεθεί γίνεται εκ νέου εκτίμηση του μοντέλου.

5. Model fit an dinterpretation(εφαρμογή του μοντέλου και ερμηνεία): Στο βήμα αυτό γίνεται έλεγχος σχετικά με το αν οι υποθέσεις που είχαν γίνει αρχικά συμφωνούν με τα πειραματικά δεδομένα.
6. Model modification - Εξέταση ισοδύναμων μοντέλων: Σε αυτό το βήμα μπορεί να γίνει κάποια αλλαγή- ρύθμιση του εκτιμώμενου μοντέλου μεπρόσθεση ή αφαίρεση παραμέτρων και δημιουργία ισοδύναμων μοντέλων. Τα ισδούναμα μοντέλα περιγράφουν τα δεδομένα το ίδιο καλά με το μοντέλο που έχει επιλέξει ο ερευνητής υπό την διαφορά ότι οι μεταβλητές έχουν διαφορετική διάταξη όμως έχουν τον ίδιο αριθμό παραμέτρων και την ίδια προσαρμογή με το μοντέλο.
7. Περιγραφή των αποτελεσμάτων: Ο ερευνητής οφείλει να εξετάσει σε κάθε βήμα τα δεδομένα που έχει και να εξάγει τα αποτελέσματα με συνέπεια και ακρίβεια.

### *5.7.1: Κανόνες προσδιορισμού μοντέλου*

Υπάρχουν κάποιοι κανόνες για τον προσδιορισμό των παραμέτρων του μοντέλου, ώστε να καθοριστούν σωστά και να μπορούν να εκτιμηθούν σε ένα προτεινόμενο μοντέλο SEM. Όταν εφαρμόζονται δεν γίνεται διάκριση μεταξύ συνδιακύμανσης και συσχέτισης μεταξύ των δύο μεταβλητών. Οι κανόνες αυτοί είναι οι εξής:

1. Όλες οι διακυμάνσεις των ανεξάρτητων μεταβλητών είναι παράμετροι του μοντέλου. Οι όροι των σφαλμάτων σε ένα διάγραμμα διαδρομής γενικά συνδέονται με κάθε εξαρτημένη μεταβλητή. Για τις λανθάνουσες μεταβλητές τα σφάλματα συμβολίζουν τους διαρθρωτικούς όρους διαταραχής παλινδρόμησης. Όλοι οι υπολειπόμενοι όροι είτε συνδέονται με παρατηρούμενες ή λανθάνουσες μεταβλητές, δεν έχουν παρατηρηθεί οντότητες επειδή δεν μπορούν να μετρηθούν και να είναι ανεξάρτητες.
2. Όλες οι συνδιασπορές μεταξύ ανεξάρτητων μεταβλητών είναι παράμετροι του μοντέλου.
3. Όλα τα φορτία, συνδετικών στοιχείων με δείκτες των λανθάνουσων μεταβλητών είναι παράμετροι του μοντέλου. Όπως μπορεί να φανεί και παρακάτω στο διάγραμμα διαδρομής στο παράδειγμα, οι παράμετροι υποδηλώνονται με τους αστερίσκους που συνδέονται με τις διαδρομές που συνδέουν κάθε μία από τις λανθάνουσες μεταβλητές.
4. Όλοι οι συντελεστές παλινδρόμησης μεταξύ των παρατηρούμενων ή λανθάνουσων μεταβλητών είναι παράμετροι του μοντέλου. Είναι σημαντικό να σημειωθεί ότι ο κανόνας 3, μπορεί να θεωρηθεί ως ειδική περίπτωση του κανόνα 4, μετά από μία διαπίστωση ότι ένα φορτίο παράγοντα μπορεί να συμπεριληφθεί ως συντελεστής παλινδρόμησης της παρατηρούμενης μεταβλητής. Ωστόσο, στην πράξη η εκτέλεση της συγκεκριμένης παλινδρόμησης είναι συνήθως αδύνατη, διότι οι παράγοντες δεν παρατηρούν

μεταβλητές για να αρχίσουν και ως εκ τούτου δεν υπάρχουν μεμονωμένες μετρήσεις.

5. Οι διακυμάνσεις των συσχετίσεων μεταξύ των εξαρτημένων μεταβλητών και των συνδιασπορών μεταξύ εξαρτώμενων και ανεξάρτητων μεταβλητών δεν είναι ποτέ οι παράμετροι του μοντέλου. Αυτό οφείλεται στο γεγονός ότι οιδιακυμάνσεις και οι συμμεταβλητές είναι ίδιοι με τους όρους των άλλων παραμέτρων του μοντέλου.
6. Για κάθε λανθάνουσα μεταβλητή που περιλαμβάνεται σε ένα μοντέλο η μετρική της λανθάνουσας κλίμακας πρέπει να ορίζεται. Ο λόγος γι αυτό είναι ότι σε αντίθεση με τις παρατηρούμενες μεταβλητές σε ένα μοντέλο δεν υπάρχει φυσική μετρική που να βασίζονται οι λανθάνουσες μεταβλητές. Για κάθε ανεξάρτητη λανθάνουσα μεταβλητή που περιλαμβάνεται σε ένα προτεινόμενο μοντέλο,ο μετρικός μπορεί να στερεωθεί σε έναν από τους δύο ισοδύναμους τρόπους. Είτε διακύμανση που τίθεται ίση με μία σταθερά ή μία διαδρομή αφήνοντας την λανθάνουσα μεταβλητή να έχει ορισθεί σε μία σταθερή.

### *5.7.2: Τύποι παραμέτρων στα SEM*

Υπάρχουν τρεις τύποι παραμέτρων του μοντέλου που είναι σημαντικοί για την διεξαγωγή SEM ανάλυσης.

1. ***Ελεύθεροι***
   Είναι οι παράμετροι που προσδιορίζονται με βάση τους παραπάνω 6 κανόνες.
2. ***Σταθεροί-Καθορισμένοι***
   Οι παράμετροι αυτοί έχουν μία σταθερή δεδομένη αξία. Οι εν λόγω παράμετροι ονομάζονται σταθεροί καθώς δεν αλλάζουν τιμή όταν το μοντέλο είναι κατάλληλο για τα παρατηρούμενα δεδομένα.
3. ***Περιορισμένοι-Βεβιασμένοι***
   Οι παράμετροι αυτοί αναφέρονται και ως κλειστοί ή συγκρατημένοι. Οι περιορισμένοι παράμετροι συνήθως περιλαμβάνονται σε ένα μοντέλο αν ο περιορισμός τους προέρχεται από τις υπάρχουσες θεωρίες ή αντιπροσωπεύει μία ουσιαστικά ενδιαφέρουσα υπόθεση που εξετάστηκε σε ένα προτεινόμενο μοντέλο.

### *5.7.3: Μέθοδοι εκτίμησης παραμέτρων*

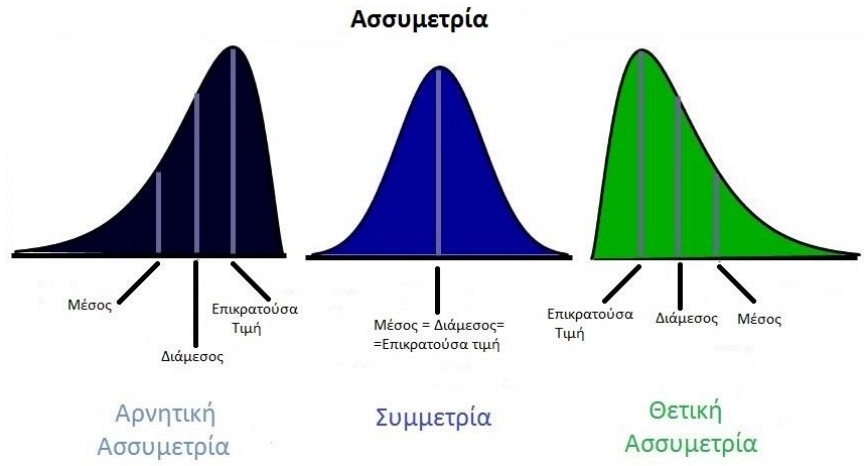
Υπάρχουν τέσσερις κύριοι μέθοδοι που χρησιμοποιούνται στα περισσότερα προγράμματα, τα οποία είναι:

1. Μη σταθμισμένων ελαχίστων τετραγωνων(ULS)
2. Μέγιστης πιθανοφάνειας(ML)
3. Γενικευμένων ελαχίστων τετραγώνων(GLMγια Μεγάλα δείγματα)
4. Σταθμισμένων ελαχίστων τετραγώνων

Η 1η μέθοδος θεωρείται ως η πιο κατάλληλη λειτουργία, το απλό μη σταθμισμένο άθροισμα των τετραγώνων των διαφορών είναι μεταξύ των αντίστοιχων στοιχείων του S και του υποδείγματος Σ του πίνακα συνδιασποράς. Οι εκτιμήσεις, επιλέγονται για τις παραμέτρους του μοντέλου όταν η μέθοδος αποκτά τηνελάχιστη τιμή. Η μέθοδος εφαρμόζεται στην πράξη όταν παρόμοιες κλίμακες μέτρησης αποτελούν την βάση των μεταβλητών που θα αναλυθούν.

Οι άλλες τρεις μέθοδοι εκτίμησης βασίζονται στο ίδιο άθροισμα των τετραγώνων αλλά σε ειδικές περιπτώσεις έχουν χρησιμοποιηθεί για να πολλαπλασιαστεί κάθε ένα από τα τετράγωνα. Η μέθοδος της μέγιστης πιθανοφάνειας και των γενικευμένων ελαχίστων τετραγώνων χρησιμοποιούνται όταν δεδομένα είναι κανονικά κατανεμημένα. Η υπόθεση της κανονικότητα είναι αρκετά συχνή στην παραγοντική ανάλυση, η οποία μπορεί να εξεταστεί χρησιμοποιώντας οποιοδήποτε στατιστικό πακέτο γενικής χρήσης. Ο απλούστερος τρόπος για να εξεταστεί η μονοπαραγοντική ομαλότητα είναι να εξεταστεί η ασσυμετρία και η κύρτωση.

Ασσυμετρία




*Εικόνα 20: Διάγραμμα Ασσυμετρίας*

Αρχικά, για το παραπάνω σχήμα πρέπει να ορίσουμε: Διάμεσος= $M_{1/2}$, Μέσος= $\bar{X}$, Επικρατούσα τιμή = $M_0$.

Ο συντελεστής ασσυμετρίας του Pearsonχρησιμοποιεί την σχέση μεταξύ του μέσου και της επικρατούσας τιμής. Ο συντελεστής αυτός συμβολίζεται:

$$S_k = \frac{\bar{X} - M_0}{S}$$

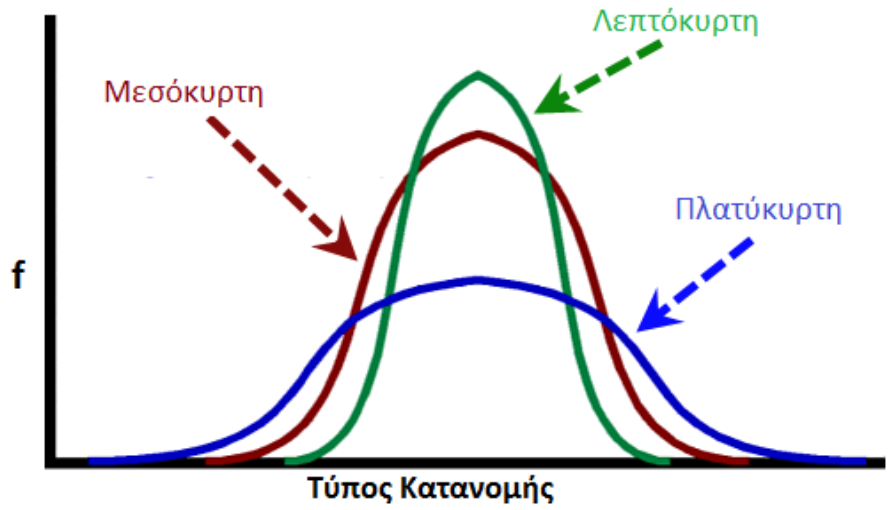
- Αν $S_k = 0$ η κατανομή είναι συμμετρική
- Αν $S_k < 0$ έχουμε αρνητική ασσυμετρία
- Αν $S_k > 0$ έχουμε θετική ασσυμετρία

Όπου,$S$: Τυπική απόκλιση

Η τιμή του $S_k$ δεν έχει συγκεκριμένα όρι αλλά σε μέτριες ασσυμετρίες κυμαίνεται μεταξύ $\pm 1$.

Κύρτωση

Η κύρτωση χαρακτηρίζει κατά το πόσο η καμπύλη μιας κατανομής είναι πεπλατισμένη ή όχι, δεχόμενοι σαν κανονική κύρτωση αυτή της κανονικής καμπύλης. Μία καμπύλη με κύρτωση μικρότερη από εκείνη της κανονικής ***(μεσόκυρτη)*** χαρακτηρίζεται ως ***πλατύκυρτη***, ενώ αν είναι μεγαλύτερη της κανονικής χαρακτηρίζεται ***λεπτόκυρτη.***




*Εικόνα 21: Διάγραμμα Κύρτωσης*

Για την μέτρηση της κύρτωσης χρησιμοποιείται συνήθως ο συντελεστής του Pearson, ο οποίος ορίζεται ως: $b_2 = \frac{m_4}{m_2^2} = \frac{m_4}{S^4}$, όπου $m_4 = \frac{\sum_{i=1}^{N} F_i (X_i - \bar{X})^4}{N}$ τέταρτη κεντρική ροπή.

Αν χαρακτηρίσουμε την κύρτωση μιας κατανομής με βάση τον συντελεστή $b_2$, θα έχουμε:

- Αν $b_2 > 3$ τότε λεπτόκυρτη κατανομή
- Αν $b_2 = 3$ τότε μεσόκυρτη κατανομή
- Αν $b_2 < 3$ τότε πλατύκυρτη κατανομή

Από στατιστικής πλευράς, οι 4 μεθοδοι εκτίμησης παραμέτρων οδηγούν σε συνεπείς εκτιμήσεις, κάτι το οποίο είναι ένα επιθυμητό αποτέλεσμα που εξασφαλίζει ότι με την αύξηση του μεγέθους του δείγματος οι εκτιμήσεις συγκλινουν προς τις πραγματικές τιμές των παραμέτρων του πληθυσμού.

Στατιστικά όταν χρησιμοποιείται η μεθοδολογία των δομικών μοντέλων εξισώσεων συνήθως δεν απορρίπτει την μηδεινική υπόθεση(η οποία δίνεται από την σχέση (1)). Επειδή το μοντέλο δοκιμής σε SEM περιλαμβάνει τον έλεγχο της μηδενικής υπόθεσης ότι το μοντέλο είναι ικανό για τέλεια προσαρμογή. Στην πραγματικότητα

όμως είναι πολύ πιθανό ότι αν το μοντέλο δεν έχει καθοριστεί σωστά αυτο μπορεί να οφείλεται σε σφάλμα δειγματοληψίας. Μπορεί να υπάρχει πληθώρα μοντέλων που να ταιριάζει εξίσου καλά στα δεδομένα.

## *5.8: Γραφική απεικόνιση δομικών μοντέλων εξισώσεων*

Ένα διάγραμμα στα συστήματα δομικών εξισώσεων αποτελείται από ορθογώνια παραλληλόγραμμα και κύκλους ή ελλείψεις που συνδέονται μεταξύ τους με απλής ή διπλής κατεύθυνσης βέλη.Τα ορθογώνια παραλληλόγραμμα αντιπροσωπεύουν τις παρατηρηθείσες μετρήσιμες μεταβλητές, ενώ οι ελλείψεις τις κρυφές ή λανθάνουσες μετρήσιμές μεταβλητές. Τα απλής κατεύθυνσης βέλη χρησιμοποιούνται για να καθορίσουν τις αιτιώδεις σχέσεις στο μοντέλο. Τα διπλής κατεύθυνσης βέλη χρησιμοποιούνται για να δείξουν συνδιακυμάνσεις ή συσχετισμούς μεταξύ παραγόντων χωρίς καμία αιτιώδη ερμηνεία.

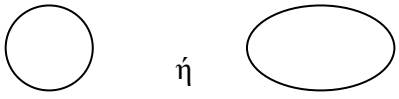
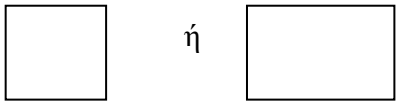
Από στατιστικής πλευράς,τα απλά βέλη αντιπροσωπεύουν τους συντελεστές παλινδρόμησης ενώ τα βέλη διπλής κατεύθυνσης δείχνουν τις συνδιακυμάνσεις μεταξύ των παραγόντων. Τα διανύσματα με φορά προς τις μεταβλητές ή τους παράγοντες δηλώνουν την υπολειπόμενη διακύμανση,δηλαδή το σφάλμα μέτρησης.

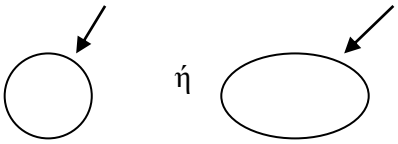
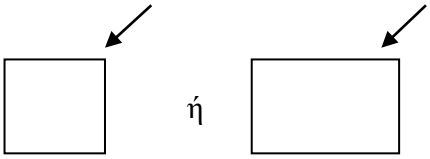
Ένας εύκολος τρόπος για να απεικονιστεί διαγραμματικά είναι μέσω ενός διαγράμματος ροής, το οποίο ουσιαστικά είναι ένα είδος μαθηματικής αναπαράστασης ενός μοντέλου υπό έρευνα.

Ένα από τα σημαντικά θέματα των δομικών μοντέλων εξισώσεων είναι η διάκριση μεταξύ των μεταβλητών αυτών.

### *5.8.1:Λανθάνουσες και παρατηρούμενες μεταβλητές*

Οι παρατηρούμενες μεταβλητές είναι οι μεταβλητές που δεν μπορούν να μετρηθούν άμμεσα, στην πραγματικότητα μετρούνται π.χ. είναι η απόδοση σε μία συγκεκριμένη δομή ή οι απαντήσεις σε συγκεκριμένα στοιχεία ή ερωτήσεις σχετικά με την καταγραφή ή μέσω ερωτηματολογίων.

| | |
|---|---|
|  ή | Λανθάνουσες μεταβλητές |
|  ή | Παρατηρούμενες μεταβλητές |

| | |
|---|---|
| ή  | Μονής κατεύθυνσης διαδρομή |
|  ή | Διαταραχή ή λάθος σε λανθάνουσας μεταβλητή |
|  ή | Μέτρηση σφαλμάτων παρατηρούμενης μεταβλητής |
| Ή  | Συσχέτιση μεταξύ των μεταβλητών |

*Πίνακας 3: Διαγραμματική απεικόνιση*

Οι παρατηρούμενες μεταβλητές έχουν μετρηθεί από τον ερευνητή κατά την διαδικασία συλλογής δεδομένων. Σε αντίθεση, οι λανθάνουσες μεταβλητές είναι υπάρχουσες όπως π.χ. η ευφυία, το άγχος, τα κίνητρα, κατάθλιψη, κοινωνική υποστήριξη και η κοινωνιολογική κατάσταση. Αν η παρατηρούμενη μεταβλητή χρησιμοποιείται ως δείκτης μια λανθάνουσας μεταβλητής το πιθανότερο είναι ότι η παρατηρούμενη μεταβλητή θα περιέχει αρκετά αναξιόπιστες πληροφορίες σχετικά με το κατασκεύασμα. Γενικά, οι ερευνητές χρησιμοποιούν πολλαπλούς δείκτες(άνω των 2) για κάθε μεταβλητή που εξετάζεται, προκειμένου να επιτευχθεί μία πολύ πιοπλήρης και αξιόπιστη εικόνα από αυτή που παρέχεται από έναν μόνο δείκτη. Φυσικά, υπάρχουν και περιπτώσεις όπου μία ενιαία παρατηρούμενη μεταβλητή μπορεί να είναι ένας πολύ καλός δείκτης της λανθάνουσας μεταβλητής.

### *5.8.2: Εξαρτημένες και ανεξάρτητες μεταβλητές*

Εξαρτημένες μεταβλητές είναι οι μεταβλητές που λαμβάνουν τουλάχιστον μία διαδρομή από μία άλλη μεταβλητή στο μοντέλο.

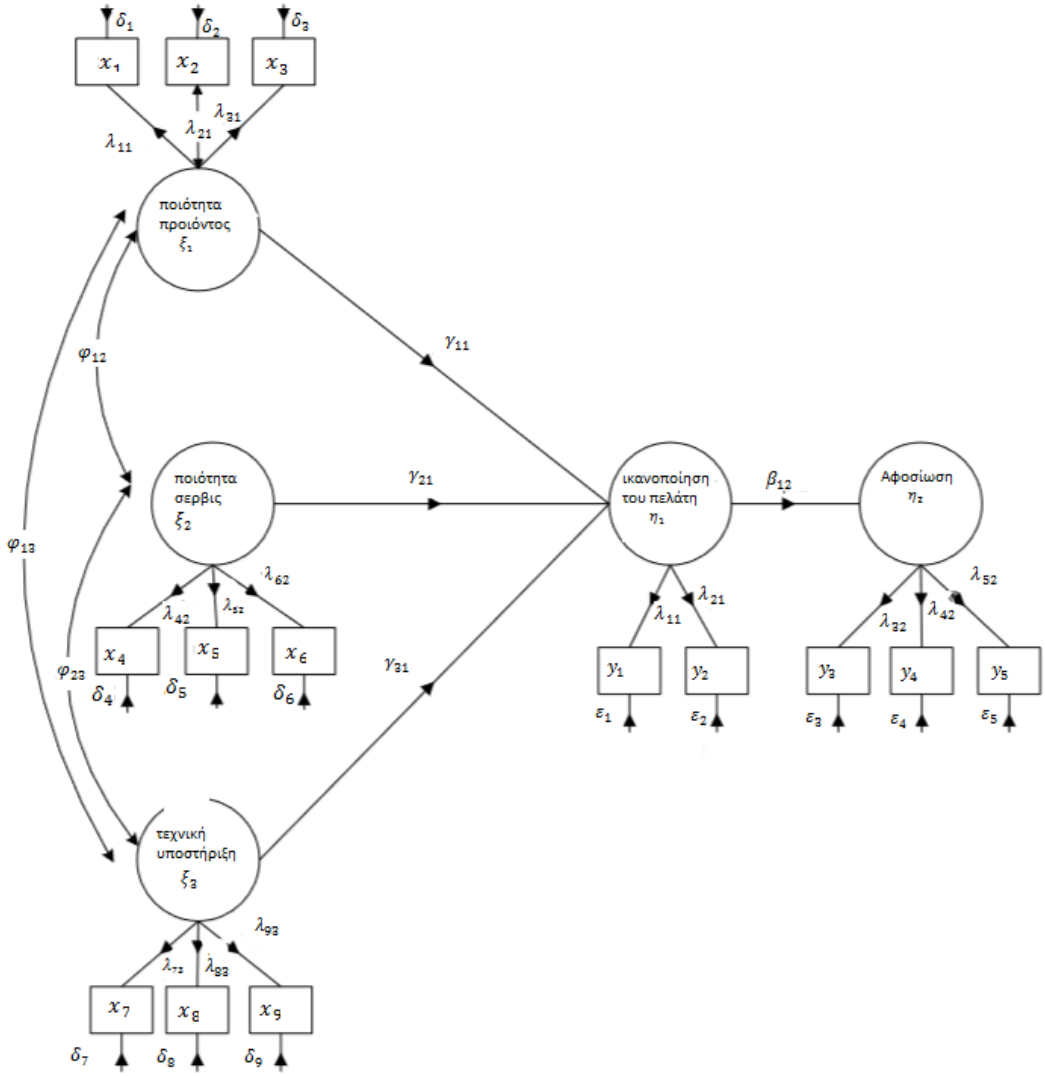
Ανεξάρτητες μεταβλητές είναι οι μεταβλητές που προέρχονται από διαδρομές αλλά ποτέ δεν λαμβάνουν κάποια διαδρομή.

Οι όροι ενδογενείς και εξωγενείς μεταβλητές χρησιμοποιούνται για να κάνουν την ίδια διάκριση μεταξύ των μεταβλητών.

Ανεξάρτητα από τους όρους που χρησιμοποιεί κανείς μια σημαντική συνέπεια της διάκρισης μεταξύ ανεξάρτητων και εξαρτημένων μεταβλητών είναι ότι δεν υπάρχουν δύο διαδρομών βέλη που συνδέουν δύο εξαρτημένες μεταβλητές σε ένα διάγραμμα μοντέλου διαδρομής.

### *5.8.3:Παράδειγμα*

Ένα απλό μοντέλο δομικών εξισώσεων με τρεις εξωγενείς μεταβλητές (ποιότητα προϊόντος, ποιότητα της εξυπηρέτησης και τεχνική υποστήριξη) και δύο ενδογενείς μεταβλητές (ικανοποίηση πελατών και καταναλωτική αφοσίωση). Οι τρεις εξωγενείς μεταβλητές επηρεάζουν την ικανοποίηση του πελάτη και ητελευταία την αφοσίωση.




*Εικόνα 22: Μοντέλο Δομικών Εξισώσεων*

Στα δεξιά του σχήματος παρατηρούμε ότι βρίσκονται δύο ενδογενείς μεταβλητές ( $\eta_1$και $\eta_2$) οι οποίες αντιπροσωπεύουν την ολική ικανοποίηση και την καταναλωτική αφοσίωση, αντίστοιχα. *Η πρώτη ενδογενής μεταβλητή*($\eta_1$)αντιπροσωπεύει την ολική ικανοποίηση και συνδέεται με δύο μετρήσιμες μεταβλητές $y_1$ και $y_2$, όπου, $\boldsymbol{y_1}$:μπορεί να είναι η ικανοποίηση που αποκομίζουν συνολικά οι πελάτες από

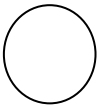
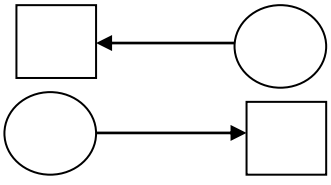
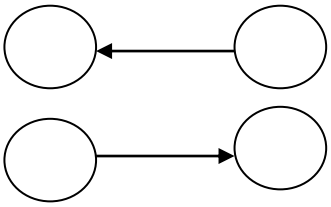
το προιόν και$\boldsymbol{y_2}$:κατά το πόσο το προιόν ανταπροκίνεται στις προσδοκίες τους. *Η δεύτερη ενδογενής μεταβλητή*($\eta_2$) συνδέεται με τρεις προφανείς μεταβλητές $y_3, y_4, y_5$ οι οποίες συνδυαζόμενες αντιπροσωπεύουν την λανθάνουσα μεταβλητή($\eta_2$). Οι μετρήσιμες μεταβλητές $y_3, y_4, y_5$ μπορεί αν είναι για παράδειγμα η πρόθεση επαναπροτίμησης του προιόντος, η πρόθεση να το προτείνουν σε γνωστούς και φίλους και η πρόθεση τους να συνεχίσουν να αγοράζουν το συγκεκριμένο προιόν ανεξάρτητα απνεξάρτητα από τις προσφορές των ανταγωνιστών.

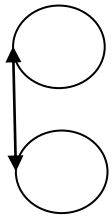
Τώρα, ως αναφορά τις εξωγενείς μεταβλητές ($\xi_1, \xi_2, \xi_3$) οι οποίες βρίσκονται αριστερά του παρπάνω σχήματος, αντιρποσωπεύουν την ποιότητα του προιόντος, την ποιότητα της εξυπηρέτησης και την τεχνική υποστήριξη.

*Η πρώτη εξωγενής μεταβλητής*($\xi_1$) αντιπροσωπεύει την ποιότητα του προιόντος και συνδέεται με τρεις μετρήσιμες μεταβλητές $x_1, x_2, x_3$ όπου: $x_1$:είναι η διάρκεια ζωής , $x_2$:η συσκευασία και $x_3$:η αντοχή του προιόντος.

*Η δεύτερη εξωγενής μεταβλητή*($\xi_2$) αντιπροσωπεύει την ποιότητα της εξυπηρέτησης και συνδέεται με τρεις μετρήσιμες μεταβλητές $x_4, x_5, x_6$.Οι μεταβλητές αυτές μπορεί για παράδειγμα να είναι η ταχύτητα εξυπηρέτησης, η ευγένια και η κατάρτιση του προσωπικού.

Και τέλος, η τρίτη εξωγενής *μεταβλητή* ($\xi_3$) αντιπροσωπεύει την τεχνική υποστήριξηπου παρέχεται στους πελάτες και συνδέεται και εκείνη αντίστοιχα με τρεις μετρήσιμες μεταβλητές $x_7, x_8, x_9$. Οι μεταβλητές αυτές μπορεί για παράδειγμα να εκφράσουν τις γνώσεις του προσωπικού στο τμήμα αυτό, την ευγένια του και την ικανότητα του να λύνει τα διάφορα προβλήματα που δημιουργούνται.

| ΣΥΜΒΟΛΟ | ΠΕΡΙΓΡΑΦΗ |
|---|---|
|  | Λανθάνουσα μεταβλητή |
|  | Μετρήσιμη μεταβλητή |
|  | Σχέση μεταξύ λανθάνουσας και μετρήσιμης μεταβλητής. Το ελληνικό γράμμα που βλέπουμε στο παραπάνω διάγραμμα χρησιμοποιείται για να υποδείξει αυτή την διαδρομή. |
|  | Σχέση μεταξύ δύο λανθάνουσων μεταβλητών. Εάν η διαδρομή είναι μεταξύ μιας εξωγενούς και μια ενδογενούς μεταβλητής χρησιμοποιείται το σύμβολο γ. Εάν οδηγείται μία ενδογενή μεταβλητή σε μία άλλη χρησιμοποιείται το σύμβολο β. |

| | |
|---|---|
|  | Το τόξο δείχνει την συνδιακύμανση μεταξύ δύο αφανών μεταβλητών. Το γράμμα φ υποδεικνύει το τόξο αυτό. |

*Πίνακας 4: Περιγραφή σύμβολων δομικών εξισώσεων*

## *5.8.4: Σχεδιασμός Δομικού μοντέλου*

### *5.8.4.1: Θεωρητικές προυποθέσεις για τον σχεδιασμό των μοντέλων*

Όταν κάποιος επιθυμήσει να εφαρμόσει στα δεδομένα του μοντέλου δομικών εξισώσεων πρέπει η έρευνα του να πληρεί τις απαραίτητες θεωρητικές προυποθέσεις ώστε τα αποτελέσματα που θα προκύψουν να είναι ερμηνεύσιμα.

Η μέθοδος αυτή θεωρείται κατάλληλη για την εκτίμηση των σχέσεων εξάρτησης που συνδέουν κάποιες μεταβλητές στην διαδρομή του χρόνου. Στην περίπτωση του ελέγχου τέτοιων θεωρητικών μοντέλων είναι απαραίτητο να υπολογίζεται κατά την μέτρηση το χρονικό διάστημα που παρενέβη ανάμεσα στη μεταβλητή που προηγείται και σε εκείνη που έπεται. Δηλαδή, αν η μέτρηση της δεύτερης μεταβλητής γίνει πολύ σύντομα ή πολύ αργά από τον χρόνο εμφάνισης της πρώτης μεταβλητής μπορεί να χαθεί το χρονικό σημείο στο οποίο η δράση της μεταβλητής είχε το μέγιστο αποτέλεσμα και να συγκαλυφθεί η αιτιώδης σχεση τους. Από ορισμένους ερευνητές οι διαχρονικές μετρήσεις θεωρούνται ως η μόνη ασφαλής μέθοδος για την διερεύνηση της αλληλεξάρτησης ανάμεσα σε μεταβλητές που συνδέονται με την παρέλευση του χρόνου.

Μία άλλη προυπόθεση είναι αν το δείγμα από το οποίο συλλέχθηκαν τα δεδομένα, σχετίζεται κατάλληλα με τις θεωρητικές ιδέες που πρόκειται να εκτιμηθούν. Αν όχι τότε τα αποτελέσματα που θα προκύψουν δεν θα έχουν πληροφόρηση για την εκτίμηση της θεωρίας.

Έπειτα, θα πρέπει ο λειτουργικός προσδιορισμός των μεταβλητών να γίνεται με τρόπο που να είναι συμβατός με την συγκεκριμένη θεωρία. Επιπλέον, θα πρέπει να είναι κανείς προσεκτικός στην χρήση των παρατηρήσιμων μεταβλητών ως δείκτες για τις λανθάνουσες μεταβλητές. Ένας παράγοντας ακόμη και αν έχει υψηλή ισχύ μπορεί να μην έχει πάντα ψυχολογικό νόημα. Γενικά, ένας παράγοντας έχει νόημα όταν εξάγεται ως λογική συνέπεια των παρατηρήσιμων μεταβλητών, οι οποίες συνδέονται μεταξύ τους όχι με αιτιώδη σχέση αλλά με αρκετά υψηλή συνάφεια.

*Για παράδειγμα*, η μόρφωση, το εισόδημα, οι οικιακές ανέσεις είναι δείκτες κοινωνικής τάξης. Η απόδοση όμως της συγκεκριμένης ταυτότητας του παράγοντα θα πρέπει να υποστηρίζεται από θεωρία, έτσι ώστε να έχει μία σωστή θέση στο μοντέλο.

Έτσι σε κάποιες περιπτώσεις η μόρφωση και το εισόδημα θα ήταν καλύτερο να μην συναποτελούν δείκτες του παράγοντα κοινωνικής τάξης, γιατί μπορεί να συνδέονται αιτιωδώς μεταξύ τους, με αποτέλεσμα παρόλο που ο παράγοντας έχει νόημα οι παρατηρήσιμες μεταβλητές μπορεί να είναι κατάλληλες ή και να προκαλούν αμφιβολίες για το πόσο καλά προσδιορίζουν εννοιολογικά την ταυτότητα του παράγοντα.

#### *5.8.4.2: Στατιστικές προυποθέσεις για τον σχεδιασμό δομικών μοντέλων εξισώσεων*

Μία προυπόθεση είναι: η μέθοδος της επιβεβαιωτικής ανάλυσης παραγόντων να μπορεί να εφαρμοστεί όταν οι εξαρτημένες παρατηρήσιμες μεταβλητές είναι συνεχείς ή μπορεί να εκληφθούν ως συνεχείς και έχουν τουλάχιστον τρεις κατηγορίες. Όταν αυτό δεν συμβαίνει γίνονται κάποιοι μετασχηματισμοί των δεδομένων ώστε να εξαλιφθεί γραμμική σχέση ανάμεσα στις μεταβλητές.

Ο αιρθμός των παρατηρήσιμων μεταβλητών που μπορεί να συμπεριλάβει ένα μοντέλοτο οποίο είναι συνάρτηση του λόγου: $\frac{\tau ο υ α ρ ι θ μ ο ύ τ ω ν μ ε τ α β λ η τ ώ ν}{\tau ο υ α ρ ι θ μ ο ύ τ ω ν υ π ο κ ε ι μ έ ν ω ν τ η ς έ ρ ε υ ν α ς}$
Για τον λόγο αυτό δεν εφαρμόζεται σε μοντέλα με περισσότερες από 20-30 μεταβλητές, ειδικά όταν το δείγμα από το οποίο προήλθαν είναι <200 άτομα. Για να ορισθεί πλήρως ένας παράγοντας χρειάζονται δύο τουλάχιστον παρατηρήσιμες μεταβλητές, διαφορετικά το μοντέλο θα βρεθεί ελλιπώς προσδιοριζόμενο.

Γενικά, συνιστάται να αποφεύγονται τα πολύ μικρά δείγματα, καθώς: πολλοί ερευνητές υποστηρίζουν ότι είναι παρακινδυνευμένο να εφαρμόζεται σε δείγματα κάτω των 100 ατόμων ακόμη οι (Adenson&Gerbing) υποστηρίζουν ότι ένα δείγμα 100 ατόμων είναι κατάλληλο όταν στο μοντέλο που ελέγχεται, ο κάθε παράγοντας προσδιορίζεται από τρεις τουλάχιστον μεταβλητές.

Από την άλλη όμως τα πολύ μεγάλα δείγματα επηρεάζουν απρόβλεπτα τα αποτελέσματα της ανάλυσης. Όμως, όσο μεγαλύτερο είναι το δείγμα τόσο πιο δύσκολο να επιτευχθεί. Για το λόγο αυτό μπορεί κανείς να οδηγηθεί στην απόρριψη του προτεινόμενου μοντέλου, ακόμη και αν αυτό είναι αρκετά κοντά στην πραγματική δομή των δεδομένων. Μπορεί κανείς να κρίνει την καταλληλότητα του μοντέλου μέσω

- των τιμών z
- αλλά και με την εξέταση του λόγου των πιθανοφανειών

## *5.9: Χρησιμότητα των δομικών μοντέλων*

Η χρησιμότητα των δομικών μοντέλων έχει αυξήσει αρκετά τα τελευταία χρόνια.Υπάρχουν τουλάχιστον δύο λόγοι για την δημοτικότητα των SEM.

