
5 ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ

5.1 Εισαγωγή

Όπως είπαμε και στην αρχή αυτών των σημειώσεων το μεγάλο πλεονέκτημα της πολυμεταβλητής ανάλυσης είναι πως μας επιτρέπει να εξετάσουμε πολλές μεταβλητές συγχρόνως λαμβάνοντας υπόψη μας και τις όποιες συσχετίσεις υπάρχουν μεταξύ τους. Για παράδειγμα έχουμε 5 μεταβλητές οι οποίες μας περιγράφουν κάποιο χαρακτηριστικό. Αν θέλουμε να εξετάσουμε αν οι παρατηρήσεις μας δεν συμφωνούν με κάποια εκ των προτέρων γνώση για τα 5 αυτά χαρακτηριστικά θα μπορούσε κανείς να δουλέψει ξεχωριστά με τις πέντε μεταβλητές, να κάνει δηλαδή έναν έλεγχο για κάθε μεταβλητή και με βάση αυτούς τους ελέγχους να αποφανθεί για το σύνολο.

Αυτή η προσέγγιση έχει δύο μειονεκτήματα:

- αν ο έλεγχος για κάθε χαρακτηριστικό γίνει σε επίπεδο 5%, τότε δεδομένου πως κάνουμε 5 ελέγχους ξεχωριστά το συνολικό επίπεδο του πολύπλοκου (πενταπλού στην περίπτωση μας) ελέγχου δεν θα είναι 5% αλλά $1 - 0.95^5$ δηλαδή μόλις 22%, δηλαδή η πιθανότητα σφάλματος πολύ μεγάλη.
- επειδή είναι πολύ πιθανό οι μεταβλητές μας να είναι μεταξύ τους συσχετισμένες κάνοντας πέντε ξεχωριστούς ελέγχους αγνοούμε αυτή τη συσχέτιση, δηλαδή αγνοούμε την πληροφορία που έχουμε για κάθε μεταβλητή από τις υπόλοιπες. Επομένως χρειαζόμαστε πολυμεταβλητούς ελέγχους υποθέσεων οι οποίοι θα εξετάζουν όλο το διάνυσμα κάθε παρατήρησης και όχι μεμονωμένες μεταβλητές.

Οι έλεγχοι που θα περιγράψουμε αφορούν παρατηρήσεις από πολυμεταβλητούς κανονικούς πληθυσμούς, αντίστοιχα με την περίπτωση της υπόθεσης της κανονικότητας που είχαμε πίσω από τους περισσότερους μονομεταβλητούς ελέγχους.

5.2 Έλεγχοι για ένα διάνυσμα μέσων τιμών

5.2.1 Γνωστός πίνακας διακύμανσης

Ας ξεκινήσουμε με την πιο απλή περίπτωση του ελέγχου για ένα διάνυσμα μέσων με γνωστή διακύμανση. Στη μονομεταβλητή περίπτωση αυτός είναι ο πιο απλός έλεγχος και γίνεται μέσω της ελεγχουσυνάρτησης Z και της κανονικής κατανομής, δηλαδή έχουμε

$$H_0: \mu = \mu_0, \quad \text{έναντι της}$$

$$H_1: \mu \neq \mu_0$$

Η ελεγχουσυνάρτηση είναι η $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$.

Θυμηθείτε πως για τον έλεγχο υποθέτουμε πως οι μεταβλητές μας είναι κανονικές αλλά αν δεν ισχύει η υπόθεση της κανονικότητας με επίκληση του κεντρικού οριακού θεωρήματος το αποτέλεσμα ισχύει για μεγάλα δείγματα.

Στην πολυμεταβλητή περίπτωση υποθέτουμε πως τα δεδομένα προέρχονται από πολυμεταβλητή κανονική κατανομή. Τώρα έχουμε να ελέγξουμε την υπόθεση

$$H_0: \underset{p \times 1}{\mu} = \underset{p \times 1}{\mu_0}, \quad \text{έναντι της εναλλακτικής}$$

$$H_1: \underset{p \times 1}{\mu} \neq \underset{p \times 1}{\mu_0}$$

Ο πίνακας διακύμανσης Σ είναι γνωστός

Η ελεγχουσυνάρτηση που θα χρησιμοποιήσουμε είναι η

$Z^2 = n(\bar{\mathbf{x}} - \underset{p \times 1}{\mu_0})' \underset{p \times p}{\Sigma}^{-1} (\bar{\mathbf{x}} - \underset{p \times 1}{\mu_0})$ η οποία με βάση αυτά που είδαμε νωρίτερα ακολουθεί χ_p^2 κατανομή. Επομένως ο έλεγχος μπορεί να γίνει είτε υπολογίζοντας το p-value είτε χρησιμοποιώντας την κριτική τιμή της κατανομής αυτής.

Μερικά ενδιαφέροντα σημεία είναι τα εξής

- Η εναλλακτική είναι δίπλευρή καθώς σε ένα διάνυσμα δεν είναι εύκολο να ορίσουμε μονόπλευρους ελέγχους. Παρόλα αυτά μονόπλευροι έλεγχοι θα μπορούσαν να οριστούν χρησιμοποιώντας ελέγχους πιθανοφάνειας.
- Αν $p=1$ τότε ο έλεγχος ταυτίζεται με τον απλό έλεγχο Z , η ελεγχουσυνάρτηση είναι απλά η ελεγχουσυνάρτηση Z^2 και για αυτό η κατανομή είναι πια χ^2 με έναν βαθμό ελευθερίας.
- Μια μορφή της ελεγχουσυνάρτησης Z^2 είχαμε πριν χρησιμοποιήσει για να δημιουργήσουμε ελλειψοειδή ίσης πιθανότητας. Επομένως μπορούμε να χρησιμοποιήσουμε την

ελεγχουσυνάρτηση για να κατασκευάσουμε από κοινού διαστήματα εμπιστοσύνης για όλο το διάνυσμα των μέσων. Πράγματι θα ισχύει πως

$$P(Z^2 < \chi^2_{1-\alpha, p}) = \\ = P(n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0) < \chi^2_{1-\alpha, p}) = 1 - \alpha$$

όπου $\chi^2_{1-\alpha, p}$ είναι το $(1-\alpha)$ ποσοστιαίο σημείο της χ^2 κατανομής με p βαθμούς ελευθερίας.

Συνεπώς η εξίσωση $n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0) = \chi^2_{1-\alpha, p}$ ορίζει ένα υπερελλειψοειδές το οποίο με πιθανότητα $1-\alpha$ περιλαμβάνει το άγνωστο διάνυσμα μέσων τιμών του πληθυσμού.

