

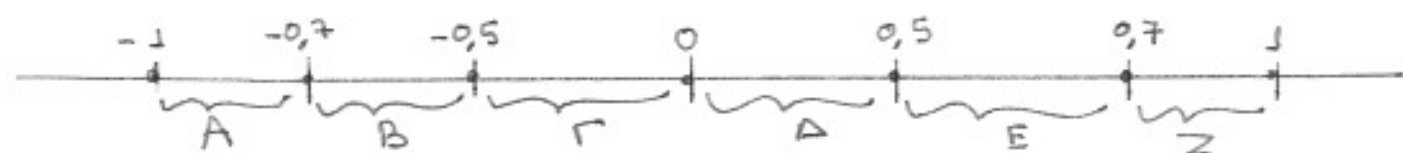
## ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΓΧΕΤΙΣΗΣ ΚΑΙ ΠΡΟΒΔΙΟΡΙΣΜΟΥ

## ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΓΧΕΤΙΣΗΣ

Ο Βαθμός συσχέτισης μεταξύ δύο μεταβλητών μετράται ποσοτικά με την χρήση του συντελεστή γραμμικής συσχέτισης του Pearson που συμβολίζεται με  $r$

$$r = \frac{n \sum_{i=1}^n x_i \cdot y_i - \sum_{i=1}^n y_i \cdot \sum_{i=1}^n x_i}{\sqrt{\left[ n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \right] \left[ n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2 \right]}}$$

Το  $r \in [-1, 1]$



Κατηγορία Α:  $-1 \leq r < -0,7$  Συσχέτιση αρνητική ισχυρή  
εως πολύ ισχυρή

Κατηγορία Β:  $-0,7 < r < -0,5$  Συσχέτιση αρνητική μέτρια  
εως ικανοποιητική

Κατηγορία Γ:  $-0,5 < r < 0$  Συσχέτιση αρνητική  
αδυνής έως μέτρια

Κατηγορία Δ:  $0 < r < 0,5$  Συσχέτιση θετική αδυνής  
εως μέτρια

Κατηγορία Ε:  $0,5 < r < 0,7$  Συσχέτιση θετική μέτρια έως  
ικανοποιητική

Κατηγορία Ζ:  $0,7 < r \leq 1$  Συσχέτιση θετική ισχυρή  
εως πολύ ισχυρή

Αν  $r > 0$  τότε η κλίση της ευθείας παλινδρόμησης είναι θετική

Αν  $r < 0$  η κλίση της ευθείας παλινδρόμησης είναι αρνητική

Αν  $r = 0$  τότε η ευθεία παλινδρόμησης είναι οριζόντια  $b = 0$

Συντελεστής προσδιορισμού

Ο συντελεστής προσδιορισμού του μοντέλου γραμμικής παλινδρόμησης αποτελεί μέτρο της προβλεπτικής του ικανότητας και εκφράζει το βαθμό εξάρτησης της εξαρτημένης μεταβλητής  $Y$  από την ανεξάρτητη μεταβλητή  $X$ .

Συμβολίζεται με  $R^2$

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

$y_i$  = η παρατήρηση  $i$  της εξαρτημένης μεταβλητής  $Y$

$\hat{y}_i$  = η τιμή που προβλέπεται από την ευθεία παλινδρόμησης  $\hat{y}_i = a + b \cdot x_i$

Το  $R^2 \in [0, 1]$

Αν  $R^2 \approx 1$  τότε η ικανότητα πρόβλεψης της πραγματικής τιμής από την γραμμή παλινδρόμησης

Αν  $R^2 \approx 0$  τότε έχουμε ελάχιστη ικανότητα <sup>(643/141)</sup> προβλεψής από την γραμμή παλινδρόμησης

Σε πρακτικές εφαρμοχές ο συντελεστής προσδιορισμού εκφράζεται σε ποσοστό επί της εκατο%.  
Οι σχετικά ικανοποιητικές θεωρούνται τιμές μεγαλύτερες του 60%.