- Αρχικά στις επιστήμες που εξετάζουν την συμπεριφορά των τόμων σε συγκεκριμένες καταστάσεις το ενδιαφέρον τους εστιάζεται κυρίως στην μελέτη των θεωρητικών κατασκευασμάτων που δεν μπορούν να παρατηρηθούν άμεσα.Μία μορφή που άμεσα διαπραγματεύονται τα δομικά μοντέλα είναι το πόσο καλά οι μετρήσεις αυτές απεικονίζουν τα προοριζόμενα δομικά μοντέλα.
- Επίσης οι ερευνητές ενδιαφέρονται κυρίως για ζητήματα πρόβέψης.Τα προβλεπτικά μοντέλα έχουν γίνει πιο σύνθετα κάτι που μέσω τεχνικών που προσφέρουν τα SEMεπιτρέπουν την χρήση ειδικών δοκιμών για χρήσιμαμοντέλα διαδρομών που ενσωματώνουν περίπλοκα εννοιολογικά διαγράμματα.
- Έπειτα τα SEM παρέχουν μία μοναδική ανάλυση που εξετάζει ταυτόχρονα θέματα και της μέτρησης αλλά και της πρόβλεψης.
- Για τα τυπικά μοντέλα με αφανείς παράγοντες τα SEMπαρέχουν ευέλικτα και ισχυρά μέσα για την ταυτόχρονη αξιολόγηση της ποιότητας της μέτρησης καθώς και την μέτρηση των σχέσεων μεταξύ των παραγόντων.
- Και τέλος,για την επίτευξη μιας επικυρωτικής παραγοντικής ανάλυσης και μίας ανάλυσης διαδρομής,τα δομικά μοντέλα εξισώσεων επιτρέπουν να πλαισιώσουν με περισσότερες ερωτήσεις τα φαινόμενα για τα οποία ενδιαφέρονται.

Τέτοιου είδους αναλύσεις προσφέρουν ιδιαίτερα πλεονεκτήματα για τις προβλεπτικές σχέσεις μεταξύ των αφανών παραγόντων.

Μπορούν να χρησιμοποιηθούν για να αντιπροσωπεύουν τη γνώση σχετικά με τα φαινόμενα που μελετήθηκαν ιδιαίτερα. Τα μοντέλα συνήθως βασίζονται στις προτεινόμενες θεωρίες που περιγράφουν και εξηγούν τα φαινόμενα υπό διερεύνηση. Όταν υπάρχει σφάλμα μέτρησης στην μελέτη ενός μοντέλου, τα δομικά μοντέλα παρέχουν ένα ελκυστικό μέσο για την επίτευξη αυτού του σφάλματος. Αυτή η διαδικασία αποκαλείται επιβεβαιωτική λειτουργία των εφαρμογών SEM.

Ένας κύριος λόγος που χρησιμοποιούνται τα δομικά μοντέλα εξισώσεων είναι ότι παρέχουν έναν μηχανισμό λήψης σφαλμάτων μέτρησης στις παρατηρούμενες μεταβλητές που θεωρούμε σε ένα μοντέλο, σε αντίθεση με την παραδοσιακή ανάλυση παλινδρόμησης η οποία αγνοεί το σφάλμα μέτρησης σε όλες τις επεξηγηματικές μεταβλητές που περιλαμβάνονται στο μοντέλο. Εκτός όμως από την αντιμετώπιση των λαθών μέτρησης τα μοντέλα δομικών εξισώσεων δίνουν την δυνατότητα να μελετήσουν τόσο τις ***άμμεσες αλλά και έμμεσες επιπτώσεις*** των διαφόρων μεταβλητών που περιλαμβάνονται σε ένα μοντέλο.

- *Άμμεσες επιδράσεις*: Είναι οι επιπτώσεις που πηγαίνουν απευθείας από μία μεταβλητή σε μία άλλη μεταβλητή.

- *Έμμεσες επιδράσεις:* Είναι οι επιδράσεις μεταξύ δύο μεταβλητών που προκαλείται από μία ή περισσότερες ενδιάμεσες μεταβλητές.

Ο συνδυασμός των δύο επιπτώσεων συνθέτουν το συνολικό αποτέλεσμα της επεξηγηματικής μεταβλητής σε μία εξαρτημένη μεταβλητή. Παρά το γεγονός ότι μπορεί να χρησιμοποιθεί για την εκτίμηση αυτών και η ανάλυση παλινδρόμησης, η προσέγγιση αυτή είναι κατάλληλη μόνο όταν δεν υπάρχουν μετρήσεις σφαλμάτων στις μεταβλητές πρόβλεψης.

## *5.10: Εκτίμηση της προσαρμοστικότητας των μοντέλων*

Ο σκοπός της εκτίμησης της συνολικής προσαρμογής ενός μοντέλου είναι ο προσδιορισμός του βαθμού στον οποίο το μοντέλο είναι συμβατό με τα εμπειρικά δεδομένα.Στα δομικά μοντέλα εξισώσεων χρησιμοποιείται ένα ευρύ φάσμα δεικτών οι οποίοι προσδιορίζουν την καταλληλότητα προσαρμογής ενός μοντέλου. Οι δείκτες αυτοί εκτιμούν την προσαρμοστικότητα και την προβλεπτικότητα του μοντέλου από τρεις πλευρές:

1. Συνολική προσαρμογή του μοντέλου(absolutefit)
2. Συγκριτική προσαρμογή του μοντέλου(incrementalfit)
3. Φειδωλότητα του μοντέλου(Parsimony)

### *5.10.1: Απόλυτοι δείκτες προσαρμογής*

Οι απόλυτοι δείκτες προσαρμογής προσδιορίζουν τον βαθμό στον οποίο το συνολικό μοντέλο προβλέπει τον παρατηρούμενο πίνακα διακύμανσης ή συσχέτισης. Μερικοί από τους απόλυτους συντελεστές σχυσχέτισης που χρησιμοποιούνται για να αξιολογηθούν τα δομικά μοντέλα εξισώσεων είναι:

- ***Η στατιστική*$x^2$**(chi-square statistic)
  Το κύριο μέτρο καθορισμού της συνολικής προσαρμογής του μοντέλου είναι το ποσοστό πιθανοφάνειας.
  Αυτό είναι και το μόνο μέτρο προσαρμοστικότητας που βασίζεται στη στατιστική και χρησιμοποιείται στα δομικά μοντέλα εξισώσεων. Μειονέκτημα αυτού είναι ότι επηρεάζεται από τις διακυμάνσεις στο μέγεθος του δείγματος. Όσο μικρότερηη τιμή του $x^2$ τόσο καλύτερος ο βαθμός προσαρμογής του μοντέλου. Αν $x^2>$ τότε η διακύμανση του μοντέλου πολύ διαφορετική από την παρατηρούμενη.
- ***Η παράμετρος της μη-κεντρικότητας*(NCP**,Non Centrality Parameter)

- ***Ο δείκτης κατάλληλης προσαρογής(*GFI**,Goodness of Fit Index)

Επινοήθηκε από τον Joreskog και τον Sorbom το 1984 και αποτελεί ένα μέτρο του συνολικού βαθμού προσαρμογής του μοντέλου. Ουσιαστικά υπολογίζει το βαθμό σύγκρισης του μοντέλου σε σύγκριση με το <<μηδενικό>> μοντέλο, δηλαδή με το να μην υπήρχε καθόλου. Μετρά την σχετική ποσότητα της διακύμανσης σε ένα δείγμα που το μοντέλο προβλέπει. Μη στατιστικός δείκτης με τιμές 0 - 1. Αναπαριστά το συνολικό βαθμό προσαρμογής αλλά δεν προσαρμόζεται ανάλογα με τους βαθμούς ελευθερίας. Αν και ο GFI είναι ανεξάρτητος του μεγέθους του δείγματος, η κατανομή των τιμών του δείκτη αυτού επηρεάζεται ισχυρά από το μέγεθος του δείγματος.

- ***Η τετραγωνική ρίζα του μέσου των υπολοίπων***
  (**RMSR**,Root Mean Square Residual)
  Είναι ένας ευρύτατα χρησιμοποιούμενος δείκτης ο οποίος παριστάνει τη διαφορά μεταξύ της διακύμανσης και της συνδιακύμανσης του δείγματος σε σχέση με τις αντίστοιχες ποσότητες όταν το μοντέλο που δοκιμάζεται θεωρείται σωστό. Ο δείκτης παίρνει τιμές μεταξύ του 0 και 1. Όσο ο δείκτης πλησιάζει το 0, τόσο πιο μεγάλος είναι ο βαθμός προσαρμογής. Αν το RMSR έχει την τιμή 0, τότε αυτό σημαίνει πως το μοντέλο ταιριάζει απόλυτα, ενώ αν είναι >1 τότε θεωρούνται υπερβολικά μεγάλες τιμές για τον δείκτη αυτόν. Γενικά, μπορούμε να πούμε ότι για ένα καλό μοντέλο ο δείκτης πρέπει να είναι μικρότερος του 0,05.
- ***Η τετραγωνική ρίζα του λάθους της εκτίμησης***
  (**RMSEA**,Root Mean Square of Approximation)
  Ο δείκτης αυτός προτάθηκε από τους Steiger και Lind το 1980 και μόλις πρόσφατα αναγνωρίσθηκε ως ένα από τα πιο σημαντικά κριτήρια για τα μοντέλα δομικών εξισώσεων. Ο δείκτης αυτός ουσιαστικά απαντά στο ερώτημα πόσο καλά θα προσαρμοζόταν ένα μοντέλο σε σχέση με ένα πρότυπο μοντέλο με άγνωστες αλλά ευνοικές τιμές. Είναι ένας δείκτης ο οποίος εκφράζει την απόκλιση ανά βαθμό ελευθερίας. Αποτελεί ένα μέτρο ασυμφωνίας ενός μοντέλου ανά βαθμούς ελευθερίας. Όταν ένα μοντέλο έχει τέλειο βαθμό προσαρμογής τότε ο δείκτης είναι 0.
  Αν 0,05<RMSEA<0.08 τότε το μοντέλο έχει μέτρια προσαρμογή, ενώ αν είναι >0,1 τότε το μοντέλο έχει κακή προσαρμογή. Το πρόγραμμα Lisrel, που επιλύει τα μοντέλα δομικών εξισώσεων υπολογίζει επίσης το διάστημα εμπιστοσύνης του δείκτη RMSEA, όσο πιο μιρό είναι το διάστημα αυτό, τόσο μεγαλύτερη ακρίβεια δηλώνεται ότι υπάρχει στο υπολογισμένο δείκτη. Επίσης, υολογίζεται η πιθανότητα η τιμή του RMSEA να είναι εντός του διαστήματος αυτού. Για παράδειγμα, αν RMSEA=0.04, τότε το μοντέλο έχει πολύ καλή προσαρμογή καθώς το RMSEA<0.05, το 90% των τιμών του RMSEA θα είναι μεταξύ 0.03 και 0.05 και η πιθανότητα να βρεθεί ο δείκτης ανάμεσα σε αυτό το διάστημα είναι αρκετά μεγάλη, 78%, έτσι το υποθετικό

μοντέλο γίνεται αποδεκτό. Εάν το δείγμα είναι μικρό(<1000) και οι παράμετροι που ορίζονται είναι πολλοί(>12) το διάστημα εμπιστοσύνης συνήθως υπολογίζεται μεγάλο.

- ***Ο αναμενόμενος δείκτης επιβεβαίωσης*(ECVI**, Expected Cross-Validation Index)
  Ο δείκτης αυτός ουσιαστικά μετράει την ασυμφωνία μεταξύ του πίνακα συνδιακύμανσης του προσαρμοσμένου μοντέλου του δείγματος που μελετάται και του αναμενόμενου πίνακα συνδιακύμανσης που θα προέκυπτε σε ένα άλλο δείγμα του ίδιου δείγματος. Για το λόγο αυτό υπολογίζονται ταυτόχρονα ο δείκτης ECVI για το κορεσμένο μοντέλο και ο δείκτης ECVI για το ανεξάρτητο μοντέλο. Κορεσμένο θεωρείται το μοντέλο που το πλήθος των εκτιμώμενων παραμέτρων είναι ίσο με το πλήθος των δεδομένων και ανεξάρτητο θεωρείται το μοντέλο όπου όλες οι μεταβλητές είναι ανεξάρτητες μεταξύ τους, δηλαδή οι συσχετίσεις μεταξύ των μεταβλητών είναι μηδενικές.

### *5.10.2: Δείκτες της επαυξητικής προσαρμογής*

Οι δείκτες της επαυξητικής προσαρμογής συγκρίνουν το προτεινόμενο μοντέλο με κάποιο μοντέλο βάσης , το οποίο συχνά αναφέρεται και ως το μηδενικό μοντέλο. (Σε πολλές περιπτώσεις το μηδενικό μοντέλο αποτελείται από μία μόνο δομή, με όλους τους δείκτες του να υπολογίζουν τέλεια τη δομή αυτή).

Δείκτες επαυξητικής προσαρμογής:

- ***Ο διορθώμενος δείκτης κατάλληλης προσαρμογής***
  (**AGFI**,Adjusted-Goodness-of-Fit-Index)
  Θεωρείται μία επέκταση του GFI. Παίρνει τιμές μεταξύ του 0 και 1, με 1 να θεωρείται η κατάλληλη προσαρμογή. Η συμπεριφορά του δείκτη αυτού γενικά είναι παρόμοια με του GFI, όπως επίσης παρουσιάζει ομοιότητες και με εκείνη του PNFI. Για πολύ μεγάλα δείγματα η τιμή είναι αυξημένη, ενώ για πολύ μικρά δείγματα υπάρχει μία περιπτωση υποτίμησης του βαθμού προσαρμογής.
- ***Κανονικοποιημένος δείκτης προσαρμογής*(NFI**,Normed-Fit-Index)
  Είναι ένας από του δημοφιλέστερους δείκτες. Παίρνει και εκείνος τιμές μεταξύ του 0 και 1, με 1 να θεωρείται η τέλεια προσαρμογή. Μετράει το βαθμό προσαρμογής του ελεγχόμενου μοντέλου με το μηδενικό μοντέλο(nullmodel) και επινοήθηκε από τους Bentler και Bonnett το 1980. Με βάση την προκύπτουσα τιμή, διαπιστώνεται η αναλογία προσαρμογής του δοκιμαζόμενου μοντέλου με το μηδενικό.
  Για παράδειγμα, αν η τιμή είναι 0,5 τότε έχει επιτευχθεί προσαρμογή κατά 50%. Αποδεκτές θεωρούνται τιμές μεγαλύτερες του 0,9.
  Εάν η τιμή είναι <0,8 τότε το μοντέλο θα πρέπει να επαναπροσδιοριστεί. Για μικρά δείγματα μπορεί να δημιουργηθεί πρόβλημα υποτιμημένου βαθμού προσαρμογής.
- ***Ο σχετικός δείκτης προσαρμογής*(RFI**,Relative-Fit-Index)

- ***Ο επαυξητικός δείκτης προσαρμογής*(IFI**,Incremental Fit Index)
  Μοιάζει αρκετά με τον συγκρτικό δείκτη προσαρμογής(CFI) των οποίων το εύρος τιμών είναι από 0(καθόλου προσαρμογή) εώς και 1(τέλεια προσαρμογή).
- ***Ο συγκριτικός δείκτης προσαρμογής*(CFI**, Comparative-Fit-Index)
  Ο δείκτης αυτός έχει ένα πλεονέκτημα σε σχέση με τους υπόλοιπους δείκτες καθώς δεν παρουσιάζει τον κίνδυνο υποεκτίμησης της προσαρμογής των δεδομένων εξαιτίας μικρού μεγέθους. Ορίστηκε από τον Bentler το 1990 και βασίστηκε σε μία σύγκριση του μοντέλου της υπόθεσης με το <<μοντέλο βάσης>> ή <<ανεξάρτητο μοντέλο>>.
  Οι τιμές κυμαίνονται απο 0-1.

Στην ουσία όλοι αυτοί οι δείκτες αναπαριστούν τα αποτελέσματα συγκρίσεων ανάμεσα στο εκτιμώμενο καιστο μηδενικό μοντέλο. Οι τιμές των δεικτών αυτών βασίζονται μεταξύ του μηδέν και του ένα, όπου οι μεγαλύτερες τιμές δείχνουν υψηλότερα επίπεδα προσαρμοστικότητας.

### *5.10.3: Δείκτες φειδωλότητας*

Οι δείκτες φειδωλότητας σχετίζουν την καταλληλότητα προσαρμογής του μοντέλου με τον αριθμό των ετκιμώμενων παραγόντων που απαιτούνται.

Βασικός στόχος τους είναι, να διαγνωσθεί εάν η προσαρμογή του μοντέλου έχει επιτευχθεί χρησιμοποιώντας υπερβολικά μεγάλο αριθμό παραγόντων για την <<περιγραφή>> των δεδομένων. Με τον όρο φειδωλότητα εκφράζεται η επίτευξη υψηλών βαθμών προσαρμογής ανά χρησιμοποιούμενο βαθμό ελευθερίας. Ουσιαστικά επιδιώκουμε να έχουμε περισσότερη φειδωλότητα, δηλαδή υψηλότερο βαθμό προσαρμογής ανά χρησιμοποιούμενο βαθμό ελευθερίας.

Τυπικοί δείκτες φειδωλότητας είναι:

- ***Ο φειδωλός ικανοποιημένος δείκτης προσαρμογής***
  (**PNFI**,Parsimonious-Normed-Fit-Index)
  Ο δείκτης αυτός προκύπτει από μία τροποποίηση του NFI. Ο δείκτης PNFI λαμβάνει υπόψην του τον αριθμότων βαθμών ελευθερίας που χρησιμοποιούνται για την επίτευξη ενός συγκεκριμένου επιπέδου προσαρμογής. Υψηλότερες τιμές του PNFI είναι καλύτερες. Η κύρια χρήση του είναι για την σύγκριση των μοντέλων με διαφορετικούς βαθμούς ελευθερίας.
- ***Ο φειδωλός δείκτης καταλληλότητας προσαρμογής***
  (**PGFI**,Parsimonious-Goodness-of-Fit-Index)
  Ο δείκτης αυτός τροποποιεί το GFI με διαφορετικό τρόπο από τον AGFI. Ο PGFI βασίζεται στην φειδωλότητα του εκτιμώμενου μοντέλου, ενώ ο AGFI βασίζεται στους βαθμούς ελευθερίας στο εκτιμώμενο και στο μηδενικό μοντέλο. Η τιμή του PGFI κυμαίνεται από 0-1 με τις υψηλότερες εξ'αυτών να φανερώνουν μεγαλύτερη φειδωλότητα για το μοντέλο.

- ***Το πληροφοριακό κριτήριο του Ak***aike(**AIC**,Akaike-Information-Criterion) Το μέτρο αυτό βασίζεται στη Στατιστική θεωρία πληροφοριών και χρησιμοποιείται για σύγκριση μεταξύ μοντέλων με διαφορετικό πλήθος δομών

$$AIC = \chi^2 + 2 - Αριθμός\ των\ εκτιμώμενων\ παραμέτρων$$

Από τους παραπάνω δείκτες αναλύθηκαν οι πιο σημαντικοί.

Οι Bentler και Bonett πρότειναν τους δείκτες NormedIndex και NonnormedIntex, οι οποίοι δεν επηρεάζονται από το μέγεθος του δείγματος για την αντιμετώπιση προβλημάτων εφραμογής του προτεινόμενου μοντέλου.Αν η τιμή είναι μεγαλύτερη του 0.9 τότε το μοντέλο έχει αρκετά καλή εφαρμογή στα δεδομένα

| ***Τρόπος αξιολόγησης μοντέλου*** | ***Ονομασίες δεικτών*** | ***Ιδιότητες*** |
|---|---|---|
| <u>Απόλυτοι Δείκτες</u> [προσδιορίζουν τον βαθμό στον οποίο το συνολικό μοντέλο προβλέπει τον παρατηρούμενο πίνακα διακύμανσης ή συσχέτισης.] | • Σττατιστικό$x^2$ (chi-square statistic)<br>• Ο δείκτης κατάλληλης προσαρμογής (GFI)<br>• Η τετραγωνική ρίζα του μέσου των υπολοίπων (RMSR)<br>• Η τετραγωνική ρίζα του λάθους της εκτίμησης (RMSEA)<br>• Ο αναμενόμενος δείκτης επιβεβαιώσης (ECVI) | • Όσο $\ll x^2$τόσο καλύτερος δείκτης προσαρμογής<br>• Υπολογίζει το βαθμό σύγκρισης του μοντέλου με το ‘’μηδενικό’’ μοντέλο<br>• Δείκτης διαφοράς διακύμανσης-συνδιακύμανσης και παίρνει τιμές 0-1.<br>• Εκφράζει την απόκλιση ανά βαθμό ελευθερίας. RMSEA=0 έχει τέλειο βαθμό προσαρμογής.<br>• Μετράει την ασυμφωνία μεταξύ του πίνακα συνδιακύμανσης του προσαρμοσμένου μοντέλου που μελετάται και του αναμενόμενου πίνακα συνδιακύμανσης. |
| <u>Δείκτες επαυξητικής προσαρμογής</u> [συγκρίνουν το προτεινόμενο μοντέλο με κάποιο μοντέλο βάσης , το οποίο συχνά αναφέρεται και ως το μηδενικό μοντέλο.] | • Ο διορθωμένος δείκτης καταλληλότητας προσαρμογής (AGFI)<br>• Κανονικοποιημένος δείκτης προσαρμογής (NFI)<br>• Ο σχετικός δείκτης προσαρμογής (RFI)<br>• Ο συγκριτικός δείκτης προσαρμογής (CFI) | • Μία επέκταση του GFIμε τιμές 0-1.<br>• Παίρνει τιμές 0-1 και μετράει το βαθμός προσαρμογής του ελεγχόμενου με το μηδενικό μοντέλο.<br>• Μοιάζει με τον CFI, με τιμές 0-1.<br>• Τιμές 0-1, και εφαρμόζεται σε μικρό μέγεθος δεδομένων. |
| <u>Δείκτες φειδωλότητας</u> [σχετίζουν την καταλληλότητα προσαρμογής του μοντέλου με τον αριθμό των ετκιμώμενων παραγόντων που απαιτούνται] | • Ο φειδωλός ικανοποιημένος δείκτης προσαρμογής (PNFI)<br>• Ο φειδωλός δείκτης καταλληλότητας προσαρμογής (RGFI)<br>• Το πληροφοριακό κριτήριο του Akaike (AIC)και BIC | • Τροποποίηση του NFI, χρησιμοποιείται κυρίως για τη σύγκριση μοντέλων με διαφορετικούς βαθμούς ελευθερίας.<br>• Με τιμές 0-1, εξετάζει την φειδωλότητα του εκτιμώμενου μοντέλου.<br>• Οι μικρότεροι δείκτες δείχνουν την υπεροχή της δομής ενός μοντέλου έναντι άλλων |

*Πίνακας 6: Πίνακας πιο σημαντικών δεικτών*

## *5.11: Λογισμικό για δομικά μοντέλα εξισώσεων*

Κάποια από τα γνωστότερα πακέτα στατιστικών προγραμμάτων που χρησιμοποιούνται στη διαδικασία ανάλυσης των μοντέλων δομικών εξισώσεων είναι τα:

- AMOS(Analysis of Moment Structures)
- LISREL(Linear Structural RELations)
- CALIS(Covariance Analysis and Linear Structural Equation)
- EQS(Equations)
- LISCOMP(Linear Structural Equations with a Comprehensive Measurement Model)
- RAMOVA(Reticular Action Model or Near Approximations)
- SEPATH(SEM and Path Analysis)

Το πιο διαδεδομένο αλλά και εκείνο που χρησιμοποιείται περισσότερο είναι το AMOS καθώς:

1. Προσφέρει ιδιαίτερα φιλικό περιβάλλον εργασίας
2. Δεν απαιτείο γνώσεις προγραμματισμού
3. Υπολογίζει περισσότερους από 20 δείκτες προσαρμοστικότητας του μοντέλου
4. Τα αποτελέσματα των αναλύσεων παρουσιάζονται ταξινομημένα με τέτοιο τρόπο ώστε να γίνονται εύκολα κατανοητά
5. Τα αποτελέσματα υποστηρίζονται πλήρως από το θεωρητικό τους υπόβαθρο

Πέρα όμως από αυτό αρκετά διαδεδομένο είναι και το πρόγραμμα Lisrel καθώς είναι το πρώτο που εμφανίστηκε για την επίλυση δομικών εξισώσεων (Byrne, 1998).

## *5.12: Σχέση ανάλυσης παραγόντων και δομικών μοντέλων εξίσωσης*

Όπως , έχει αναφερθεί και σε προηγούμενο κεφάλαιο η ανάλυση παραγόντων ή αλλιώς παραγοντική ανάλυση χωρίζεται:

- Στην διερευνητική ανάλυση παραγόντων
- Στην επιβεβαιωτική ανάλυση παραγόντων

Στην $1^{η}$ περίπτωση υποθέτουμε ότι η διακύμανση μιας παρατηρήσιμης μεταβλητής είναι συνάρτηση ενός αριθμού από παράγοντες, οι οποίοι αντιστοιχούν στις ποικίλες διαστάσεις της επίδοσης που εκπροσωπεί αυτή η μεταβλητή. Εφαρμόζεται όταν ο τρόπος με τον οποίο οι παρατηρήσιμες μεταβλητές ανάγονται σε λανθάνουσες δομές είναι άγνωστος.Στην $2^{η}$ περίπτωση εφαρμόζεται στον έλεγχο της υπόθεσης ότι υπάρχει ένα συγκεκριμένο πρότυπο μεταξύ των παρατηρίσιμων μεταβλητών και των παραγόντων.Η επιβεβαιωτική ανάλυση παραγόντων συνίσταται στον έλεγχο μοντέλων δομικών εξισώσεων που γίνεται πάνω σε ένα δείγμα δεδομένων. Το δομικό μοντέλο στην περίπτωση αυτή περιλαμβάνει έναν αριθμό από παρατηρήσιμες και λανθάνουσες μεταβλητές.

### *5.12.1: Πλεονεκτήματα των SEM υπερ της παλινδρόμησης*

Ένα από σημαντικότερα μέρη μιας έρευνας είναι η κατάλληλη επιλογή της μεθολογίας. Μετά την πολυμεταβλητή τεχνική, το μέτρο που εφαρμόστηκε για την εγκυρότητα και την αξιοπιστία αξιολόγησης των μέτρων είναι τα δομικά μοντέλα εξισώσεων. Η πρώτη γενιά μεθόδων, όπως οι πολλαπλές παλινδρομήσεις σκοπό είχαν την πρόβλεψη, ενώ η πρόθεση της συσχέτισης ήταν να αξιολογηθεί η σχέση μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών.

Ένα από τα καλύτερα προγνωστικά διακύμανσης σε ένα διάστημα όπου οι εξαρτώμενες μεταβλητές είναι από πολλαπλές παλινδρομήσεις οι οποίες αποτελούν μία μέθοδο για να τον προσδιορισμό του μοντέλου της σχέσης μεταξύ των εξαρτώμενων μεταβλητών όπως $Y$και ανεξάρτητων μεταβλητών όπως $X$. Εάν η επεξηγηματική μεταβλητή αποτελείται από μία μεταβλητή, αυτή θεωρείται απλή παλινδρόμηση, ενώ για περισσότερες επεξηγηματικές μεταβλητές θωρείται πολλαπλή παλινδρόμηση.

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i}$$

Όπου,

$\boldsymbol{\beta_0, \beta_1}$**:** είναι παράμετροι εκτίμησης στη γραμμική εξίσωση.

Όταν η ανεξάρτητη μεταβλητή $X$αλλάζει κατά 1 μονάδα, οι τιμές του $\beta_1$ δείχνει την ποσότητα της εξαρτημένης μεταβλητής $Y$, η οποία αλλάζει ενώ οι άλλες ανεξάρτητες μεταβλητές παραμένουν σταθερές.

Στην στατιστική, η γραμμική παλινδρόμηση είναι μία προσέγγιση για την σχέση μεταξύ βαθμωτών μεταβλητών $y$και μιας ή περισσοτέρων επεξηγηματικών μεταβλητών. Η αξιολόγηση θα πρέπει να γίνεται με λογισμικά όπως, LISREL, AMOS, EQS, RAMOVA. Ακόμη, για την ανάλυση πολλών μεταβλητών στο ίδιο χρονικό διάστημα είναι δυνατή η εφαρμογή δομικών μοντέλων εξισώσεων.

## *5.13: Παραδείγματα δομικών μοντέλων εξισώσεων*

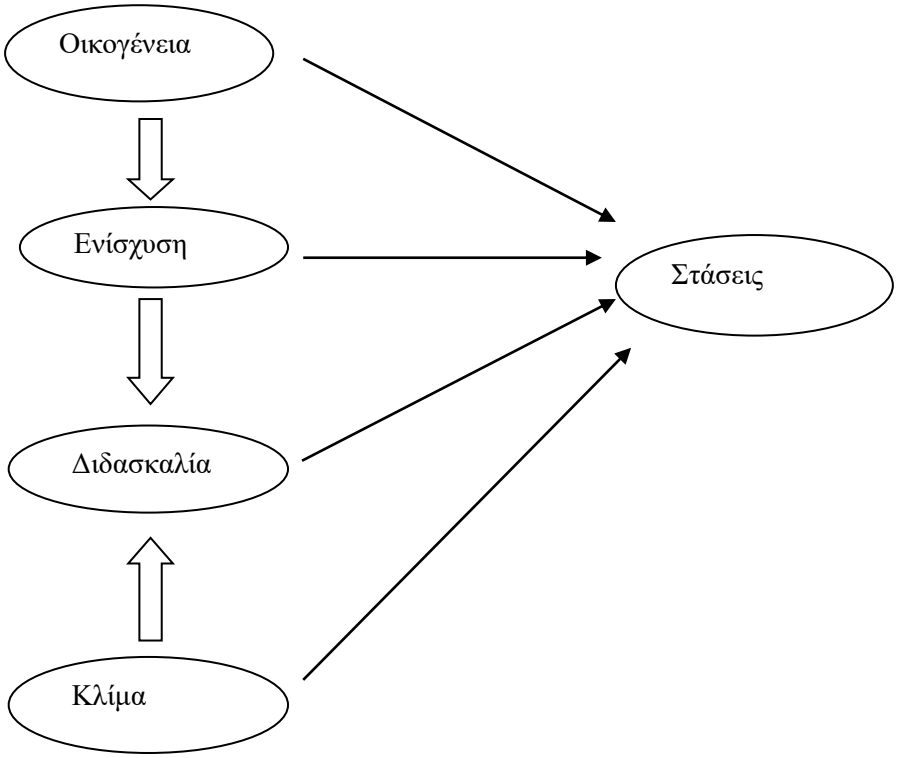
### *1ο Παράδειγμα(Θεωρητικό)*

Παράγοντες που επηρεάζουν τις στάσεις των μαθητών για τις φυσικές επιστήμες. Το δομικό μοντέλο δείχνει ότι η μεγαλύτερη άμεση επίδρση στην διαμόρφωση των στάσεων απέναντι στις φυσικές επιστήμες προέρχονται από την διδασκαλία και ακολουθεί το σχολικό κλίμα και η ενίσχυση των μαθητών.

Η μελέτη έγινε γύρω από 3 διαφορετικούς πληθυσμούς,ο πληθυσμός 1 περιλάμβανε μαθητές των δύο διαδοχικών τάξεων που η πλειοψηφία τους ήταν 9 ετών,ο πληθυσμός 2 περιλάμβανε μαθητές των δύο διαδοχικών τάξεων με πλειοψηφία παιδιών ηλικίας δεκατριών ετών και ο πληθυσμός 3 τελειόφοιτος της μέσης εκπαίδευσης.

*Δομικό μοντέλο:* Τα περισσότερα μοντέλα τονίζουν 2 είδη μεταβλητών,τις μεταβλητέςπου αναφέρονται στο περιβάλλον και εκείνες που αναφέρονται στο μαθητή.

*Θεωρητικό δομικό μοντέλο*




*Εικόνα 23: Διαγραμματική απεικόνιση του προβλήματος*

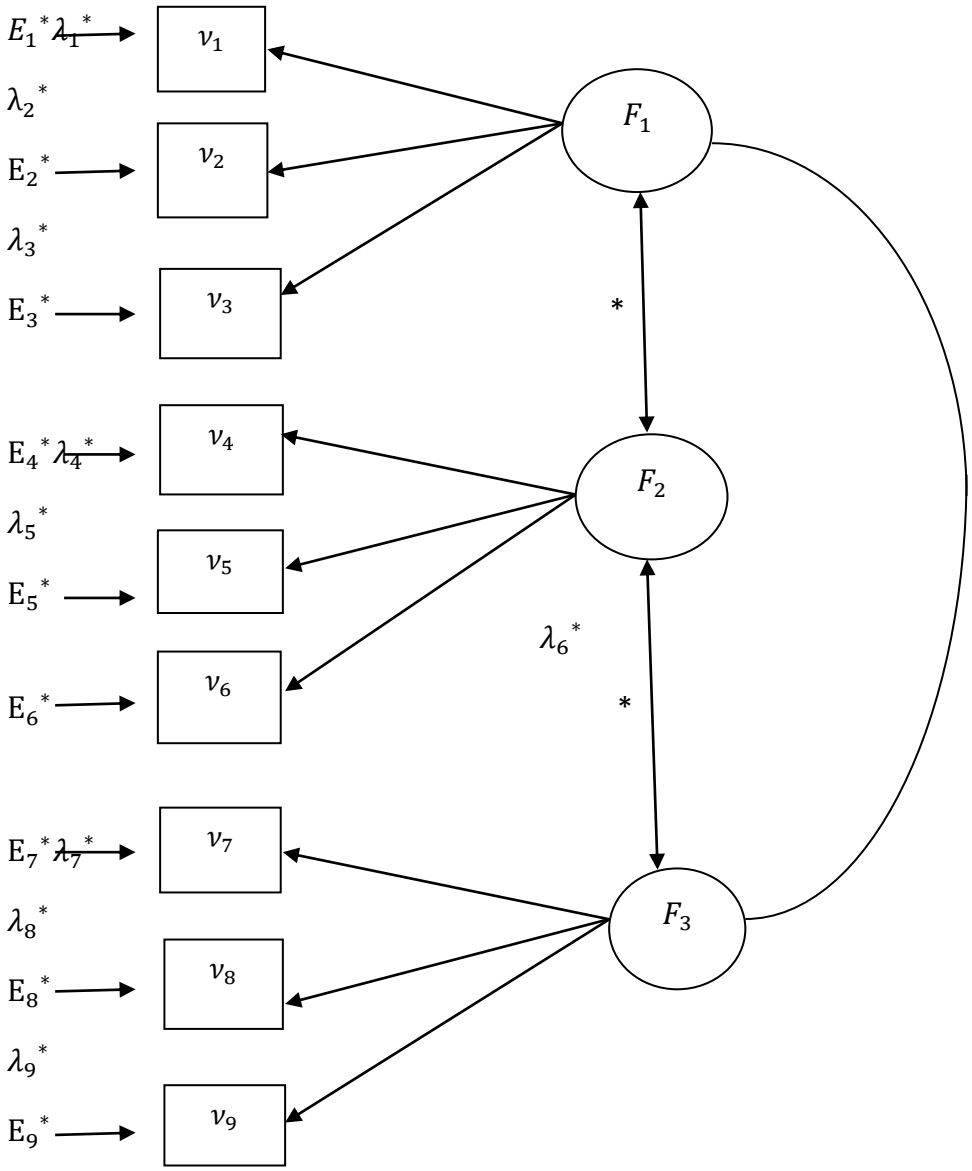
Το μορφωτικό επίπεδο της οικογένειας χρησιμοποιήθηκε ως εξωγενής παράγοντας από το γεγονός ότι οι πιο μορφωμένοι γονείς είναι φυσικό να δίνουν μεγαλύτερηαξία στην μόρφωση και σαν αποτέλεσμα να επιδρούν θετικά στις στάσεις των παιδιών τους προς τα σχολικά μαθήματα.

Αφού εδώ έχουμε πολλές μεταβλητές χρησιμοποιούμε τα δομικά μοντέλα εξισώσεων,τα οποία συχνά χρησιμοποιούνται στην ανάλυση σχέσεων ανάμεσα σε μεταβλητές και άλλους επιστημονικούς κλάδους όπως την κοινωνιολογία,την ψυχολογία,την ιατρική,τα οικονομικά και την εκπαίδευση.

***2ο παράδειγμα (παράδειγμα διαγράμματος διαδρομής)***

Είναι ένα παράδειγμα το οποίο αντιπροσωπεύει τις σχέσεις μεταξύ γονικής κυριαρχίας, της ευφυίας των παιδιών και τα κίνητρα επιτεύγματος. Το μοντέλο αυτοαποτελείται από 9 παρατηρούμενες μεταβλητές, οι οποίες αντιπροσωπεύουν εννέα βαθμολογίες κλιμάκων που ελήφθησαν από ένα δείγμα 245 μαθητών δημοτικού σχολείου. Οι μεταβλητές συμβολίζονται με τις ετικέτες $\nu_1 - \nu_9$(παρατηρούμενες μεταβλητές).

Οι λανθάνουσες μεταβλητές είναι:η γονιακή κυριαρχία($F_1$), η παιδική νοημοσύνη($F_2$) και τα κίνητρα επιτεύγματος($F_3$).Οι λανθάνουσες αυτές μεταβλητές για το καθένα μετράται με τρεις δείκτες με κάθε διαδρομή που αντιπροσωπεύει το συντελεστή επιβάρυνσης της παρατηρούμενης μεταβλητής.