Αν θυμηθούμε τη σχέση διπλευρων ελέγχων και διαστημάτων εμπιστοσύνης μπορούμε να κάνουμε έλεγχο κατασκευάζοντας το διάστημα εμπιστοσύνης και ελέγχοντας αν η τιμή κάτω από τη μηδενική υπόθεση περιέχεται σε αυτό

5.2.2 Άγνωστος πίνακας διακύμανσης

Ας υποθέσουμε τώρα πως ο πίνακας διακύμανσης δεν είναι γνωστός. Στην μονομεταβλητή περίπτωση χρησιμοποιούσαμε τη δειγματική διακύμανση αντί για τη διακύμανση του πληθυσμού και ο έλεγχος γινόταν με τη βοήθεια της κατανομής t , δηλαδή η ελεγχουσυνάρτηση ήταν η

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

Στην πολυμεταβλητή περίπτωση θα χρησιμοποιήσουμε παρόμοια τεχνική. Αντικαθιστούμε τον πίνακα διακύμανσης του πληθυσμού με το δειγματικό πίνακα διακύμανσης και προκύπτει η συνάρτηση του Hotelling

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)$$

για την οποία μιλήσαμε πριν και γνωρίζουμε πως ακολουθεί την T^2 κατανομή. Χρησιμοποιώντας τη σχέση της T^2 με την F μπορούμε να χρησιμοποιήσουμε την ελεγχουσυνάρτηση

$$\frac{n}{n-1} \frac{n-p}{p} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim F_{p, n-p}$$

Και πάλι αν $p=1$ ο έλεγχος του Hotelling συμπίπτει με τον απλό μονομεταβλητό έλεγχο.

Διαστήματα εμπιστοσύνης μπορούν να κατασκευαστούν και πάλι αλλά τώρα με τη χρήση κριτικών τιμών της κατανομής F . Συγκεκριμένα το υπερελλειψοειδές θα σχηματίζεται από την ανισότητα

$$n(\boldsymbol{\mu} - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\boldsymbol{\mu} - \bar{\mathbf{x}}) < \frac{p(n-1)}{n-p} F_{\alpha(p, n-p)}$$

Ο έλεγχος που μόλις περιγράψαμε συνήθως ονομάζεται έλεγχος T^2 για μια μέση τιμή (one sample T^2 test).

5.3 Έλεγχος για διαφορά δύο μέσων

Ας δούμε τώρα την περίπτωση που έχουμε δύο δείγματα και άρα θέλουμε να ελέγξουμε αν υπάρχουν διαφορές ανάμεσα στα διανύσματα των μέσων για τις δύο ομάδες. Θα δούμε απευθείας την ρεαλιστική περίπτωση της άγνωστης διακύμανσης. Παράλληλα υποθέτουμε πως η διακύμανση είναι κοινή και επομένως αυτό είναι κάτι που θα πρέπει ουσιαστικά να ελέγξουμε. Θα δούμε αργότερα πως.

Στη μονομεταβλητή περίπτωση ο έλεγχος είχε τις υποθέσεις

$$H_0: \mu_1 = \mu_2 \quad \text{έναντι της εναλλακτικής}$$

$$H_1: \mu_1 \neq \mu_2$$

Η ελεγχουσυνάρτηση είναι η

$$T = \frac{\bar{X}_1 - \bar{X}_2}{S_{pooled} / \sqrt{n_1 + n_2 - 2}} \sim t_{n_1 + n_2 - 2} \quad \text{που ακολουθεί κάτω από τη μηδενική υπόθεση την κατανομή } t.$$

Η ποσότητα S_{pooled}^2 είναι μια εκτίμηση της κοινής διακύμανσης και υπολογίζεται ως σταθμικός μέσος των διακυμάνσεων των δύο ομάδων.

Στην πολυμεταβλητή περίπτωση τα πράγματα είναι ανάλογα.

Οι υποθέσεις που θέλουμε να ελέγξουμε είναι οι

$$H_0: \underset{p \times 1}{\boldsymbol{\mu}}_1 = \underset{p \times 1}{\boldsymbol{\mu}}_2, \quad \Sigma \text{ κοινός και άγνωστος έναντι της}$$

$$H_1: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

Θυμηθείτε πως θα ισχύει πως

$$(n_1 - 1)S_1 \sim W_p(n_1 - 1, \Sigma)$$

$$(n_2 - 1)S_2 \sim W_p(n_2 - 1, \Sigma)$$

όπου S_i , $i=1,2$ είναι οι αμερόληπτοι πίνακες διακύμανσης για τα δύο γκρουπ. Επομένως από τις ιδιότητες της κατανομής Wishart

$$(n_1 - 1)S_1 + (n_2 - 1)S_2 \sim W_p(n_1 + n_2 - 2, \Sigma).$$

Επομένως οι προϋποθέσεις για να δημιουργήσουμε μια ανάλογη της T^2 συνάρτησης υπάρχουν (έχουμε ένα διάνυσμα από πολυμεταβλητή κανονική κατανομή και έναν πίνακα από Wishart) και άρα ορίζουμε την ελεγχουσυνάρτηση

$$\frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \sim F_{p, n_1 + n_2 - p - 1}$$

με $\mathbf{S}_{pooled} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$ η οποία έχει μετασχηματιστεί ώστε να ακολουθεί την κατανομή

F.

Και σε αυτή την περίπτωση μπορούμε να φτιάξουμε από κοινού διαστήματα εμπιστοσύνης καθώς και να προκύψει η ειδική περίπτωση για $p=1$.

Τελειώνοντας με τους ελέγχους για μέσες τιμές πρέπει να τονιστεί πως όταν απορριφθεί η μηδενική υπόθεση, μπορεί ο ερευνητής να εστιάσει το ενδιαφέρον του στο ποια(ες) μεταβλητή(ές) είναι υπεύθυνη(ες) για την απόρριψη της υπόθεσης. Δεν θα ασχοληθούμε εδώ με αυτό το πρόβλημα.

Αν θέλουμε να ελέγξουμε την ισότητα περισσότερων από δύο δειγμάτων τότε υπάρχει μια αντίστοιχη μεθοδολογία με αυτή της ανάλυσης διακύμανσης την οποία θα δούμε σε λίγο.

5.4 Έλεγχος Ισότητας Πινάκων Διακύμανσης

Στην περίπτωση του ελέγχου για δύο διανύσματα μέσω χρειαστήκαμε την υπόθεση ότι οι πίνακες διακύμανσης στον πληθυσμό είναι ίσοι. Αυτή την υπόθεση μπορούμε να την ελέγξουμε με μια διαδικασία που θα περιγράψουμε στην πιο γενική της μορφή ώστε να μας φανεί χρήσιμη και στην πολυμεταβλητή ανάλυση διακύμανσης που θα ακολουθήσει.