### Παρατηρήσεις

α) Ισχύει  $R^2 = r^2$  μόνο στην περίπτωση της απλής γραμμικής παλινδρόμησης με μια μόνο ανεξάρτητη μεταβλητή

β) Η ποσότητα  $1 - R^2$  εκφράζει το ποσοστό της συνολικής μεταβλητότητας της  $Y$  που οφείδεται στο τυχαίο σφάλμα

γ)  $r > R^2$  και υπάρχει κίνδυνος υπερεκτίμησης του βαθμού συσχέτισης αν υιοθετηθεί η αποκλειστική χρήση του  $r$ .

Παράδειγμα αν  $r = 0,7$  έχουμε ισχυρή γραμμική συσχέτιση ενώ  $R^2 = r^2 = (0,7)^2 = 0,49 = 49\%$  μικρότερο του 60% που δείχνει ότι η αξιοπιστία του μοντέλου είναι μέτρια

8) Ο  $R^2$  αποτελεί μεν μέτρο της ερμηνείας (σελ 142)  
της εξαρτημένης μεταβλητής από την ανεξάρτητη  
αλλά δεν μετρά την κλίση  $\beta$  της γραμμικής  
παλινδρόμησης. Χαμηλή τιμή του  $R^2$  δεν σημαίνει  
ότι δεν υπάρχει καμία γραμμική συσχέτιση (μπορεί  
να είναι αδύναμη ή ακόμη και ισχυρή που οφείλε-  
ται σε άλλους παράγοντες)

ε) Η ποσότητα  $1-R^2$  εκφράζει το ποσοστό της  
ενοχικής διακύμανσης που οφείλεται στο τυχαίο  
σφάλμα. Όσο μικρότερο είναι τόσο πιο αξιόπιστο  
καθίσταται το μοντέλο της παλινδρόμησης

Παράδειγμα Αν υπολογίσουμε το  $r$  και το  $R^2$   
του προηγούμενου παραδείγματος

Αν  $X$  διαφημίση (έσοδα) και  $Y$  έσοδα βρίσκουμε με  
χρήση των τύπων  $r = 0,95$  και  $R^2 = 0,90$

Αφού η τιμή του  $r$  είναι πολύ κοντά στη μονάδα  
τότε προκύπτει ότι υπάρχει πολύ ισχυρή θετική  
αλληλεξάρτηση των εσόδων από πωλήσεις με την  
διαφημιστική δαπάνη

Το  $R^2 = 90\%$  δηλώνει ότι το 90% των πωλ<sup>των πωλ</sup>  
από την παλινδρόμηση ενώ το υπόλοιπο 10%  
ερμηνεύεται από άλλους παράγοντες

Αν  $X$  αριθμός πωλητών και  $Y$  έσοδα πωλήσεων  
βρίσκουμε με εφαρμογή των τύπων ότι  $r = 0,72$

και  $R^2 = 0,52$

Η τιμή του  $r$  δηλώνει ισχυρή θετική συσχέτιση ωστόσο από την τιμή του  $R^2$  συνάχεται ότι το μοντέλο παλινδρόμησης που κατασκευάστηκε ερμηνεύει μόνο λίγες περισσότερες από τις μισές τιμές 52%

Συγκρίνοντας τις δύο ευθείες παλινδρόμησης είναι φανερό ότι για την αύξηση των εσόδων από τις πωλήσεις η εταιρεία πρέπει να στραφεί κατά προτεραιότητα στην αύξηση δαπανών διαφήμισης που είναι πιο προοδοφόρες από τη χρήση επιπλέον δαπανών.