*Εικόνα 24: Διάγραμμα διαδρομής*

Στο παραπάνω σχήμα υπάρχουν συνολικά 12 ανεξάρτητες μεταβλητές. Αυτές είναι οι 3 λανθάνουσες μεταβλητές($F_1, F_2, F_3$) και οι 9 όροι σφάλματος($E_1{}^* - E_9{}^*$).
Όπως μπορούμε να παρατηρήσουμε υπάρχουν αμφίδρομα βέλη.
Οι εξαρτημένες μεταβλητές είναι οι 9 παρατηρούμενες μεταβλητές($\nu_1 - \nu_9$). Αυτές οι εξαρτημένες μεταβλητές λαμβάνουν 2 διαδρομές, μία διαδρομή από την λανθάνουσα μεταβλητή(που αντιπροσωπεύουν τον παράγοντα) και μία από της εναπομένουσας διάρκειας (που αντιπροσωπεύει τον όρο του σφάλματος).

Οι σχέσεις μεταξύ παρατηρούμενων και μη-παρατηρούμενων μεταβλητών καθορίζουν το προτεινόμενο μοντέλο.

Η εξίσωση τώρα για κάθε μία από τις εξαρτώμενες μεταβλητές στο μοντέλο λαμβάνεται γράφοντας μία εξίσωση για κάθε μεταβλητή που παρατηρείται σε σχέση με το πώς εξηγείται στο μοντέλο.Έτσι, έχουμε το ακόλουθο σύστημα εξισώσεων:

$$\nu_1 = \lambda_1 F_1 + E_1$$
$$\nu_2 = \lambda_2 F_2 + E_2$$
$$\nu_3 = \lambda_3 F_3 + E_3$$
$$\nu_4 = \lambda_4 F_4 + E_4$$
$$\nu_5 = \lambda_5 F_5 + E_5$$
$$\nu_6 = \lambda_6 F_6 + E_6$$
$$\nu_7 = \lambda_7 F_7 + E_7$$
$$\nu_8 = \lambda_8 F_8 + E_8$$
$$\nu_9 = \lambda_9 F_9 + E_9$$

Όπου $\lambda_1, \dots, \lambda_9$:είναι οι φορτώσεις του παράγοντα που θα πρέπει να εκτιμάται με βάση τα παρατηρούμενα δεδομένα. Όπως παρατηρούμε από τις παραπάνω εξισώσεις, στο αριστερό της μέλος υπάρχει μόνο μία μεταβλητή η εξαρτημένη αντί για συνδυασμό μεταβλητών και δεν εμφανίζεται καμία ανεξάρτητη μεταβλητή.

Σε αυτό το μοντέλο:

1. Στην μελέτη πληθυσμού ο μέσος όρος κάθε υπολλειπόμενης μεταβλητής εξαφανίζεται
2. Οι υπολλειματικοί όροι είναι ανεξάρτητοι από τις λανθάνουσες μεταβλητές.

Οι παραπάνω παραδοχές δεν ισχύουν μόνο στα δομικά μοντέλα εξισώσεων αλλά για κάθε μοντέλο.

Ένα άλλο σημαντικό χαρακτηριστικό των διαγραμμάτων διαδρομής είναι οι (*) οι οποίοι είναι:

- Σύμβολα των άγνωστων παραμέτρων
- Χρήσιμα για την ορθή λειτουργία της διαδικασίας αλλά και εκτίμησης των περισσότερων προγραμμάτων SEM.

***3ο παράδειγμα***

Έγινε έρευνα για τον προσδιορισμό της επίδρασης που έχει η αντίληψη του υποστηρικτικού περιβάλλοντος και η συναισθηματική νοημοσύνη στην επιχειρηματική συμπεριφορά των εργαζομένων,συνεπώς απαιτείται να εξεταστούν εργαζόμενοι ώστε να εξαχθούν συμπεράσματα.

Η έρευνα έγινε μέσω ερωτηματολογίου που αποτελούνταν από συνολικά 35 ερωτήσεις,που αφορούσαν τις έννοιες προς διερεύνηση και κάποια δημογραφικά στοιχεία.Οι έννοιες προς διερεύνηση είναι:

- Αντίληψη του υποστηρικτού περιβάλλοντος και η επιχειρηματική συμπεριφορά
- Συναισθηματικήνοημοσύνη και η επιχειρηματική συμπεριφορά

Τα ερωτηματολόγια δόθηκαν σε εργαζομένους 8 διαφορετικά είδη εταιρειών και προσπάθησε η έρευνα να είναι όσο πιο αντικειμενική γίνεται και το δείγμα όσο πιο αντιπροσωπευτικό για να μας οδηγήσει σε ασφαλή συμπεράσματα.Απαντήθηκαν από 232 άτομα εκ των οποίων οι 113 γυναίκες και 119 άντρες με εύρος ηλικιών 20-61 ετών.

Σε θεωρητικό επίπεδο (σύμφωνα με τη θεωρία του υποστηρικτικού περιβάλλοντος) τρεις είναι οι βασικοί τύποι ευνοικής μεταχείρησης που αναλαμβάνουν οι εργαζόμενοι από την επιχείρηση(δίκαια μεταχείρηση,εποπτιακή υποστήριξη,ανταμοιβές-συνθήκες εργασίας).

Για την εκτίμηση της ισχύος,της προσαρμοστικότητας και της προβλεπτικότητας του ερευνητικού μοντέλου επιλέχθηκε κυρίως η πολυμεταβλητή ανάλυση(διερευνητική και επιβεβαιωτική παραγοντική).Έγινε χρήση των SEMλόγω αναγκών της επιβεβαιωτικής παραγοντικής ανάλυσης.Καθώς οι διαδικασίες ανάκλυσης που προσφέρουν επιτρέπουν στους ερευνητές να εξετάσουν την προτεινόμενη υποθετική δομή του μοντέλου συνολικά για το σύνολο των σχέσεων μεταξύ εξαρτημένων και ανεξάρτητων μεταβλητών ταυτόχρονα.

Υπολογίζουν εκτιμήσεις για τους παράγοντες του μοντέλου(διακυμάνσεις,συνδιακυμάνσεις των παραγόντων,υπολογισμό διακύμανσης των καταλοίπων και λαθών) και τον βαθμό προσαρμογής τους με τα δεδομένα.

Στην συγκεκριμένη έρευνα μετρήθηκαν συγκεκριμένες πτυχές της συναισθηματικής συμπεριφοράς,η ικανότητα έκφρασης των ατομικών συναισθημάτων ενός ατόμου,η κατανόηση των συναισθημάτων των άλλων,η χρήση των συναισθημάτων μας,η διαχείρηση των συναισθημάτων μας και συγκεκριμένα σημεία τόσο της επιχειρηματικής συμπεριφοράς αλλά και της αντίληψης του υποστηρικτικού περιβάλλοντος.

Για περιορισμό τυχόν σφαλμάτων μέτρησης και βελτίωση των ψυχομετρικών ιδιοτήτων των μεταβλητών χωρίσαμε τις ερωτήσεις που αφορούν κάθε έννοια σε κατηγορίες ανάλογα με τις πτυχές τις εκάστοτε έννοιας που θέλουμε να εκτιμήσουμε.Έτσι δημιουργήσαμε 4 κατηγορίες για την συναισθηματική νοημοσύνη και την αντίληψη του υποστηρικτού περιβάλλοντος και 3 για την επιχειρηματική συμπεριφορά.Πρώτα προσαρμόσαμε ένα μετρικό μοντέλο,το οποίο παρείχε τις συνδέσεις μεταξύτων ερωτήσεων του μοντέλου και των παραγόντων που έχει ορισθεί

ότι καθορίζουν και έπειτα επιλέξαμε το δομικό μοντέλο που καθορίζει την προσαρμοστικότητα με τα δεδομένα.

Επιπλέον,για λόγους εγκυρότητας,συγκρίναμε το μετρικό μοντέλο με ένα ίδιο μοντέλο στο οποίο θέσαμε τις συσχετίσεις ίσες μεταξύ τους και ίσες με την μονάδα.Το μοντέλο με τους μεγαλύτερους δείκτες προσαρμογής θα είναι και το πλέον έγκυρο.

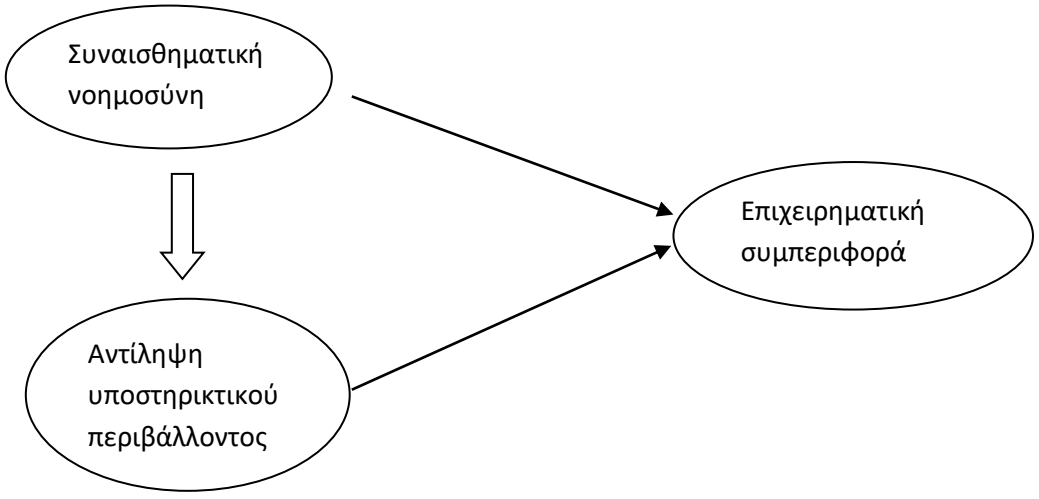
Σύμφωνα με την θεωρία των δομικών μοντέλων θα πρέπει τα στοιχεία που χρησιμοποιούνται για εργασία να ακολουθούν κανονική κατανομή ή αν δεν ακολουθούν υποχρεωτικά θα πρέπει να εφαρμοστεί κάποια διαδικασία κανονικοποίησης.Αν παραβλέψουμε αυτό το στοιχείο η μη-κανονικότητα θα επηρεάσει την ακρίβεια των αποτελεσμάτων,δηλαδή ο ερευνητής θα έχει λανθασμένη άποψη σε σχέση με την προσαρμογή του μοντέλου στα δεδομένα.Γίνεται διάγραμμα μέσων όρων και τυπικών αποκλίσεων των τριών μεταβλητών.

| | 1 | 2 | 3 |
|---|---|---|---|
| Σύνολο | 232 | 232 | 232 |
| Άκυρες μετρήσεις | 0 | 0 | 0 |
| Μέσος όρος | 3,8533 | 2,9841 | 3,6153 |
| Τυπικές αποκλίσεις | 0,48181 | 0,61784 | 0,51752 |

*Πίνακας 6: Πίνακας περιγραφικών*

Από διαγράμματα παρατηρούμε ότι και οι τρεις έννοιες ακολουθούν κανονική κατανομή.Το μοντέλο που έχουμε κατασκευάσει υποθέτει ότι η συναισθηματική νοημοσύνη και η αντίληψη για το υποστηρικτό περιβάλλον επηρεάζουν και μάλιστα θετικά την επιχειρηματική συμπεριφορά,δηλαδή θελουμε να αποδείξουμε ότι οι δύο πρώτες έννοιες ευθύνονται ίσως κατά ένα μεγάλο ποσοστό για την διαφοροποίηση που παρατηρείται στην επιχειρηματική συμπεριφορά.

Μέσω του προγράμματος AMOS (του SPSS) το μοντέλο που προκύπτει είναι το ακόλουθο:




*Εικόνα 25: Διαγραμματική απεικόνιση του προβλήματος*

Το ζητούμενο είναι να αποδείξουμε πως πράγματι αυτές οι μεταβλητές συσχετίζονται, κάτι που θα προκύψει από τους σχετικούς συντελεστές. Οι συντελεστές συσχέτισης προέκυψαν από το λογισμικό SPSS και δείχνουν εάν οι έννοιες που εξετάζουμε όντως έχουν κάποια σχέση μεταξύ τους και σε ποιο βαθμό συμβαίνει αυτό.

| | Συναισθηματική Νοημοσύνη | Αντίληψη Υποστηρικτικού περιβάλλοντος | Επιχειρηματική συμπεριφορά |
|---|---|---|---|
| Συναισθηματική Νοημοσύνη | Pearsoncorr.<br>sig.<br>N | 1<br><br>232 | 0,003<br>0,965<br>232 | 0,873<br>0,001<br>232 |
| Αντίληψη Υποστηρικτικού Περιβάλλοντος | Pearsoncorr.<br>sig.<br>N | 0,003<br>0,966<br>232 | 1<br><br>232 | 0,189<br>0,004<br>232 |
| Επιχειρηματική Συμπεριφορά | Pearsoncorr.<br>sig.<br>N | 0,873<br>0,001<br>232 | 0,189<br>0,004<br>232 | 1<br><br>232 |

*Πίνακας 7: Συντελεστές συσχέτισης*

Ο συντελεστής συσχέτισης παίρνει την μέγιστη τιμή του για την συσχέτιση των δύο υποθέσεων.Οι συσχετίσεις που προέκυψαν είναι σημαντικές στο επίπεδο 0,01, δηλαδή εάν επαναλαμβάναμε την συγκεκριμένη έρευνα στα ίδια άτομα κάτω από τις ίδιες συνθήκες η πιθανότητα να μην βρίσκαμε τα ίδια αποτελέσματα είναι 1%.Αφού οι συντελεστές συσχέτισης που προέκυψαν εμφανίζουν θετικά πρόσημα,οι δύο αρχικές υποθέσεις συνδέονται και μάλιστα προς την ίδια κατεύθυνση,δηλαδή η αύξηση του ενός επιφέρει αύξηση του άλλου και το αντίθετο.

Η αξιοπιστία μιας έρευνας,δηλαδή κατά το πόσο οι μετρήσεις είναι απαλλαγμένες από σφάλματα μετράται μέσω του συντελεστή α του Cronbach(δείκτης αξιοπιστίας).Για να θεωρηθεί αξιόπιστη η κλίμακα μέτρησης πρέπει α>0,70. Το Cronbachα υπολογίζεται από το SPSSεάν αθροίσουμε τα σκορ των ερωτήσεων για κάθε χαρακτηριστικό(8 ερωτήσεις υποστηρικτού περιβάλλοντος,16 ερωτήσεις συναισθηματικής νοημοσύνης,6 ερωτήσεις επιχειρηματικής συμπεριφοράς).

α1=(Σqi/8)i=1-8

α2=(Σqi/6)i=9-14

α3=(Σqi/16)i=15-30

1)Για την αντίληψη του υποστηρικτικού περιβάλλοντος:

*Πίνακας 8:* Συντελεστής αξιοπιστίας για την αντίληψη του υποστηρικτικού περιβάλλοοντος

| Cronbach's alpha | N of items |
|---|---|
| 0.816 | 8 |

2)Για την επιχειρηματική συμπεριφορά

*Πίνακας 8* : Συντελεστής αξιοπιστίας για την επιχειρηματική συμπεριφορά

| Cronbach'salpha | Nofitems |
|---|---|
| 0,735 | 6 |

3)Για την συναισθηματική νοημοσύνη

*Πίνακας 8:* Συντελεστής αξιοπιστίας για την συναισθηματική νοημοσύνη

| Cronbach'salpha | Nofitems |
|---|---|
| 0,843 | 16 |

Παρατηρούμε ότι όλοι οι συντελεστές είναι >0,7,άρα οι ερωτήσεις με τις οποίες κατασκευάσαμε το ερωτηματολόγιο μετρούν τα χαρακτηριστικά για τα οποία τις έχουμε προορίσει με αρκετά μεγάλη αξιοπιστία.

Για να δείξει τώρα κατά το πόσο τα αποτελέσματα προσεγγίζουν την πραγματικότητα συγκρίνει το μοντέλο με ένα ίδιο στο οποίο θα έχουμε συσχετίσεις ίσες μεταξύ τους και ίσες με την μονάδα.Το μοντέλο με μεγαλύτερους δείκτες προσαρμογής είναι το πιο έγκυρο.

Οι σημαντικότεροι δείκτες προσαρμογής του μοντέλου μέτρησης στα δεδομένα είναι οι :

**CFI,TLI,IFI**:με εύρος τιμών 0-1

**RMSEA**: Αν RMSEA=0 έχουμε τέλεια προσαρμογή

Αν 0.05<RMSEA<0.08 έχουμε μέτρια προσαρμογή

Αν RMSEA>0,1 έχουμε κακή προσαρμογή

**AIC**:χρησιμοποιείται για σύγκριση μοντέλων

Α)Για το μοντέλο μέτρησης με τις συσχετίσεις όπως προέκυψαν από την στατιστική επεξεργασία,παρατηρούμε ότι εμφανίζει σχεδόν τέλεια προσαρμογή αφού οι δείκτες CFI,TLI,IFIπαίρνουν τιμές >0,95 και ο RMSEA=0.059.

Β)Το υποθετικό μοντέλο με ίσες συσχετίσεις εμφανίζει κακή προσαρμογή στα δεδομένα,με τιμές των δεικτών CGI,TLI,IFI κοντά στο 0,78 και RMSEA=0,142.

Επιπλέον το πρώτο μοντέλο εμφανίζει χαμηλότερες τιμές του δείκτη AIC σε σχέση με το υποθετικό, όπως επίσης και χαμηλότερες τιμές του $\chi^2$,επομένως το μοντέλο που έχει προκύψει από την επεξεργασία θεωρείται έγκυρο, δηλαδή όντως μετράει διαφορετικές έννοιες.

***ΕΚΤΙΜΗΣΗ ΔΟΜΙΚΟΥ ΜΟΝΤΕΛΟΥ***

Ως προς την εκτίμηση του δομικού μοντέλου γενικά μπορούμε να πούμε ότι: Γίνεται κατανοητό πως η συναισθηματική νοημοσύνη κατέχει σημαντικό ρόλο στην επιχειρηματική συμπεριφορά. Εργαζόμενοι με υψηλούς δείκτες συναισθηματικής νοημοσύνης είναι πιο ικανοί να αναγνωρίζουν και να ρυθμίζουν τα συναισθήματά τους αλλά και των άλλων, το οποίο μπορεί να τους οδηγήσει να συμπεριφερθούν επιχειρηματικά.

Αντίστοιχα, εργαζόμενοι που απολαμβάνουν υψηλά επίπεδα υποστήριξης από το εργασιακό τους περιβάλλον, εμφανίζουν υψηλή απόδοση, νιώθουν ικανοποιημένοι από την εργασία τους, αισθάνονται περισσότερο ελεύθεροι να πάρουν ρίσκα προς όφελος της εργασίας τους και επομένως έχουν τη δυνατότητα να αναπτύξουν επιχειρηματική συμπεριφορά.

*ΣΥΝΟΨΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ*

Τα αποτελέσματα της ανάλσυσης δίνουν έμφαση σε δύο κύρια σημεία. **Πρώτον** σε συνδυασμό με την κεντρική έννοια της θεωρίας του υποστηρικτικού περιβάλλοντος,οι εργαζόμενοι που πιστεύουν ότι απολαμβάνουν υψηλά επίπεδα υποαστήριξης από την επιχείρησή τους είναι πολύ πιθανό να απόδώσουν αυτή την υποστήριξη με θετικές για την εργασία συμπεριφορές.Έτσι,η αντίληψη του υποστηρικτικού περιβάλλοντος συνδέεται με την επιχειρηματική συμπεριφορά. Το δεύτερο σημείο αναφέρεται στην επίδραση της συναισθηματικής νοημοσύνης στην επιχειρηματική συμπεριφορά.Η συναισθηματική νοημοσύνη των εργαζομένων επιδρά και μάλιστα με θετικό τρόπο στην επιχειρηματική τους συμπεριφορά,αφού τα άτομα αυτά είναι πιο ικανά στο να αναγνωρίζουν και να ρυθμίζουν τα συναισθηματά τους αλλά και των άλλων.

Τα αποτελέσματα φανερώνουν ότι η συναισθηματική νοημοσύνη εν μέρει είναι υπεύθυνη για την διαφοροποίηση που υπάρχει στην επιχειρηματική συμπεριφορά των εργαζομένων, το οποίο τονίζει την αναγκαιότητα της γνώσης των συναισθηματικών χαρακτηριστικών του ατόμου ώστε να προαχθεί η επιχειρηματική συμπεριφορά.

Η αντίληψη του υποστηρικτικού περιβάλλοντος και η συναισθηματική νοημοσύνη, είναι αρκετά χρήσιμες για την κατανόηση, την προώθηση και τη βελτίωση της επιχειρηματικής συμπεριφοράς στο εσωτερικό επιχειρήσεων και οργανισμών. Οι εργαζόμενοι, όταν αφενός έχουν υψηλούς δείκτες συναισθηματικής νοημοσύνης και αφετέρου έχουν υψηλή αντίληψη του υποστηρικτικού περιβάλλοντος, είναι πιο πιθανό να δράσουν επιχειρηματικά.

## *5.14: Σχέση δομικών μοντέλων με παλινδρόμηση, πολυμεταβλητή ανάλυση και γενικευμένα γραμμικά μοντέλα*

### *5.14.1: Σχέση SEM με παλινδρόμηση και πολυμεταβλητή ανάλυση*

Τα SEMμπορούν να θεωρηθούν μία επέκταση της παλινδρόμησης και της παραγοντικής ανάλυσης, η οποία όμως εξετάζει ταυτόχρονα τις σχέσεις μιας ή περισσοτέρων εξαρτημένων μεταβλητών και μεταξύ δύο ή περισσοτέρων ανεξάρτητων μεταβλητών.

Στην πολλαπλή παλινδρόμηση, ο αναλυτής θεωρεί ότι οι μεταβλητές που συμμετέχουν στο σχηματισμό του μοντέλου έχουν μετρηθεί απόλυτα, ακόμα και αν αυτό δεν ισχύει. Επίσης, εκτιμά μόνο άμεσες επιδράσεις. Από την άλλη το μοντέλο διαδρομής (pathmodel) που δημιουργείται στα SEMεπιτρέπει τη μελέτη της επίδρασης των επιμέρους μεταβλητών στις οποίες μπορεί να αποσυντεθεί μία αρχική μεταβλητή. Δηλαδή, μία ανεξάρτητη μεταβλητή μπορεί να έχει άμεση αλλά και έμμεση επίδραση σε μια εξαρτημένη μεταβλητή.

Ουσιαστικά, τα δομικά μοντέλα εξισώσεων είναι μία πολυμεταβλητή τεχνική η οποία έχει την δυνατότητα να επεξεργάζεται μια σειρά εξαρτημένων γραμμικών σχέσεων, όπου μία ή περισσότερες πολύπλοκες μεταβλητές μπορούν να είναι είτε εξαρτημένες είτε ανεξάρτητες, ανάλογα με την σχέση στην οποία ανήκουν.

Μπορούμε να πούμε ότι, τα μοντέλα δομικών εξισώσεων έχουν τρεις κοινές υποθέσεις με τα άλλα μοντέλα πολυμεταβλητής ανάλυσης:

a. Ανεξάρτητες παρατηρήσεις
b. Τυχαία επιλογή των ερωτώμενων
c. Γραμμικότητα όλων των σχέσεων

Η έλλιψη πολυμεταβλητής κανονικότητας στα μοντέλα αυτά, δημιουργεί πρόβλημα καθώς διογκώνει το $\chi^2$ στατιστικό και δημιουργεί πρόβλημα στις κρίσιμες τιμές για τον καθορισμό της σημασίας των συντελεστών.

### *5.14.2: Σχέση SEM με γενικά γραμμικά μοντέλα*

Τα μοντέλα δομικών εξισώσεων υπόκεινται σε εμπειρικές αναλύσεις οι οποίες χρησιμοποιούν το γενικό γραμμικό μοντέλο και τις επεκτάσεις του.

Τα δομικά μοντέλα εξισώσεων σε μεγάλο βαθμό περιλαμβάνουν και χρησιμοποιούν τα γενικευμένα γραμμικά μοντέλα, τόσο σε θεωρητικό αλλά και στο πρακτικό τους μέρος. Για τον λόγο αυτό έγινε και η πλήρης ανάλυση των γενικευμένων γραμμικών μοντέλων σε ένα από τα παραπάνω κεφάλαια.

## *5.15: Συνοπτικά μερικές παρατηρήσεις για τα δομικά μοντέλα εξισώσεων*

Τα δομικά μοντέλα εξισώσεων θεωρούνται μια καινούργια στατιστική μέθοδος η οποία είναι δυνατόν να μας δώσει μία διαφορετική εικόνα της δομής ή της αλλαγής των μεταβλητών από ότι οι συνήθεις μέθοδοι στατιστικής ανάλυσης. Είναι μία μέθοδος η οποία παρέχει μοναδικές δυνατότητες για την αξιοποίηση των δεδομένων.

Ο σύγχρονος ερευνητής χρειάζεται στατιστικές μεθόδους ή τεχνικές οι οποίες να είναι κατάλληλες για να περιγράψουν με ακρίβεια τη δυναμική κατάσταση των χαρακτηριστικών της συμπεριφοράς σε διαδοχικές φάσεις της ανάπτυξης, να προσδιορίσουν τις σχέσεις μεταξύ χαρακτηριστικών που υπάρχουν σε διαφορετικές χρονικές φάσεις και να επιτρέπουν προβλέψεις για τις μελλοντικές αλλαγές. Τα μοντέλα δομικών εξισώσεων θεωρείται μία μέθοδος η οποία εφαρμόζωντάς την μπορεί να ανταποκριθεί σε πολλές από αυτές τις ανάγκες, καθώς η εφαρμογή της προσφέρει αρκετά ακριβή έλεγχο των χαρακτηριστικών και των ικανοτήτων , δίνουν ακόμη την δυνανότητα προσδιορισμού του είδους και του βαθμού της σχέσης ανάμεσα σε χαρακτηριστικά και ικανότητες, στη διαδρομή του χρόνου και επιτρέπουν τη διατύπωσξ προβλέψεων για την εξέλιξή τους.

Πέρα όμως από τα πλεονεκτήματα τους μπορούμε να πούμε ότι τα δομικά μοντέλα εξισώσεων έχουν και αρκετούς περιορισμούς κατά τη εφαρμογή τους. Οι περιορισμοί αυτοί είναι οι εξής:

- Επειδή εφαρμόζωνται σε δείγματα δεκάδων ή εκατοντάδων ατόμων τα αποτελέσματα της προσαρμογής αφορούν μία ομάδα και όχι ένα άτομο.
- Επιχειρώντας τον έλεγχο λανθάνουσας δομής ενός συνόλου από πραγματικά δεδομένα μπορεί να <<παγιδευτούν>> από τεχνητές συναθροίσες ικανοτήτων που μπορεί να απέχουν από την πραγματική δομή του συστήματος στο οποίο απευθύνονται.
- Έχουν αυστηρές απαιτήσεις ως προς την μεθοδολογική και στατιστική καθαρότητα και ακρίβεια των μετρήσεων.

***Στάδια ανάπτυξης δομικού μοντέλου***

Η διαδικασία – στρατηγική για την ανάπτυξη ενός δομικού μοντέλου είναι η εξής: (Byrne, 1998):

1. Καθορίζεται ένα αρχικό μοντέλο αντλώντας πληροφορίες από την υπάρχουσα θεωρία ή τις μελέτες που υπάρχουν στην βιβλιογραφία και αφορούν το θέμα που προσπαθούμε να μοντελοποιήσουμε. Εάν η βιβλιογραφία είναι ελλιπής, τότε διατυπώνονται από τον εκάστοτε ερευνητή κάποιες υποθέσεις οι οποίες εξετάζονται στην πορεία.
2. Γίνεται εκτίμηση του μοντέλου μέτρησης, στην αρχή ξεχωριστά για κάθε υποθετική δομή του μοντέλου και έπειτα, αν χρειαστεί για κάθε ζεύγος

συνδεόμενων μεταβλητών. Τελικά κάνουμε την εκτίμηση του μοντέλου δομικών εξισώσεων για κάθε δομή, από κοινού με το μοντέλο μέτρησης.

3. Ελέγχεται και κρίνεται ο βαθμός προσαρμογής κάθε μοντέλου που εκτιμήθηκε στο προηγούμενο βήμα. Ιδιαίτερη βάση δίνεται στο Chi-Square και στους δείκτες RMSEA,ECVI,CFI και GFI. Σε κάθε έλεγχο προσαρμογής εάν το μοντέλο που προκύπτει δεν έχει υψηλό βαθμό προσαρμογής και προτείνονται κάποιοι δείκτες τροποποίησης, και επαναλαμβάνεται η διαδικασία από την αρχή αφού πρώτα γίνουν οι τροποποιήσεις αυτές.
4. Ακολουθώντας τα παραπάνω βήματα, καταλήγουμε τελικά σε ένα μοντέλο στο οποίο τα δεδομένα προσαρμόζονται σχετικά καλά και του οποίου οι παράμετροι μπορούν να ερμηνευτούν δίνοντας λογικά συμπεράσματα. Ωστόσο, το γεγονός ότι τα δεδομένα προσαρμόζονται καλά στο μοντέλο δε σημαίνει ότι αυτό το μοντέλο είναι το καλύτερο. Υπάρχει περίπτωση τα ίδια τα δεδομένα να μην είναι λογικά και συνεπή, ιδίως σε μία έρευνα ικανοποίησης με ερωτηματολόγια κατά την συμπλήρωση του ερωτηματολογίου υπάρχουν πολλοί παράγοντες που μπορεί να προκαλέσουν σύγχυση στο ερωτώμενο, όπως βιασύνη, παρανόηση ή ελλιπή κατανόηση των ερωτήσεων και διάφορα άλλα εξωτερικά ερεθίσματα που επηρεάζουν την εγκυρότητα και συνέπεια των συλλεγόμενων δεδομένων. Για να ξεπεραστεί το πρόβλημα των <<μη έγκυρων>> δεδομένων, η λύση είναι το δομικό μοντέλο που δημιουργείται να εξετάζεται όχι μόνο <<μαθηματικά>> αλλά και <<λογικά>>. Ένας άλλος τρόπος για να εκτιμηθεί η εγκυρότητα των δεδομένων και του μοντέλου είναι να μοιραστεί με τυχαίο τρόπο το δείγμα των δεδομένων και να εφαρμοστεί το προσαρμοσμένο δομικό μοντέλο σε κάθε τμήμα του δείγματος ξεχωριστά.

# *Κεφάλαιο 6:Στατιστική συμπερασματολογία και εκτίμηση παραμέτρων*

Με τον όρο στατιστική συμπερασματολογία (StatisticalInference) εννοούμε τη διαδικασία χρησιμοποίησης πληροφοριών από το δείγμα μας με σκοπό της εξαγωγή συμπερασμάτων για τον πληθυσμό ή την ικανότητα διαδικασίας. Για παράδειγμα, η εκτιμητική όπου χρησιμοποιούμε μία στατιστική συνάρτηση για να εκτιμήσουμε την τιμή μιας παραμέτρου ύπο μελέτη πληθυσμού. Στατιστική συμπερασματολογία εννοούμε: μέθοδοι πιθανοφάνειας, έλεγχοι υποθέσεων, στατιστικά τεστ, διαστήματα εμπιστοσύνης.

## *6.1: Τύποι παραμέτρων στα SEM*

Όπως αναφέραμε και στο προηγούμενο κεφάλαιο υπάρχουν κάποιοι τύποι παραμέτρων στα SEM,οι οποίοι έχουν αναλυθεί και περισσότερο*(στο 5.7.2)* αυτοί είναι:

- Ελεύθεροι
- Σταθεροί-Καθορισμένοι
- Περιορισμένοι-Βεβιασμένοι

## *6.2: Βήματα ελέγχου δομικών μοντέλων*

- ***Έλεγχος εκτίμησης των παραμέτρων***
  Δηλαδή, εξετάζετε: αν οι τιμές έχουν το σωστό πρόσημο και το σωστό μέγεθος και αν τα αποτελέσματα αυτά συμβαδίζουν με το θεωρητικό μοντέλο. Ωστόσο μπορεί να υπάρξουν προβλήματα αν εμφανίζονται συσχετίσεις μεγαλύτερες από την μονάδα ή αν κάποιες βγαίνουν αρνητικές. Επιπλέον πρέπει να υπάρχουν σχετικά μικρές τιμές τυπικών σφαλμάτων(standarderrors).
  Μία τυποποίηση των τιμών που λαμβάνουμε γίνεται από το CriticalRatio(CR), το οποίο είναι ένα πηλίκο της μετρούμενης μεταβλητής προς το τυπικό σφάλμα. Για να είναι μία παράμετρος στατιστικά διάφορη του μηδενός θα πρέπει αυτό το πηλίκο σε απόλυτη τιμή να είναι μεγαλύτερο του 1,96.

  Ακόμη, οι τιμές της στατιστικής σημαντικότητας πρέπει να είναι διάφορες του μηδεν.

- ***Αξιολόγηση εγκυρότητας του μοντέλου***
  1. *Εξέτασητωνδεικτών GOF(Goodness of fit)*

Έχουμε τους δείκτες:

| | | |
|---|---|---|
| **Absolute Fit** | Chi-Square | $x^2$ |
| | Akaike Information Criterion | AIC |
| | Browne-Cudeck Criterion | BCC |
| | Bayes Information Criterion | BIC |
| | Consistent AIC | CAIC |
| | Expected Cross-Validation Index | ECVI |

*Πίνακας 8:Πίνακας δεικτών*

Οι παραπάνω δείκτες αξιολογούν πόσο καλά το μοντέλο που κατασκευάστηκε αναπαράγει τα δεδομένα που προέκυψαν από την παρατήρηση. Από τους παραπάνω δείκτες οι πιο σηματικοί είναι:

- **Chi-Square:** $\frac{\boldsymbol{x^2}}{\boldsymbol{df}}$**,**όπου$df$:βαθμοί ελευθερίας του μοντέλου
- $\boldsymbol{AIC = \chi^2 + N(N+1) - 2df}$, όπου N: είναι το μέγεθος του δείγματος και $df$: είναι οι βαθμοί ελευθερίας του μοντέλου. Το κριτήριο αυτό αποτελεί ένα συγκριτικό μέτρο ανάμεσα σε διαφορετικές δομές μοντέλων, οι οποίες καθορίζονται από τις διάφορες πιθανές διασυνδέσεις μεταξύ των διαθέσιμων μεταβλητών. Χαμηλότερες τιμές υποδεικνύουν μια καλύτερη εφαρμογή και έτσι το μοντέλο με χαμηλότερο AICείναι εκείνο με καλύτερη εφαρμογή των δεδομένων.
- $\boldsymbol{BIC = \chi^2 + ln(N)[\frac{k(k+1)}{2} - df]}$, όπου $\boldsymbol{ln(N)}$: είναι ο φυσικός λογάριθμος του αριθμού των περιπτώσεων του δείγματος. Και το κριτήριο αυτό αποτελεί ένα μέτρο σύγκρισης ανάμεσα σε διάφορες δομές μοντέλων και χαμηλότερες τιμές του, δείχνουν το μοντέλο με την καλύτερη προσαρμογή στα δεδομένα.

| | | |
|---|---|---|
| **Incremental Fit** | Normal Fit Index | NFI |
| | Incremental Fit Index | IFI |
| | Tucker-Lewis Index | TLI |
| | Comparative Fit Index | CFI |
| | Relative Non-centrality Fit Index | RNI |

*Πίνακας 9: Πίνακας δεικτών*

Οι παραπάνω δείκτες συγκρίνουν το προς αξιολόγηση μοντέλο με το μοντέλο ανεξαρτησίας, με εκείνο δηλαδή στο οποίο θεωρούμε ότι δεν υπάρχει καμία σχέση μεταξύ των μεταβλητών.

- $TLI = \frac{\left[\frac{\chi^2}{df}(\mu\eta\delta\varepsilon\nu\iota\kappa\acute{o}\ \mu o\nu\tau\acute{\varepsilon}\lambda o\upsilon) - \frac{\chi^2}{df}(\pi\rho o\tau\varepsilon\iota\nu\acute{o}\mu\varepsilon\nu o\upsilon\ \mu o\nu\tau\acute{\varepsilon}\lambda o\upsilon)\right]}{\frac{\chi^2}{df}(\mu\eta\delta\varepsilon\nu\iota\kappa\acute{o}\ \mu o\nu\tau\acute{\varepsilon}\lambda o\upsilon - 1)}$,θεωρείται ένας άλλος δείκτης καλής προσαρμογής που επηρεάζεται με την

προσθήκη παραμέτρων στο μοντέλο και υπολογίζεται από την παραπάνω σχέση. Παίρνει τιμές 0-1, με όσο η τιμή του TLIπλησιάζει την μονάδαμ συνεπάγεται καλύτερη προσαρμογή του μοντέλου.

| | | |
|---|---|---|
| **Parsimony Fit** | Parsimony adjusted NFI | PNFI |
| | Parsimony adjusted CFI | PCFI |
| | Parsimony adjusted GFI | PGFI |

*Πίνακας 10: Πίνακας δεικτών*

Οι παραπάνω δείκτες εκτιμούν την απλότητα του μοντέλου και μπορεί να υποδείξουν στον ερευνητή την ανάγκη να συμπεριλάβει λιγότερες παραμέτρους ώστε να καταλήξει σε απλούστερο μοντέλο.

| | | |
|---|---|---|
| **Other** | Goodnes-of- Fit Index | GFI |
| | Adjusted GFI | AGFI |
| | Root Mean Square Residual | RMR |
| | Standarized RMR | SRMR |
| | Root Mean Square Error of Approximation | RMSEA |

*Πίνακας 11: Πίνακας δεικτών*

- RMSEA=$\frac{\sqrt{\chi^2-df}}{\sqrt{[df(N-1)]}}$, όπου N: το μέγεθος του δείγματος και $df$:οι βαθμοί ελευθερίας του μοντέλου. Είναι η ρίζα των μέσων τετραγώνων των σφαλμάτων εκτίμησης (RMSEA) είναι το απόλυτο μέτρο καταλληλότητας που βασίζεται στην παράμετρο μη κεντρικότητας. Αν το $\chi^2$ είναι μικρότερο του $df$, τότε το RMSEAμηδενίζεται.