Ας ξεκινήσουμε και πάλι από την περίπτωση μιας μεταβλητής. Έλεγχος περί ισότητας διακυμάνσεων (συντά αποκαλούνται έλεγχοι ομοιογένειας διακύμανσης, homogeneity of variance tests) υπάρχουν αρκετοί. Οι πιο γνωστοί είναι ο έλεγχος του Levene και ο έλεγχος του Bartlett και φυσικά για 2 δείγματα ο κλασσικός έλεγχος F. Θα περιγράψουμε την πολυμεταβλητή γενίκευση του ελέγχου του Bartlett στις πολλές διαστάσεις. Ο έλεγχος δεν είναι τίποτα διαφορετικό από έλεγχο λόγου πιθανοφαινιών.

Συγκεκριμένα οι υποθέσεις που θα εξετάσουμε είναι

$$H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_k$$

$$H_1: \text{Τουλάχιστον 2 διαφέρουν}$$

όπου k ο αριθμός των ομάδων που έχουμε. Ο έλεγχος, που ονομάζεται Box-M, χρησιμοποιεί την ελεγχοσυνάρτηση

$$M = \phi \sum_{i=1}^k \left[(n_i - 1) \ln |S_i^{-1} S_{pooled}| \right],$$

$$\text{όπου } S_{pooled} = \frac{\sum_{i=1}^k (n_i - 1) S_i}{n - k},$$

$$\phi = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \sum_{i=1}^k \frac{1}{(n_i - k)(n - k)},$$

$$n = \sum_{i=1}^k n_i, \text{ και}$$

p είναι ο αριθμός των μεταβλητών (επομένως οι πίνακες διακύμανσης είναι $p \times p$).

Ο έλεγχος βασίζεται στο λόγο πιθανοφάνειας κάτω από τις δύο υποθέσεις. Μελετώντας την ελεγχοσυνάρτηση παρατηρούμε πως αν όλοι οι πίνακες διακύμανσης είναι σχεδόν ίσοι, δηλαδή ισχύει η μηδενική υπόθεση, τότε και ο S_{pooled} είναι σχεδόν ίσος με αυτούς και επομένως ο πίνακας $S_i^{-1} S_{pooled}$ θα μοιάζει πολύ με τον μοναδιαίο πίνακα. Συνεπώς η οριζούσα του θα είναι κοντά στο 1, αυτό με τη σειρά του σημαίνει πως ο λογάριθμος θα είναι σχεδόν 0 και άρα η τιμή της ελεγχοσυνάρτησης θα είναι πολύ μικρή.

Δεδομένου πως πρόκειται για έλεγχο πιθανοφανειών η κατανομή της ελεγχοσυνάρτησης κάτω από τη μηδενική υπόθεση και ασυμπτωτικά είναι η $\chi^2_{p(p+1)(k-1)/2}$. Η ποσότητα ϕ έχει σκοπό να κάνει την ασυμπτωτική προσέγγιση πιο καλή για πεπερασμένα δείγματα.

Παράδειγμα 5.1: Σε ένα δείγμα 15 μαθητών μιας τάξης μετρήθηκαν το ύψος (X_1), η περιφέρεια στήθους (X_2), η περιφέρεια βραχίονα (X_3), και καταγράφηκε και το φύλο (0 αγόρι, 1 κορίτσι). Τα δεδομένα μπορείτε να τα δείτε στον πίνακα που ακολουθεί.

Παρατηρείστε πως με την παραπάνω παρουσίαση των δεδομένων που είναι αυτή που όλα τα στατιστικά πακέτα ακολουθούν κάθε παρατήρηση είναι μια γραμμή. Όμως στη θεωρία χρησιμοποιήσαμε ως παρατήρηση το διάνυσμα στήλη. Επομένως ο παραπάνω πίνακας περιέχει τις παρατηρήσεις μας ως διανύσματα στήλη, δηλαδή ισχύει (λαμβάνοντας τις 3 πρώτες μεταβλητές υπόψη) πως $x_1' = [78, 60.6, 16.5]$ κλπ

Υψος	Περίφερεια Στήθους	Περίφερεια Βραχίονα	Φύλο
78	60.6	16.5	0
76	58.1	12.5	0
92	63.2	14.5	0
1	59	14	0
81	60.8	15.5	0
84	59.5	14	0
80	58.4	14	1
75	59.2	15	1
78	60.3	15	1
75	57.4	13	1
79	59.5	14	1
78	58.1	14.5	1
75	58	12.5	1
64	55.5	11	1
80	59.2	12.5	1

Πίνακας 5.1 Δεδομένα για το παράδειγμα 5.1 (Πηγή: Chatfield and Colling, 1984.)

Έστω ότι θέλουμε να ελέγξουμε την υπόθεση πως το διάνυσμα των μέσων για τα αγόρια είναι το

$$\begin{bmatrix} 90 \\ 58 \\ 16 \end{bmatrix} \text{ δοθέντος πως ο πίνακας διακύμανσης είναι γνωστός και ίσος με } \Sigma = \begin{bmatrix} 30 & 6 & 0.5 \\ 6 & 3 & 1.5 \\ 0.5 & 1.5 & 2 \end{bmatrix}.$$

Συνεπώς οι υποθέσεις που μας ενδιαφέρουν είναι οι

$$H_0: \mu' = [90 \quad 58 \quad 16] \quad \text{έναντι της}$$

$$H_1: \mu' \neq [90 \quad 58 \quad 16]$$

Για τα αγόρια έχουμε 6 παρατηρήσεις ($n=6$). Και επειδή ο πίνακας διακύμανσης είναι γνωστός θα κάνουμε έλεγχο για γνωστό πίνακα διακύμανσης. Το διάνυσμα των μέσων είναι

$$\bar{x} = \begin{bmatrix} 82 \\ 60.2 \\ 14.5 \end{bmatrix} \text{ και η ελεγχοσυνάρτηση που θα χρησιμοποιήσουμε είναι η}$$

$$Z^2 = n(\bar{x} - \mu_0)' \Sigma^{-1}(\bar{x} - \mu_0)$$

Ο αντίστροφος του πίνακα διακύμανσης που χρειαζόμαστε είναι ο

$$\Sigma^{-1} = \begin{bmatrix} 0.07692 & & \\ -0.23077 & 1.22564 & \\ 0.15385 & -0.86154 & 1.170769 \end{bmatrix},$$

και το διάνυσμα με τις αποκλίσεις των δειγματικών μέσων από αυτές της μηδενικής υπόθεσης είναι

$$(\bar{x} - \mu_0) = \begin{bmatrix} -8 \\ 2.2 \\ -1.5 \end{bmatrix}.$$

Συνεπώς η τιμή της ελεγχοσυνάρτησης υπολογίζεται ως

$$Z^2 = 6 \begin{bmatrix} -8 & 2.2 & -1.5 \end{bmatrix} \begin{bmatrix} 0.07692 & -0.23077 & 0.15385 \\ -0.23077 & 1.22564 & -0.86154 \\ 0.15385 & -0.86154 & 1.17077 \end{bmatrix} \begin{bmatrix} -8 \\ 2.2 \\ -1.5 \end{bmatrix} = 185.094,$$

Δεδομένου πως $p=3$, η κατανομή της ελεγχοσυνάρτησης είναι η χ^2 με 3 βαθμούς ελευθερίας και βρίσκουμε πως $p\text{-value} < 0.0001$, δηλαδή απορρίπτουμε τη μηδενική υπόθεση σχεδόν σε κάθε επίπεδο στατιστικής σημαντικότητας.