Τυπικό σφάλμα εκτίμησης  $S$

$$S = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}$$

Αν  $S$  έχει μικρή τιμή αυτό σημαίνει ότι οι παρατηρούμενες και οι εκτιμούμενες τιμές διαφέρουν πολύ λίγο μεταξύ τους και συνεπώς η γραμμή παλινδρόμησης περιγράφει με ικανοποιητική ακρίβεια την πραγματική σχέση μεταξύ των μεταβλητών  $X$  και  $Y$ . Αντίθετα αν  $S$  έχει μεγάλη τιμή τότε η σχέση των δύο μεταβλητών δεν περιγράφεται ικανοποιητικά από την επίδωξη παλινδρόμησης.

## ΑΛΛΗΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Είναι συχνά χρήσιμο να ελεγχθούν στατιστικά μέσω ελέγχων υποθέσεων οι τιμές των βασικών παραμέτρων του μοντέλου της γραμμικής παλινδρόμησης  $Y = a + b \cdot X + \varepsilon$

Έλεγχος της κλίσης της γραμμής παλινδρόμησης

Ελέγχεται στατιστικά η ισότητα της κλίσης  $b$  με μια δεδομένη τιμή

Η μηδενική υπόθεση είναι:

$$H_0: b = b^* \text{ ή } H_0: b > b^* \text{ ή } H_0: b < b^*$$

όπου  $b$  η κλίση της γραμμής παλινδρόμησης του πληθυσμού

$b^*$  μια συγκεκριμένη τιμή ελέγχου της κλίσης  $b$

Η εναλλακτική υπόθεση είναι

$$H_1: b = b^* \text{ ή } H_1: b > b^* \text{ ή } H_1: b < b^*$$

Ορίζοντας ως τιμή ελέγχου το μηδέν ελέγχεται η υπόθεση  $H_0: b = 0$ . Αν προκύψει από το στατιστικό έλεγχο ότι ισχύει η μηδενική υπόθεση τότε η ευθεία παλινδρόμησης είναι παράλληλη με τον οριζόντιο άξονα άρα η μεταβλητή  $Y$  δεν εξαρτάται από την  $X$

Αντίθετα αν απορριφθεί η υπόθεση  $H_0$  οπότε  $b \neq 0$  τότε η μεταβλητή  $Y$  επηρεάζεται από την μεταβλητή  $X$ . Με την ίδια λογική αν  $H_1: b > b_1$  ελέγχεται αν η κλίση της ευθείας είναι θετική ενώ αν  $H_1: b < b_1$  ελέγχεται αν η κλίση είναι αρνητική

1. Μεγάλο μέγεθος δείγματος ( $n > 30$ )

Το στατιστικό  $Z = \frac{\hat{\beta} - \beta^*}{s_{\hat{\beta}}}$  ακολουθεί την κανονική κατανομή

$\hat{\beta}$ : η εκτιμώμενη τιμή της κλίσης της γραμμής παλινδρόμησης

$\beta^*$  η τιμή ελέγχου της κλίσης  $\beta$

$s_{\hat{\beta}}$  η τυπική απόκλιση της εκτιμώμενης τιμής  $\hat{\beta}$

$$s_{\hat{\beta}} = \frac{\sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\frac{\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}}{n}}}$$

Όταν ο έλεγχος είναι αμφίπλευρος (δηλ  $H_1: \beta \neq \beta^*$ ) οι περιοχές απόρριψης της  $H_0$  προδιορίζονται

όταν  $|Z| > Z_{\alpha/2}$ . Οι τιμές  $Z_{\alpha/2}$  και  $-Z_{\alpha/2}$

προδιορίζονται από τον πίνακα της τυποποιημένης κανονικής κατανομής για διάφορες τιμές του επιπέδου σημαντικότητας  $\alpha$

Όταν ο έλεγχος είναι δεξιά μονόπλευρος

$H_0: \beta > \beta^*$  τότε η περιοχή απόρριψης προδιορίζεται από την σχέση  $Z > Z_{\alpha}$

και αν  $H_0: \beta < \beta^*$  αριστερά μονόπλευρος

τότε  $Z < -Z_{\alpha}$

Μικρό μέγεθος δείγματος ( $n < 30$ )