***2.*** *Εξέταση πιθανών λαθών στους ορισμούς*

- Από τα τυποποιημένα κατάλοιπα.
  Οι τιμές αυτών δεν πρέπει να ξεπερνούν σε απόλυτη τιμή το 2,58, ενώ κάποιοι θεωρούν ότι οι τιμές δεν πρέπει να υπερβαίνουν κατά απόλυτη τιμή το 4. Σε περίπτωση που υπάρχει απόκλιση η οποία υπερβαίνει τις ανώτερες τιμές ο ερευνητής πρέπει να αποφασίσει αν θα διατηρήσει ή όχι τα στοιχεία αυτά.
- Τους δείκτες διαμόρφωσης(ModificationIndices)
  Τιμές διασπορών, τιμές συνδιασπορών, βάρη παλινδρόμησης

***3.*** *Έλεγχος εννοιολογικής εγκυρότητας και αξιοπιστίας*

Εγκυρότητα σύγκλισης (ConvergentValidity): Εξετάζεται από τρεις παράγοντες:

1.FactorLoadings: τα τυποποιημένα φορτία θα πρέπει να έχουν τιμές πάνω από 0,5 και στην ιδανική περίπτωση τιμές μεγαλύτερες από 0,7.
2.TheaverageVarianceExtracted(AVE): πρέπει να έχει τιμή μεγαλύτερη ή ίση του 0,5.

3.CompositeReliability: Εξετάζει την αξιοπιστία του μοντέλου και πρέπει να έχει τιμές μεγαλύτερες από 0,7.

Εγκυρότητα διάκρισης (DiscriminantValidity): Εξετάζει κατά το πόδο διαφέρουν οι εννοιολογικές κατασκευές μεταξύ τους.
Νομολογική εγκυρότητα (NomologicalValidity): Ελέγχει κατά πόσο οι *συσχετίσεις μεταξύ των εννοιολογικών κατασκευών έχουν νόημα.*

***Αξιοπιστία***
*Ο δείκτης του Cronbach*

Οι George&Mallery (2003) προτείνουν την εξής κλίμακα
>0,90: εξαιρετικός

>0,80: Καλός

>0,70: Απαράδεκτος

>0,60: Αμφισβητήσιμος

>0,50: Φτωχός

>0,40: Απορριπτέος

*Μονοδιαστικότητα(Unidimentionality)*
Μπορεί να ελεχθεί μέσω της ExplanatoryFactorAnalysis (EFA). Κάθε ομάδα πρέπει να φορτώνεται σε μία μόνο εννοιολογική κατασκευή.

## *6.3: Μέθοδοι εκτίμησης παραμέτρων στα SEM*

Υπάρχουν τέσσερις κύριοι μέθοδοι που χρησιμοποιούνται στα περισσότερα προγράμματα, τα οποία είναι:

- Μη σταθμισμένων ελαχίστων τετραγωνων(ULS)
- Μέγιστης πιθανοφάνειας(ML)
- Γενικευμένων ελαχίστων τετραγώνων(Μεγάλα δείγματα)
- Σταθμισμένων ελαχίστων τετραγώνων

Η 1$^{η}$ μέθοδος θεωρείται ως η πιο κατάλληλη λειτουργία, το απλό μη σταθμισμένο άθροισμα των τετραγώνων των διαφορών είναι μεταξύ των αντίστοιχων στοιχείων του S και του υποδείγματος Σ του πίνακα συνδιασποράς. Οι εκτιμήσεις που επιλέγονται για τις παραμέτρους του μοντέλου όταν η μέθοδος αποκτά την ελάχιστη τιμή. Η μέθοδος εφαρμόζεται στην πράξη όταν παρόμοιες κλίμακες μέτρησης αποτελούν την βάση των μεταβλητών που θα αναλυθούν.

Οι άλλες τρεις μέθοδοι εκτίμησης βασίζονται στο ίδιο άθροισμα των τετραγώνων αλλά σε είδικές περιπτώσεις έχουν χρησιμοποιηθεί για να πολλαπλασιαστεί κάθε ένα από τα τετράγωνα. Η μέθοδος της μέγιστης πιθανοφάνειας και των γενικευμένων ελαχίστων τετραγώνων χρησιμοποιούνται όταν δεδομένα είναι κανονικά κατανεμημένα. Η υπόθεση της κανονικότητα είναι αρκετά συχνή στην παραγοντική ανάλυση, η οποία μπορεί να εξεταστεί χρησιμοποιώντας οποιοδήποτε στατιστικό

πακέτο γενικής χρήσης. Ο απλούστερος τρόπος για να εξεταστεί η μονοπαραγοντική ομαλότητα είναι να εξεταστεί η ασσυμετρία και η κύρτωση.

Τα τελευταία χρόνια έχει δειχθεί ότι η μέθοδος μέγιστης πιθανότητας(ML) μπορεί επίσης να χρησιμοποιηθεί με μικρές αποκλίσεις από την κανονικότητα. Θεωρείται μία μέθοδος η οποία σε γενικές γραμμές καθορίζει τις εκτιμήσεις για τις παραμέτρους του μοντέλου που μεγιστοποιούν την πιθανότητα της παρατήρησης των διαθέσιμων δεδομένων, ακόμη και αν κάποιος ήθελε να συλλεγούν δεδομένα από τον ίδιο πληθυσμό πάλι το ίδιο θα γινόταν. Αυτή η μεγιστοποίηση επιτυγχάνεται με την επιλογή των παραμέτρων του μοντέλου κατά τέτοιο τρόπο ώστε να ελαχιστοποιηθεί η λειτουργία προσαρμογής που περιγράφηκε νωρίτερα.

Με περισσότερες αποκλίσεις από το φυσιολογικό, *η μέθοδος σταθμισμένων ελαχίστων τετραγώνων* μπορεί να χρησιμοποιηθεί εφ'όσον το μέγεθος του αναλυθέντος δείγματος είναι μεγάλο. Το μέγεθος του δείγματος παίζει σημαντικό ρόλο σχεδόν σε κάθε στατιστική τεχνική που εφαρμόζεται στην πράξη. Μπορούμε να τονίσουμε ότι παρόλο που δεν υπάρχει κάποια συμφωνία μεταξύ των ερευνητών, δηλαδή ότι όσο μεγαλύτερο είναι το δείγμα τόσο πιο σταθερή είναι η εκτίμηση των παραμέτρων, δεν υπάρχει συμφωνία ως προς το τι συνιστά μεγάλο. Είναι ένα συχνό θέμα που όμως δεν έχουν δοθεί καποιοι συγκεκριμένοι κανόνες γύρω από αυτό. Μία προσεκτική προσπάθεια-κανόνας δείχνει ότι το μέγεθος του δείγματος θα πρέπει πάντα να είναι περισσότερο από 10 φορές του αριθμού των ελεύθερων παραμέτρων του μοντέλου(Bentler, 1995, Hu, Bentler, &Kano 1992). Σε αντίθετη περίπτωση τα αποτελέσματα από τη μέθοδο σταθμισμένων ελαχίστων τετραγώνων δεν πρέπει να είναι αξιόπιστα. Εάν από την άλλη, το μέγεθος του δείγματος είναι μικρότερο, οι ερευνητές ενθαρρύνονται να χρησιμοποιήσουν την ισχυρή μέθοδο SatorraBentler η οποία είναι διαθέσιμη στο πρόγραμμα EQS.

Μία άλλη εναλλακτική λύση για την αντιμετώπιση μη-κανονικών δεδομένων είναι να κάνει τα δεδομένα εισάγοντας κάποια ομαλοποίηση μετασχηματισμού στα ανεπεξέργαστα δεδομένα. Όταν τα δεδομένα έχουν μετασχηματιστεί, η κανονική ανάλυση θεωρίας μπορεί να πραγματοποιηθεί. Σε γενικές γραμμές, οι μετασχηματισμοί είναι απλά μία επανέκφραση των δεδομένων σε διαφορετικές μονάδες μεγέθους. Οι πιο δημοφιλής μετασχηματισμοί που περιλαμβάνονται στα περισσότερα στατιστικά πακέτα γενικής χρήσης είναι μετασχηματισμοί ενέργειας, μετασχηματισμοί τετραγωνικής ρίζας, αμοιβαίοι εμτασχηματισμοί και λογαριθμικοί μετασχηματισμοί. Επίσης, κάποιος μπορεί να θέλει να εξετάσει τα στοιχεία σχετικά με άλλα μέτρα δομών που εμπλέκονται στο προτεινόμενο μοντέλο, αν τέτοια είναι άμεσα διαθέσιμα.

Με τα δεδομένα που προκύπτουν από σχέδια με μόνο μερικές πιθανές κατηγορίες απαντήσεων, η μέθοδος σταθμισμένων ελαχίστων τετραγώνων χρησιμοποιείται συνήθως με πολυχωρικούς ή πολυσειριακούς συσχετισμούς. Για παράδειγμα, έστω ότι ένα ερωτηματολόγιο περιλαμβάνει το στοιχείο, ΄΄Πόσο ικανοποιημένοι είστε με την πρόσφατη αγορά του αυτοκινήτου σας;΄΄ με κατηγορίες απαντήσεων να είναι:

<<πολύ ικανοποιημένοι>>, <<κάπως ικανοποιημένοι>>, <<δεν είναι ικανοποιημένοι>>. Ένα σημαντικό μέρος της έρευνας έχει δείξει ότι αγνοεί τα κατηγορικά χαρακτηριστικά των δεδομένων που λαμβάνονται από τα στοιχεία όπως αυτά μπορεί να οδηγήσουν σε μεροληπτικά αποτελέσματα SEM. Για το λόγο αυτό, οι ερευνητές προτείνουν τη χρήση του συντελεστή πολυχρονικής-συσχέτισης(για την αξιολόγηση του βαθμού συσχέτισης μεταξύ ordinalμεταβλητών) και τον συντελεστή πολυσειριακής συσχέτισης(για την αξιολόγηση του βαθμού συσχέτισης μεταξύ μιας ordinal και συνεχής μεταβλητής). Έρευνες έχουν δείξει ότι όταν υπάρχουν πέντε ή περισσότερες κατηγορίες απαντήσεων, τότε τα προβλήματα αγνοώντας την κατηγορηματική φύση των απαντήσεων είναι πιθανό να ελαχιστοποιείται(Rigdon, 1998).

Από στατιστικής πλευράς, οι 4 μεθοδοι εκτίμησης παραμέτρων οδηγούν σε συνεπείς εκτιμήσεις, κάτι το οποίο είναι ένα επιθυμητό αποτέλεσμα που εξασφαλίζει ότι με την αύξηση του μεγέθους του δείγματος οι εκτιμήσεις συγκλινουν προς τις πραγματικές τιμές των παραμέτρων του πληθυσμού.

## *6.4: Στατιστικοί έλεγχοι*

Τα χαρακτηριστικά των στατιστικών δοκιμών είναι ιδιαίτερα σημαντικά για τα δομικά μοντέλα εξισώσεων και θα τονιστούν παρακάτω.

### *6.4.1: Τυπικά σφάλματα*

Ίσως και η πιο βασική μορφή μιας στατιστικής δοκιμής είναι ο κρίσιμος λόγος, ο οποίος είναι ο λόγος του ενός στατιστικού δείγματος επί το τυπικό σφάλμα. Το τυπικό σφάλμα είναι η τυπική απόκλιση μιας κατανομής δειγματοληψίας, η οποία ουσιαστικά είναι μία κατανομή πιθανότητας μιας στατιστικής με βάση όλα τα πιθανά τυχαία δείγματα.

Δεδομένης της σταθερής μεταβλητότητας μεταξύ των περιπτώσεων πληθυσμού, το τυπικό σφάλμα μεταβάλλεται αντιστρόφως ανάλογα με το μέγεθος του δείγματος. Αυτό σημαίνει ότι οι κατανομές των στατιστικών από μεγαλύτερα δείγματα είναι γενικά πιο περιορισμένες (δηλαδή έχουν λιγότερες μεταβλητές) σε σχέση με κατανομές του ίδιου στατιστικού στοιχείου από μικρότερα δείγματα.

Υπάρχουν τυπικά σφάλματα της στατιστικής με απλές κατανομές, με τον όρο ‘’απλό’’ εννοούμε ότι:

1. Η στατιστική υπολογίζει μία μόνο παράμετρο
2. Από το σχήμα της κατανομής δεν είναι συνάρτηση της εν λόγω παραμέτρου

Για παράδειγμα, ο τύπος για την εκτίμηση του τυπικού σφάλματος του μέσου είναι:

$$SE_M = \frac{SD}{\sqrt{N}}$$

Όπου, $SD$: Τυπικές αποκλίσεις και $N$: Μέγεθος δείγματος.

Είναι αρκετά πιο δύσκολο να εκτιμηθούν τυπικά σφάλματα για στατιστικά στοιχεία που δεν έχουν απλές κατανομές. Υπάρχουν κατά προσέγγιση μέθοδοι υπολογισμού χειροκίνητα για μερικά στατιστικά στοιχεία, όπως το δείγμα αναλογιών, όπου το σχήμα της κατανομής και της μεταβλητότητας εξαρτάται από την αξία της αναλογίας του πληθυσμού. Τέτοιες μέθοδοι παράγουν αυμπτωτικά τυπικά σφάλματα που αναλαμβάνουν ένα μεγάλο δείγμα. Ωστόσο εάν το δείγμα δεν είναι μεγάλο εκτιμάται ότι τα τυπικά σφάλματα μπορεί να μην είναι ακριβές. Όμως και κάποια άλλα στατιστικά στοιχεία όπως η πολλαπλή συσχέτιση R, έχουν κατανομές τόσο περίπλοκες που μπορεί να υπάρχει κατά προσέγγιση κανονικό τύπος σφάλματος για τον χειροκίνητο υπολογισμό. Στα SEM, τα τυπικά σφάλματα για τις επιπτώσεις των παρατηρούμενων μεταβλητών υπολογίζονται από τον υπολογιστή και οι εκτιμήσεις αυτές είναι απλές. Αυτό σημαίνει ότι οι τιμές θα μπορούσαν να αλλάξουν αν ας πούμε χρησιμοποιούνταν μία διφορετική μέθοδος εκτίμησης.

### *6.4.2: Στατιστικοί έλεγχοι στα SEM*

Εδώ έχουμε ένα κρίσιμο σημείο για τις στατιστικές δοκιμές SEM: Κατά την εκτίμηση πολλαπλής παλινδρόμησης, τα τυπικά σφάλματα γενικά υπολογίζονται για την μη τυποποιημένη μόνη λύση. Αυτό διαπιστώνεται όταν μπορούμε να δούμε μέσα από την έξοδο αποτελεσμάτων ενός εργαλείου υπολογιστή SEM και βρίσκουμε τυπικά σφάλματα τα οποία τυπώνονται για μη-τυποποιημένες εκτιμήσεις. Αυτό σημαίνει ότι τα αποτελέσματα των στατιστικών ελέγχων είναι διαθέσιμα μόνο για μη-τυποποιημένες εκτιμήσεις. Οι ερευνητές συνήθως υποθέτουν ότι τα αποτελέσματα των στατιστικών ελέγχων της μη-τυποποιημένες εκτίμησης ισχύουν και για τις αντίστοιχες τυποποιημένες εκτιμήσεις. Για δείγματα μεγάλα και αντιπροσωπευτικά αυτή η υπόθεση μπορεί να μην είναι προβληματική.

<u>*Παράδειγμα*</u>
Έστω, ότι κατά την εκτίμηση μέγιστης πιθανοφάνειας οι τιμές της μη-τυποποιημένης εκτίμησης, τυπικού σφάλματος και τυποποιημένης εκτίμησης είναι αντίστοιχα:4,20 , 2,00 , 0,60. Σε ένα μεγάλο δείγμα, η μη τυποποιημένη εκτίμηση θα είναι στατιστικά θα είναι στατιστικά σημαντική στο επίπεδο 0,05 επειδή z=4,20/2,00 η οποία υπερβαίνει την κρίσιμη τιμή (1,96) στο p<0.5. Το αν η τυποποημένη εκτίμήση 0,60 είναι στατιστικά σημαντική σε p<0,5 δεν το γνωρίζουμε καθώς δεν έχει τυπικό σφάλμα. Έτσι δεν θα ήταν σκόπιμο να αναφέρεται η τυποποιημένη εκτίμηση από μόνη της ως 0,60*, όπου ο (*):υποδεικνύει ότι p<0.5.

Έχουμε, τις μη-τυποποιημένες και τυποποιημένες εκτιμήσεις και επίσης το προηγούμενο τυπικό σφάλμα, δηλαδή

4,20* (2,10) 0,60

όπου, το τυπικό σφάλμα δίνεται σε παρενθέσεις και ο στερίσκος συνδέεται με την μη-τυποποιημένη εκτίμηση (4,20).

## *6.5 : Έλεγχοι Υποθέσεων*

Ο έλεγχος υποθέσεων στα δομικά μοντέλα εξισώσεων περιλαμβάνει την επιβεβαίωση ότι τα δεδομένα του δείγματος ταιριάζουν σε ένα θεωρητικό μοντέλο.Αυτό ελέγχεται με την χρήση του $x^2$ και των βαθμών ελευθερίας με βάρη την μη-κεντρική κατανομή αποδίδει την παράμτερο μη-κεντρικότητας (NCP) και τη ρίζα, το μέσο και το τετραγωνικό σφάλμα των τιμών προσέγγισης (RMSEA).Ορίζουμε λοιπόν μία μηδενική υπόθεση ($H_0$) και την εναλλακτική υπόθεση ($H_1$) ότι είναι το αντίθετο από ότι χρησιμοποιείται σε μία παραμετρική στατιστική, η οποία βασίζεται στην κεντρική κατανομή. Αυτό που θέλουμε είναι να διατηρήσουμε στην μηδενική υπόθεση τα ακριβή στοιχεία για την μοντελοποίηση και ναπορρίψουμε την εναλλακτική υπόθεση, η οποία είναι το αντίθετο από που θα προσδιορίζει τις παραδοσιακές στατιστικές δοκιμές μας. Ένα $x^2$ τεστ συχνά αναφέρεται ως μία κακή προσαρμογή του δείκτη, διότι επιδιώκουμε μία μη-σημαντική $x^2$ τιμή για να μας δείξει την ομοιότητα του πίνακα της διακύμανσης του δείγματος, της συνδιακύμανσης και του μοντέλου. Εάν η τιμή του $x^2$ τεστ είναι στατιστικά σημαντική αυτό σημαίνει ότι τα δεδομένα δεν έχουν καλή προσαρμογή στο μοντέλο.

Ο δείκτηςRMSEA( η ρίζα μέσων τετραγώνων του σφάλματος εκτίμησης) μπορεί να χρησιμοποιηθεί για να καθορίσει την μηδενική αλλά και την εναλλακτική υπόθεση. Μερικοί συγγραφείς έχουν προτείνει μία σειρά από $\varepsilon$ τιμές να εκπροσωπεί την μηδενική υπόθεση του αποδεκτού μοντέλου προσαρμογής ($\varepsilon \leq 0{,}5$) και μη αποδεκτό μοντέλο προσαρμογής ($\varepsilon \geq 0{,}5$). Άλλες έρευνες έχουν ερμηνεύσει στενή εφαρμογή του RMSEA μεταξύ 0,05 – 0,08 με δεδομένη την πολυπλοκότητα του θεωρητικού μοντέλου. Όταν *ε=0* θα έχουμε ακριβή δεδομένα για το μοντέλο εφαρμογής. Έτσι, ένας ερευνητής θα μπορούσε να καθορίσει τη μηδενική και την εναλλακτική υπόθεση ως εξής:

$$H_0: \varepsilon \leq 0{,}5$$
$$H_1: \varepsilon > 0{,}5$$

Όπου $\varepsilon$ υπολογίζεται με την χρήση $F_{ML}$ και$df$ τιμών από την ανάλυση του μοντέλου:

$$\varepsilon = \sqrt{\frac{F_{ML}}{df}}$$

Η προσέγιση των SEM είναι να υποθέσει ένα θεωρητικό μοντέλο, τη συλλογή δεδομένων του δείγματος και να ελέγξει αν τα στοιχεία ταιριάζουν με το μοντέλο. Σε αυτό το σημείο μπορούμε να πούμε ότι έχουν συζητηθεί αρκετοί δείκτες για το αν είναι ικανοί ώστε τα δεδομένα να προσδιορίζουν το θεωρητικό μοντέλο. Όταν τα δεδομένα δεν ταιριάζουν με το θεωρητικό μοντέλο, εξετάζεται το ενδεχόμενο τροποποίησης των δεικτών ώστε να τροποποιηθεί το μοντέλο για μία βελτιωμένη

εφαρμογή. Έχουμε τελικά την δομική ενός μηδενικού μοντέλου ενάντια σε ένα εναλλακτικό μοντέλο και ο έλεγχος υποθέσεων περιλαμβάνει την κατανόηση ‘’εξουσίας’’, με το όρο ‘’εξουσία’’ στα παραδοσιακά παραμετρικά τεστ εννοούμε την ικανότητα να ανιχνεύται η διαφορά μεταξύ των δύο υποθέσεων και έτσι αν απορρίπτουμε την μηδενική και να δεχόμαστε την εναλλακτική υπόθεση.

Στα δομικά μοντέλα εξισώσεων πρέπει να εντάξουμε το σκεπτικό της ‘’εξουσίας’’ και τη δυνατότητα να διατηρήσουμε την μηδενική υπόθεση και να απορρίψουμε την εναλλακτική υπόθεση. Ισχύς και συναφή θέματα (όπως μέγεθος δείγματος, βαθμοί ελευθερίας, κατευθυντική φύση της υπόθεσης ) επηρεάζουν την απόφαση σχετικά με το αν τα δεδομένα του δείγματος ταιριάζουν με το θεωρητικό μοντέλο.

[Γιαέλεγχουποθέσεωνείναιαπότοηλεκτρονικόβιβλίο, ‘’A beginner’s Guide to Structural Equation Modeling’’ από Randall E. Schumacker & Richard G. Lomax]

## *6.6: Διαστήματα εμπιστοσύνης*

Το διάστημα εκτίμησης είναι μία εναλλακτική λύση για την σημασία του ελέγχου. Περιλαμβάνει αποτελέσματα μεγεθών με διάστημα εμπιστοσύνης που δείχνει μία σειρά από αποτελέσματα τα οποία θεωρούνται ισοδύναμα εντός των ορίων σφάλαμτος δείγματοληψίας για το συγκεκριμένο αποτελέσμα. Από στατιστικής πλευράς με απλές κατανομές το πλάτος κάθε πλευράς είναι:

$$100 \times (1 - a)\%$$

Διάστημα εμπιστοσύνης προσδιορίζεται από το προιόν του τυπικού σφάλματος και την κρίσιμη τιμή ενός κεντρικού στατιστικού αποτελέσματος της δοκιμής σε ένα επιπέδο στατιστικής σημασίας για δύο εναλλακτικές υποθέσεις. Για παράδειγμα δεδομένου ότι:

$$M = 100.00, SD = 9.00, N = 25, and SE_m = 1.80$$

Το 95% διάστημα εμπιστοσύνης θα είναι:

$$100.00 \pm (1.80) t_{2-tail, a=0.05}$$

Όπου, $t_{2-tail,a=0.05}$ είναι δίπλευρη κρίσιμη τιμή για μία κεντρική κατανομή t σε επίπεδο σημαντικότητα 0,5, η οποία για $df = 24$ είναι 2.064.
Το 95% διάστημα εμπιστοσύνης είναι:

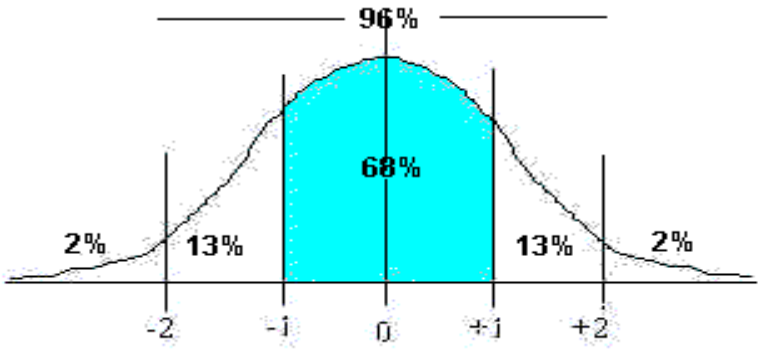
$$100.00 \pm 1.80(2.06), or\ 100.00 \pm 3.72$$

το οποίο ορίζει το διάστημα [96.28 , 103.72].

Το διάστημα αυτό καθορίζει ένα εύρος τιμών το οποίο θεωρείται ισοδύμανο με την παρατηρούμενη μέση τιμή εντός των ορίων σφάλματος δειγματοληψίας σε επίπεδο εμπιστοσύνης 95%. Η εκτίμηση σε ένα σημείο π.χ. 100.00 πέφτει στο ακριβές κέντρο του διαστήματος και το όλο διάστημα μεταφέρει ρητά την ιδέα ότι ένα περιθώριο σφάλματος σχετίζεται με την αντίστοιχη στατιστική. Το παραπάνω διάστημα [96.28 , 103.72] βασίζεται σε μία ενιαία εκτίμηση του $SE_m$=180. Αλλά η ποσότητα αυτή είναι ίδια ακριβώς με την εκτίμηση του σημείου και η τιμή του $SE_m$ σε ένα διαφορετικό δείγμα σχεδόν σίγουρα δεν θα είναι 1,80. Αυτό σημαίνει ότι το διάστημα [96.28 , 103.72] είναι πράγματι πολύ περιορισμένο αν λάβουμε υπόψη και το σφάλμα δειγματοληψίας στο $SE_m$.

Υποθέτουμε ένα διάστημα εμπιστοσύνης 95% με βάση $M = 2.50$ είναι [0 , 500], το οποίο περιλαμβάνει το μηδέν. Το γεγονός αυτό μπορεί να παρερμηνευθεί , οπότε εσφαλμένα καταλήγει στο συμπέρασμα ότι μ=0. Αλλά το μηδέν είναι μόνο μία τιμή εντός του εύρους των εκτιμήσεων γι αυτό δεν έχει ιδιαίτερο καθεστώς. Αυτό σημαίνει ότι η υπόθεση ότι $\mu = 0$ δεν ευνοείται από την υπόθεση ότι $\mu = 5{,}00$. Τα διάστηματα εμπιστοσύνης υπόκεινται σε δειγματοληπτικό σφάλμα, έτσι το μηδέν δεν μπορεί να εμπίπτει στο διάστημα εμπιστοσύνης 95% σε ένα δείγμα αντιγραφής.

Η πιο κοινή χρήση των δομικών μοντέλων (SEM) είναι η δημιουργία διαστημάτων εμπιστοσύνης. Τα SEM είναι μία εκτίμηση του πόσο λάθος υπάρχει σε μία δοκιμή, ενώ επίσης μπορούν να εξεταστούν με τον ίδιο τρόπο όπως και οι τυπικές αποκλίσεις.




*Εικόνα 26 : Διάστημα εμπιστοσύνης για τις βαθμολογίες των μαθητών*

Έστω ότι στο παραπάνω σχήμα απεικονίζονται οι βαθμολογίες μαθητών. Όπως μπρούμε να διακρίνουμε και από το παραπάνω διάγραμμα το 96% των μαθητών έχει βαθμολογίες που κυμαίνονται στην κλίμακα (-2,+2). Θα μπορούσαμε να είμαστε 68% σίγουροι ότι η πραγματική βαθμολογία των μαθητών θα είναι μεταξύ (-1,1). Ή αλλιώς αν ο μαθητής έκανε τη δοκομασία 100 φορές, 64 φορές η πραγματικά βαθμολογία θα κυμαίνεται από (-1,1).

# *Κεφάλαιο 7: Εφαρμογή των Δομικών Μουτέλων Εξίσωσης(SEM)*

## *7.1: Εισαγωγικά*

Στα μέχρι τώρα κεφάλαια δόθηκε η πλήρης παρουσίαση της θεωρητικής έκτασης των δομικών μοντέλων εξίσωσης. Παρουσιάστηκε η πλήρης ανάλυση των δομικών μοντέλων εξίσωσης και αναλύθηκαν οι προυποθέσεις χρήσης. Στην πράξη όμως, κατά την επεξεργασία των στοιχείων μιας έρευνας, μέσω των υπολογιστικών προγραμμάτων που προσφέρονται (όπως AMOS, LISREL, EQS , LISCOMP, RAMOVA, SEPATH) τα πιο δημοφιλή και εύχρηστα εκ των οποίων είναι το AMOS και το LISREL όπως έχει αναφερθεί και σε προηγούμενο κεφάλαιο.

Η σταστική ανάλυση των δεδομένων επιτελέστηκε με την χρήση του προγράμματος LISREL, επιπροσθέτως χρησιμοποιηθήκε ελάχιστα τοστατιστικό πρόγραμμα SPSS για περαιτέρω εφαρμογή.

### *7.1.1: Πρόγραμμα LISREL*

Το πρόγραμμα LISREL θεωρείται το πρώτο υπολογιστικό πρόγραμμα που εμφανίστηκε για την επίλυση των δομικών εξισώσεων και παραμένει μέχρι και σήμερα το πιο δημοφιλές (Byrne, 1998). Η λέξη LISREL προέρχεται από τα αρχικά των λέξεων LinearStructuralRELations και βασίζεται στην προσέγγιση των Joreskog-Keesling και Wiley που παρουσιάζει συστήματα δομικών εξισώσεων (Bentler, 1980). Η γλώσσα εντολών του βασίζεται στη χρήση αλγεβρικών πινάκων για την παρουσίαση της επιβεβαιωτικής παραγοντικής ανάλυσης (ConfirmatoryFactorAnalysis – CFA) ή των πλήρων μοντέλων δομικών εξισώσεων.

Το πρόγραμμα LISREL εισάγει εξωτερικά δεδομένα σε μορφές όπως SPSS, SAS, STATA, STATISTICA, το MicrosoftExcel, SYSTAT κλπ, ως αρχείο PRELIS συστήματος (PSF).

Ένα μοντέλο δομικών εξισώσεων μπορεί να καθοριστεί με την βοήθεια ενός διαγράμματος διαδρομής, απαιτείται SIMPLIS αρχείο έργου, ένα αρχείο έργου LISREL, ένα αρχείο σύνταξης SIMPLIS ή ένα αρχείο σύνταξης LISREL. Το πρόγραμμα για τα windows χρησιμοποιεί ένα αρχείο γραφικών με την PTH προεπιλεγμένη επέκταση για να καταγράψει ένα διάγραμμα διαδρομής. Οι επεκτάσεις SPJ και LPJ χρησιμοποιούνται για SIMPLUS και LISREL αρχεία αντίστοιχα. Τα αρχεία σύνταξης SIMPLIS και LISREL είναι αρχεία κειμένου με προκαθορισμένες επεκτάσεις SPL και LS8 αντίστοιχα. Αυτοί λοιπόν είναι οι πέντε τύποι αρχείων που μπορούν να έχουν πρόσβαση στα δεδομένα από το PSF. Εάν ένας χρήστης έχει ετοιμάσει οποιοδήποτε από αυτά τα αρχεία τότε το LISREL μπορεί να χρησιμοποιηθεί για να ταιριάζει με το καθορισμένο μοντέλο για τα δεδομένα που καθορίζουν το αντίστοιχο PSF.

### *7.1.2: Σκοπός της έρευνας*

Στόχος της εφαρμογής, είναι ο προσδιορισμός των συστατικών που είναι ικανά να περιγράψουν τις ποικιλίες κρασιών αλλά και ο διαχωρισμός αυτών των συστατικών σε μικρότερες ομάδες ώστε να να αναλυθεί καλύτερα η διαμόρφωση των συστατικών σε όλες τις ποικιλίες.

### *7.1.3: Λίγα λόγια για τα δεδομένα*

Τα στοιχεία που χρησιμοποιήθηκαν για την εφαρμογή των δομικών μοντέλων εξίσωσης βασίζονται σε αποτελέσματα χημικής ανάλυσης 178 κρασιών που καλλιεργούνται στην ίδια περιοχή στην Ιταλία και παράγουν 3 διαφορετικές ποικιλίες της ίδια καλλιέργειας. Η κάθε ποικιλία χαρακτηρίζεται από 9 συστατικά τα οποία αντιστοιχούν στις ποσοτικές μεταβλητές και περιγράφονται παρακάτω. Η σειρά παρουσίασης είναι ανάλογη με την σειρά ταξινόμησης που έγινε στο πρόγραμμα (spss).

- Αλκοόλ
- Τέφρα
- Αλκαλικότητα της τέφρας
- Μαγνήσιο
- Φαινόλες
- Φλαβανοειδή
- Ανθοκυανίνες
- Ένταση χρώματος
- Απόχρωση

Η μοναδική κατηγορική μεταβλητή αντιστοιχεί στην ποικιλία των κρασιών και όπως ανέφερα και πιο πάνω αποτελείται από 3 κατηγορίες

- Ποικιλία 1
- Ποικιλία 2
- Ποικιλία 3

Στην Ποικιλία 1 περιέχονται 59 κρασία, στην Ποικιλία 2 περιέχονται 71 κρασιά και στην Ποικιλία 3 περιέχονται 48 κρασιά.

*Ορισμοί (ως αναφορά τα συστατικά του κρασιού)*

***Τέφρα***

Τέφρα καλείται το σύνολο των προιόντων αποτέφρωσης του στερεού υπολείμματος που λαμβάνεται μετά την εξάτμιση του οίνου, όταν αυτή πραγματοποιείται κατά τέτοιο τρόπο ώστε να λαμβάνεται τελικά το σύνολο των κατιόντων υπό μορφή ανθρακικών και άλλων ανύδρων μεταλλικών αλάτων.

Η ποσότητα της τέφρας, δηλαδή των ανόργανων υλών του κρασιού, ευρίσκεται συνήθως σε σταθερή αναλογία με το άνευ σακχάρου υπόλειμμα κρασιού.

***Αλκαλικότητα της τέφρας***

Αλκαλικότητα της τέφρας ή ολική αλκαλικότητα της τέφρας, καλείται το σύνολο των κατιόντων, εκτός από το αμμώνιο, τα οποία είναι ενωμένα με τα οργανικά οξέα του κρασιού. Κατά την αποτέφρωση τα άλατα των οργανικών οξέων του κρασιού από τα οποία υπερτερεί το όξινο τρυγικό κάλιο, μετατρέπονται σε ανθρακικά άλατα, τα οποία έχουν αλκαλική αντίδραση. Η αλκαλικότητα της τέφρας εκφράζεται συνήθως $K_2CO_3$, από το κυριότερο συστατικό της. Ο προσδιορισμός της αλκαλικότητας της τέφρας παρουσιάζει ενδιαφέρον, γιατί βοηθάει στην απόδειξη κατεργασιών του κρασιού ή και αλλοιώσεων του.

(http://ecourse.uoi.gr/course/view.php?id=867 )

***Φαινολικά οξέα – Ανθοκυάνες – Ταννίνες***

**Τα φαινολικά οξέα** δηλαδή, **οι φλαβανοειδείς** και **μη φλαβανοειδείς φαινόλες**, είναι υπεύθυνες για το χρώμα στα κράσια, την λιπαρότητα της γεύσης και τα οργανοληπτικά χαρακτηριστικά τους. Στα φλαβανοειδή οφείλεται το κιτρινωπό χρώμα των λευκών κρασιών και στις ανθοκυάνες ή ανθοκυανίνες το ερθυροιώδες έως ερυθρό χρώμα των κόκκινων.

**Οι ανθοκυάνες** έχουν την τάση να δίνουν αδιάλυτα συσσωματώματα με το χρόνο και να σχηματίζουν ίζημα στη φιάλη, προκαλώντας μείωση της έντασης του χρώματος του οίνου.