Ας υποθέσουμε τώρα πως η διακύμανση είναι άγνωστη. Σε αυτή την περίπτωση θα πρέπει να χρησιμοποιήσουμε το δειγματικό πίνακα διακύμανσης.

Η ελεγχοσυνάρτηση θα είναι η T^2 του Hotelling δηλαδή η

$$\frac{n}{n-1} \frac{n-p}{p} (\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0)$$

Ο δειγματικός πίνακας διακύμανσης και ο αντίστροφός του είναι οι

$$S = \begin{bmatrix} 31.6 & 8.04 & 0.5 \\ 8.04 & 3.172 & 1.31 \\ 0.5 & 1.31 & 1.9 \end{bmatrix}, \quad S^{-1} = \begin{bmatrix} 0.1836 & -0.63 & 0.38 \\ & 2.584 & -1.615 \\ & & 1.53 \end{bmatrix}$$

και επομένως υπολογίζουμε την ελεγχοσυνάρτηση ως

$$\frac{6}{5} \frac{3}{3} \begin{bmatrix} -8 & 2.2 & -1.5 \end{bmatrix} S^{-1} \begin{bmatrix} -8 \\ 2.2 \\ -1.5 \end{bmatrix} = 84.09$$

Επειδή η ελεγχοσυνάρτηση ακολουθεί την κατανομή F με βαθμούς ελευθερίας 3 και 3 βρίσκουμε πως $p\text{-value} = 0.00003$ και άρα απορρίπτουμε σχεδόν σε κάθε επίπεδο στατιστικής σημαντικότητας τη μηδενική υπόθεση.

Έστω πως τώρα θέλουμε να ελέγξουμε την υπόθεση πως τα διανύσματα των μέσων θα είναι τα ίδια για τα αγόρια και τα κορίτσια. Δηλαδή η μηδενική μας υπόθεση είναι πως

$$H_0: \mu_A = \mu_K \quad \text{έναντι της εναλλακτικής}$$

$$H_1: \mu_A \neq \mu_K$$

Ο πίνακας διακυμάνσης Σ υποθέτουμε πως είναι άγνωστος αλλά κοινός. Θα ελέγξουμε αργότερα αν η υπόθεση για ίσες διακυμάνσεις είναι λογική.

$$\text{Ελεγχοςυνάρτηση} \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{\text{pooled}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

$$\text{Είναι } \bar{\mathbf{x}}_1 = \begin{bmatrix} 82 \\ 60.2 \\ 14.5 \end{bmatrix}, \bar{\mathbf{x}}_2 = \begin{bmatrix} 76 \\ 58.4 \\ 13.5 \end{bmatrix}$$

Με τη βοήθεια των πινάκων $\mathbf{S}_1$ και $\mathbf{S}_2$ θα υπολογίσουμε

$$\mathbf{S}_{\text{pooled}} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2} = \begin{bmatrix} 27.23 & 6.56 & 2.84 \\ 6.56 & 2.43 & 1.4 \\ 2.84 & 1.4 & 1.84 \end{bmatrix}$$

Άρα η τιμή της ελεγχοςυνάρτησης είναι $1.489 < F_{0.95,3,11}$, ενώ $p\text{-value} = 0.24$ και συνεπώς δεν απορρίπτουμε H_0 .

Όπως είδαμε προηγουμένως, για να κάνουμε τον έλεγχο δύο διανυσμάτων μέσω τιμών χρειαζόμαστε την υπόθεση ισότητας των πινάκων διακυμάνσεων. Θα χρησιμοποιήσουμε τον έλεγχο Box-M.

Συγκεκριμένα θα ελέγξουμε την υπόθεση

$$H_0: \Sigma_1 = \Sigma_2 \quad \text{έναντι της εναλλακτικής}$$

$$H_1: \Sigma_1 \neq \Sigma_2$$

Για τα δεδομένα μας έχουμε πως $p=3$, $k=2$, $n=15$, $n_1=6$, $n_2=9$ ενώ βρίσκουμε

$$\mathbf{S}_1 = \begin{bmatrix} 31.6 & 8.04 & 0.5 \\ 8.04 & 3.17 & 1.31 \\ 0.5 & 1.31 & 1.9 \end{bmatrix}, \mathbf{S}_2 = \begin{bmatrix} 24.5 & 5.63 & 4.31 \\ 5.63 & 1.97 & 1.45 \\ 4.31 & 1.45 & 1.31 \end{bmatrix}$$

και άρα

$$\mathbf{S}_{\text{pooled}} = \frac{5\mathbf{S}_1 + 8\mathbf{S}_2}{13} = \begin{bmatrix} 27.23 & 6.56 & 2.84 \\ 6.56 & 2.43 & 1.4 \\ 2.84 & 1.4 & 1.84 \end{bmatrix}, |\mathbf{S}_1^{-1}\mathbf{S}_{\text{pooled}}| = 0.95, |\mathbf{S}_2^{-1}\mathbf{S}_{\text{pooled}}| = 1.82$$

$$\phi = 1 - \frac{18+9-1}{24} \left(\frac{1}{5 \cdot 13} + \frac{1}{8 \cdot 13} \right) = 0.963$$

Χρησιμοποιώντας λοιπόν τον τύπο για την ελεγχοσυνάρτηση του ελέγχου βρίσκουμε πως

$M = 0.963 \cdot (5 \ln 0.95 + 8 \ln 1.82) = 4.37 < \chi_{5,0.95}^2$, ενώ το $p\text{-value} = 0.63$ οπότε και δεν απορρίπτουμε την H_0 .