Το στατιστικό  $t = \frac{\hat{\beta} - \beta^*}{S_{\hat{\beta}}}$  ακολουθεί την κατανομή  $t$

(Student) με  $n-2$  βαθμούς ελευθερίας

Περιοχές απορρίψης της μηδενικής υπόθεσης είναι  
αμφίπλευρος έλεγχος  $|t| > t_{n-2, \alpha/2}$

Δεξιά μονόπλευρος  $t > t_{n-2, \alpha}$

Αριστερά μονόπλευρος  $t < -t_{n-2, \alpha}$

Εκτίμηση της τιμής του συντελεστή συσχέτισης

Αυτό που ελέγχεται είναι αν η γραμμική σχέση  
μεταξύ της εξαρτημένης και της ανεξάρτητης  
μεταβλητής είναι στατιστικά σημαντική

Το στατιστικό  $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$  ακολουθεί την

κατανομή  $t$

$H_0: r=0$  απορρίπτεται όταν  $|t| > t_{n-2, \alpha/2}$

Εκτίμηση της τιμής της εξαρτημένης μεταβλητής

Η μέση τιμή της μεταβλητής  $Y$  έστω  $E(\hat{y})$  εκτιμά-  
ται από την επίλυση παλινδρόμησης ( $y = a + b \cdot x + \varepsilon$ )  
έχοντας θεωρήσει ότι το σφάλμα  $\varepsilon$  ακολουθεί  
κανονική κατανομή με μέση τιμή το μηδέν.

Για μια δεδομένη τιμή  $x^*$  της ανεξάρτητης μεταβλη-  
της  $X$  η μέση εκτιμώμενη τιμή της  $Y$  υπολογίζεται

από την σχέση  $E(\hat{Y}) = \hat{Y} = \hat{a} + \hat{b} \cdot x^*$

Προσδιορισμός διαστημάτων εμπιστοσύνης  
Για τον καθορισμό διαστήματος εμπιστοσύνης  
της εξαρτημένης μεταβλητής χρησιμοποιούμε την  
κατανομή  $t$  σε επίπεδο σημαντικότητας  $\alpha$  και  $n-2$   
βαθμούς ελευθερίας

Διπλευρό διάστημα εμπιστοσύνης

$$\left[ \hat{Y}^* - t_{n-2, \alpha/2} \cdot S_Y, \hat{Y}^* + t_{n-2, \alpha/2} \cdot S_Y \right]$$

Δεξιά μονόπλευρο δ.ε

$$(-\infty, \hat{Y}^* + t_{n-2, \alpha} \cdot S_Y]$$

Αριστερά μονόπλευρο δ.ε

$$[\hat{Y}^* - t_{n-2, \alpha} \cdot S_Y, +\infty)$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^m (Y_i - \hat{Y}_i)^2}{n-2}} \sqrt{\frac{1}{n} + \frac{(X^* - \bar{X})^2}{\sum_{i=1}^m X_i^2 - \frac{(\sum_{i=1}^m X_i)^2}{n}}}$$

Παράδειγμα Συνέχεια του παραδείγματος (687) <sup>(688)</sup> για τα μηνιαία έσοδα πωλήσεων και τις αντίστοιχες διαφημιστικές δαπάνες στο διάστημα των τελευταίων 16 μηνών να ελεγχθούν στατιστικά σε στάθμη σημαντικότητας 5% και 1% τα ακόλουθα

- 1) Η κλίση της γραμμής παλινδρόμησης
- 2) Ο συντελεστής συσχέτισης
- 3) Να κατασκευαστούν τα διαστήματα εμπιστοσύνης της κλίσης της γραμμής παλινδρόμησης και της εξαρτημένης μεταβλητής