**Οι τανίνες**, περιέχονται στα λευκά και κυρίως στα κόκκινα κρασιά, η ύπαρξη της τανίνης είναι απαραίτητη στο κόκκινο κρασί, τόσο για να εξισορροπεί τη δομή του, σε σχεσή με όλα τα χαρακτηριστικά του, όσο και για να αυξάνει τη δυναμική παλαίωσης, έτσι ώστε το κρασί να διατηρήσει τη ζωντάνια και το χαρακτήρα του σε βάθος χρόνου. Είναι επιστημονικά αποδεδειγμένο ότι οι ταννίνες προσαρτώνται στις λιπαρές πρωτείνες, και αυτό το μυστικό του επιτυχημένου συνδυασμού του κόκκινου κρέατος με το κόκκινο κρασί.

(http://proionta-tis-fisis.com/ta-flavonoidi-kai-pou-ofeloun/)

***Ένταση χρώματος – Απόχρωση***

Το χρώμα των οινών είναι αποτέλεσμα εκλεκτικής απορρόφησης ορισμένων ακτινοβολιών του ηλιακού φάσματος και οφείλεται στις φαινολικές ενώσεις. Σύμφωνα με τις επίσημες μεθόδους του OIV, χρωματικά χαρακτηριστικά ενός οίνου ονομάζονται η φωτεινότητα του και η χρωματικότητά του. Η φωτεινότητα αντιστοιχεί στη διαπερατότητα. Είναι αντιστρόφως ανάλογη προς την ένταση χρώματος του οίνου. Η χρωματικότητα ανταποκρίνεται στο επικρατούν μήκος κύματος (που χαρακτηρίζει την απόχρωση) και την καθαριότητα. Η μέθοδος

αναφοράς είναι μία φασματοφωτομετρική μέθοδος που επιτρέπει τον υπολογισμό των τρισερεθιστικών τιμών και των τριχωματικών συντελεστών που απαιτούνται για τον καθορισμό του χρώματος σύμφωνα με τους κανόνες της διεθνούς επιτροπής φωτισμού (CIE).

https://www.google.gr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwjHsbrjh5HOAhUjK8AKHeumD4EQFggaMAA&url=http%3A%2F%2Ftdd.aua.gr%2Fsites%2Fdefault%2Ffiles%2FIIX.%2520%25CE%25A0%25CE%25A1%25CE%259F%25CE%25A3%25CE%2594%25CE%2599%25CE%259F%25CE%25A1%25CE%2599%25CE%25A3%25CE%259C%25CE%259F%25CE%25A3%2520%25CE%25A7%25CE%25A1%25CE%25A9%25CE%259C%25CE%2591%25CE%25A4%25CE%2599%25CE%259A%25CE%25A9%25CE%259D%2520%25CE%25A7%25CE%2591%25CE%25A1%25CE%2591%25CE%259A%25CE%25A4%25CE%2597%25CE%25A1%25CE%2599%25CE%25A3%25CE%25A4%25CE%2599%25CE%259A%25CE%25A9%25CE%259D.doc&usg=AFQjCNEi1rrYsEaxkw6iQG3aBEwHG-M02g&sig2=3X-rfG7Gub2V1EH0fBOjrA&bvm=bv.127984354,d.ZGg&cad=rja

***Μαγνήσιο***

Το μαγνήσιο απορροφάται δια μέσου της ρίζας του φυτού και επηρεάζεται από την υψηλή περιεκτικότητα σε κάλιο ή ασβέστιο στο έδαφος.

*Το μαγνήσιο*:

- Είναι το κεντρικό άτομο της χλωροφύλλης
- Είναι υπεύθυνο για την μεταφορά και την αφομοίωση
- Βελτιώνει την δημιουργία σακχάρων και αρωμάτων
- Μειώνει την εμφάνιση ξηρής ράχης και τον μαρασμί των σταφυλιών

*Έλλειψη μαγνησίου:*

- Κιτρίνισμα των φύλλων
- Πρόωρη πτώση των φύλλων
- Χαμηλή περιεκτικότητα σε σάκχαρα
- Υψηλή εξάτμιση
- Χειρότερη εκμετάλλευση του φωσφόρου
- Κίνδυνος ξηρής ράχης

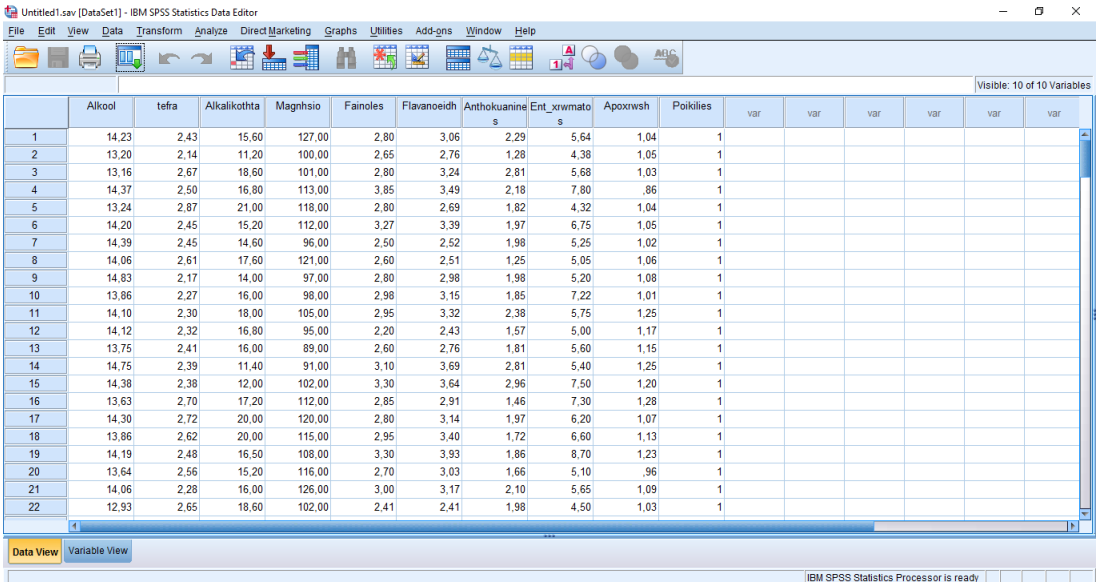
Επίσης το μαγνήσιο λειτουργεί κατά την ξήρανσης.

***Αλκοόλ***

Το κρασί έχει 8-14% περιεκτικότητα σε αλκοόλ, κατά μέσο όρο 12% για τα λευκά κρασιά και 14% για τα κόκκινα.

## *7.2: Παρουσίαση δεδομένων*

Τα δεδομένα μου αρχικά τα πέρασα στο πρόγραμμα SPPS για να μπορέσω στην συνέχεια να τα περάσω στο πρόγραμμα LISREL (καθώς όπως ανέφερα και παραπάνω πρέπει να είναι σε μορφή SPPSή και EXCEL) για να εισαχθούν στο πρόγραμμα.




*Εικόνα 27: Παρουσίαση δεδομένων σε SPSS*
*(όπου 1=Ποικιλία 1, 2=Ποικιλία 2, 3=Ποικιλία 3)*

Τα δεδομένα βρίσκονται στον ιστότοπο:http://archive.ics.uci.edu/ml/datasets/Wine

### *7.2.1: Παρουσίαση περγραφικών μέτρων*

Ένα από τα πρώτα πράγματα που πραγματοποιείται σε κάθε στατιστική ανάλυση είναι η περιγραφή των στατιστικών μέτρων των μεταβλητών για να αποκτήσουμε μια γενική εικόνα γύρω από τα δεδομένα μας. Στο παρακάτω πίνακα περιλαμβάνονται κάποια βασικά μέτρα θέσης και διασποράς:

| Ποικιλίες<br>Μέγεθος δείγματος 178 | | Ποικιλία 1<br>59 | Ποικιλία 2<br>71 | Ποικιλία 3<br>48 |
|---|---|---|---|---|
| Αλκοόλ | *Μέση τιμή* | 13.7368 | 12.2825 | 13.1538 |
| | *Διακύμανση* | 0.216 | 0.296 | 0.281 |
| Τέφρα | *Μέση τιμή* | 2.4556 | 2.2477 | 2.4371 |
| | *Διακύμανση* | 0.052 | 0.099 | 0.034 |
| Αλκαλικότητα τέφρας | *Μέση τιμή* | 17.0373 | 20.2380 | 21.1458 |
| | *Διακύμανση* | 6.484 | 11.221 | 5.276 |
| Μαγνήσιο | *Μέση τιμή* | 106.339 | 94.5493 | 99.3125 |
| | *Διακύμανση* | 110.228 | 280.680 | 118.602 |
| Φαινόλες | *Μέση τιμή* | 2.8402 | 2.2504 | 1.6544 |
| | *Διακύμανση* | 0.115 | 0.287 | 0.142 |
| Φλαβανοειδή | *Μέση τιμή* | 2.9824 | 2.0808 | 0.7806 |
| | *Διακύμανση* | 0.158 | 0.498 | 0.085 |
| Ανθοκυανίνες | *Μέση τιμή* | 1.8937 | 1.6303 | 1.1535 |

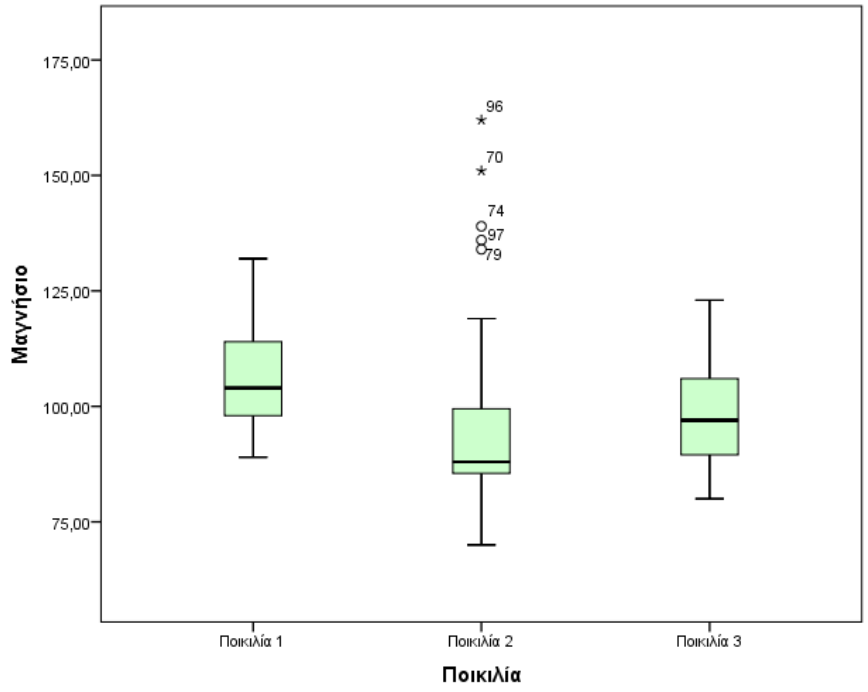
| | | | | |
|---|---|---|---|---|
| | *Διακύμανση* | 0.174 | 0.362 | 0.167 |
| Ένταση χρώματος | *Μέση τιμή* | 5.5283 | 3.0866 | 7.4065 |
| | *Διακύμανση* | 1.534 | 0.855 | 5.272 |
| Απόχρωση | *Μέση τιμή* | 1.0620 | 1.0545 | 0.6827 |
| | *Διακύμανση* | 0.014 | 0.041 | 0.013 |

*Πίνακας 12: Παρουσίαση περιγραφικών μέτρων*

Όπως, μπορούμε να διακρίνουμε στον παραπάνω πίνακα φαίνονται τόσο οι μέσες τιμές αλλά και οι διακυμάνσεις των ανεξάρτητων μεταβλητών για κάθε κατηγορία της εξαρτημένης μεταβλητής. Μεγαλύτερες περιεκτικότητες στις μεταβλητές του αλκοόλ, της τέφρας, του μαγνησίου, των φαινολών, των φλαβανοειδών, των ανθοκυανίνων και της απόχρωσης φαίνεται να παρουσιάζουν τα κρασιά της ποικιλίας 1, ενώ μεγαλύτερες περιεκτικότητες στις μεταβλητές αλκαλικότητας της τέφρας και ένταση χρώματος φαίνεται να παρουσιάζουν τα κρασιά της ποικιλίας 3.

Τώρα, ως αναφορά τις διακυμάνσεις παρατηρούμε διακρίνουμε μικρές τιμές για όλες τις μεταβλητές πέραν του μαγνησίου που είναι αρκετά αυξημένες, κάτι που δηλώνει πως ίσως σε αυτή την κατηγορία να υπάρχουν κρασιά που περιέχουν μεγαλύτερες ποσότητες μαγνησίου σε σχέση με τα υπόλοιπα κρασιά.

Για μια καλύτερη κατανόηση, θα ακολουθήσει η γραφική απεικόνιση των μεταβλητών. Αρχικά, θα γίνει γραφική απεικόνιση του μαγνησίου όπου παρατηρήθηκαν αρκετά μεγάλες τιμές διακυμάνσεων.

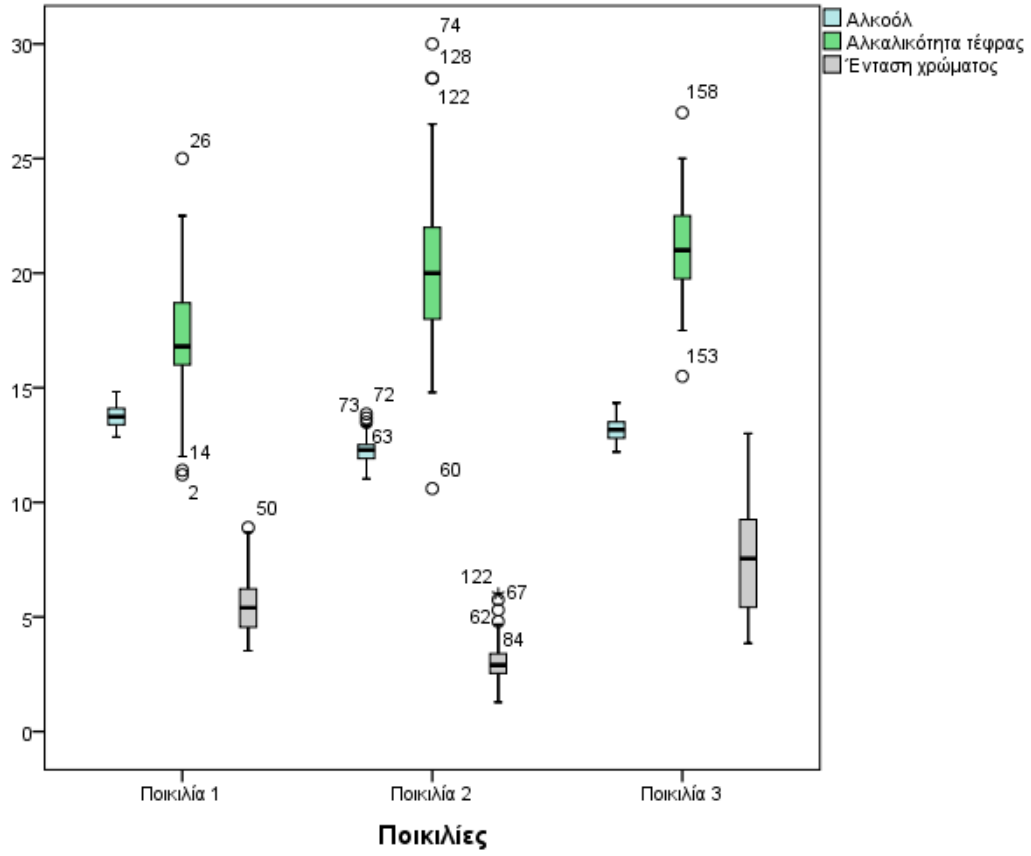



*Εικόνα 28: Θηκόγραμμα μεταβλητής μαγνήσιο*

Όπως παρατηρήσαμε και από τον παραπάνω πίνακα, η τιμή της διακύμανσης του μαγνησίου στην ποικιλία 2 είναι αρκετά μεγάλη, κάτι που φαίνεται και από το παραπάνω διάγραμμα καθώς υπαρχουν αρκετά απομακρυσμένες τιμές από το μέσο.

Ακόμη, μπορούμε να παρατηρήσουμε ότι η ποικιλία 1 περιέχει κρασιά με μεγαλύτερα ποσά περιεκτικότητας μαγνησίου από τις άλλες δύο ποικιλίες.

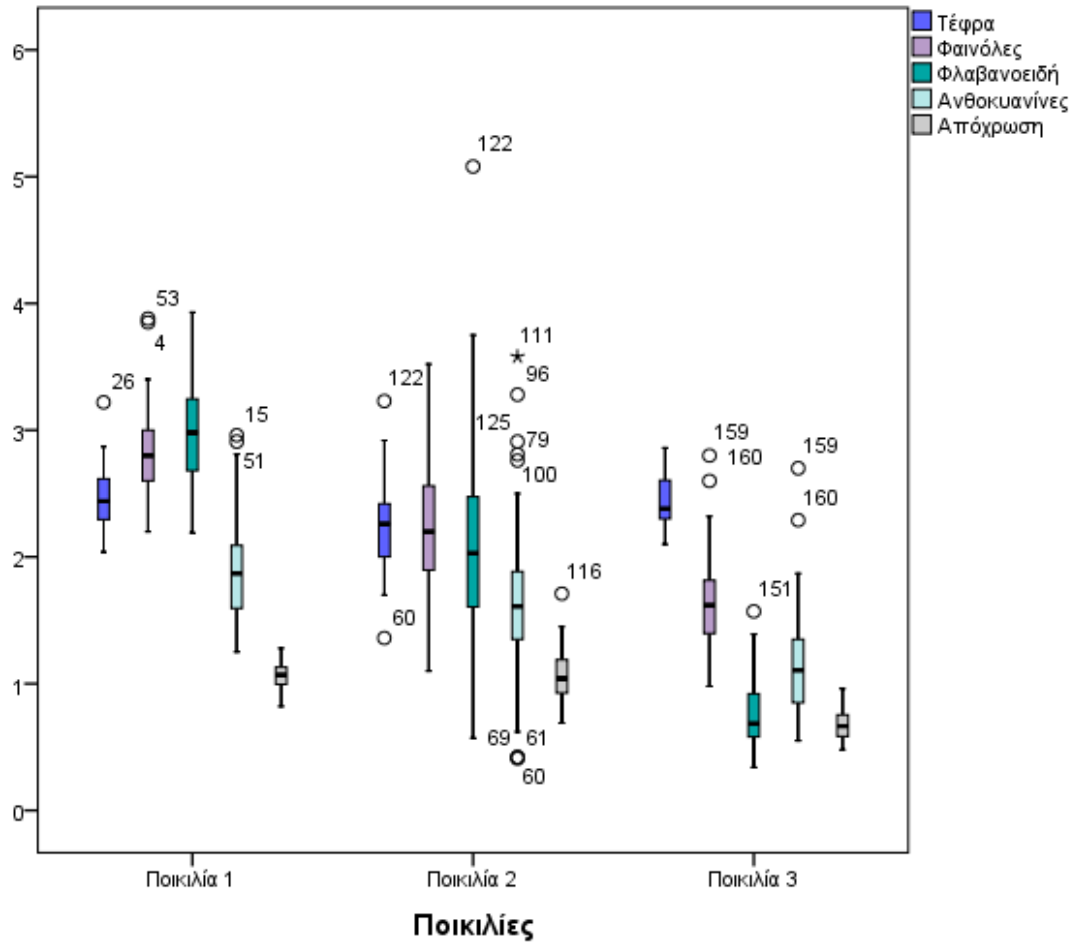
Στη συνέχεια θα γίνει γραφική απεικόνιση σύνθετων θηκογραμμάτων. Επιλέξαμε, για την παρουσίαση των αποτελεσμάτων των περιγραφικών να χωρίσουμε το θηκόγραμμα σε ομάδες συστατικών για να είναι πιο ευδιάκριτες οι διαφορές. Ο διαχωρισμός των μεταβλητών δεν έγινε με βάση κάποιο συγκεκριμένο κριτήριο αλλά με βάση τα ύψη των τιμών, απλά και μόνο για λόγους ευκολίας του αναγνώστη. Έτσι γίνεται γραφική απεικόνιση των μεταβλητών αλκοόλ-αλκαλικότητα τέφρας-ένταση χρώματος και γραφική απεικόνιση των μεταβλητών τέφρας–φαινόλες-φλαβανοειδή-Ανθοκυανίνες-Απόχρωση.




*Εικόνα 29: Θηκόγραμμα μεταβλητών Αλκόολ-Αλκαλικότητα τέφρας-Ένταση χρώματος*

Από το παραπάνω θηκόγραμμα παρατηρούμε πως οι τιμές του αλκοόλ είναι αρκετά κοντά και για τις τρεις ποικιλίες. Η μεταβλητή της αλκαλικότητας της τέφρας έχει μεγαλύτερες τιμές για την ποικιλία 3, μικρότερες για την ποικιλία 2 και ακόμη πιο μικρές τιμές για την ποικιλία 1. Επίσης, παρατηρείται ότι στην πρώτη και στην δεύτερη ποικιλία υπάρχουν κάποιες απομακρυσμένες τιμές αλλά όχι ιδιαίτερα απομακρυσμένες.

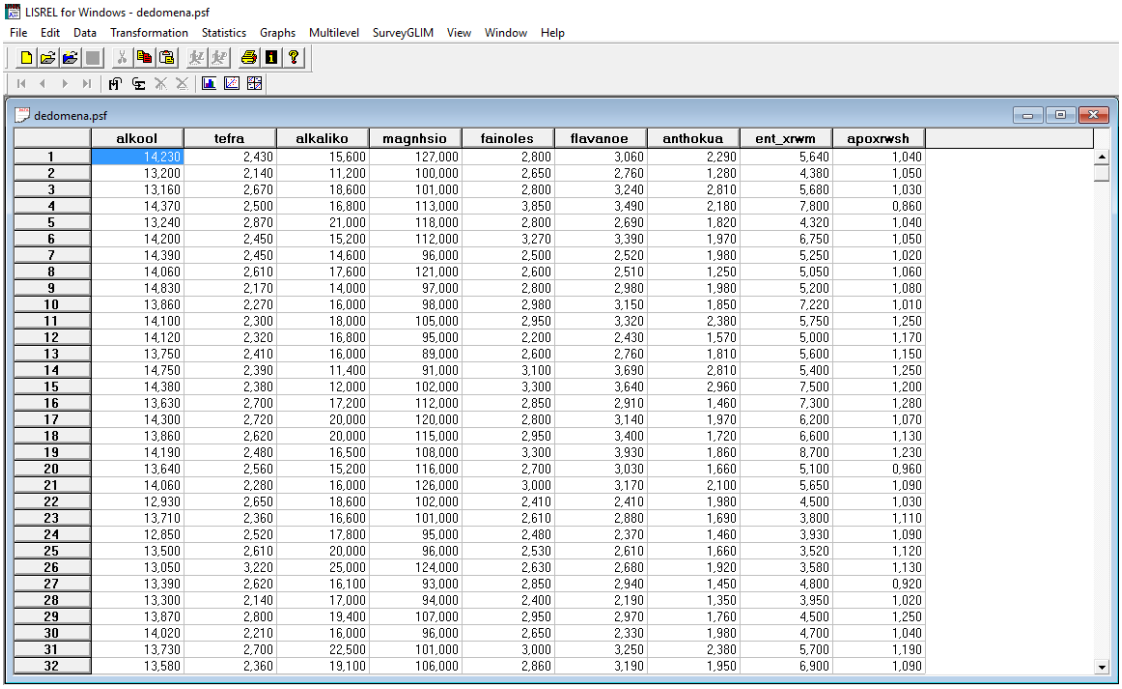
Και τέλος, ως αναφορά την ένταση του χρώματος, μεγαλύτερες τιμές παρατηρούνται στην τρίτη ποικιλία με μαγλύτερη διακύμανση τιμών και μεγαλύτερη ανομοιογένεια τιμών από τις άλλες δύο ποικιλίες μιας και το έυρος της θήκης του γραφήματος είναι μεγαλύτερο. Μερικές ακραίες τιμές, όχι όμως ιδιαίτερα απομακρυσμένες εμφανίζονται στην κατηγορία κρασιών της δεύτερης ποικιλίας με μεγαλύτερη όμως ομοιογένεια τιμών και χαμηλότερες τιμές συγκριτικά με τις άλλες δύο ποικιλίες.




*Εικόνα 30: Θηκόγραμμα μεταβλητών Τέφρας-Φαινόλες-Ανθοκυανίνες-Φλαβανοειδή-Απόχρωση*

Σε αυτό το διάγραμμα σαν γενικότερη εικόνα μπορούμε να πούμε ότι δεν υπάρχουν ιδαίτερες μεγάλες διαοφορές μεταξύ των μεταβλητών στις τρεις ποικιλίες. Βέβαια, πιο συγκεκριμένα μπορούμε να διακρίνουμε ότι υπάρχει μικρότερη ποσότητα φαινολών και ανθοκυανίνων στην τρίτη ποικιλία συγκριτικά με τις άλλες δύο. Ακόμη, για την μεταβλητή φλαβανοειδών στην δεύτερη ποικιλία παρουσιάζεται μεγάλη διασπορά τιμών και ανομοιογένεια, ενώ για την ίδια μεταβλητή η τρίτη ποικιλία περιέχει εξαιρετικά μικρές τιμές. Τέλος, διακρίνονται σε όλες τις μεταβλητές μερικές ακραίες τιμές που όμως δεν επηρεάζουν τα αποτελέσματά μας.

## *7.3: Εφαρμογή στο πρόγραμμα LISREL*


LISREL for Windows - dedomena.psf

File Edit Data Transformation Statistics Graphs Multilevel SurveyGLIM View Window Help

dedomena.psf

| | alkool | tefra | alkaliko | magnhsio | fainoles | flavanoe | anthokua | ent_xrwm | apoxrwsh |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 14,230 | 2,430 | 15,600 | 127,000 | 2,800 | 3,060 | 2,290 | 5,640 | 1,040 |
| 2 | 13,200 | 2,140 | 11,200 | 100,000 | 2,650 | 2,760 | 1,280 | 4,380 | 1,050 |
| 3 | 13,160 | 2,670 | 18,600 | 101,000 | 2,800 | 3,240 | 2,810 | 5,680 | 1,030 |
| 4 | 14,370 | 2,500 | 16,800 | 113,000 | 3,850 | 3,490 | 2,180 | 7,800 | 0,860 |
| 5 | 13,240 | 2,870 | 21,000 | 118,000 | 2,800 | 2,690 | 1,820 | 4,320 | 1,040 |
| 6 | 14,200 | 2,450 | 15,200 | 112,000 | 3,270 | 3,390 | 1,970 | 6,750 | 1,050 |
| 7 | 14,390 | 2,450 | 14,600 | 96,000 | 2,500 | 2,520 | 1,980 | 5,250 | 1,020 |
| 8 | 14,060 | 2,610 | 17,600 | 121,000 | 2,600 | 2,510 | 1,250 | 5,050 | 1,060 |
| 9 | 14,830 | 2,170 | 14,000 | 97,000 | 2,800 | 2,980 | 1,980 | 5,200 | 1,080 |
| 10 | 13,860 | 2,270 | 16,000 | 98,000 | 2,980 | 3,150 | 1,850 | 7,220 | 1,010 |
| 11 | 14,100 | 2,300 | 18,000 | 105,000 | 2,950 | 3,320 | 2,380 | 5,750 | 1,250 |
| 12 | 14,120 | 2,320 | 16,800 | 95,000 | 2,200 | 2,430 | 1,570 | 5,000 | 1,170 |
| 13 | 13,750 | 2,410 | 16,000 | 89,000 | 2,600 | 2,760 | 1,810 | 5,600 | 1,150 |
| 14 | 14,750 | 2,390 | 11,400 | 91,000 | 3,100 | 3,690 | 2,810 | 5,400 | 1,250 |
| 15 | 14,380 | 2,380 | 12,000 | 102,000 | 3,300 | 3,640 | 2,960 | 7,500 | 1,200 |
| 16 | 13,630 | 2,700 | 17,200 | 112,000 | 2,850 | 2,910 | 1,460 | 7,300 | 1,280 |
| 17 | 14,300 | 2,720 | 20,000 | 120,000 | 2,800 | 3,140 | 1,970 | 6,200 | 1,070 |
| 18 | 13,860 | 2,620 | 20,000 | 115,000 | 2,950 | 3,400 | 1,720 | 6,600 | 1,130 |
| 19 | 14,190 | 2,480 | 16,500 | 108,000 | 3,300 | 3,930 | 1,860 | 8,700 | 1,230 |
| 20 | 13,640 | 2,560 | 15,200 | 116,000 | 2,700 | 3,030 | 1,660 | 5,100 | 0,960 |
| 21 | 14,060 | 2,280 | 16,000 | 126,000 | 3,000 | 3,170 | 2,100 | 5,650 | 1,090 |
| 22 | 12,930 | 2,650 | 18,600 | 102,000 | 2,410 | 2,410 | 1,980 | 4,500 | 1,030 |
| 23 | 13,710 | 2,360 | 16,600 | 101,000 | 2,610 | 2,880 | 1,690 | 3,800 | 1,110 |
| 24 | 12,850 | 2,520 | 17,800 | 95,000 | 2,480 | 2,370 | 1,460 | 3,930 | 1,090 |
| 25 | 13,500 | 2,610 | 20,000 | 96,000 | 2,530 | 2,610 | 1,660 | 3,520 | 1,120 |
| 26 | 13,050 | 3,220 | 25,000 | 124,000 | 2,630 | 2,680 | 1,920 | 3,580 | 1,130 |
| 27 | 13,390 | 2,620 | 16,100 | 93,000 | 2,850 | 2,940 | 1,450 | 4,800 | 0,920 |
| 28 | 13,300 | 2,140 | 17,000 | 94,000 | 2,400 | 2,190 | 1,350 | 3,950 | 1,020 |
| 29 | 13,870 | 2,800 | 19,400 | 107,000 | 2,950 | 2,970 | 1,760 | 4,500 | 1,250 |
| 30 | 14,020 | 2,210 | 16,000 | 96,000 | 2,650 | 2,330 | 1,980 | 4,700 | 1,040 |
| 31 | 13,730 | 2,700 | 22,500 | 101,000 | 3,000 | 3,250 | 2,380 | 5,700 | 1,190 |
| 32 | 13,580 | 2,360 | 19,100 | 106,000 | 2,860 | 3,190 | 1,950 | 6,900 | 1,090 |

*Εικόνα 31: Παρουσίση δεδομέων στο LISREL*

Για να γίνει εφαρμογή και ανάλυση των δεδομένων μας στο παραπάνω πρόγραμμα, πέρα από τον ορισμό των παρατηρήσιμων μεταβλητών πρέπει να οριστούν και οι λανθάνουσες μεταβλητές.

Οι λανθάνουσες μεταβλητές τις περισσότερες φορές είναι αποτέλεσμα της ομαδοποίησης μετά από την ανάλυση πολυμεταβλητής ανάλυσης. Κατά την διαδικασία αυτή ορίζονται οι ομάδες και κάθε ομάδα έχει κάθε όνομα. Το όνομα της ομάδας που συνδέει τις εκάστοτε μεταβλητές ονομάζεται στα δομικά μοντέλα λανθάνουσα μεταβλητή.

Στα δικά μας δεδομένα λοιπόν, επειδή δεν είναι εμφανής η ομαδοποίσηση των μεταβλητών, πρέπει να γίνει πρώτα πολυμεταβλητή ανάλυση (μέσω του προγράμματος SPSS) για να γίνει αυτή η διάκριση.

Θα γίνει λοιπόν εφαρμογή της παραγοντικής ανάλυσης (μιας από τις πολυμεταβλητές μεθόδους). Αρχικά όμως πρέπει να τηρούνται οι προυποθέσεις για να γίνει η εφαρμογή της.

## *7.4: Εφαρμογή παραγοντικής ανάλυσης*

### *7.4.1: Προυποθέσεις*

Υπάρχουν κάποιες αναγκαίες αλλά όχι ικανές προυποθέσεις για την εφαρμογή της παραγοντικής ανάλυσης. Παρακάτω αναφέρονται μερικές από αυτές, όμως στη συνέχεια γίνεται πλήρης ανάλυση της παραγοντικής ανάλυσης.

***Μέγεθος δείγματος***:Γενικά το μέγεθος του δείγματος πρέπει να είναι ικανοποιητικό, (κατά προτίμηση >100) και όλες οι ματαβλητές να είναι ποσοτικές (συνεχείς ή διακριτές). Οι μεταβλητές στο παράδειγμα μας που συμμετέχουν στην παραγοντική ανάλυση είναι 9 συνολικά και είναι όλες συνεχείς αφού αποδίδουν τα συστατικά στις ποικιλίες κρασιών. Το πλήθος των τριών ποικιλιών είναι 178.

***Αμεροληψία στην επιλογή των μεταβλητών:*** Στις μεταβλητές που χρησιμοποιούμε δεν υπάρχουν μεταβλητές με το ίδιο νόημα. Η κάθε μία αντιπροσωπεύει ένα ξεχωριστό συστατικό των κρασιών.

***Μέτριες προς υψηλές συσχετίσεις χωρίς να παρουσιάζουν πολυσυγγραμικότητας:***Στον παρακάτω πίνακα συσχετίσεων των μεταβλητών θα παρατήρουμε ότι αν υπάρχουν συσχετίσεις μεταξύ των μεταβλητών μεγαλύτερες από 0,3, αυτό μας εξασφαλίζει την ύπαρξη κάποιου βαθμού πολυσυγγραμικότητας.

*Τα δεδομένα πρέπει να ακολουθούν την διμεταβλητή κανονική κατανομή για κάθε ζεύγος μεταβλητών* και *οι παρατηρήσεις να είναι ανεξάρτητες*.

#### *7.4.1.1: Κανονικότητα των τιμών καθεμιάς των ανεξάρτητων μεταβλητών*

Αρχικά πρέπει να γίνει έλεγχος κανονικότητας των τιμών όλων των ανεξάρτητων μεταβλητών, κάτι το οποίο θα πραγματοποιηθεί μέσω του στατιστικού ελέγχου του Kolmogorov- Smirnov.

| *Έλεγχος κανονικότητας Kolmogorov-Smirnov* | | | |
|---|---|---|---|
| | N | Τιμές ελεγχοσυνάρτησης | p-value |
| Αλκοόλ | 178 | 0,907 | 0,384 |
| Τέφρα | 178 | 0,827 | 0,501 |
| Αλκολικότητα τέφρας | 178 | 0,798 | 0,547 |
| Μαγνήσιο | 178 | 1,195 | 0,115 |
| Φαινόλες | 178 | 0,932 | 0,350 |
| Φλαβανοειδή | 178 | 1,138 | 0,150 |
| Ανθοκυανίνες | 178 | 0,815 | 0,521 |
| Ένταση χρώματος | 178 | 1,213 | 0,853 |
| Απόχρωση | 178 | 0,105 | 0,460 |

*Πίνακας 13: Έλεγχος κανονικότητας Komologorov-Smirnov*

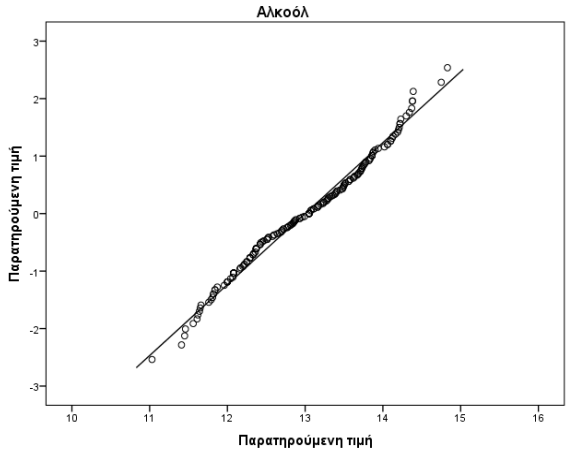
Όπως προκύπτει από τον παραπάνω πίνακα τα αποτελέσματα του ελέγχου επιβεβαιώνουν την υπόθεση κανονικότητας των δεδομένων αφού οι τιμές του στατιστικού p-valueείναι μεγαλύτερες του επιπέδου σημαντικότητας 5% (α=0,05). Εκτός από τις τιμές του στατιστικού p-valueμπορούμε να επιβεβαιώσουμε την κανονικότητας και από την ελεγχοσυνάρτηση του ελέγχου των Kolmogorov-

Smirnovτης οποία οι τιμές συγκρίνονται με το στατιστικό $D_{n;a}$, οι οποίες προκύπτουν από τον πίνακα που δίνεται μετά την βιβλιογραφία.

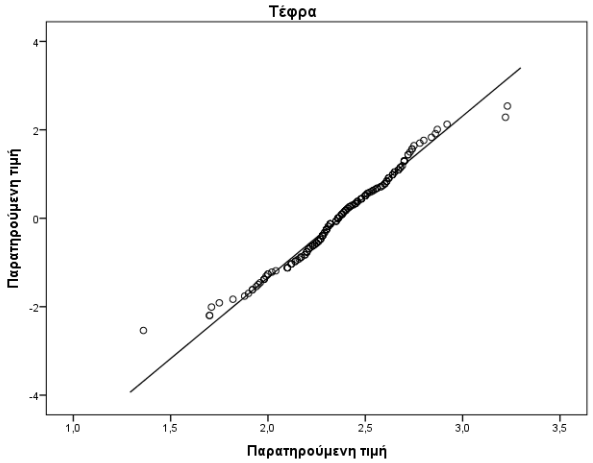
Ελεγχοσυνάρτηση Kolmogorov–Smirnov: $D_n = max|F(X) - F_n(X)$, απορρίπτεται η μηδενική υπόθεση όταν $D_n \leq D_{n,a}$. Για μέγεθος δείγματος 178 παρατηρήσεων και επίπεδο σημαντικότητας 0,05 η τιμή του στατιστικού που προκύπτει είναι $D_{178;0.05} = \frac{1.36}{\sqrt{178}} = 0.1019$.