5.5 Συμπεράσματα

Στον παρακάτω πίνακα εμφανίζουμε συνοπτικά τους ελέγχους για τη μονοδιάστατη και την πολυδιάστατη περίπτωση για να γίνει σαφής η σχέση τους

Μονοδιάστατη περίπτωση	Πολυδιάστατη περίπτωση
$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	$n(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim \chi_p^2$
$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$	$\frac{n}{n-1} \frac{n-p}{p} (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}_0) \sim F_{p, n-p}$
$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{pooled}/\sqrt{n_1+n_2-2}} \sim t_{n_1+n_2-2}$	$\frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \frac{n_1 n_2}{n_1+n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_{pooled}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \sim F_{p, n_1+n_2-p-1}$
$M = \phi \cdot \sum_{i=1}^k \left[(n_i - 1) \ln \left(\frac{s_{pooled}^2}{s_i^2} \right) \right]$	$M = \phi \sum_{i=1}^k \left[(n_i - 1) \ln \mathbf{S}_i^{-1} \mathbf{S}_{pooled} \right]$

Ο μονομεταβλητός έλεγχος ομοιογένειας διακυμάνσεων είναι ο έλεγχος του Bartlett. Η σταθερά ϕ^* υπολογίζεται με πολύ παρόμοιο τρόπο όπως και η σταθερά ϕ της πολυμεταβλητής περίπτωσης

Ολοκληρώνοντας σχετικά με τους ελέγχους υποθέσεων θα πρέπει να αναφέρουμε τα εξής:

- Όπως και στη μονομεταβλητή περίπτωση οι έλεγχοι είναι βασισμένοι στην υπόθεση της πολυμεταβλητής κανονικότητας. Επειδή και για τα διανύσματα μέσω υπάρχουν αντίστοιχα κεντρικά οριακά θεωρήματα τα αποτελέσματα μπορούν να χρησιμοποιηθούν αν το μέγεθος του δείγματος είναι σχετικά μεγάλο. Μόνο ο έλεγχος Box-M έχει πρόβλημα και πολύ κακές ιδιότητες αν η υπόθεση της κανονικότητας δεν ισχύει

- Αν στον έλεγχο για μια μέση τιμή (ή για διαφορά δύο μέσων τιμών) απορρίψουμε τη μηδενική υπόθεση, το ενδιαφέρον είναι να βρούμε πια ποια ή ποιες μεταβλητές διαφέρουν και γενικά οδηγούν στην απόρριψη της μηδενικής υπόθεσης. Υπάρχουν στη βιβλιογραφία τέτοιες τεχνικές αλλά δεν θα ασχοληθούμε μαζί τους σε αυτές τις σημειώσεις.
- Προφανώς οι έλεγχοι που συζητήσαμε είναι απλά κάποια δείγματα για το πώς ιδέες πολύ γνωστές στη μονοδιάστατη περίπτωση μεταφέρονται ή αν θέλετε γενικεύονται σε περισσότερες διαστάσεις. Υπάρχει μια τεράστια ποικιλία από ελέγχους υποθέσεων για πολυμεταβλητά δεδομένα.

5.6 Έλεγχοι για την πολυμεταβλητή κανονική κατανομή

Ένα πρόβλημα το οποίο εμφανίζεται πολύ συχνά, αλλά δυστυχώς δεν έχει εύκολη και ξεκάθαρη λύση είναι το εξής: πως μπορούμε να ελέγξουμε αν τα δεδομένα μας προέρχονται από μια πολυμεταβλητή κανονική κατανομή. Αυτή η υπόθεση είναι βασική σε αρκετές διαδικασίες στατιστικής συμπερασματολογίας.

Στη μονομεταβλητή περίπτωση για παράδειγμα υπάρχει μια ποικιλία ελέγχων για να ελέγξει κανείς την κανονικότητα των δεδομένων. Αυτό που χρειαζόμαστε είναι μια γενίκευση της ιδέας σε μεγαλύτερες διαστάσεις.

Δυστυχώς τα πράγματα δεν είναι τόσο εύκολα στην πολυμεταβλητή περίπτωση. Στην πράξη οι έλεγχοι για την κανονικότητα βασίζονται είτε σε κάποιες χαρακτηριστικές ιδιότητες της κανονικής κατανομής είτε (οι περισσότεροι γνωστοί) στην εμπειρική συνάρτηση κατανομής. Στην πολυμεταβλητή περίπτωση δεν είναι εύκολο να ορίσει κανείς την εμπειρική συνάρτηση κατανομής. Αυτό αποτελεί επομένως ένα πρόσθετο πρόβλημα εκτός από την πολυπλοκότητα των πολλών διαστάσεων των δεδομένων.

Στην πράξη υπάρχουν πολλοί έλεγχοι για την πολυμεταβλητή κανονικότητα των δεδομένων. Οι περισσότεροι από αυτούς είναι ιδιαίτερα πολύπλοκοι και για αυτό δεν θα περιγραφούν σε αυτές τις σημειώσεις. Θα περιγράψουμε απλά κάποιες διαδικασίες περισσότερο διαγνωστικού χαρακτήρα.

Ένας πρώτος τρόπος είναι να δούμε τι γίνεται με τις περιθωριακές κατανομές. Ξέρουμε πως αν τα δεδομένα προέρχονται από πολυμεταβλητή κανονική τότε και κάθε μια μεταβλητή θα ακολουθεί από μόνη της μια κανονική κατανομή. Προσοχή, το αντίστροφο δεν ισχύει δηλαδή αν όλες οι περιθωριακές κατανομές είναι κανονικές αυτό δεν μας αποδεικνύει πως η από κοινού είναι πολυμεταβλητή και αυτή κανονική.

Επομένως μια πρώτη καθαρά διαγνωστική μέθοδος είναι να δούμε αν όλες οι περιθώριες κατανομές είναι κανονικές. Αν δεν είναι έχουμε αρκετά στοιχεία να πιστεύουμε πως και όλες μαζί δεν ακολουθούν πολυμεταβλητή κανονική. Αν όμως δεν απορρίψουμε την κανονικότητα για καμία μεταβλητή τότε δεν μπορούμε με σιγουριά να πούμε ότι η από κοινού κατανομή είναι πολυμεταβλητή κανονική.

Παρατηρείστε επίσης πως αν έχουμε πολλές μεταβλητές, δηλαδή η διάσταση της πολυμεταβλητής κανονικής είναι μεγάλη, τότε έχουμε το γνωστό πρόβλημα των πολλαπλών ελέγχων υποθέσεων και έτσι η πιθανότητα σφάλματος τύπου I είναι αρκετά μεγάλη. Για αυτό το λόγο αυτή η προσέγγιση καλό είναι να αποφεύγεται.

Για το παράδειγμα με τις 20 αμερικανικές πόλεις, κάνοντας τον έλεγχο κανονικότητας των Anderson-Darling βρήκαμε τα εξής p-value για τις 5 μεταβλητές: 0.400, 0.633, 0.586, 0.525, 0.165. Δηλαδή σε επίπεδο στατιστικής σημαντικότητας 5% δεν απορρίπτουμε την υπόθεση της κανονικότητας για καμία μεταβλητή. Δυστυχώς αυτό δεν είναι αρκετό για να μας δείξει και την πολυμεταβλητή κανονικότητα των δεδομένων μας.