Λύση 1) Έχουμε  $n=16 < 30$  άρα χρησιμοποιείται η κατανομή  $t$

$$t = \frac{\hat{\beta} - \beta^*}{S_{\hat{\beta}}}$$

Η μηδενική υπόθεση  $H_0: \hat{\beta} = 0$

Τότε  $t = \frac{3,11}{0,265} = 11,74$

Η τιμή  $t_{n-2, \alpha/2}$

$n=16$   $n-2=16-2=14$  βαθμοί ελευθερίας

$\alpha = 5\% = 0,05$   $\frac{\alpha}{2} = 0,025$   $1 - \alpha/2 = 0,975$

Τότε  $t_{14, 0,025} = 2,145$  Άρα  $|t| = 11,74 > 2,145$  ιχνεί

άρα η υπόθεση  $H_0: \hat{\beta} = 0$  απορρίπτεται.

Αν  $\alpha = 1\% = 0,01$  τότε  $\frac{\alpha}{2} = 0,005$   $1 - \alpha/2 = 0,995$

Τότε  $t_{14, 0,005} = 2,977$   $|t| = 11,44 > 2,977$  ικνεί (681 149)

άρα η υπόθεση  $H_0: \hat{\beta} = 0$  απορρίπτεται

2) Το στατιστικό  $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,95}{\sqrt{\frac{1-(0,95)^2}{16-2}}} = 11,38$

Επίσης  $|t| > t_{14, 0,025} = 2,145$  και  $|t| > t_{14, 0,005} = 2,977$

ικνεί συνεπώς η μηδενική υπόθεση  $H_0: r = 0$  απορρίπτεται άρα η γραμμική σχέση των δύο μεταβλητών είναι στατιστικά σημαντική

3) Το διάνυσμα διάστημα εμπιστοσύνης για την κλίση της γραμμής παλινδρόμησης είναι

$$[\hat{\beta} - s_{\hat{\beta}} \cdot t_{n-2, \alpha/2}, \hat{\beta} + s_{\hat{\beta}} \cdot t_{n-2, \alpha/2}] = [2,32, 3,90]$$

Για  $\alpha = 0,05$  έχουμε  $\hat{\beta} = 3,11$   $s_{\hat{\beta}} = 0,265$

$$t_{n-2, \alpha/2} = t_{14, 0,025} = 2,145$$

Για  $\alpha = 0,01$  έχουμε  $t_{14, 0,005} = 2,977$  άρα το

διάστημα εμπιστοσύνης είναι

$$[3,11 - 0,265 \cdot 2,977, 3,11 + 0,265 \cdot 2,977] = [2,321, 3,89]$$

Το δ.ε. της ετήσιμης μεταβολής  $\hat{y}^n$  (6Δ  
150)

$$S_y = 272,7$$

Αν  $\alpha = 5\%$

$$\begin{aligned} & [\hat{y}^* - t_{14, 0,025} \cdot S_y, \hat{y}^* + t_{14, 0,025} \cdot S_y] = \\ & = [\hat{y}^* - 2,145 \cdot 272,7, \hat{y}^* + 2,145 \cdot 272,7] = \\ & = [\hat{y}^* - 584,9, \hat{y}^* + 584,9] \quad (*) \end{aligned}$$

Αντίστοιχα αν  $\alpha = 1\%$  τότε το δ.ε. είναι

$$[\hat{y}^* - 811,83, \hat{y}^* + 811,83] \quad (*) (*)$$

Αν η μοναία διαφημιστική δαπάνη είναι 1500 € τα προβλεπόμενα έσοδα από πωλήσεις είναι  $\hat{y} = \hat{a} + \hat{b} \cdot x = 677,4 + 3,11 \cdot 1500 = 5342,4 \text{ €}$