Παρατηρούμε πως οι ελεγχοσυναρτήσεις όλων των μεταβλητών έχουν τιμές μεγαλύτερες του στατιστικού $D_{178;0.05}$ οπότε, επιβεβαιώνεται και με αυτόν τον τρόπο η κανονικότητα των μεταβλητών.

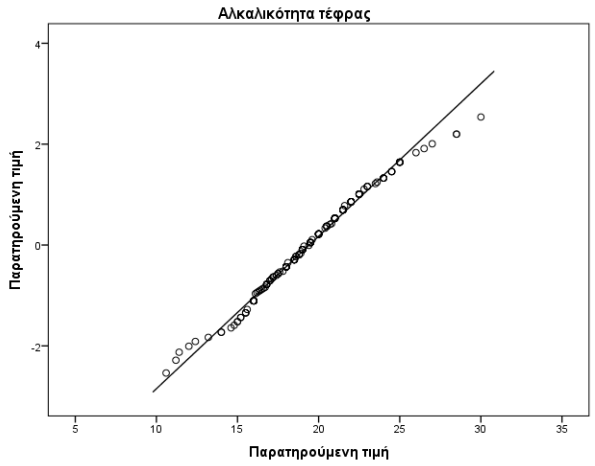
Για οπτική κιόλας εξακρίβωση της κανονικότητας παρουσιάζονται διαγράμματα σημείων με βάση τις πιθανότητες, όπως παρουσιάζονται παρακάτω:



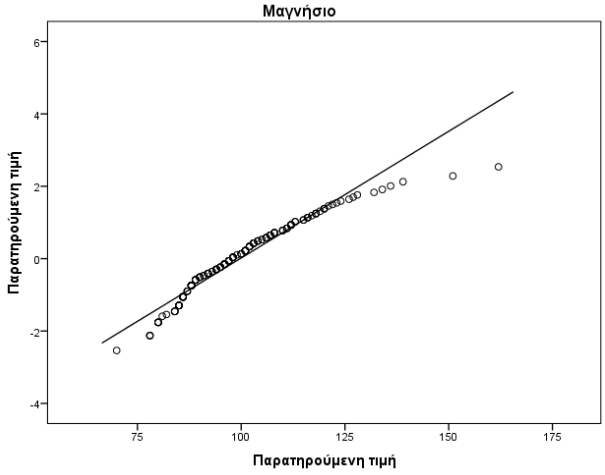

*Εικόνα 32: Διάγραμμα κανονικότητας των τιμών της μεταβλητής αλκοόλ*



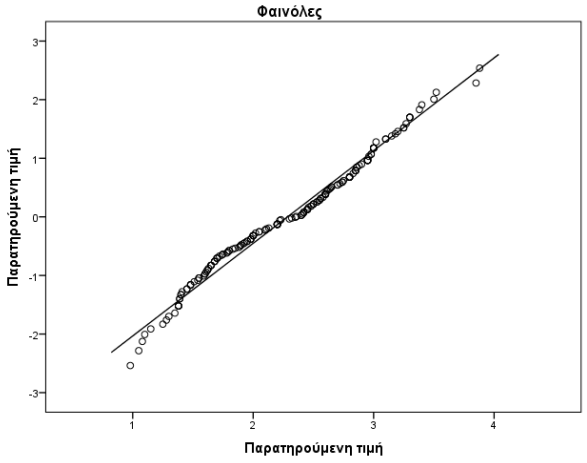

*Εικόνα 33: Διάγραμμα κανονικότητας των τιμών της μεταβλητής τέφρα*



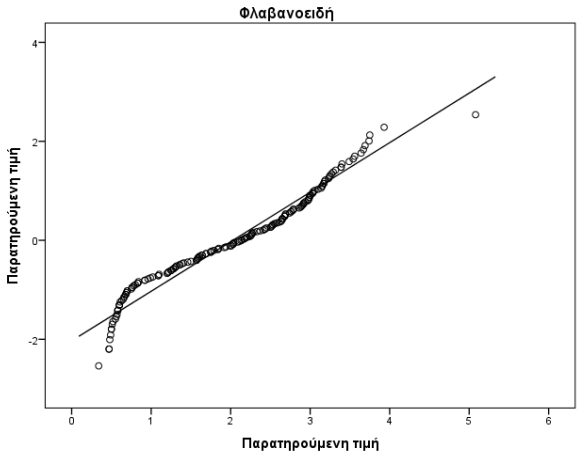

*Εικόνα 34: Διάγραμμα κανονικότητας των τιμών της μεταβλητής αλακλικοτητα τέφρας.*



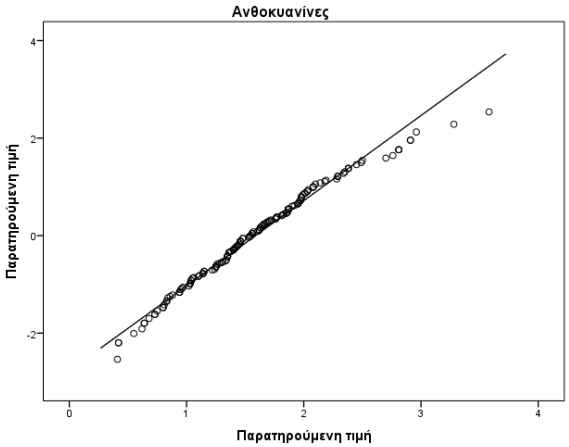

*Εικόνα 35: Διάγραμμα κανονικότητας των τιμών της μεταβλητής μαγνήσιο*



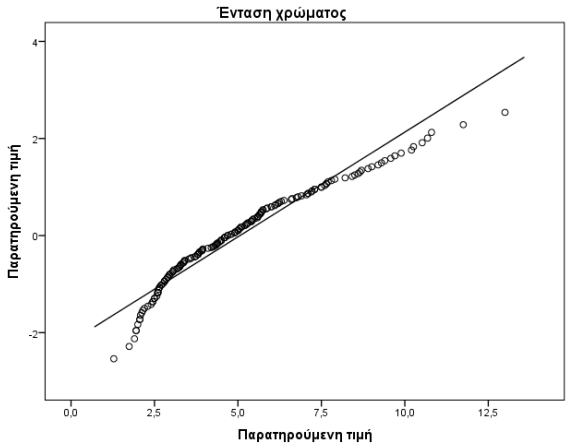

*Εικόνα 36: Διάγραμμα κανονικότητας των τιμών της μεταβλητής φαινόλες*



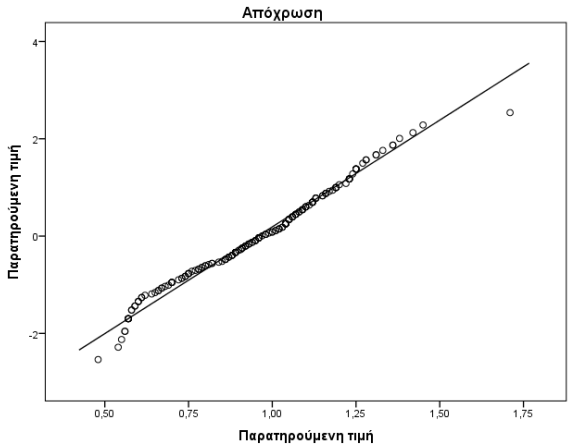

*Εικόνα 37: Διάγραμμα κανονικότητας των τιμών της μεταβλητής φλαβανοειδή*




Εικόνα 38: Διάγραμμα κανονικότητας των τιμών της μεταβλητής ανθοκυανίνες




Εικόνα 39: Διάγραμμα κανονικότητας των τιμών της μεταβλητής ένταση χρώματος




Εικόνα 40: Διάγραμμα κανονικότητας των τιμών της μεταβλητής απόχρωση

### *7.4.2: Έλεγχος συσχετίσεων*

Εξετάζουμε το τον έλεγχο της απουσίας συσχέτισης έναντι της υπόθεσης ύπαρξης συσχέτισης. Για τα δεδομένα του παραδείγματος μας με βάση τις τιμές των συντελεστών συσχέτισης των ανεξάρτητων μεταβλητών από τους πίνακες παρατηρούμε αρκετές συσχετίσεις.

| Συσχέτιση<br>p-value | Αλκόολ | Τέφρα | Αλκολικότητα τέφρας | Μαγνήσιο | Φαινόλες | Φλαβανοειδή | Ανθοκυανίνες | Ένταση χρώματος | Απόχρωση |
|---|---|---|---|---|---|---|---|---|---|
| Αλκοόλ | 1 | 0,204 | -0,316 | 0,261 | 0,285 | 0,224 | 0,113 | 0,550 | -0,062* |
| | | 0,06 | 0,000 | 0,000 | 0,000 | 0,003 | 0,134 | 0,000 | 0,409 |
| Τέφρα | 0,204 | 1 | 0,422 | 0,284* | 0,140 | 0,117 | 0,008 | 0,255 | -0,084 |
| | 0,06 | | 0,000 | 0,000 | 0,062 | 0,119 | 0,917 | 0,001 | 0,264 |
| Αλκολικότητα τέφρας | -0,316 | 0,422 | 1 | -0,112 | -0,320* | -0,340* | -0,193 | -0,003 | -0,244 |
| | 0,000 | 0,000 | | 0,137* | 0,000 | 0,000 | 0,010 | 0,964 | 0,001* |
| Μαγνήσιο | 0,261 | 0,284* | -0,112 | 1 | 0,208* | 0,195 | 0,236 | 0,200 | 0,055 |
| | 0,000 | 0,000 | 0,137* | | 0,005 | 0,009 | 0,002 | 0,007 | 0,467 |
| Φαινόλες | 0,285 | 0,140 | -0,320* | 0,208* | 1 | 0,867* | 0,588 | -0,056 | 0,452* |
| | 0,000 | 0,062 | 0,000 | 0,005 | | 0,000 | 0,000 | 0,455 | 0,000 |
| Φλαβανειδή | 0,224 | 0,117 | -0,340* | 0,195 | 0,867* | 1 | 0,649* | -0,174* | 0,545 |
| | 0,003 | 0,119 | 0,000 | 0,009 | 0,000 | | 0,000 | 0,020 | 0,000 |
| Ανθοκυανίνες | 0,113 | 0,008 | -0,193 | 0,236 | 0,588 | 0,649* | 1 | -0,027 | 0,298* |
| | 0,134 | 0,917 | 0,010 | 0,002 | 0,000 | 0,000 | | 0,719* | 0,000 |
| Ένταση χρώματος | 0,550 | 0,255 | -0,03 | 0,200 | -0,056 | -0,174* | -0,027 | 1 | -0,524 |
| | 0,000 | 0,001 | 0,964 | 0,007 | 0,455 | 0,020 | 0,719* | | 0,000 |
| Απόχρωση | -0,062* | -0,084 | -0,244 | 0,055 | 0,452* | 0,545 | 0,298* | -0,524 | 1 |
| | 0,409 | 0,264 | 0,001* | 0,467 | 0,000 | 0,000 | 0,000 | 0,000 | |

*Πίνακας 14: Πίνακας Συσχετίσεων*

Από τον παραπάνω πίνακα παρατηρούνται κάποιες μεμονωμένες συσχετίσεις μεταξύ των μεταβλητών.Πέραν όμως από τις μεμονωμένες συσχετίσεις που παρατηρούνται από τον παραπάνω πίνακα η διαπίστωση της επιβαβαιώνεται και από τον στατιστικό έλεγχο του Bartlettαπό τον παρακάτω πίνακα έχει τιμή $X^2_{(r-1)(c-1);\alpha} = X^2_{(9-1)(9-1);0,05} = 792.915$ και το στατιστικό

p-value=0,0001<0.05(=α) οπότε οδηγούμαστε στην απόρριψη της υπόθεσης απουσίας συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών.

Όπως μπορούμε να διαπιστώσουμε οι μεταβλητές μας συσχετίζονται κάτι που όπως αναφέραμε δεν θέλουμε να συμβεί.Παρόλα αυτά θα συνεχίσουμε την διαδικασία της παραγοντικής ανάλυσης.

### *7.4.3: Εφαρμογή παραγοντικής ανάλυσης*

Επειδή θέλουμε οι τιμές των μεταβολών (communalities) της κάθε μεταβλητής να είναι μεγαλύτερες από 0,7 και οι τιμές του **μαγνησίου** και των **ανθοκυανίνων**είναι μικρότερες, αφαιρέθηκαν από την παραγοντική ανάλυση. Οπότε, έχουμε:

*Πίνακας συσχετίσεων:* Στον παρακάτω πίνακα εμφανίζονται οι συσχετίσεις όλων των ζευγών των μεταβλητών.

| Συσχέτιση | | Αλκοόλ | Τέφρα | Αλκαλικότητα τέφρας | Μαγνήσιο | Φαινόλες | Φλαβανοειδή | Ανθοκυανίνες | Ένταση χρώματος | Απόχρωση |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Αλκοόλ** | **1,000** | **0,204** | **-0,316** | **0,261** | **0,285** | **0,224** | **0,113** | **0,550** | **-0,062** |
| | **Τέφρα** | 0,204 | **1,000** | **0,422** | **0,284** | **0,140** | **0,117** | **0,08** | **0,255** | **-0,084** |
| | **Αλκαλικότητα τέφρας** | -0,316 | 0,422 | **1,000** | **-0,112** | **-0,320** | **-0,340** | **-0,193** | **-0,003** | **-0,244** |
| | **Μαγνήσιο** | 0,261 | 0,284 | **-0,112** | **1,000** | **0,208** | **0,195** | **0,236** | **0,200** | **0,055** |
| | **Φαινόλες** | 0,285 | 0,140 | -0,320 | **0,208** | **1,000** | **0,867** | **0,588** | **-0,056** | **0,452** |
| | **Φλαβανοειδή** | 0,224 | 0,117 | -0,340 | 0,195 | 0,867 | **1,000** | **0,649** | **-0,174** | **0,545** |
| | **Ανθοκυανίνες** | 0,113 | 0,008 | -0,193 | 0,236 | **0,588** | **0,649** | **1,000** | | **0,298** |
| | **Ένταση χρώματος** | 0,550 | 0,255 | -0,03 | 0,200 | -0,056 | -0,174 | **-0,027** | **1,000** | **-0,524** |
| | **Απόχρωση** | -0,062 | -0,084 | -0,244 | 0,055 | 0,452 | 0,545 | **0,298** | -0,524 | **1,000** |

*Πίνκας 15: Πίνακας συσχετίσεων*

Από τον παραπάνω πίνακα, παρατηρούμε ότι στην διαγώνιο υπάρχει η τιμή 1, η οποία είναι ο συντελεστής συσχέτισης της κάθε μεταβλητής με την ίδια. Οι τιμές κάτω από την διαγώνιο είναι ίδιες με τις τιμές πάνω από την διαγώνιο. Από τον παραπάνω πίνακα μπορούμε να εντοπίσουμε μεταξύ ποιων μεταβλητών υπάρχει μεγάλη εξάρτηση. Στον ίδιο πίνακα, στη γραμμή sig.(1-tailed), εμφανίζονται οι σημαντικότητες αυτών των συσχετίσεων.

*Πίνακας συμμετοχικότητας(Communalities)*: Ο πίνακας αυτός περιέχει τιμές οι οποίες εκφράζουν το ποσοστό της μεταβολής της κάθε μεταβλητής το οποίο ερμηνεύεται από τους παράγοντες. Στον παρακάτω πίνακα υπάρχουν τιμές πριν και μετά την επιλογή του πλήθους των παραγόντων της ανάλυσης.Στην μέθοδο Principalcomponents–κύριες συνιστώσες, οι τιμές Initialείναι πάντα 1, ενώ οι τιμές communalitiesκυμαίνονται από 0-1. Η τιμή 0 δηλώνει ότι οι παράγοντες δεν ερμηνεύουν κανένα ποσοσό μεταβολής της μεταβλητής, ενώ η τιμή 1 δηλώνει ότι το 100% των μεταβολών της μεταβλητής ερμηνεύεται από τους παράγοντες. Θέλουμε οι βαρύτητες να έχουν τιμές >0,7 κάτι το οποίο συμβαίνει στον παρακάτω πίνακα, εκτός από την ***μεταβλητή-μαγνήσιο*** και την ***μεταβλητή ανθοκυανίνες*** οι οποίες είχαν αρκετά χαμηλή βαρύτητα και έτσι αποκλείστηκαν από την ανάλυσή μας.

***Communalities***

| | Initial | Βαρύτητες |
|---|---|---|
| Αλκόολ | 1,000 | 0,787 |
| Τέφρα | 1,000 | 0,857 |
| Αλκακικότητα τέφρας | 1,000 | 0,845 |
| Φαινόλες | 1,000 | 0,847 |
| Φλαβανοειδή | 1,000 | 0,884 |
| Ένταση χρώματος | 1,000 | 0,845 |
| Απόχρωση | 1,000 | 0,722 |

*Πίνακας 16: Πίνακα βαρυτήτων*

Πέρα όμως από τον παραπάνω πίνακα που όπως ανέφερα πρέπει οι βαρύτητες να είναι >0,7 για να συνεχισεί η μεταβλητή να είναι σημαντικής για την ανάλυσης μας ρόλο σε αυτό παίζει και από παρακάτω πίνακα, ο δείκτης MSA.

*Τώρα,* για να έλεγχο του δείκτη ***MSA(MeasuresofSamplingAdeaquacy, Μέτρα Δειγματικής Επάρκειας)*** της κάθε μεταβλητής ξεχωριστά, δημιουργούμε τον πίνακα ***Anti-imageMatrices***.Οι μεταβλητές που έχουν τιμές ***MSA*** μη αποδεκτές δηλαδή >0,5 θα αποκλειστούν από την συνέχεια της διαδικασίας και οι υπόλοιπες θα δώσουν ***KMO***αποδεκτό έτσι ώστε να συνεχιστεί η διαδικασία.Στον πίνακα anti-imageοι 8 από τις 9 τιμές της κύριας διαγωνίου είναι μεγαλύτερες από 0,5, οπότε στη συνέχεια της αρχικής εφαρμογής της παραγοντικής ανάλυσης η μία αυτή μεταβλητή θα αποκλειστεί από την ανάλυση αφού το μέτρο επάρκειας της δειγματοληψίας για κάθε μία μεταβλητή πρέπει να είναι MSA> 0,5. Η μεταβλητή αυτή που έχει τιμή < 0,5 είναι η τέφρα.

| | | Αλκοόλ | Τέφρα | Αλκολικότητα Τέφρας | Φαινόλες | Φλαβανοειδή | Ένταση χρώματος | Απόχρωση |
|---|---|---|---|---|---|---|---|---|
| Anti-image Correlation | Αλκοόλ | 0,570 | -0,198 | 0,326 | -0,067 | -0,034 | -0,526 | -0,099 |
| | Τέφρα | -0,198 | 0,447 | -0,563 | -0,041 | -0,139 | -0,160 | -0.012 |
| | Αλκολικότητα Τέφρας | 0,326 | -0,563 | 0,550 | 0,035 | 0,148 | 0,039 | 0,084 |
| | Φαινόλες | -0,067 | -0,041 | 0,035 | 0,641 | -0,794 | -0,103 | -0,046 |
| | Φλαβανοειδή | -0,034 | -0,139 | 0,148 | -0,794 | 0,629 | 0,127 | -0,226 |
| | Ένταση χρώματος | -0,526 | -0,160 | 0,039 | -0,103 | 0,127 | 0,545 | 0,484 |
| | Απόχρωση | -0,099 | 0,012 | 0,084 | -0,046 | -0,226 | 0,484 | 0,735 |

*Πίνακας 17: Έλεγχος δειγματικής επάρκειας*

Αν από τον παραπάνω πίνακα αν αφαιρεθεί και η τέφρα, τότε επηρεάζεται και η βαρύτητα της αλκαλικότητα της τέφρας και είναι αρκετά μικρότερη του 0,7, οπότε πρέπει και εκείνη να αφαιρεθεί από την ανάλυση μας.

Άρα μετά την αφαίρεση των μεταβλητών **μαγνήσιο**, **ανθοκυανίνες**, **τέφρα** και **αλκαλικότητα τέφρας** η παραγοντική ανάλυση εφαρμόζεται παρακάτω:

Αρχικά, δίνεται ο πίνακας συσχέτισης μεταξύ των πέντε αυτών μεταβλητών:

| Συσχέτιση | | Αλκοόλ | Φαινόλες | Φλαβανοειδή | Ένταση χρώματος | Απόχρωση |
|---|---|---|---|---|---|---|
| | Αλκοόλ | 1,000 | 0,285 | 0,224 | 0,550 | -0,62 |
| | Φαινόλες | 0,285 | 1,000 | 0,867 | -0,056 | 0,452 |
| | Φλαβανοειδή | 0,224 | 0,867 | 1,000 | -0,174 | 0,545 |
| | Ένταση χρώματος | 0,550 | -0,056 | -0,174 | 1,000 | -0,524 |
| | Απόχρωση | -0,062 | 0,452 | 0,545 | -0,524 | 1,000 |

*Πίνακας 18: Πίνακας συσχέτισης*

***Έλεγχος KMO:***Όπως μπορούμε να παρατηρήσουμε και στον παρακάτω πίνακα σχετικά με τον έλεγχο ***KMO*** ( Kaiser-Meyer—OlkinMeasureofSamplingAdaquary) δηλαδή είναι το μέτρο επάρκειας της δειγματοληψίας του Kaiser-Meyer-Olkein. Μπορούμε σε αυτό το σημείο να τονίσουμε ότι ο λόγος σφαιρικότητας του Bartlettκαι το ***KMO***testδείχνουν εάν η ανάλυση θα αποφέρει διακριτούς και αξιόπιστους παράγοντες. Ουσιαστικά είναι ένας δείκτης σύγκρισης του σχετικού μεγέθους των συντελεστών συσχέτισης με τους μερικούς συντελεστές συσχέτισης. Οι τιμές του δείκτη αυτού κυμαίνονται από 0-1, τιμές <0,5 θεωρούνται μη αποδεκτές και δεν συνισταται η συνέχιση της παραγοντικής διαδικασίας.

Γενικά μπορούμε να πούμε ότι ο δείκτης ***KMO***μεγαλώνει όταν :

- Το μέγεθος του δείγματος μεγαλώνει
- Ο μέσος όρος των συσχετίσεων μεγαλώνει
- Το πλήθος των μεταβλητών αυξάνει ή
- Το πλήθος των παραγόντων ελαττώνεται

Στον ίδιο πίνακα υπάρχει και το ***Barlett'sTestofsphericity***, το οποίο ουσιαστικά μας δίνει τη πιθανότητα ο πίνακας συσχέτισης να έχει σημαντικές συσχετίσεις μεταξύ κάποιων μεταβλητών. Έτσι αν η τιμή του δείκτη αυτού είναι <0,05 απορρίπτεται η υπόθεση της μη ύπαρξης σημαντικών συσχετίσεων σε επίπεδο σημανικότητας 5%. .

| Έλεγχοι KMO και Barlett | | |
|---|---|---|
| Kaiser-Mayer-Olkin | | 0,594 |
| Σφαιρικότητατου Barlett | $x^2$ | 486,376 |
| | β.ε. | 10 |
| | p-value | 0,000 |

*Πίνακας 19: Έλεγχος KMO*

Ο έλεγχος του ***Kaiser-Mayer-Olkin***δίνει τιμή 0,594.Οπότε μπορεί να συνεχιστεί η εφαρμογή της παραγοντικής ανάλυσης. Και η τιμή του ***Barlett***είναι <0,005 που σημαίνει ότι υπάρχουν σημαντικές συσχετίσεις μεταξύ των μεταβλητών.Το ότι η τιμή του είναι 0,594 (σχετικά μικρή αλλά αποδεκτή) μπορεί να οφείλεται όπως θα φανεί και παρακάτω στον πίνακα ότι οι τιμές του δείκτη MSAείναι σχετικά χαμηλοί.

Στον παρακάτω πίνακα φαίνονται οι τιμές του δείκτηMSA, και όπως παρατηρούμε όλες είναι ≤ 0,5 οπότε είναι όλες αποδεκτές.

| | | Αλκοόλ | Φαινόλες | Φλαβανοειδή | Ένταση χρώματος | Απόχρωση |
|---|---|---|---|---|---|---|
| Anti-image Correlation | Αλκοόλ | 0,541 | -0,083 | -0,090 | -0,582 | -0,134 |
| | Φαινόλες | -0,083 | 0,604 | -0,813 | -0,110 | -0,048 |
| | Φλαβανοειδή | -0,090 | -0,813 | 0,604 | 0,114 | -0,239 |
| | Ένταση χρώματος | -0,582 | -0,110 | 0,114 | 0,500 | 0,497 |
| | Απόχρωση | -0,134 | -0,048 | -0,239 | 0,497 | 0,706 |

*Πίνακας 20:Έλεγχος δειγματικής επάρκειας*

Έπειτα δίνεται ο πίνακας των βαρυτήτων των μεταβλητων.

| | Intial | Βαρύτητες |
|---|---|---|
| Αλκοόλ | 1,000 | 0,774 |
| Φαινόλες | 1,000 | 0,862 |
| Φλαβανοειδή | 1,000 | 0,889 |
| Ένταση χρώματος | 1,000 | 0,856 |
| Απόχρωση | 1,000 | 0,724 |

*Πίνακας 21: Πίνακας βαρυτήτων*

Παρακάτω δίνεται ο πίνακας ***TotalVarianceExplained – Συνολική ερμηνευθείσα διακύμανση***, όπου παρουσιάζονται αρκετές στήλες με αρκετές πληροφορίες.

| | Initial Eigenvalues | | | Extraction Sums of squared Loadings | | | Rotation Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|---|---|---|
| Component | Total | % of variance | Cumulative % | Total | % of variance | Cumulative % | Total | % of variance | Cumulative % |
| 1 | 2,393 | 47,850 | 47,850 | 2,393 | 47,850 | 47,850 | 2,369 | 47,385 | 47,385 |
| 2 | 1,713 | 34,260 | 82,111 | 1,713 | 34,260 | 82,110 | 1,736 | 34,726 | 82,111 |
| 3 | 0,502 | 10,042 | 92,152 | | | | | | |
| 4 | 0,269 | 5,381 | 97,533 | | | | | | |
| 5 | 0,123 | 2,467 | 100,000 | | | | | | |

*Πίνακας 22: Συνολική εξήγηση διακύμανσης*

Η 1η στήλη δηλώνει το πλήθος των παραγόντων οι οποίοι είναι όσοι και οι μεταβλητές

*Για InitialEingenvalues (αρχικές ιδιοτιμές) έχουμε:*

- Η στήλη ***Total***δίνει τις ιδιοτιμές –eingenvalues για κάθε παράγοντα και είναι ταξηνομημένες κατά τάξη μεγέθους. Το άθροισμα των τιμών αυτών είναι ίσο με το πλήθος των παραγόντων.
- Η στήλη ***% ofvariance***δίνει το ποσοστό της διακύμανσης το οποίο ερμηνεύεται από τον παράγοντα. Είναι το πηλίκο της κάθε ιδιοτιμής της στήλης ***Total***διά του συνολικού πλήθους των παραγόντων.

- Η στήλη ***Cumulative %*** περιέχει αθροιστικά τα δύο ποσοστά της προηγούμενης στήλης (περιέχει αθροιστικά τα δύο ποσοστά της προηγούμενης στήλης. (47,850+34,260=82,110)κ.ο.κ.

*Για ExtractionSumsofsquaredLoadings (τα ποσά εξαγωγής τετραγωνικών φορτίσεων) έχουμε:*

- Η στήλη ***Total***δίνει τους παράγοντες των οποίων οι ιδιοτιμές είναι μεγαλύτερες του 1. Είναι ουσιαστικά ένας δείκτης ο οποίος καθορίζει τον αριθμό των παραγόντων οι οποίοι θα προκύψουν από την παραγοντική ανάλυση.
- Ηστήλη***% ofvariance***δίνειτο ποσοστό της διακύμανσης το οποίο ερμηνεύεται από τους δύο παράγοντες, με ιδιοτιμή μεγαλύτερη του 1,
- Η στήλη ***Cumulative %***περιέχει αθροιστικά τα δύο ποσοστά της προηγούμενης στήλης. (47,850+34,260=82,110).

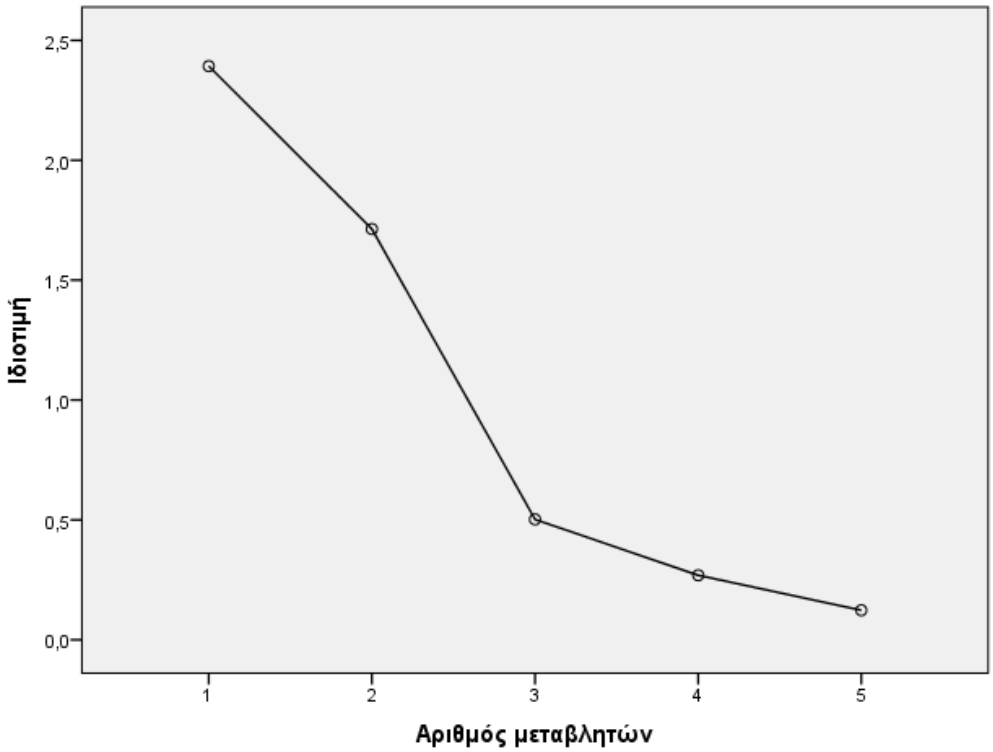
*Για RotationSumsofSquaredLoadings ( τα ποσά περιστροφής των τετραγωνικών φορτίσεων) έχουμε:*

- Η στήλη ***Total***δίνει τους παράγοντες οι οποίοι έχουν ιδιοτιμές μεγαλύτερες του 1 μετά την περιστροφή. Παρατηρούμε ότι οι τιμές μετά την περιστροφή έχουν αλλάξει.
- Ηστήλη***% ofvariance***δίνει το ποσοτό της διακύμανσης που ερμηνεύεται από τους δύο παράγοντες, με ιδιοτιμή μεγαλύτερη του 1, μετά την περιστροφή.
- Η στήλη ***Cumulative %***είναι η στήλη των αθροιστικών ποσοστών της προηγούμενης στήλης. Στην πρώτη γραμμή γράφεται το ποσοστό του $1^{ου}$ παράγοντα 47,385, στην $2^{η}$ γραμμή προστίθεται σε αυτό και το ποσοστό του $2^{ου}$ παράγοντα (47,385+34,726=82,111)

Υπάρχουν κάποια κριτήρια:

1. Συνήθως η τιμή της ιδιοτιμής πρέπει να είναι μεγαλύτερη του 1. Έτσι, κάθε παράγοντας με ιδιοτιμή μεγαλύτερη του 1 θεωρείται σημαντικός, ενώ κάθε παράγοντας με ιδιοτιμή μικρότερη του 1 θεωρείται η σημαντικός και αγνοείται.
2. Το $2^{ο}$ κριτήριο είναι η εκ των προτέρων βούληση για συγκεκριμένο αριθμό παραγόντων με βάση κάποια δεδομένα. Για παράδειγμα, αν ο ερευνητής θέλει να κάνει σύγκριση των αποτελεσμάτων της δικής του έρευνας με κάποια άλλη η οποία έδωσε 5 παράγοντες, θα πρέπει και αυτός να ζητήσει 5 παράγοντες.
3. Το $3^{ο}$ κριτήριο θεωρείται το άθροισμα των διακυμάνσεων το οποίο θεωρείται ικανοποιητικό. Δηλαδή, αν θέλουμε το ποσοστό της διακύμανσης, το οποίο περιγράφεται από τους παράγοντες να είναι 85%, θα επιλέξουμε τόσους παράγοντες ώστε να φτάσουμε σε αυτό το αποτέλεσμα.

4. Το 4ο κριτήριο στηρίζεται στο ScreePlot, με βάση την γραφική απεικόνιση, μετά από σημείο το οποίο η καμπύλη τείνει να γίνει ευθεία πρέπει να απορρίπτονται οι παράγοντες.




*Εικόνα 41: Διάγραμμα σημείων*

Από το παραπάνω διάγραμμα παρουσιάζονται οι μεταβλητές, που περιλαμβάνονται και επηρεάζουν το διάγραμμα το οποίο βοηθά στην εξαγωγή των παραγόντων. Από το παραπάνω διάγραμμα, βλέποντας την ιδιοτιμή του κάθε παράγοντα, μπορούμε εύκολα να προσδιορίσουμε αυτούς που υπερβαίνουν το 1 και επομένων αποτελούν τους παράγοντες που πληρούν τον περιορισμό που θέσαμε με βάση το 1ο κριτήριο. Επομένως δημιουργούμε 2 παράγοντες.
Από τον παρακάτω πίνακα έχουμε:

| | Συνιστώσες | |
|---|---|---|
| | 1 | 2 |
| Φαινόλες | 0,916 | 0,154 |
| Φλαβανοειδή | 0,942 | 0,031 |
| Απόχρωση | 0,709 | -0,471 |
| Αλκόολ | 0,295 | 0,829 |
| Ένταση χρώματος | -0,229 | 0,896 |

*Πίνακας 23: Factorloadings παραγόντων σε πίνακα με περιστροφή*

Οι παραπάνω τιμές θέλουμε να είναι μεγαλύτερες από το 0,3 κάτι που παρατηρούμε ότι ισχύει. Με μεγαλύτερη την τιμή του συστατικού φλαβανοειδή και μικρότερη την τιμή του συστατικού μαγνησίου. Με μωβ είναι οι τιμές των μεταβλητών που

περιλαμβάνονται στον παράγοντα 1, με καφέ οι τιμές των μεταβλητών που περιλαμβάνονται στον παράγοντα 2. Οι τιμές για τον κάθε παράγοντα επιλέγονται με βάση του ότι σε κάθε παράγοντα επιλέγουμε τις υψηλότερες τιμές των συστατικών. Η αρνητική τιμή σημαίνει πως όσο πιο μικρές είναι οι τιμές για παράδειγμα της απόχρωσης τόσο πιθανότερο είναι να ανήκει στον άλλον παράγοντα.

Εφαρμόζοντας την διαδικασία της παραγοντικής ανάλυσης, προέκυψαν δύο παράγοντες. Δυστυχώς δεν μπορώ να δώσω ονόματα στους παράγοντες που προέκυψαν μιας και από τις μεταβλητές που συμπεριλαμβάνουν δεν μπορούμε να διακρίνουμε κάποια εμφανές κοινά χαρακτηριστικά, οπότε θα τα αφήσουμε ως ΄΄F1΄΄,΄΄F2΄΄.

***Παράγοντες***

Ο 1ος παράγοντας αποτέλειται από τις μεταβλητές <<Φλαβανοειδή>>, <<Φαινόλες>>, <<Απόχρωση>>.

| Μεταβλητές | F1 |
|---|---|
| Φαινόλες | 0,916 |
| Φλαβανοειδή | 0,942 |
| Απόχρωση | 0,709 |

*Πίνακας 24: FactorLoading του 1ου παράγοντα*

Ο 2ος παράγοντας αποτέλειται από τις μεταβλητές <<Ενταση χρώματος>>, <<Αλκοόλ>>.

| Μεταβλητές | F2 |
|---|---|
| Αλκοόλ | 0,829 |
| Ένταση χρώματος | 0,896 |

*Πίνακας 25: FactorLoading του 2ου παράγοντα*

Με βάση τα παραπάνω, εφαρμόστηκε παραγοντική ανάλυση με τη μέθοδο κύριων αξόων (principalaxisfactoring) και με περιστροφή(varimax) για την διερεύνηση των λανθάνουσων μεταβλητών. Η ανάλυσή μας έδειξε ότι υπάρχουν σαφείς λανθάνουσες δομές προτίμησης.