Παράδειγμα: Έστω ότι η από κοινού κατανομή των x, y , δίνεται από την

$$f(x, y) = \frac{1}{2} f(x, y | \rho_1) + \frac{1}{2} f(x, y | \rho_2)$$

όπου

$$f(x, y | \rho_1) = \frac{1}{2\pi(1-\rho_1^2)^{1/2}} \exp \left\{ -\frac{1}{2(1-\rho_1^2)} (x^2 + y^2 - 2\rho_1 xy) \right\} \quad \text{και}$$

$$f(x, y | \rho_2) = \frac{1}{2\pi(1-\rho_2^2)^{1/2}} \exp \left\{ -\frac{1}{2(1-\rho_2^2)} (x^2 + y^2 - 2\rho_2 xy) \right\}$$

δηλαδή διμεταβλητές κανονικές κατανομές με μέσους μηδέν, διακυμάνσεις 1 και διαφορετικές συσχετίσεις. Μπορεί κανείς εύκολα να επαληθεύσει πως οι περιθώριες κατανομές είναι κανονικές αλλά η από κοινού ξεκάθαρα δεν είναι διμεταβλητή κανονική.

Ο δεύτερος έλεγχος βασίζεται στην ιδιότητα που είδαμε σχετικά με την κατανομή T^2 του Hotelling. Συγκεκριμένα επειδή οι παρατηρήσεις προέρχονται από $N_p(\mu, \Sigma)$ και ο δειγματικός πίνακας διακύμανσης ακολουθεί μια Wishart κατανομή μπορεί κανείς να δει πως η ποσότητα

$$\frac{n-p}{np} (x_i - \mu)' S^{-1} (x_i - \mu) \sim F(p, n-p)$$

Συνεπώς το πολυμεταβλητό πρόβλημα έχει μετασχηματισθεί σε ένα πρόβλημα να ελέγξουμε αν οι παρατηρήσεις μας προέρχονται από την κατανομή F. Εδώ θα πρέπει να σημειώσουμε τα εξής:

- Επειδή δεν γνωρίζουμε το διάνυσμα των μέσων χρησιμοποιούμε το διάνυσμα των δειγματικών μέσων. Αυτό οδηγεί σε κάποια απόκλιση από το γενικό αποτέλεσμα, η κατανομή δεν είναι πια F , αλλά, ευτυχώς, πολύ όμοια με την F .
- Μπορεί η παραπάνω σχέση να ισχύει για κάθε παρατήρηση αλλά οι μετασχηματισμένες τιμές δεν είναι ανεξάρτητες. Συνεπώς δεν έχουμε ένα ανεξάρτητο δείγμα από παρατηρήσεις.
- Ακόμα και για ανεξάρτητα δείγματα έλεγχοι για την καλή προσαρμογή της κατανομής F δεν είναι τόσο γνωστοί. Ουσιαστικά κάποιος μπορεί να καταφύγει σε ελέγχους Monte Carlo, είτε απλά σε κάποιο P-P plot για να δει αν οι μετασχηματισμένες τιμές ακολουθούν την κατανομή F .

Συνεπώς ο γραφικός τρόπος ελέγχου της πολυμεταβλητής κανονικότητας είναι ο εξής:

Έστω n το μέγεθος του δείγματος και p ο αριθμός των μεταβλητών

- Για κάθε παρατήρηση δημιουργήσε την τιμή

$$r_i = \frac{n-p}{np} (x_i - \bar{x})' S^{-1} (x_i - \bar{x}), i=1,2,\dots,n$$

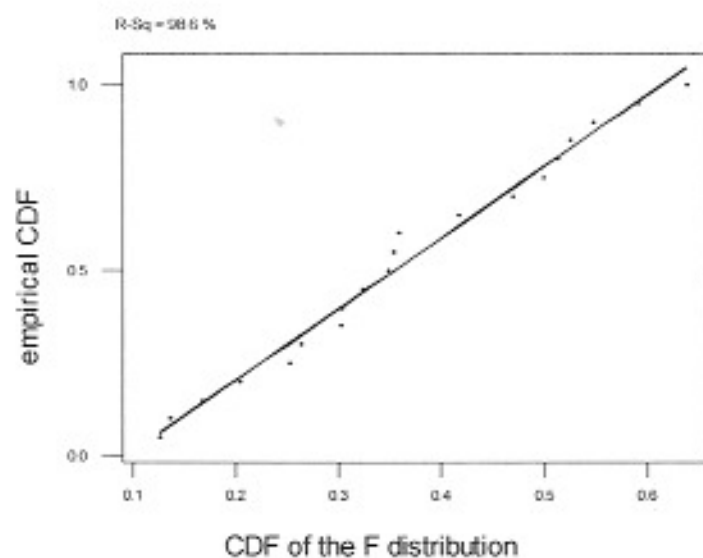
- Φτιάξε ένα P-P plot (ή εναλλακτικά ένα Q-Q plot) για τα r_i χρησιμοποιώντας την κατανομή F με βαθμούς ελευθερίας p και $n-p$. Για να γίνει αυτό αρκεί να βρεθεί η εμπειρική συνάρτηση κατανομής των r_i και οι αντίστοιχες θεωρητικές τιμές της κατανομής F .
- Αν τα σημεία είναι κοντά σε μια νοητή ευθεία στη διαγώνιο του γραφήματος (με κάποιες μικροαποκλίσεις φυσικά) τότε υπάρχουν ισχυρές ενδείξεις πως τα δεδομένα ακολουθούν πολυμεταβλητή κανονική κατανομή.

Θα πρέπει να σημειωθούν τα εξής:

- Η μέθοδος αυτή δεν δουλεύει καλά για μικρά μεγέθη δείγματος (κυρίως γιατί τότε η εξάρτηση ανάμεσα στις τιμές r_i είναι υψηλή). Επίσης δεν δουλεύει καλά για μεγάλες τιμές του p .
- Επειδή η κατανομή F όταν οι βαθμοί ελευθερίας του παρονομαστή είναι πολύ μεγάλος αριθμός προσεγγίζεται καλά από την κατανομή χ^2 , αυτό σημαίνει πως για μεγάλα δείγματα μπορεί να χρησιμοποιηθεί η κατανομή χ^2 με p βαθμούς ελευθερίας.

Παράδειγμα Ας δούμε πάλι το παράδειγμα των αμερικάνικων πόλεων. Θέλουμε να ελέγξουμε αν τα δεδομένα προέρχονται από πολυμεταβλητή κανονική κατανομή. Στην περίπτωση μας $p=5$, $n=20$. Βρίσκουμε τις τιμές και κατασκευάζουμε το P-P plot που βλέπετε στο γράφημα 5.1

Μπορεί να δει κανείς πως τα σημεία είναι πολύ κοντά στη διαγώνιο και επομένως από το γράφημα μοιάζει λογικό να υποθέσουμε την πολυμεταβλητή κανονικότητα των δεδομένων μας



Γράφημα 5.1. P-P plot για τα μετασχηματισμένα δεδομένα. Η διαγώνιος φαίνεται να προσεγγίζει καλά στα σημεία και αυτό αποτελεί μια καλή ένδειξη πως η υπόθεση της πολυμεταβλητής κανονικότητας είναι λογική