Αν  $\alpha = 5\%$  τα προβλεπόμενα έσοδα είναι αντικατάσταση στο (\*)

$$[5342,4 - 584,9, 5342,4 + 584,9] = [4757,5, 5927,3]$$

Αν  $\alpha = 1\%$  αντικαθιστούμε στο (\*) (\*)

$$[5342,4 - 811,83, 5342,4 + 811,83] = [4530,57, 6154,23]$$

## 6.5 Πολυμεταβλητή Παλινδρόμηση

Η πολυμεταβλητή Παλινδρόμηση αποτελεί μια γενίκευση της απλής παλινδρόμησης όπου εξετάζονται η σχέση μιας μεταβλητής  $Y$  με μια σειρά επεξηγηματικών μεταβλητών  $X_1, X_2, \dots, X_k$ . Στην πολυμεταβλητή παλινδρόμηση μπορούμε να εξετάσουμε περισσότερες από μια μεταβλητές. Με αυτό κερδίζουμε στο γεγονός ότι μας επιτρέπει να λάβουμε υπόψη τη συνδιακύμανση που έχουν οι μεταβλητές κάτι που αν χρησιμοποιούσαμε πολλά μονομεταβλητά μοντέλα θα το αγνοούσαμε. Το μοντέλο μπορεί να γραφτεί ως

$$Y = XB + E$$

όπου

$Y$  είναι ένας  $n \times 1$  πίνακας που περιέχει τις  $n$  μεταβλητές για τις  $n$  παρατηρήσεις μας

$X$  είναι ένας  $n \times p$  πίνακας που περιέχει τις  $p$  επεξηγηματικές μεταβλητές για τις  $n$  παρατηρήσεις μας (η πρώτη στήλη περιέχει μονάδες που αντιστοιχούν στη σταθερά, άρα έχουμε στην ουσία  $p-1$  μεταβλητές)

$B$  είναι ένας  $p \times 1$  πίνακας με τους συντελεστές της παλινδρόμησης όπου ο  $ij$  συντελεστής αναφέρεται στο συντελεστή της  $j$  επεξηγηματικής μεταβλητής στην  $i$  μεταβλητή και

$E$  είναι ένας πίνακας  $n \times 1$  με τα τυχαία σφάλματα.

Για το μοντέλο αυτό υποθέτουμε πως το διάνυσμα των σφαλμάτων  $\varepsilon_i$  ακολουθεί πολυμεταβλητή κανονική κατανομή με  $E(\varepsilon_i)=0$  και  $Var(\varepsilon_i)=\Sigma$ .

Είναι ιδιαίτερα σημαντικό να παρατηρήσουμε πως η εκτιμήτρια ελαχίστων τετραγώνων για τον πίνακα των συντελεστών  $B$  δίνεται και στην πολυμεταβλητή παλινδρόμηση από τον τύπο

$$\hat{B} = (X'X)^{-1} X'Y$$

Μπορεί επίσης να δείχτεί ότι οι συντελεστές είναι οι ίδιοι που θα έπαιρνε κάποιος αν έκανε τις παλινδρομήσεις μια προς μια. Εκεί που διαφέρει και αυτό είναι σημαντικό είναι πως επιτρέπει έναν όρο συνδιακύμανσης στα σφάλματα τον οποίο δεν θα μπορούσαμε να πάρουμε κάνοντας μια τις παλινδρομήσεις. Κάτω από αυτό το γενικό μοντέλο μπορεί κανείς να δει τόσο την πολυμεταβλητή ανάλυση διακύμανσης όσο και την πολυμεταβλητή ανάλυση συνδιακύμανσης

## 6.6 Η MANOVA ως Γραμμικό Μοντέλο

Η MANOVA μπορεί να γραφτεί ως ένα γραμμικό μοντέλο πολυμεταβλητής παλινδρόμησης με όμοιο τρόπο όπως και στην μονομεταβλητή περίπτωση δηλαδή το μοντέλο είναι το

$$x_{ij} = \mu + r_j + \varepsilon_{ij}, \quad i=1,2,\dots,n_j, j=1,\dots,k$$

όπου

$\mu$  είναι το διάνυσμα των γενικών μέσων

$r_j$  είναι το διάνυσμα με την επίδραση του  $j$  παράγοντα και

$\varepsilon_{ij}$  είναι τα τυχαία σφάλματα που ακολουθούν  $N_p(0, \Sigma)$ .