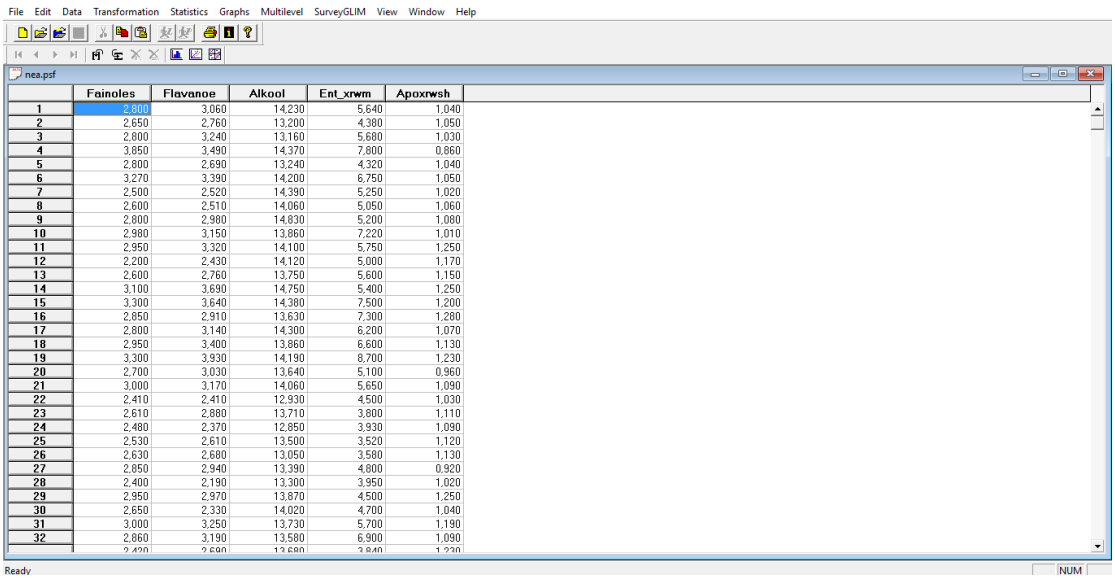
Για κάθε ομάδα και για κάθε περίπτωση του δείγματος υπολογίζεται ο μέσος όρος του βαθμού προτίμησης.

Καταλήγουμε έτσι σε 2 νέες μεταβλητές (λανθάνουσες μεταβλητές). Μετά την ανάδειξη λοιπόν των μεταβλητών αυτών, μπορεί να γίνει τώρα εφαρμογή του προγράμματος LISREL.

## *7.5: Εφαμογή του προγράμματος LISREL*

### *7.5.1: Εφαρμογή με βάση τους παράγοντες της παραγοντικής ανάλυσης*

Έτσι τώρα λοιπόν εισάγω τις μεταβλητές μου στο πρόγραμμα χωρίς τις τιμές των λανθάνουσων, και η μορφή τους είναι η εξής:

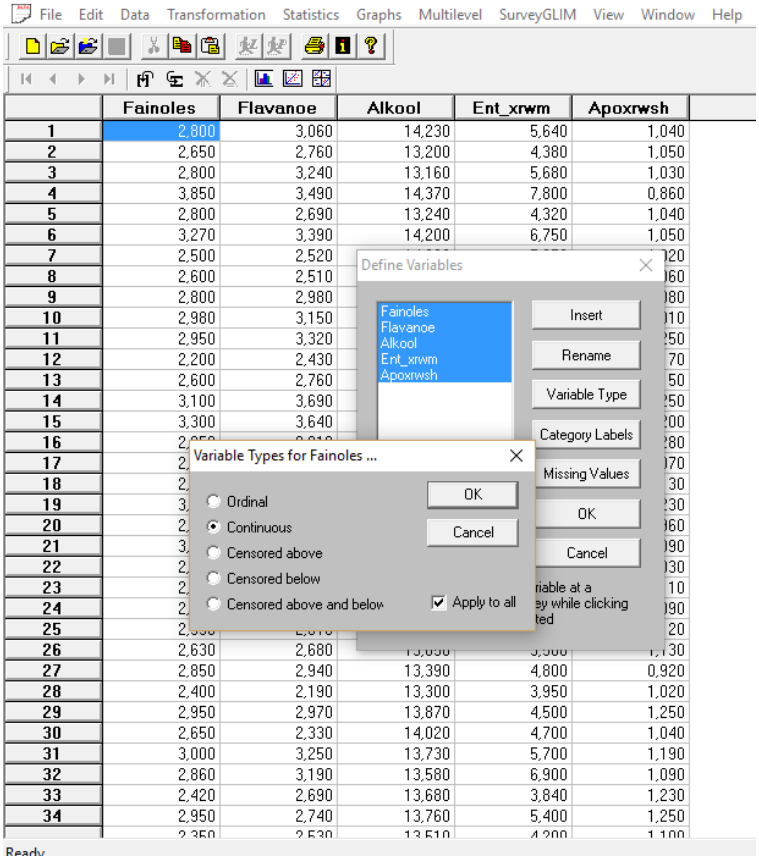



| | Fainoles | Flavanoe | Alkool | Ent_xrwm | Apoxrwsh |
|---|---|---|---|---|---|
| 1 | 2,800 | 3,060 | 14,230 | 5,640 | 1,040 |
| 2 | 2,650 | 2,760 | 13,200 | 4,380 | 1,050 |
| 3 | 2,800 | 3,240 | 13,160 | 5,680 | 1,030 |
| 4 | 3,850 | 3,490 | 14,370 | 7,800 | 0,860 |
| 5 | 2,800 | 2,690 | 13,240 | 4,320 | 1,040 |
| 6 | 3,270 | 3,390 | 14,200 | 6,750 | 1,050 |
| 7 | 2,500 | 2,520 | 14,390 | 5,250 | 1,020 |
| 8 | 2,600 | 2,510 | 14,060 | 5,050 | 1,060 |
| 9 | 2,800 | 2,980 | 14,830 | 5,200 | 1,080 |
| 10 | 2,980 | 3,150 | 13,860 | 7,220 | 1,010 |
| 11 | 2,950 | 3,320 | 14,100 | 5,750 | 1,250 |
| 12 | 2,200 | 2,430 | 14,120 | 5,000 | 1,170 |
| 13 | 2,600 | 2,760 | 13,750 | 5,600 | 1,150 |
| 14 | 3,100 | 3,690 | 14,750 | 5,400 | 1,250 |
| 15 | 3,300 | 3,640 | 14,380 | 7,500 | 1,200 |
| 16 | 2,850 | 2,910 | 13,630 | 7,300 | 1,280 |
| 17 | 2,800 | 3,140 | 14,300 | 6,200 | 1,070 |
| 18 | 2,950 | 3,400 | 13,860 | 6,600 | 1,130 |
| 19 | 3,300 | 3,930 | 14,190 | 8,700 | 1,230 |
| 20 | 2,700 | 3,030 | 13,640 | 5,100 | 0,960 |
| 21 | 3,000 | 3,170 | 14,060 | 5,650 | 1,090 |
| 22 | 2,410 | 2,410 | 12,930 | 4,500 | 1,030 |
| 23 | 2,610 | 2,880 | 13,710 | 3,800 | 1,110 |
| 24 | 2,480 | 2,370 | 12,850 | 3,930 | 1,090 |
| 25 | 2,530 | 2,610 | 13,500 | 3,520 | 1,120 |
| 26 | 2,630 | 2,680 | 13,050 | 3,580 | 1,130 |
| 27 | 2,850 | 2,940 | 13,390 | 4,800 | 0,920 |
| 28 | 2,400 | 2,190 | 13,300 | 3,950 | 1,020 |
| 29 | 2,950 | 2,970 | 13,870 | 4,500 | 1,250 |
| 30 | 2,650 | 2,330 | 14,020 | 4,700 | 1,040 |
| 31 | 3,000 | 3,250 | 13,730 | 5,700 | 1,190 |
| 32 | 2,860 | 3,190 | 13,580 | 6,900 | 1,090 |

*Εικόνα 42: Εισαγωγή δεδομένων στο πρόγραμμα*

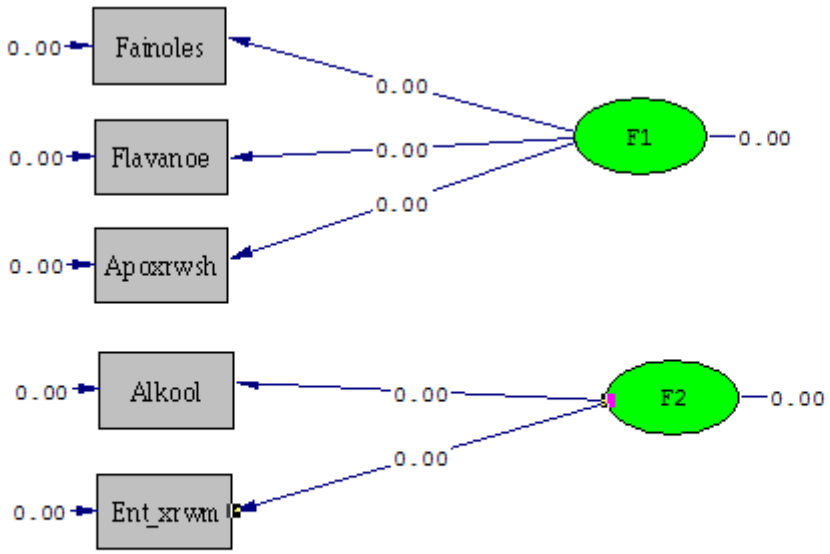
Ο τρόπος εισαγωγής των μεταβλητών στο πρόγραμμα έγινε μέσω του προγράμματος SPSSκαι με την σειρά που η εφαρμογή της παραγοντικής ανάλυσης, χώρισε τις μεταβλητές σε δύο παράγοντες. Η μέθοδος που εφαρμόζεται είναι η επιβεβαιωτική παραγοντική ανάλυση καθώς εκείνη συνιστάται στον έλεγχο δομικών μοντέλων εξισώσεων που γίνεται πάνω σε ένα σε ένα δείγμα δεδομένων. Στην περίπτωση αυτή, όπως και στη δική μας περίπτωση το μοντέλο περιλαμβάνει ένα από παρατηρήσιμες και λανθάνουσες μεταβλητές.

Αυτό που εφαρμόζεται αρχικά στο πρόγραμμα είναι να ορίσουμε ότι οι μεταβλητές μου είναι συνεχείς (continuous).



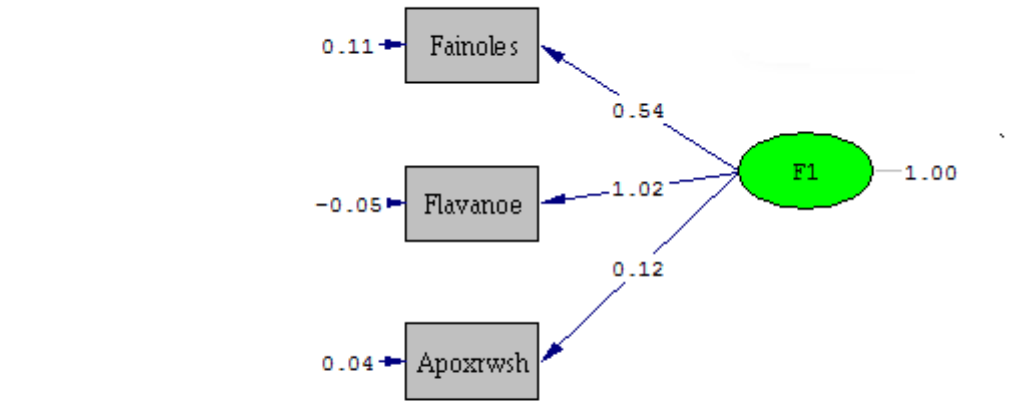

*Εικόνα 43: Ορισμός συνεχών μεταβλητών*

Στη συνέχεια, θα γίνει το διάγραμμα διαδρομής με την εισαγωγή και των παρατηρήσιμων αλλά και των λανθάνουσων μεταβλητών, για να γίνουν εμφανής οι σχέσεις μεταξύ τους.

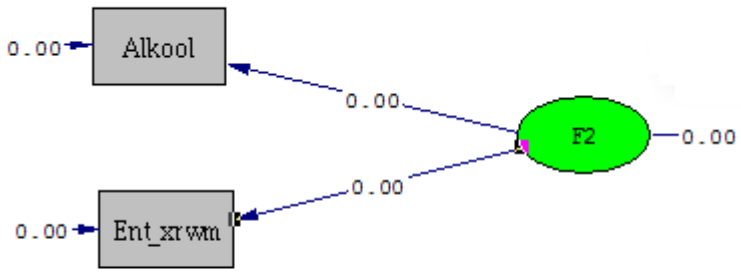



*Εικόνα 44: Διάγραμμα διαδρομής*

Στην παραπάνω εικόνα μπορούμε να πούμε ότι υπάρχει κάποιο πρόβλημα, δηλαδή η σύνδεση ενός από τους δύο παράγοντες με τις αντίστοιχες μεταβλητές, δημιουργεί πρόβλημα και για τον λόγο αυτό δεν εμφανίζεται τόσο οι τιμές των φορτίσεων (λ) αλλά και των σφαλμάτων στο σχήμα. Δίπλα ακριβώς από το διάγραμμα διαδρομής βλέπουμε τις αντίστοιχες εντολές σε μορφή κώδικά. Αν τρέξουμε τους 2 παράγοντες ξεχωριστά παρατηρούμε το εξής:




*Εικόνα 45: Διάγραμμα με τον 1ο παράγοντα*




*Εικόνα 46: Διάγραμμα για τον 2ο παράγοντα.*

Όπως μπορούμε να διακρίνουμε από τις εικόνες 43,44 το διαγραμμα διαδρομής μεταξύ του παράγοντα F1 και των μεταβλητών λειτουργεί και εμφανίζει αποτελέσματα, σε αντίθεση με το διάγραμμα διαδρομής του παράγοντα F2 και των μεταβλητών το οποίο δεν λειτουργεί καθώς δεν εμφανίζει αποτελέσματα.

Το διάγραμμα διαδρομής στην εικόνα 44, αφού δεν λειτουργεί μου βγάζει ως πρόβλημα ότι το μοντέλο δεν συγκλίνει και ακόμη ότι βαθμοί ελευθερίας είναι αρνητικοί, οπότε για το λόγω αυτό δημιουργεί και πρόβλημα στο συνολικό μας διάγραμμα διαδρομής. Ακόμη από τον πίνακα συσχετίσης των μεταβλητών που μας εμφανίζει το πρόγραμμα, όπως εμφανίζεται παρακάτω μπορούμε να παρατηρήσουμε ότι η συσχέτιση μεταξύ αλκοόλ και φαινόλων ή φλαβανοειδών είναι αρκετά μικρή και μάλιστα αρνητική, ακόμα όμως και με το αλκοόλ που η συσχέτιση είναι αρκετά μεγάλη, παρατηρούμε πάλι ότι υπάρχει πρόβλημα.

| | Φαινόλες | Φλαβανοειδή | Αλκοόλ | Ένταση χρώματος | Απόχρωση |
|---|---|---|---|---|---|
| Φαινόλες | 0,400 | | | | |
| Φλαβανοειδή | 0,548 | 0,992 | | | |
| Αλκοόλ | 0,146 | 0,181 | 0,656 | | |
| Ένταση χρώματος | -0,083 | -0,403 | 1,033 | 5,369 | |
| Απόχρωση | 0,065 | 0,124 | -0,011 | -0,277 | 0,052 |

*Πίνακας 26: Συσχέτιση μεταξύ των μεταβλητών από LISREL*

Επίσης,το πρόγραμμα μας εμφανίζει σε στήλες τις τιμές τόσο των λανθάνουσων μεταβλητών F1,F2 αλλά και των καταλοίπων για την κάθε μεταβλητή, όπως φαίνονται παρακάτω:

File Edit Data Transformation Statistics Graphs Multilevel SurveyGLIM View Window Help

| | Alkool | Ent_xrwm | Apoxrwsh | F1 | F2 | R_Fai_es | R_Fla_oe | R_Alkool | R_Ent_wm | R_Apo_sh |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 14,230 | 5,640 | 1,040 | 0,734 | -1,564 | 0,113 | 0,297 | 1,446 | 13,788 | -0,008 |
| 2 | 13,200 | 4,380 | 1,050 | 0,616 | -0,281 | 0,027 | 0,115 | 0,239 | 1,696 | 0,017 |
| 3 | 13,160 | 5,680 | 1,030 | 1,284 | -0,161 | -0,189 | -0,072 | 0,183 | 1,977 | -0,086 |
| 4 | 14,370 | 7,800 | 0,860 | 1,684 | -1,635 | 0,642 | -0,223 | 1,596 | 16,547 | -0,306 |
| 5 | 13,240 | 4,320 | 1,040 | 0,542 | -0,337 | 0,218 | 0,120 | 0,287 | 2,108 | 0,016 |
| 6 | 14,200 | 6,750 | 1,050 | 1,137 | -1,469 | 0,362 | 0,224 | 1,403 | 14,097 | -0,048 |
| 7 | 14,390 | 5,250 | 1,020 | 0,064 | -1,801 | 0,180 | 0,427 | 1,639 | 15,395 | 0,055 |
| 8 | 14,060 | 5,050 | 1,060 | 0,046 | -1,380 | 0,290 | 0,435 | 1,251 | 11,645 | 0,098 |
| 9 | 14,830 | 5,200 | 1,080 | 0,364 | -2,371 | 0,315 | 0,587 | 2,157 | 20,159 | 0,078 |
| 10 | 13,860 | 7,220 | 1,010 | 1,038 | -1,007 | 0,126 | 0,083 | 0,999 | 10,666 | -0,076 |
| 11 | 14,100 | 5,750 | 1,250 | 0,584 | -1,393 | 0,345 | 0,708 | 1,292 | 12,448 | 0,221 |
| 12 | 14,120 | 5,000 | 1,170 | -0,337 | -1,467 | 0,100 | 0,738 | 1,323 | 12,325 | 0,255 |
| 13 | 13,750 | 5,600 | 1,150 | 0,230 | -0,949 | 0,189 | 0,501 | 0,881 | 8,555 | 0,165 |
| 14 | 14,750 | 5,400 | 1,250 | 0,855 | -2,250 | 0,346 | 0,806 | 2,061 | 19,338 | 0,187 |
| 15 | 14,380 | 7,500 | 1,200 | 1,039 | -1,672 | 0,445 | 0,572 | 1,611 | 16,557 | 0,114 |
| 16 | 13,630 | 7,300 | 1,280 | 0,150 | -0,719 | 0,483 | 0,731 | 0,730 | 8,308 | 0,305 |
| 17 | 14,300 | 6,200 | 1,070 | 0,746 | -1,630 | 0,106 | 0,365 | 1,525 | 14,906 | 0,021 |
| 18 | 13,860 | 6,600 | 1,130 | 1,050 | -1,035 | 0,089 | 0,321 | 1,003 | 10,277 | 0,043 |
| 19 | 14,190 | 8,700 | 1,230 | 1,391 | -1,365 | 0,253 | 0,510 | 1,379 | 15,170 | 0,100 |
| 20 | 13,640 | 5,100 | 0,960 | 1,052 | -0,815 | -0,162 | -0,051 | 0,753 | 6,923 | -0,128 |
| 21 | 14,060 | 5,650 | 1,090 | 0,797 | -1,341 | 0,278 | 0,345 | 1,245 | 11,917 | 0,034 |
| 22 | 12,930 | 4,500 | 1,030 | 0,308 | 0,070 | -0,044 | 0,073 | -0,079 | -1,156 | 0,035 |
| 23 | 13,710 | 3,800 | 1,110 | 0,470 | -0,973 | 0,067 | 0,381 | 0,845 | 6,960 | 0,095 |
| 24 | 12,850 | 3,930 | 1,090 | 0,130 | 0,146 | 0,124 | 0,211 | -0,170 | -2,367 | 0,117 |
| 25 | 13,500 | 3,520 | 1,120 | 0,167 | -0,717 | 0,153 | 0,414 | 0,599 | 4,517 | 0,142 |
| 26 | 13,050 | 3,580 | 1,130 | 0,357 | -0,126 | 0,149 | 0,294 | 0,068 | -0,418 | 0,129 |
| 27 | 13,390 | 4,800 | 0,920 | 1,105 | -0,502 | -0,041 | -0,194 | 0,460 | 3,980 | -0,174 |
| 28 | 13,300 | 3,950 | 1,020 | -0,050 | -0,441 | 0,142 | 0,211 | 0,361 | 2,617 | 0,069 |
| 29 | 13,870 | 4,500 | 1,250 | 0,202 | -1,155 | 0,554 | 0,739 | 1,030 | 9,192 | 0,268 |
| 30 | 14,020 | 4,700 | 1,040 | -0,121 | -1,346 | 0,431 | 0,422 | 1,206 | 11,009 | 0,098 |
| 31 | 13,730 | 5,700 | 1,190 | 0,746 | -0,911 | 0,306 | 0,475 | 0,856 | 8,328 | 0,141 |
| 32 | 13,580 | 6,900 | 1,090 | 0,969 | -0,658 | 0,044 | 0,192 | 0,671 | 7,398 | 0,013 |
| 33 | 13,680 | 3,840 | 1,230 | -0,048 | -0,941 | 0,161 | 0,709 | 0,810 | 6,725 | 0,279 |
| | 13,760 | 5,400 | 1,250 | -0,043 | -0,976 | 0,689 | 0,754 | 0,895 | 8,577 | 0,299 |

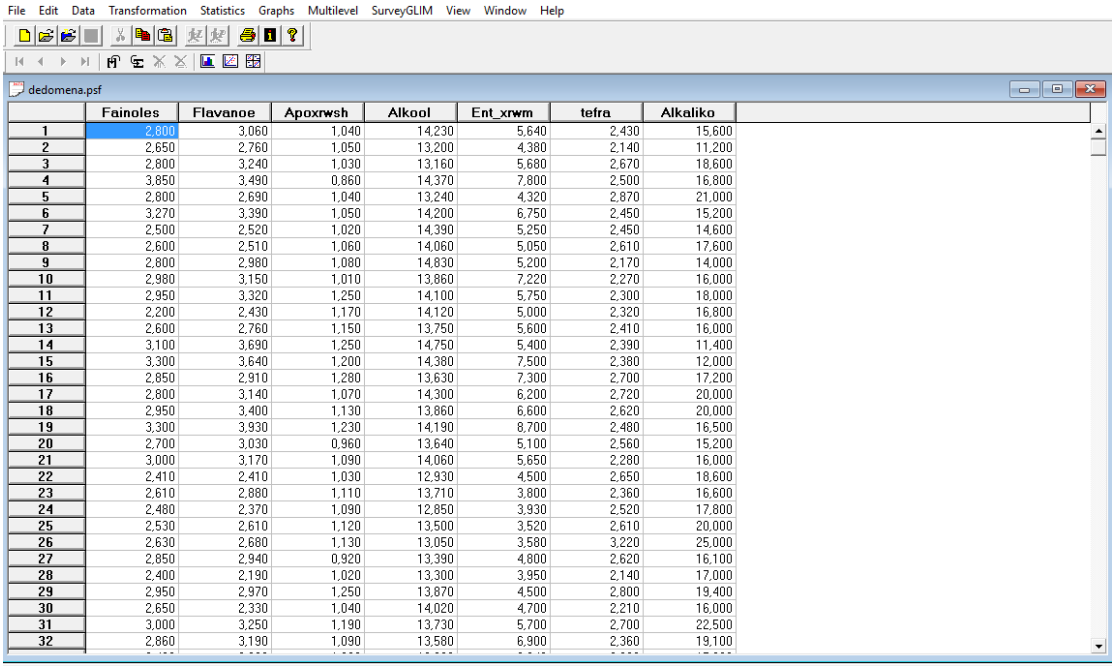
*Εικόνα 47: Τιμές καταλοίπων και παραγόντων*

### *7.5.2: Εφαρμογή με βάση κάποιον άλλον διαχωρισμό ανάμεσα στις μεταβλητές*

Μιας και με την εφαρμογή των παραπάνω διαχωρισμών ανάμεσα στις μεταβλητές δεν μπορούμε να βγάλουμε αρκετά αποτελέσματα, δοκιμάζοντας διάφορους διαχωρισμούς ανάμεσα τους για την δημιουργία του διαγράμματος διαδρομής μου έβγαλε αποτελέσματα για τις παρακάω μεταβλητές:

Φαινόλες, Φλαβαονειδή, Απόχρωση, Αλκοόλ, ένταση χρώματος, τέφρα, αλκαλικότητα τέφρας.

Πρόσθεσα τις δύο μεταβλητές που με βάση την παραγοντική ανάλυση αφαίρεσα τελευταίες από την εφαρμογή και παρατήρησα πως έτσι το πρόγραμμα και δουλευεί αλλά και μας αποδίδει αρκετά αποτελέσματα.

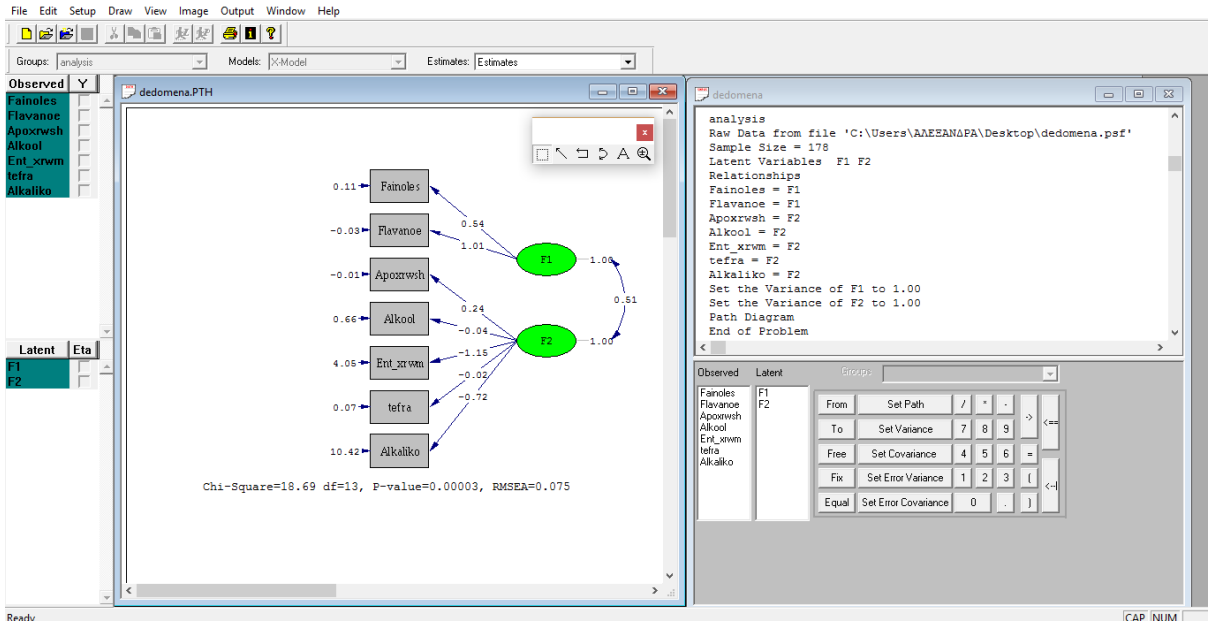

File Edit Data Transformation Statistics Graphs Multilevel SurveyGLIM View Window Help

dedomena.psf

| | Fainoles | Flavanoe | Apoxrwsh | Alkool | Ent_xrwm | tefra | Alkaliko |
|---|---|---|---|---|---|---|---|
| 1 | 2,800 | 3,060 | 1,040 | 14,230 | 5,640 | 2,430 | 15,600 |
| 2 | 2,650 | 2,760 | 1,050 | 13,200 | 4,380 | 2,140 | 11,200 |
| 3 | 2,800 | 3,240 | 1,030 | 13,160 | 5,680 | 2,670 | 18,600 |
| 4 | 3,850 | 3,490 | 0,860 | 14,370 | 7,800 | 2,500 | 16,800 |
| 5 | 2,800 | 2,690 | 1,040 | 13,240 | 4,320 | 2,870 | 21,000 |
| 6 | 3,270 | 3,390 | 1,050 | 14,200 | 6,750 | 2,450 | 15,200 |
| 7 | 2,500 | 2,520 | 1,020 | 14,390 | 5,250 | 2,450 | 14,600 |
| 8 | 2,600 | 2,510 | 1,060 | 14,060 | 5,050 | 2,610 | 17,600 |
| 9 | 2,800 | 2,980 | 1,080 | 14,830 | 5,200 | 2,170 | 14,000 |
| 10 | 2,980 | 3,150 | 1,010 | 13,860 | 7,220 | 2,270 | 16,000 |
| 11 | 2,950 | 3,320 | 1,250 | 14,100 | 5,750 | 2,300 | 18,000 |
| 12 | 2,200 | 2,430 | 1,170 | 14,120 | 5,000 | 2,320 | 16,800 |
| 13 | 2,600 | 2,760 | 1,150 | 13,750 | 5,600 | 2,410 | 16,000 |
| 14 | 3,100 | 3,690 | 1,250 | 14,750 | 5,400 | 2,390 | 11,400 |
| 15 | 3,300 | 3,640 | 1,200 | 14,380 | 7,500 | 2,380 | 12,000 |
| 16 | 2,850 | 2,910 | 1,280 | 13,630 | 7,300 | 2,700 | 17,200 |
| 17 | 2,800 | 3,140 | 1,070 | 14,300 | 6,200 | 2,720 | 20,000 |
| 18 | 2,950 | 3,400 | 1,130 | 13,860 | 6,600 | 2,620 | 20,000 |
| 19 | 3,300 | 3,930 | 1,230 | 14,190 | 8,700 | 2,480 | 16,500 |
| 20 | 2,700 | 3,030 | 0,960 | 13,640 | 5,100 | 2,560 | 15,200 |
| 21 | 3,000 | 3,170 | 1,090 | 14,060 | 5,650 | 2,280 | 16,000 |
| 22 | 2,410 | 2,410 | 1,030 | 12,930 | 4,500 | 2,650 | 18,600 |
| 23 | 2,610 | 2,880 | 1,110 | 13,710 | 3,800 | 2,360 | 16,600 |
| 24 | 2,480 | 2,370 | 1,090 | 12,850 | 3,930 | 2,520 | 17,800 |
| 25 | 2,530 | 2,610 | 1,120 | 13,500 | 3,520 | 2,610 | 20,000 |
| 26 | 2,630 | 2,680 | 1,130 | 13,050 | 3,580 | 3,220 | 25,000 |
| 27 | 2,850 | 2,940 | 0,920 | 13,390 | 4,800 | 2,620 | 16,100 |
| 28 | 2,400 | 2,190 | 1,020 | 13,300 | 3,950 | 2,140 | 17,000 |
| 29 | 2,950 | 2,970 | 1,250 | 13,870 | 4,500 | 2,800 | 19,400 |
| 30 | 2,650 | 2,330 | 1,040 | 14,020 | 4,700 | 2,210 | 16,000 |
| 31 | 3,000 | 3,250 | 1,190 | 13,730 | 5,700 | 2,700 | 22,500 |
| 32 | 2,860 | 3,190 | 1,090 | 13,580 | 6,900 | 2,360 | 19,100 |

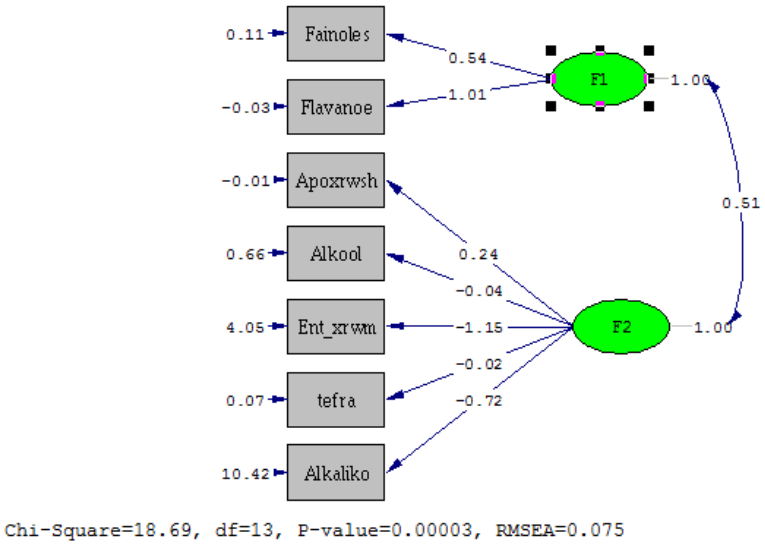
*Εικόνα 48: Εισαγωγή 7 μεταβλητών στο πρόγραμμα*

Το διάγραμμα διαδρομής και οι δύο λανθάνουσες μεταβλητές για τις παραπάνω μεταβλητές, αλλά και ο ορισμός τους σε μορφή κώδικα εμφανίζονται παρακάτω:




*Εικόνα 49: Διάγραμμα διαδρομής*

Στη συνέχεια γίνεται απεικόνιση του διαγράμματος διαδρομής με τις τυποποιημένες τιμές (standardizedsolution), η οποία είναι ως εξής:




*Εικόνα 50: Διάγραμμα διαδρομής (τυποποιημένες τιμές)*

Αυτό που μπορούμε να διακρίνουμε από το παραπάνω διάγραμμα και αξίζει να σχολιάσουμε είναι ότι οι φορτίσεις (λ) μεταξύ των παραγόντων και των μεταβλητών αλλά και τα κατάλοιπα που φαίνονται δίπλα από κάθε μεταβλητή.

| F1 | F2 |
|---|---|
| $X_1 = \lambda_{11} + e_{11}$ | $X_3 = \lambda_{31} + e_{32}$ |
| $X_2 = \lambda_{21} + e_{21}$ | $X_4 = \lambda_{41} + e_{42}$ |
| | $X_5 = \lambda_{51} + e_{52}$ |
| | $X_6 = \lambda_{61} + e_{62}$ |
| | $X_7 = \lambda_{71} + e_{72}$ |

*Πίνακας 27:Σχέσεις μετρικού μοντέλου X*

Με βάση λοιπόν το παραπάνω πίνακα αλλά και την εικόνα μπορούμε να πούμε ότι

$\lambda_{11} = 0{,}54$ και $e_{11} = 0{,}11$

$\lambda_{21} = 1{,}01$ και $e_{21} = -0{,}03$

$\lambda_{32} = 0{,}24$ και $e_{32} = -0{,}01$

$\lambda_{42} = -0{,}04$ και $e_{42} = 0{,}66$

$\lambda_{52} = -1{,}15$ και $e_{52} = 4{,}05$

$\lambda_{62} = -0{,}02$ και $e_{62} = 0{,}07$

$\lambda_{72} = -0{,}72$ και $e_{72} = 10{,}42$

Οι τυποποιημένες αυτές φορτώσεις αντιπροσωπεύουν τη συσχέτιση μεταξύ κάθε παρατηρηθείσας μεταβλητής με τον αντίστοιχο παράγοντα.

Από την παραπάνω εικόνα παρατηρούμε ότι το στατιστικό $\chi^2$ με τιμή 18,69 και βαθμούς ελευθερίας 13 είναι στατιστικά σημαντικό, καθώς η τιμή p-value του ελέγχου είναι 0,00003<0,05. Από τον άλλη όμως ο δείκτης RMSEA=0.07 (τετραγωνική ρίζα του λάθους της εκτίμησης), ο δείκτης αυτός ουσιαστικά απαντά στο ερώτημα πόσο καλά θα προσαρμοζόταν ένα μοντέλο σε σχέση με ένα πρότυπο μοντέλο με άγνωστες αλλά ευνοικές τιμές. Είναι ένας δείκτης ο οποίος εκφράζει την απόκλιση ανά βαθμό ελευθερίας.Αν 0,05<RMSEA<0.08 τότε το μοντέλο έχει μέτρια προσαρμογή, ενώ αν είναι >0,1 τότε το μοντέλο έχει κακή προσαρμογή.

Ακόμη κάτι άλλο που μας εμφανίζει και μπορούμε να από κει να αποδώσουμε μία καλή ή κακή προσαρμογή του μοντέλου είναι οι τιμές των $R^2$για κάθε παρατηρηθείσα μεταβλητή με την αντίστοιχη λανθανουσα.Έχουμε:

| Μεταβλητές | Τιμή $R^2$ |
|---|---|
| Φαινόλες | 0,73 |
| Φλαβανοειδή | 1,03 |
| Απόχρωση | 1,14 |
| Αλκοόλ | 0,0019 |
| Ένταση χρώματος | 0,25 |
| Τέφρα | 0,0037 |
| Αλακαλικότητα | 0,047 |

*Πίνκας 28: Τιμές των$R^2$*

Αυτό που μπορούμε να σχολιάσουμε είναι πως ενώ όλες οι μεταβλητές έχουν αρκετά μεγάλες τιμές για το $R^2$, το αλκοόλ και η τέφρα έχουν σχετικά χαμηλές, κάτι που σημαίνει ότι δεν ταιριάζουν καλά με τον μοντέλο, κάτι που σημαίνει ότι θα μπορούσαν να μην χρησιμοποιηθούν στην ανάλυση μας, αλλά παρόλα αυτά δεν δημιουργούν και κάποιο πρόβλημα σε αυτή. Η περίπτωση που το αλκοόλ και η τέφρα δεν ταιριάζουν καλά με το μοντέλο μας μπορεί να είναι η περίπτωση που αυτές οι δύο μεταβλητές δεν ταιριάζουν με τον F2 παράγοντα.