6 ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

6.1 Εισαγωγή

Η μέθοδος της Πολυμεταβλητής Ανάλυσης Διακύμανσης (Multivariate Analysis of Variance, MANOVA) αποτελεί γενίκευση της μονομεταβλητής ανάλυσης διακύμανσης όταν εξετάζουμε συγχρόνως περισσότερες από μια μεταβλητές. Αποτελεί επομένως μια μέθοδο ελέγχου του αν οι μέσοι δύο ή περισσότερων ομάδων διαφέρουν και γενικεύοντας σε περιπτώσεις πολλών παραγόντων αν οι παράγοντες αυτοί επιδρούν στη μέση τιμή (μιλάμε πια για διάνωση μέσων τιμών). Όπως θα δούμε και στη συνέχεια πολλά πράγματα που ισχύουν στη μονομεταβλητή περίπτωση μεταφέρονται με ανάλογο τρόπο και στην πολυμεταβλητή περίπτωση, όπως για παράδειγμα η διάσπαση της συνολικής διακύμανσης στην μεταξύ των γκρουπ και εντός των γκρουπ διακύμανση (between και within). Επίσης η MANOVA μπορεί να ειπωθεί σαν ένα πολυμεταβλητό γραμμικό μοντέλο. Στις σημειώσεις αυτές θα γίνει προσπάθεια να περιγραφεί η μέθοδος σε ένα επίπεδο εφαρμογών και όχι θεωρίας, αποφεύγοντας πολύπλοκες μαθηματικές αποδείξεις.

6.2 MANOVA ως προς έναν Παράγοντα

Θα αρχίσουμε εξετάζοντας την απλούστερη περίπτωση, ανάλυσης με έναν παράγοντα (one way).

Εστω ότι έχουμε ένα δείγμα $X_{i,j}$, $j=1,\dots,k$, και $i=1,\dots,n_j$, όπου γενικά j συμβολίζει το γκρουπ και i την παρατήρηση μέσα στο γκρουπ. Η παρατήρηση $X_{i,j}$, είναι πια ένα διάνωση με p στοιχεία και όχι μια απλή τιμή. Παρατηρείστε ότι το μέγεθος του δείγματος δεν είναι αναγκαστικά το ίδιο για όλα τα γκρουπ. Υποθέτουμε ότι οι πληθυσμοί για κάθε γκρουπ είναι $N_p(\mu_j, \Sigma)$, όπου p είναι η διάσταση (μιλάμε για μια πολυμεταβλητή κανονική κατανομή p

διαστάσεων), και μ και Σ είναι το διάνυσμα των μέσων και ο πίνακας διακύμανσης συνδιακύμανσης αντίστοιχα. Υποθέτουμε επίσης πως ο πίνακας διακύμανσης συνδιακύμανσης είναι ο ίδιος για όλα τα γκρουπ. Επομένως οι δύο υποθέσεις που χρειάζεται να κάνουμε στην MANOVA είναι

- Κανονικότητα των δεδομένων
- Ίδιος πίνακας διακύμανσης συνδιακύμανσης για όλα τα γκρουπ.

Η μηδενική υπόθεση που θέλουμε να ελέγξουμε είναι

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k \text{ έναντι της εναλλακτικής}$$

$$H_1: \text{τουλάχιστον δύο μέσοι διαφέρουν}$$

Προχωρώντας με το συνηθισμένο τρόπο κατασκευής ελέγχων με τη μέθοδο του λόγου πιθανοφανειών μπορούμε να βρούμε την πιθανοφάνεια κάτω από τις δύο υποθέσεις. Κάτω από τη μηδενική υπόθεση ο λογάριθμος της πιθανοφάνειας των δεδομένων είναι

$$-\frac{nkp}{2} \ln(2\pi) - \frac{nkp}{2} \ln|\Sigma| - \frac{1}{2} \sum_j \sum_i (x_{ij} - \mu_j)' \Sigma^{-1} (x_{ij} - \mu_j)$$

το οποίο αγνοώντας τον πρώτο όρο γράφεται ως

$$-\frac{nkp}{2} \ln|\Sigma| - \frac{1}{2} \text{trace} \Sigma^{-1} \left[\sum_j \sum_i (x_{ij} - \bar{x}_{\cdot j})(x_{ij} - \bar{x}_{\cdot j})' + n \sum_j (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})' \right]$$

Παρατηρείστε ότι τώρα μπορεί να γραφτεί στη μορφή

$$-\frac{nkp}{2} \ln|\Sigma| - \frac{1}{2} \text{trace} \Sigma^{-1} [B + W] \text{ όπου}$$

$$W = \sum_j \sum_i (x_{ij} - \bar{x}_{\cdot j})(x_{ij} - \bar{x}_{\cdot j})'$$

$$B = n \sum_j (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})'$$

$$\text{Όπως και στη μονομεταβλητή περίπτωση } \bar{x}_{\cdot j} = \frac{\sum_i x_{ij}}{n_j}, \bar{x}_{\cdot\cdot} = \frac{\sum_j \sum_i x_{ij}}{n}, \text{ δηλαδή ο μέσος}$$

του j γκρουπ και ο γενικός μέσος αντίστοιχα.

Δηλαδή παρατηρούμε ότι σπάσαμε σε δύο όρους τις τετραγωνικές αποκλίσεις όπως κάναμε και στην μονομεταβλητή περίπτωση. Ο πρώτος όρος που τον αποτελεί ο πίνακας W μας δίνει τις αποκλίσεις μέσα στο κάθε γκρουπ και αντιστοιχεί στο within sum of squares. Ο

δεύτερος όρος μας δίνει τις αποκλίσεις ανάμεσα στους μέσους των γκρουπ από τον γενικό μέσο και αντιστοιχεί στο between sum of squares. Θα πρέπει να σημειωθεί πως τα στοιχεία της διαγωνίου είναι τα γνωστά από τη μονομεταβλητή ανάλυση διακύμανσης αθροίσματα τετραγώνων που χρησιμοποιούνται για τους μονομεταβλητούς ελέγχους.

Δουλεύοντας ομοίως κάτω από την εναλλακτική υπόθεση βρίσκουμε πως ο λογάριθμος της πιθανοφάνειας είναι

$$-\frac{nkp}{2} \ln|\Sigma| - \frac{1}{2} tr \Sigma^{-1} [W]$$

και αυτό γιατί διασπώντας τις συνολικές τετραγωνικές αποκλίσεις παίρνουμε πως ο λογάριθμος της πιθανοφάνειας είναι

$$-\frac{nkp}{2} \ln|\Sigma| - \frac{1}{2} trace \Sigma^{-1} \left[\sum_j \sum_i (x_{ij} - \bar{x}_{.j})(x_{ij} - \bar{x}_{.j})' + n \sum_j (\bar{x}_{.j} - \bar{x}_{..})(\bar{x}_{.j} - \bar{x}_{..})' \right]$$

Ο δεύτερος όρος μέσα στην αγκύλη όμως είναι 0 οπότε παίρνουμε το αποτέλεσμα που θέλουμε.