Όπως και στην μονομεταβλητή περίπτωση η MANOVA μπορεί να γίνει με τη χρήση του γραμμικού μοντέλου και τη χρήση ψευδομεταβλητών (dummy variables). Ομοίως μπορούμε να γενικεύσουμε σε μοντέλα με περισσότερους από έναν παράγοντες και με αλληλεπιδράσεις. Βάζοντας στο παραπάνω μοντέλο και άλλες συνεχής μεταβλητές καταλήγουμε σε μοντέλα πολυμεταβλητής ανάλυσης συνδιακύμανσης

## 6.7 Άλλα Θέματα

Ένα πρόβλημα που μπορεί να εμφανιστεί στην MANOVA και που είναι άγνωστο στη μονομεταβλητή περίπτωση είναι το γεγονός ότι ο πίνακας διακύμανσης συνδιακύμανσης μπορεί να μην είναι πλήρους βαθμού. Αυτό συμβαίνει όταν κάποιες μεταβλητές είναι πολύ ισχυρά συσχετισμένες. Για λόγους τυχαίου σφάλματος, στην πράξη κάποιος δεν θα δει μια μηδενική ορίζουσα αλλά μια ορίζουσα πολύ κοντά στο 0. Για αυτό καλό θα ήταν να εξεταστεί πρώτα η ύπαρξη τέτοιων μεταβλητών. Δουλεύοντας πιο αυστηρά θα μπορούσε κάποιος να εργαστεί με τις λεγόμενες κανονικές μεταβλητές (canonical variates). Η μέθοδος αυτή που μπορεί να ειπωθεί ως μια πολυμεταβλητή γενίκευση της ανάλυσης σε κύριες συνιστώσες μειώνει τις διαστάσεις του προβλήματος δημιουργώντας διανύσματα μεταβλητών (και όχι απλές μεταβλητές όπως η ανάλυση σε κύριες συνιστώσες) που έχουν δομή συνδιακύμανσης όμοια με την αρχική.

## 6.8 Παράδειγμα (Fisher's Iris Data).

Τα δεδομένα αυτά αποτελούνται από  $k=3$  είδη τριαντάφυλλων (iris setosa, versicolour, virginica). Από κάθε είδος ο ερευνητής πήρε 50 ( $n_j = 50$ ) τριαντάφυλλα και κατέγραψε τα εξής χαρακτηριστικά ως μεταβλητές:

$x_1$  : το μήκος του στέπαλου,                       $x_2$  : το πλάτος του στέπαλου  
 $x_3$  : το μήκος του πετάλου,                       $x_4$  : το πλάτος του πετάλου

Έχουμε επίσης  $n=150$ .

Οι μέσοι για κάθε δείγμα βρέθηκαν

$$\bar{x}_{\bullet 1} = \begin{pmatrix} 5.01 \\ 3.43 \\ 1.46 \\ 0.25 \end{pmatrix}, \quad \bar{x}_{\bullet 2} = \begin{pmatrix} 5.94 \\ 2.77 \\ 4.26 \\ 1.33 \end{pmatrix}, \quad \bar{x}_{\bullet 3} = \begin{pmatrix} 6.59 \\ 2.97 \\ 5.55 \\ 2.03 \end{pmatrix}$$

Δουλεύοντας με τους τύπους βρήκαμε πως οι πίνακες που χρειαζόμαστε είναι οι

$$B = \begin{bmatrix} 63.21 & -19.95 & 165.25 & 71.28 \\ & 11.35 & -57.24 & -22.93 \\ & & 437.11 & 186.78 \\ & & & 80.41 \end{bmatrix} \text{ και}$$

$$W = \begin{bmatrix} 38.96 & 13.63 & 24.62 & 5.64 \\ & 16.96 & 8.12 & 4.81 \\ & & 27.22 & 6.27 \\ & & & 6.16 \end{bmatrix}$$