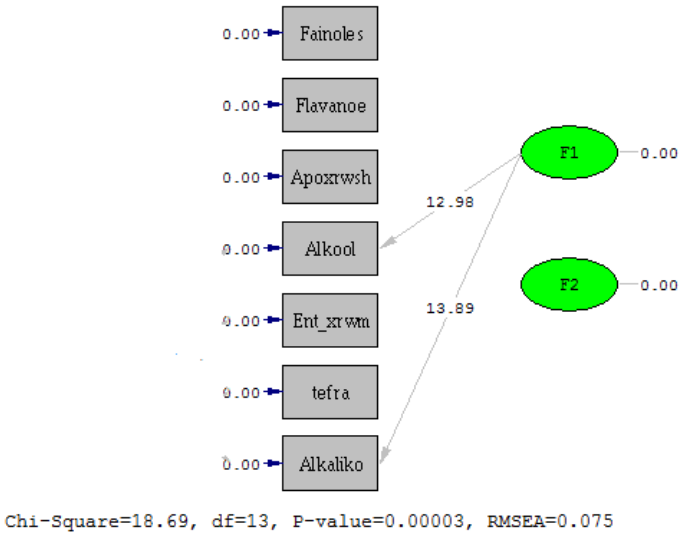
Ακόμη για να κρίνουμε κατά το πόσο το μοντέλο μας είναι αξιόλογο θα πρέπει να αναφέρουμε του δείκτες προσαρμογής οι οποίοι μας το δείχνουν. Η ανάλυση του παραπάνω προγράμματος μας έδειξε τις τιμές των παρακάτων δεικτών:

| Δείκτες προσαρμογής | Τιμές |
|---|---|
| $\chi^2$ | 18,69 |
| RMSEA | 0.007 |
| GFI | 0.55 |
| NFI | 0.54 |
| IFI | 0.56 |
| CFI | 0.55 |
| PNFI | 0.34 |

*Πίνακας 29: Δείκτες προσαρμογής*

Από τον παραπάνω πίνακα που αναφέρεται στους δείκτες προσαρμογής, αυτό που αξίζει να σχολιαστεί είναι πως ναι μεν από την μία οι τιμές είναι μέσα επιτρεπόμενα (0-1) για όλους τους δείκτες αλλά από την άλλη είναι σχετικά χαμηλές. Κάτι το οποίο με άλλα λόγια μπορεί να εκφραστεί ότι το μοντέλο μας δεν έχει πάρα πολύ καλή προσαρμογή.

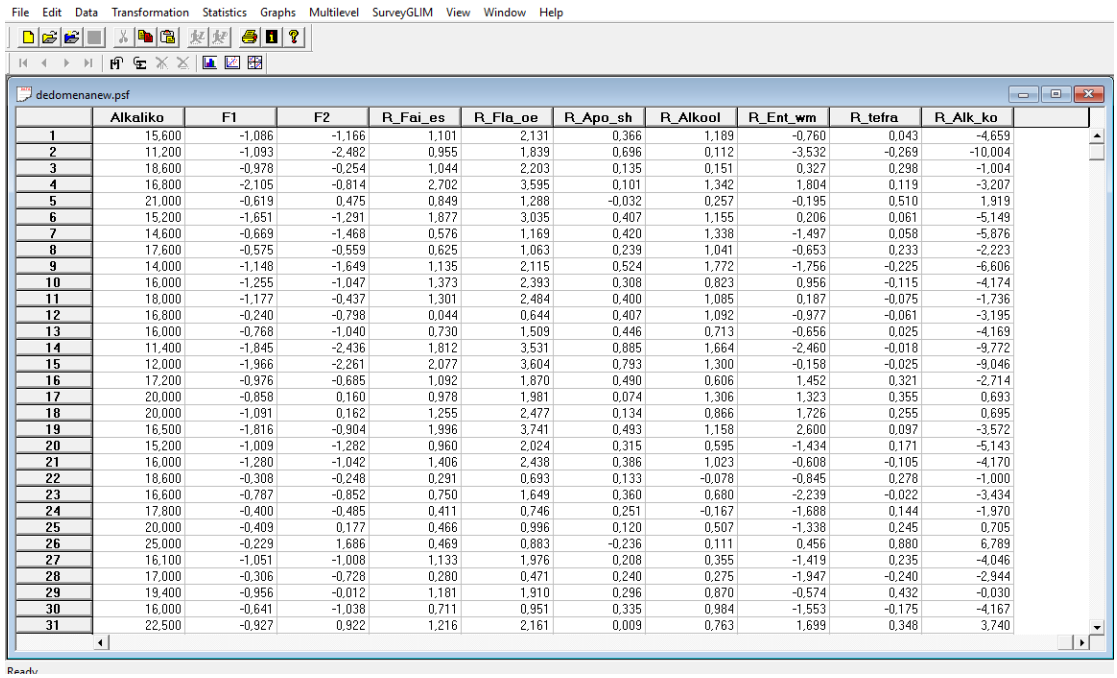
Περά από το παραπάνω που αναφέραμε το πρόγραμμα μας αναφέρει και δείκτες τροποποίσησης (modificationindices) και εμφανίζεται ένα καινούργιο διάγραμμα διαδρομής. Οι δείτκες αυτοί ουσιαστικά προσφέρουν προτάσεις για την βελτίωση της συνολικής προσαρμογής του μοντέλου.




*Εικόνα 51: Διάγραμμα διαδρομής δεικτών τροποποίησης*

Από την παραπάνω εικόνα παρατηρούμε ότι ο παράγοντας F1 δείχνει μία διαδρομή προς το αλκοόλ και την αλκαλικότητα και μάλιστα με πολύ υψηλές φορτίσεις μεταξύ τους, κάτι το οποίο σημαίνει ότι υπάρχει κάτι κοινό μεταξύ του παράγοντα F1 και των δύο αυτών παρατηρηθείσων τιμών. Αυτό μπορεί να επαληθευθεί τόσο από το διάγραμαμ διαδρομής των τυποποιημένων φοτίσεων (εικόνα 47) που παρατηρούμε ότι οι φορτίσεις από τον παράγοντα F1 προς τις δύο μεταβλητές είναι σχετικά μικρές και μάλιστα αρνητικές αλλά και από τον πίνακα 30 που εμφανίζει τις τιμές των $R^2$όπου και εκείνες είναι πολύ χαμηλές.

Κάτι τελευταίο το οποίο μπορούμε να αναφέρουμε και να σχολιάσουμε σχετικά με τα αποτελέσματα που μας δίνει το πρόγραμμα , είναι πως μας εμφανίζει και τις στήλες με τις τιμές τόσο των παραγόντων F1 και F2 αλλά και τις τιμές των καταλοίπων για κάθε μία μεταβλητή.




| | Alkaliko | F1 | F2 | R_Fai_es | R_Fla_oe | R_Apo_sh | R_Alkool | R_Ent_wm | R_tefra | R_Alk_ko |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 15,600 | -1,086 | -1,166 | 1,101 | 2,131 | 0,366 | 1,189 | -0,760 | 0,043 | -4,659 |
| 2 | 11,200 | -1,093 | -2,482 | 0,955 | 1,839 | 0,696 | 0,112 | -3,532 | -0,269 | -10,004 |
| 3 | 18,600 | -0,978 | -0,254 | 1,044 | 2,203 | 0,135 | 0,151 | 0,327 | 0,298 | -1,004 |
| 4 | 16,800 | -2,105 | -0,814 | 2,702 | 3,595 | 0,101 | 1,342 | 1,804 | 0,119 | -3,207 |
| 5 | 21,000 | -0,619 | 0,475 | 0,849 | 1,288 | -0,032 | 0,257 | -0,195 | 0,510 | 1,919 |
| 6 | 15,200 | -1,651 | -1,291 | 1,877 | 3,035 | 0,407 | 1,155 | 0,206 | 0,061 | -5,149 |
| 7 | 14,600 | -0,669 | -1,468 | 0,576 | 1,169 | 0,420 | 1,338 | -1,497 | 0,058 | -5,876 |
| 8 | 17,600 | -0,575 | -0,559 | 0,625 | 1,063 | 0,239 | 1,041 | -0,653 | 0,233 | -2,223 |
| 9 | 14,000 | -1,148 | -1,649 | 1,135 | 2,115 | 0,524 | 1,772 | -1,756 | -0,225 | -6,606 |
| 10 | 16,000 | -1,255 | -1,047 | 1,373 | 2,393 | 0,308 | 0,823 | 0,956 | -0,115 | -4,174 |
| 11 | 18,000 | -1,177 | -0,437 | 1,301 | 2,484 | 0,400 | 1,085 | 0,187 | -0,075 | -1,736 |
| 12 | 16,800 | -0,240 | -0,798 | 0,044 | 0,644 | 0,407 | 1,092 | -0,977 | -0,061 | -3,195 |
| 13 | 16,000 | -0,768 | -1,040 | 0,730 | 1,509 | 0,446 | 0,713 | -0,656 | 0,025 | -4,169 |
| 14 | 11,400 | -1,845 | -2,436 | 1,812 | 3,531 | 0,885 | 1,664 | -2,460 | -0,018 | -9,772 |
| 15 | 12,000 | -1,966 | -2,261 | 2,077 | 3,604 | 0,793 | 1,300 | -0,158 | -0,025 | -9,046 |
| 16 | 17,200 | -0,976 | -0,685 | 1,092 | 1,870 | 0,490 | 0,606 | 1,452 | 0,321 | -2,714 |
| 17 | 20,000 | -0,858 | 0,160 | 0,978 | 1,981 | 0,074 | 1,306 | 1,323 | 0,355 | 0,693 |
| 18 | 20,000 | -1,091 | 0,162 | 1,255 | 2,477 | 0,134 | 0,866 | 1,726 | 0,255 | 0,695 |
| 19 | 16,500 | -1,816 | -0,904 | 1,996 | 3,741 | 0,493 | 1,158 | 2,600 | 0,097 | -3,572 |
| 20 | 15,200 | -1,009 | -1,282 | 0,960 | 2,024 | 0,315 | 0,595 | -1,434 | 0,171 | -5,143 |
| 21 | 16,000 | -1,280 | -1,042 | 1,406 | 2,438 | 0,386 | 1,023 | -0,608 | -0,105 | -4,170 |
| 22 | 18,600 | -0,308 | -0,248 | 0,291 | 0,693 | 0,133 | -0,078 | -0,845 | 0,278 | -1,000 |
| 23 | 16,600 | -0,787 | -0,852 | 0,750 | 1,649 | 0,360 | 0,680 | -2,239 | -0,022 | -3,434 |
| 24 | 17,800 | -0,400 | -0,485 | 0,411 | 0,746 | 0,251 | -0,167 | -1,688 | 0,144 | -1,970 |
| 25 | 20,000 | -0,409 | 0,177 | 0,466 | 0,996 | 0,120 | 0,507 | -1,338 | 0,245 | 0,705 |
| 26 | 25,000 | -0,229 | 1,686 | 0,469 | 0,883 | -0,236 | 0,111 | 0,456 | 0,880 | 6,789 |
| 27 | 16,100 | -1,051 | -1,008 | 1,133 | 1,976 | 0,208 | 0,355 | -1,419 | 0,235 | -4,046 |
| 28 | 17,000 | -0,306 | -0,728 | 0,280 | 0,471 | 0,240 | 0,275 | -1,947 | -0,240 | -2,944 |
| 29 | 19,400 | -0,956 | -0,012 | 1,181 | 1,910 | 0,296 | 0,870 | -0,574 | 0,432 | -0,030 |
| 30 | 16,000 | -0,641 | -1,038 | 0,711 | 0,951 | 0,335 | 0,984 | -1,553 | -0,175 | -4,167 |
| 31 | 22,500 | -0,927 | 0,922 | 1,216 | 2,161 | 0,009 | 0,763 | 1,699 | 0,348 | 3,740 |

*Εικόνα 52: Τιμές καταλοίπων και των παραγόντων*

## *7.6: Συμπεράσματα – Αποτελέσματα της εφαρμογής*

Αρχικά με τον μέθοδος διαχωρισμού της παραγοντικής ανάλυσης αναφέραμε και παραπάνω ότι ο διαχωρισμός που έγινε ήταν ο εξής:

| Παράγοντες | Μεταβλητές |
|---|---|
| 1ος παράγοντας | Φαινόλες |
| | Φλαβανοειδή |
| | Απόχρωση |
| 2ος παράγοντας | Αλκοόλ |
| | Ένταση χρώματος |

Πίνακας 30: Πίνακας παραγόντων

Η εφαρμογή όμως αυτού του διαχωρισμού στο πρόγραμμα παρατηρήσαμε ότι δεν λειτούργει, καθώς δεν εμφανίζει ούτε τις φορτίσεις ούτε τα σφάλματα στο διάγραμμα διαδρομής. Στον έλεγχο όμως ξεχωριστά των 2 παραγόντων παρατηρούμε ότι ο 1ος παράγοντας σε συνδυασμό με τις μεταβλητές λειτουργεί κανόνικα, ενώ ο 2ος παράγοντας σε συνδυασμό με τις μεταβλητές δεν λειτούργει, εμφανίζοντας ως πρόβλημα ότι έχει αρνητικούς βαθμούς ελευθερίας αλλά και χαμηλή συσχετίση μεταξύ των δύο παρατηρήσιμων μεταβλητών (αλκοόλ, ένταση χρώματος).

Έτσι για την διεξαγωγή ενός πιο συγκεντρωτικού συμπεράσματος, σε δοκιμές μας στο πρόγραμμα LISREL, παρατηρούμε ότι κάνει τον εξής διαχωρισμό:

| Παράγοντες | Μεταβλητές |
|---|---|
| 1ος παράγοντας | Φαινόλες |
| | Φλαβανοειδή |
| 2ος παράγοντας | Αλκοολ |
| | Ένταση χρώματος |
| | Απόχρωση |
| | Τέφρα |
| | Αλκαλικότητα τέφρας |

*Πίνακας 31: Πίνακας μεταβλητών*

Με τον διαχωρισμό παρατηρούμε ότι παίρνουμε αποτελέσματα και μάλιστα η εκτίμηση προσαρμογής του μοντέλου μέσω των δεικτών προσαρμογής που έχω αναφέρει και πιο πάνω, μας δείχνει ότι δεν υπάρχει τέλεια προσαρμογή του μοντέλου, δηλαδή μεταξύ του διαχωρισμού των μεταβλητών που έγινε, αλλά η προσαρμογή είναι αρκετά καλή.

# *Κεφάλαιο 8: Βιβλιογραφία*

## *8.1: Ελληνική Βιβλιογραφία*


- Γ.Δονάτος-Μ.Αδάμ (Αθήνα 2008):''Γραμμική Άλγεβρα, Θεωρία και εφαρμογές''
- Ιωάννης μαρούλας (2005):''Γραμμική Άλγεβρα''
- Π. Νικήτας (Θεσσαλονική 2010): ''Σημειώσεις γραμμικής άλγεβρας (αποσπάσαμε Ιδιοτιμές και ιδιοδιανύσματα)
- Καρλής (2005): Πολυμεταβλητή Ανάλυση
- Παπαγεωργίου Α (2009-10):''Παραγοντική Ανάλυση και ανάλυση σε κύριες συνιστώσες''
- Πλαστίδου Μ. (2001): ''Η επιβεβαιωτική ανάλυση παραγόντων στην ψυχολογική έρευνα.
- Δρ. Βασίλης και Π. Αγγελίδης: ''Ανάλυση δεδομένων – Παραγοντική Ανάλυση'' Διαθέσιμο στον ιστότοπο https://eclass.duth.gr/modules/document/file.php/TME179/%CE%A0%CE%B1%CF%81%CE%BF%CF%85%CF%83%CE%B9%CE%AC%CF%83%CE%B5%CE%B9%CF%82/lecture7_paragontiki.pdf
- Κωνσταντινίδης Θ. (Θεσσαλονίκη 2009): ''Η μέθοδος της πολυδιάστατης κλιμακοποίησης – εφαρμογές'' Πανεπιστήμι Θεσσαλονικής , Μεταπτυχιακό Τμήμα Ειδίκευση Στατιστική και Επιχειρησιακή έρευνα''
- Εμμανουηλίδης Ι. (2012-13): ''Εφαρμοσμένη στατιστική έρευνα''
- Καρλής (2005): ''Πολυμεταβλητή Ανάλυση''
- Ντζούφρας Ι. (2008): '' Ανάλυση πολυμεταβλητών δεδομένων'' Σημειώσεις τμήματος διοίκησης επιχειρήσεων, πανεπιστήμιο Αιγαίου
- Ντύκεν Μ.Ν. & Τσιάπα Μ. : ''Το γενικευμένο γραμμικό μοντέλο (Α)'' Πανεπιστήμιο Θεσσαλίας. Διαθέσιμο στον ιστότοπο: http://eclass.uth.gr/eclass/modules/document/file.php/MHXA197/LECTURE%204/LECTURE_04.pdf
- Νταιλανάς Χρίστος (Αθήνα 2012): ΄Γενικευμένα γραμμικα μοντέλα με χρήση του στατιστικού πακέτου R
- Αθανάσιος Τατσιος (Αθήνα 2009): ΄Γενικευμένα γραμμικά μοντέλα με χρήση του στατιστικού πακέτου R
- Μάριος Κούτρας & Χ. Ευαγγελάρας (Αθήνα 2010): Ανάλυση παλινδρόμησης (Θεωρία και Εφαρμογές) εκδόσεις: Αθ. Σταμούλης
- Μιχάλης Θαλασσιάς (Αθήνα 2011): Γενικευμένα γραμμικά μοντέλα και παραγοντικοί σχεδιασμοί
- Γ. Τζαβέλας:Γενικευμένα γραμμικά μοντέλα (ΜΕΡΟΣ Α) , Εισαγωγή λογιστική παλινδρόμηση Διαθέσιμο στον ιστότοπο: http://www.unipi.gr/faculty/tzafor/glm_notes_part_A.pdf

- Φουσκάκης Δ. : Εκτιμητική Στατιστική, Διαθέσιμο στον ιστότοπο: http://www.math.ntua.gr/~fouskakis/estimation.pdf

- Ε. Ιωαννίδης (Φεβρουάριος 2015): Γενικευμένα γραμμικά μοντέλα (Περιληπτική απόδοση στα ελληνικά των σημειώσεων J. Foster και Π. Δελλαπόρτα)
- Χαράλαμπος Δαμιανού, Μάρκος Κούτρας: Εισαγωγή στην Στατιστική (Μέρος 1) , Εκδόσεις Συμμετρία- Αθήνα 2013
- Φωκιανός Κ. Και Χαραλάμπους Χαράλαμπος: Εισαγωγή στην R πρόχειρες σημειώσεις, τμήμα μαθηματικών και στατιστικής πανεπιστήμιου Κύπρου (Εκδόσεις 2008 και 2010) Διαθέσιμο στον ιστότοπο: ftp://cran.r-project.org/pub/R/doc/contrib/mainfokianoscharalambous.pdf
- Ζερβαλάκη Θεονύμφη (2007):΄΄Μελέτη ικανοποίησης αφοσίωσης πελατών βασισμένη σε ένα μοντέλο δομικών εξισώσεων:Εμπειρική εφαρμογή σε μία αλυσίδα Supermarkets - Πολυτεχνείο Κρήτης
- Μπελδέκος Παναγιώτης (2008): ΄΄Η επίδραση της συναισθηματικής Νοημοσύνης και της αντίληψης υποστηρικτικού περιβάλλοντος στην επιχειρηματική συμπεριφορά εργαζομένων στο εσωτερικό οργανισμό΄΄ – Πολυτεχνείο Κρήτης
- Κωστάκη Σταυρούλα (Χανιά 2008): ΄΄Επίδραση της αντίληψης του υφιστάμενου για τη συναισθηματική νοημοσύνη των προισταμένων στην ικανοποίηση από την εργασία και στην ανάπτυξη της επιχειρηματικής συμπεριφοράς του υφιστάμενου
- Παπαγεωργίου Δ. (Χανιά 2014): ΄΄ Οργανωσιακή Αλλαγή: Οι παράγοντες που επηρεάζουν την εισαγωγή της στους οργανισμούς΄΄
- Σπυρίδωνος Βερονίκη (Κέρκυρα 2010): ΄΄Μελέτη παραμέτρων σχεδίασης σχημάτων πανταχού παρόντος υπολογιστή για την ανάπτυξη και αξιολόγηση υπηρεσιών σε περιβάλλοντα υβριδικών βιβλιοθηκών΄΄
- Για διάγραμμα ασσυμετρία στον παρακάτω ιστότοπο: http://slideplayer.gr/slide/2645550/
- http://androulakis.bma.upatras.gr/mediawiki/index.php/%CE%A3%CF%84%CE%BF%CE%B9%CF%87%CE%B5%CE%B9%CF%8E%CE%B4%CE%B5%CE%B9%CF%82_%CE%AD%CE%BD%CE%BD%CE%BF%CE%B9%CE%B5%CF%82_%CF%84%CE%B7%CF%82_%CF%83%CF%84%CE%B1%CF%84%CE%B9%CF%83%CF%84%CE%B9%CE%BA%CE%AE%CF%82
- Παναγιώτα Καπουλέα (2014-2015): ΄΄Εφαρμογή Θεωρίας Παραγνίβν στο προσδιορισμό της Επικυνδινότητας κατά το δαινόμενο της Προσπέρασης σε Αστικές Αρτηρίες (βλέπε σελ. 27 σχετικά με τους δείκτες προσαρμογής)
- Σίλια Βιτωράτου – Ευγενία Τσομπανάκη: Μοντέλα λανθάνουσων μεταβλητών
- Ευστάθιος Δημητριάδης (Εκδόσεις ΚΡΙΤΙΚΗ 2012): Στατιστική Επιχειρήσεων σε SPSS και LISREL
- Ρεκούτη Αγγελική (Πάτρα 2001): ‘’ Εφαρμογή τηα παραγοντικής στατιστικής ανάλυσης για την ανίχνευση &περιγραφή της κατανάλωσης αλκοολούχων ποτών του ελληνικού πληθυσμού

- Α.Δ.ΜΠΑΤΣΙΔΗΣ: ‘’Παραγοντική ανάλυση και SPSS (Πρόχειρες Σημειώσεις)’’
- Δημήτρης Φουσκάκης: ‘’Ανάλυση Δεδομένων με χρήση του Στατιστικού Πακέτου R’’, Σχολή εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών, Εθνικό Μετσόβιο Πολυτεχνείο.
- Από εργαστήριο Ανάλυσης και τεχνολογία οίνου, Διδάσκοντες: Κ.Ρηγανάκος, Κ.Ακρίδα (13 Οκτωμβρίου 2013), από τον παρακάτω ιστότοπο: http://ecourse.uoi.gr/course/view.php?id=867
- http://www.agrotypos.gr/images/stories/file/kali/patentkali_ampeli.pdf
- Αικατερίνη Σπανού: ‘’Αιτιώδης συμπερασματολογία και εφαρμογές στη βιοστατιστική΄΄, Εθνικό Μετσόβιο Πανεπιστήμιο (Δ.Π.Μ.Σ: Εφαρμοσμένες μαθηματικές επιστήμες, Πιθανότητες – Στατιστική), Ιούλιος 2014.


## *8.2:Αγγλική Βιβλιογραφία*


- Για διάγραμμα μέγιστης πιθανοφάνειας στον παρακάτω ιστότοπο: http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_3.html
- Για διάγραμμα κύρτωσης στον παρακάτω ιστότοπο: http://slideplayer.com/slide/4666344/
- Richard A.Johson & Dean W.Wichern (2007): ‘’Applied multivariate statistical analysis’’
- Alan Agresti, WILEY (2015): ‘’Foundations of Linear and Generalized Linear Models’’
- Tenko Ravkov – George A. Marcoulides (2000): ‘’ A first course in structural equation modeling’’
- Path Analysis and example, στοπαρακάτωιστότοπο: http://faculty.cas.usf.edu/mbrannick/regression/Pathan.html
- Path analysis pdf.
- Michael Friendly ( 2008): Exploratory and Confirmatory Factor Analysis (Part 3: CFA&SEM models)
- An introduction to Pth Analysis, Elazar J. Pendhazur (Multiple Regression in Behavioral Research, 2nd edition, Holt, Rinehard and Winston, 1982)
- Richard Buxton (2008): ‘’Statistics: Multilivel modelling’’
- FactorAnalysisUsingSPSS (2005). Δίνεται στον παρακάτων ιστότοπο: http://www.statisticshell.com/docs/factor.pdf
- Sabie Landau and Brian S. Everitt (2003): ‘’ A Hanbook of Statistical Analyses using SPSS
- 5. Structural Equation models, Διαθέσιμοστονπαρακάτωνιστότοπο: http://www.ssicentral.com/lisrel/techdocs/compsem.pdf
- Using LISREL: SIMPLIS and LISREL Language, δίνεταιστονπαρακάτωιστότοπο: file:///C:/Users/%CE%91%CE%9B%CE%95%CE%9E%CE%91%CE%9D%CE%94%CE%A1%CE%91/Downloads/foli4.pdf

- Rendall E. Schumacker – Ricahrd G. Lomax (2010): ‘’A Beginner’s Guide to Structural Equation Modeling, Third Edition
- http://home.apu.edu/~bsimmerok/WebTMIPs/Session6/TSes6.html (βλέπε σχετικά με τα διαστήματα εμπιστοσύνης και το διάγραμμα στο κεφάλαιο 6)
- Gerhard Mels: ‘’LISREL for Windows Getting Started Guide’’ από SSI(Scientific Software International) 2006
- Structural Equation Modeling, 2010, Copyright: Taylor & Francis Group, LIC, Testing Inequality Constrained Hypotheses in SEM Models
- Jeremy J. Albright, 2006-2008: Confirmatory factor analysis usinh AMOS, LISREL AND MPLUS
- Mels, G. 2006: LISREL for Windows: Getting Started Guide, Lincolnwood, IL: Scientific Software International, Inc
  Copyright 2001-2010, Scientific Software International, Inc.
- LISREL for Windows: PRELIS User’s Guide, SSI:Scientific Software International


## *8.3: Πίνακες*

**$a$-ποσοστιαία σημεία του στατιστικού $D_n$ του ελέγχου Kolmogorov**

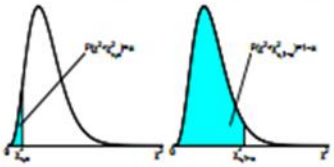
| $n$ | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 |
|---|---|---|---|---|---|
| 1 | .900 | .950 | .975 | .990 | . 995 |
| 2 | .684 | .776 | .842 | .900 | . 929 |
| 3 | .565 | .636 | .708 | .785 | . 829 |
| 4 | .493 | .565 | .624 | .689 | . 734 |
| 5 | .447 | .509 | .563 | .627 | . 669 |
| 6 | .410 | .468 | .519 | .577 | . 617 |
| 7 | .381 | .436 | .483 | .538 | . 576 |
| 8 | .358 | .410 | .454 | .507 | . 542 |
| 9 | .339 | .387 | .430 | .480 | . 513 |
| 10 | .323 | .369 | .409 | .457 | . 489 |
| 11 | .308 | .352 | .391 | .437 | . 468 |
| 12 | .296 | .338 | .375 | .419 | . 449 |
| 13 | .285 | .325 | .361 | .404 | . 432 |
| 14 | .275 | .314 | .349 | .390 | . 418 |
| 15 | .266 | .304 | .338 | .377 | . 404 |
| 16 | .258 | .295 | .327 | .366 | . 392 |
| 17 | .250 | .286 | .318 | .355 | . 381 |
| 18 | .244 | .279 | .309 | .346 | . 371 |
| 19 | .237 | .271 | .301 | .337 | . 361 |
| 20 | .232 | .265 | .294 | .329 | . 352 |

| $n$ | 0.20 | 0.10 | 0.05 | 0.02 | 0.01 |
|---|---|---|---|---|---|
| 21 | .226 | .259 | .287 | .321 | .344 |
| 22 | .221 | .253 | .281 | .314 | .337 |
| 23 | .216 | .247 | .275 | .307 | .330 |
| 24 | .212 | .242 | .269 | .301 | .323 |
| 25 | .208 | .238 | .264 | .295 | .317 |
| 26 | .204 | .233 | .259 | .290 | .311 |
| 27 | .200 | .229 | .254 | .284 | .305 |
| 28 | .197 | .225 | .250 | .279 | .300 |
| 29 | .193 | .221 | .246 | .275 | .295 |
| 30 | .190 | .218 | .242 | .270 | .290 |
| 31 | .187 | .214 | .238 | .266 | .285 |
| 32 | .184 | .211 | .234 | .262 | .281 |
| 33 | .182 | .208 | .231 | .258 | .277 |
| 34 | .179 | .205 | .227 | .254 | .273 |
| 35 | .177 | .202 | .224 | .251 | .269 |
| 36 | .174 | .199 | .221 | .247 | .265 |
| 37 | .172 | .196 | .218 | .244 | .262 |
| 38 | .170 | .194 | .215 | .241 | .258 |
| 39 | .168 | .191 | .213 | .238 | .255 |
| 40 | .165 | .189 | .210 | .235 | .252 |
| $n>40$ | $\frac{1.07}{\sqrt{n}}$ | $\frac{1.22}{\sqrt{n}}$ | $\frac{1.36}{\sqrt{n}}$ | $\frac{1.52}{\sqrt{n}}$ | $\frac{1.63}{\sqrt{n}}$ |

## Στατιστικός Πίνακας Κατανομής $\chi^2$

$\nu$: βαθμοί ελευθερίας

$\alpha$ ή $1-\alpha$: τιμή αθροιστικής συνάρτησης



Παράδειγμα:

$\nu = 10, \alpha = 0.05 \implies \chi^2_{10,0.05} = 3.94$, $\chi^2_{10,0.95} = 18.31$

$\nu = 18, \alpha = 0.025 \implies \chi^2_{18,0.025} = 8.23$, $\chi^2_{18,0.975} = 31.53$

| | $\alpha$ | | | | | $1-\alpha$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\nu$ | 0.001 | 0.005 | 0.010 | 0.025 | 0.050 | 0.950 | 0.975 | 0.990 | 0.995 | 0.999 |
| 1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 3.84 | 5.02 | 6.63 | 7.88 | 10.83 |
| 2 | 0.00 | 0.01 | 0.02 | 0.05 | 0.10 | 5.99 | 7.38 | 9.21 | 10.60 | 13.82 |
| 3 | 0.02 | 0.07 | 0.11 | 0.22 | 0.35 | 7.81 | 9.35 | 11.34 | 12.84 | 16.27 |
| 4 | 0.09 | 0.21 | 0.30 | 0.48 | 0.71 | 9.49 | 11.14 | 13.28 | 14.86 | 18.47 |
| 5 | 0.21 | 0.41 | 0.55 | 0.83 | 1.15 | 11.07 | 12.83 | 15.09 | 16.75 | 20.52 |
| 6 | 0.38 | 0.68 | 0.87 | 1.24 | 1.64 | 12.59 | 14.45 | 16.81 | 18.55 | 22.46 |
| 7 | 0.60 | 0.99 | 1.24 | 1.69 | 2.17 | 14.07 | 16.01 | 18.48 | 20.28 | 24.32 |
| 8 | 0.86 | 1.34 | 1.65 | 2.18 | 2.73 | 15.51 | 17.53 | 20.09 | 21.95 | 26.12 |
| 9 | 1.15 | 1.73 | 2.09 | 2.70 | 3.33 | 16.92 | 19.02 | 21.67 | 23.59 | 27.88 |
| 10 | 1.48 | 2.16 | 2.56 | 3.25 | 3.94 | 18.31 | 20.48 | 23.21 | 25.19 | 29.59 |
| 11 | 1.83 | 2.60 | 3.05 | 3.82 | 4.57 | 19.68 | 21.92 | 24.72 | 26.76 | 31.26 |
| 12 | 2.21 | 3.07 | 3.57 | 4.40 | 5.23 | 21.03 | 23.34 | 26.22 | 28.30 | 32.91 |
| 13 | 2.62 | 3.57 | 4.11 | 5.01 | 5.89 | 22.36 | 24.74 | 27.69 | 29.82 | 34.53 |
| 14 | 3.04 | 4.07 | 4.66 | 5.63 | 6.57 | 23.68 | 26.12 | 29.14 | 31.32 | 36.12 |
| 15 | 3.48 | 4.60 | 5.23 | 6.26 | 7.26 | 25.00 | 27.49 | 30.58 | 32.80 | 37.70 |
| 16 | 3.94 | 5.14 | 5.81 | 6.91 | 7.96 | 26.30 | 28.85 | 32.00 | 34.27 | 39.25 |
| 17 | 4.42 | 5.70 | 6.41 | 7.56 | 8.67 | 27.59 | 30.19 | 33.41 | 35.72 | 40.79 |
| 18 | 4.90 | 6.26 | 7.01 | 8.23 | 9.39 | 28.87 | 31.53 | 34.81 | 37.16 | 42.31 |
| 19 | 5.41 | 6.84 | 7.63 | 8.91 | 10.12 | 30.14 | 32.85 | 36.19 | 38.58 | 43.82 |
| 20 | 5.92 | 7.43 | 8.26 | 9.59 | 10.85 | 31.41 | 34.17 | 37.57 | 40.00 | 45.31 |
| 21 | 6.45 | 8.03 | 8.90 | 10.28 | 11.59 | 32.67 | 35.48 | 38.93 | 41.40 | 46.80 |
| 22 | 6.98 | 8.64 | 9.54 | 10.98 | 12.34 | 33.92 | 36.78 | 40.29 | 42.80 | 48.27 |
| 23 | 7.53 | 9.26 | 10.20 | 11.69 | 13.09 | 35.17 | 38.08 | 41.64 | 44.18 | 49.73 |
| 24 | 8.08 | 9.89 | 10.86 | 12.40 | 13.85 | 36.42 | 39.36 | 42.98 | 45.56 | 51.18 |
| 25 | 8.65 | 10.52 | 11.52 | 13.12 | 14.61 | 37.65 | 40.65 | 44.31 | 46.93 | 52.62 |
| 26 | 9.22 | 11.16 | 12.20 | 13.84 | 15.38 | 38.89 | 41.92 | 45.64 | 48.29 | 54.05 |
| 27 | 9.80 | 11.81 | 12.88 | 14.57 | 16.15 | 40.11 | 43.19 | 46.96 | 49.64 | 55.48 |
| 28 | 10.39 | 12.46 | 13.56 | 15.31 | 16.93 | 41.34 | 44.46 | 48.28 | 50.99 | 56.89 |
| 29 | 10.99 | 13.12 | 14.26 | 16.05 | 17.71 | 42.56 | 45.72 | 49.59 | 52.34 | 58.30 |
| 30 | 11.59 | 13.79 | 14.95 | 16.79 | 18.49 | 43.77 | 46.98 | 50.89 | 53.67 | 59.70 |
| 32 | 12.81 | 15.13 | 16.36 | 18.29 | 20.07 | 46.19 | 49.48 | 53.49 | 56.33 | 62.49 |
| 34 | 14.06 | 16.50 | 17.79 | 19.81 | 21.66 | 48.60 | 51.97 | 56.06 | 58.96 | 65.25 |
| 36 | 15.32 | 17.89 | 19.23 | 21.34 | 23.27 | 51.00 | 54.44 | 58.62 | 61.58 | 67.99 |
| 38 | 16.61 | 19.29 | 20.69 | 22.88 | 24.88 | 53.38 | 56.90 | 61.16 | 64.18 | 70.70 |
| 40 | 17.92 | 20.71 | 22.16 | 24.43 | 26.51 | 55.76 | 59.34 | 63.69 | 66.77 | 73.40 |
| 50 | 24.67 | 27.99 | 29.71 | 32.36 | 34.76 | 67.50 | 71.42 | 76.15 | 79.49 | 86.66 |
| 60 | 31.74 | 35.53 | 37.48 | 40.48 | 43.19 | 79.08 | 83.30 | 88.38 | 91.95 | 99.61 |
| 100 | 61.92 | 67.33 | 70.06 | 74.22 | 77.93 | 124.34 | 129.56 | 135.81 | 140.17 | 149.45 |
| 120 | 77.76 | 83.85 | 86.92 | 91.57 | 95.70 | 146.57 | 152.21 | 158.95 | 163.65 | 173.62 |
| 150 | 102.11 | 109.14 | 112.67 | 117.98 | 122.69 | 179.58 | 185.80 | 193.21 | 198.36 | 209.26 |
| 200 | 143.84 | 152.24 | 156.43 | 162.73 | 168.28 | 233.99 | 241.06 | 249.45 | 255.26 | 267.54 |
| 300 | 229.96 | 240.66 | 245.97 | 253.91 | 260.88 | 341.40 | 349.87 | 359.91 | 366.84 | 381.43 |
| 400 | 318.26 | 330.90 | 337.16 | 346.48 | 354.64 | 447.63 | 457.31 | 468.72 | 476.61 | 493.13 |
| $\infty$ | 867.48 | 888.56 | 898.91 | 914.26 | 927.59 | 1074.68 | 1089.53 | 1106.97 | 1118.95 | 1143.92 |