από τους οποίους βρίσκουμε πως  $|W|=19976.4$  και  $|B+W|=871968$  και επομένως  $\Lambda = 0.023$ . Δεδομένου πως έχω 3 γκρουπ μπορώ να χρησιμοποιήσω τους ακριβείς τύπους που περιγράφηκαν. Έτσι ξέρω πως η ποσότητα

$$\frac{(n-k-p+1)(1-\sqrt{\Lambda})}{p\sqrt{\Lambda}} = \frac{(150-3-4+1)(1-\sqrt{0.023})}{4\sqrt{0.023}} = 201.25$$

ακολουθεί την κατανομή  $F$  με 8 και 288 βαθμούς ελευθερίας  $2p$  και  $2(n-k-p+1)$ . Επειδή η κριτική τιμή της κατανομής  $F$  με 8 και 288 βαθμούς ελευθερίας αντίστοιχα είναι 1.94 ( $\alpha=5\%$ ) απορρίπτουμε τη μηδενική υπόθεση περί ισότητας των μέσων.

Χρησιμοποιώντας το ασυμπτωτικό αποτέλεσμα του Bartlett (παρατηρείστε πως επειδή στην περίπτωση μας έχουμε κάποιο ακριβές αποτέλεσμα δεν είναι σωστό να χρησιμοποιήσουμε το ασυμπτωτικό), βρίσκουμε πως η ελεγχουσυνάρτηση έχει τιμή 548.86 η οποία συγκρινόμενη με την κριτική τιμή από μια  $\chi^2$  κατανομή με 8 βαθμούς ελευθερίας οδηγεί και πάλι στην απόρριψη της μηδενικής υπόθεσης.

Παρατηρείστε ότι τα διαγώνια στοιχεία των πινάκων B και W είναι αυτά που χρησιμοποιούνται για τους μονομεταβλητούς ελέγχους.

Έτσι έχουμε

|                               | Variable sum<br>of squares | Error Sum<br>of Squares | F       | p-value |
|-------------------------------|----------------------------|-------------------------|---------|---------|
| $x_1$ : το μήκος του σέπαλου, | 63.21                      | 38.96                   | 119.26  | 0.000   |
| $x_2$ : το πλάτος του σέπαλου | 11.35                      | 16.96                   | 49.16   | 0.000   |
| $x_3$ : το μήκος του πετάλου, | 437.11                     | 27.22                   | 1180.16 | 0.000   |
| $x_4$ : το πλάτος του πετάλου | 80.41                      | 6.16                    | 960.01  | 0.000   |
| βαθμοί ελευθερίας (για όλα)   | 2                          | 147                     |         |         |

κάνοντας δηλαδή μονομεταβλητούς ελέγχους θα απορρίψαμε για όλες τις μεταβλητές την ισότητα των μέσων.

Ο πίνακας ανάλυσης διακύμανσης για τα παραπάνω δεδομένα είναι ο

| Πηγή                               | SSP πίνακας                                                                                                                            | B <sub>e</sub> | Test  |
|------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------|----------------|-------|
| Ανάμεσα στα<br>γروهύπ<br>(between) | $B = \begin{bmatrix} 63.21 & -19.95 & 165.25 & 71.28 \\ & 11.35 & -57.24 & -22.93 \\ & & 437.11 & 186.78 \\ & & & 80.41 \end{bmatrix}$ | 2              | 0.023 |
| Μεσα στα<br>γروهύπ<br>(within)     | $W = \begin{bmatrix} 38.96 & 13.63 & 24.62 & 5.64 \\ & 16.96 & 8.12 & 4.81 \\ & & 27.22 & 6.27 \\ & & & 6.16 \end{bmatrix}$            | 147            |       |
| Συνολική                           | T=W+B                                                                                                                                  | 1491           |       |

Πίνακας 6.1 Ο πίνακας ανάλυσης διακύμανσης για τα δεδομένα του Fisher