Παρατήρηση: Στην πρώτη περίπτωση επειδή υποθέτουμε πως όλοι οι μέσοι είναι ίδιοι ο εκτιμητής μεγίστης πιθανοφάνειας του κοινού μέσου είναι ο μέσος όλου του δείγματος. Στην δεύτερη περίπτωση όμως, που τέτοια υπόθεση δεν κάνουμε, χρησιμοποιούμε τον εκτιμητή για κάθε γκρουπ και έτσι οδηγούμαστε στο αποτέλεσμα που είδαμε.

Έτσι οι μέγιστες τιμές των λογαριθμοποιημένων πιθανοφανειών είναι

$$-\frac{nk}{2} \ln|B+W| - \frac{np}{2} \quad \text{και} \quad -\frac{nk}{2} \ln|W| - \frac{np}{2}$$

για τη μηδενική και την εναλλακτική υπόθεση αντίστοιχα, οπότε καταλήγουμε πως ο λόγος των πιθανοφανειών LR είναι

$$LR = kn \ln \frac{|W|}{|B+W|}$$

Είναι επομένως λογικό να βασίσουμε τη συμπερασματολογία μας στο λόγο των δύο οριζουσών

και συγκεκριμένα στην ποσότητα $\Lambda = \frac{|W|}{|B+W|}$.

6.3 Έλεγχοι Υποθέσεων

Αν συμβολίσουμε με $T=B+W$ και θυμηθούμε πως η ορίζουσα ενός πίνακα διακύμανσης συνδιακύμανσης είναι ένα μέτρο της γενικής μεταβλητότητας μπορούμε να δούμε πως η ποσότητα αυτή είναι παρόμοια με το λόγο F που χρησιμοποιούμε στη μονομεταβλητή περίπτωση.

Βασισμένοι στην υπόθεση ότι τα δεδομένα μας προέρχονται από μια πολυμεταβλητή κανονική κατανομή και χρησιμοποιώντας τις ιδιότητες της κατανομής Wishart και κυρίως του τρόπου που αυτή προκύπτει από την πολυμεταβλητή κανονική κατανομή, έχειδειχτεί ότι η στατιστική συνάρτηση Λ ακολουθεί την κατανομή Λ -Wilks. Πιο συγκεκριμένα μπορεί ναδειχτεί ότι

$$W \sim W_p(\Sigma, n-k)$$

$$B \sim W_p(\Sigma, k-1)$$

Αν οι πίνακες B, W είναι ανεξάρτητοι (κάτι που ισχύει στην περίπτωση μας) τότε η συνάρτηση Λ ακολουθεί την κατανομή $\Lambda(p, n-k, k-1)$, χρησιμοποιώντας τα αποτελέσματα που είδαμε στο κεφάλαιο 4. συνεπώς ο έλεγχος θα στηριχτεί στην κατανομή Λ του Wilks και τις σχέσεις που τη συνδέουν με γνωστές κατανομές.

Στην πράξη, απορρίπτουμε την μηδενική υπόθεση όταν η τιμή του Λ είναι κοντά στο 0. Αυτό είναι λογικό, αν αναλογισθεί κανείς ότι μικρή τιμή Λ συνεπάγεται για τον πίνακα W των 'within διαφορών' ότι αυτές είναι μικρές σε σχέση με τις 'between διαφορές' του πίνακα B και άρα οι διαφορές ανάμεσα στις ομάδες είναι μεγάλες.

Παρατηρείστε ότι

$$\Lambda = \frac{1}{|W^{-1}| |B+W|} = \frac{1}{|I+W^{-1}B|} = \prod_{i=1}^p \frac{1}{\lambda_i}$$

όπου λ_i είναι οι ιδιοτιμές του πίνακα $I+W^{-1}B$.

Όμως

$$\begin{aligned} |I+W^{-1}B-\lambda I| &= 0 \Rightarrow |W^{-1}B-(\lambda-1)I| = 0 \Rightarrow \\ &\Rightarrow I = (\lambda-1)I \Rightarrow \lambda = I+1 \end{aligned}$$

όπου I είναι οι ιδιοτιμές του πίνακα $W^{-1}B$. Συνεπώς μπορούμε να γράψουμε πως

$\Lambda = \prod_{i=1}^p \frac{1}{I_i+1}$ και να βασιστούμε στις ιδιοτιμές αυτές για την κατασκευή εναλλακτικών ελέγχων.

Άλλοι έλεγχοι που έχουν προταθεί στη βιβλιογραφία είναι

1. Έλεγχος Roy. Θεωρούμε τη στατιστική συνάρτηση

$$F = \frac{n-k-p+1}{p} I_1 \quad \text{όπου } I_1 \text{ είναι η μεγαλύτερη ιδιοτιμή του πίνακα } W^{-1}B. \text{ Η στατιστική}$$

συνάρτηση F ακολουθεί την κατανομή F με βαθμούς ελευθερίας $|k-p|+1$ και $n-k-p+1$ αντίστοιχα. Απορρίπτουμε για μεγάλες τιμές της F .

2. Έλεγχος Pillai. Η στατιστική συνάρτηση είναι η $\text{tr}(B+W)^{-1}B = \sum_{j=1}^k \frac{1}{\lambda_j}$.

3. Έλεγχος ίχνους Lawley-Hotelling. Η στατιστική συνάρτηση είναι η $\text{tr}(W^{-1}B) = \sum_{j=1}^k I_j$.

Ενώ για τον έλεγχο του Roy καταλήγουμε σε γνωστή κατανομή, αυτό δεν ισχύει για τους άλλους δυο ελέγχους όπου η κατανομή δεν είναι γενικά γνωστή αν και υπάρχουν πίνακες.

Θα πρέπει να σημειωθεί πως αν οι υποθέσεις του μοντέλου μας είναι γενικά σωστές (πολυμεταβλητή κανονικότητα και σταθερή διακύμανση) όλοι οι έλεγχοι θα πρέπει να μας δίνουν ίδιο αποτέλεσμα, και p-value τα οποία να είναι κοντά. Για αυτό σε περίπτωση που παρατηρούμε μεγάλες αποκλίσεις στα αποτελέσματα χρησιμοποιώντας τους ελέγχους αυτό είναι μια ισχυρή ένδειξη πως θα πρέπει να κοιτάξουμε τα δεδομένα για το κατά πόσο η MANOVA μπορεί να εφαρμοστεί. Σε σχέση με την ισχύ των ελέγχων το ποιος από αυτούς είναι καλύτερος εξαρτάται από το σχεδιασμό του πειράματος. Σαν συμβουλή έλεγχοι που βασίζονται στη συνάρτηση Λ καλό είναι να αποφεύγονται όταν η υπόθεση της ισότητας των διακυμάνσεων δεν ισχύει. Ο έλεγχος Pillai έχει βρεθεί να είναι πιο ανθεκτικός (robust) όταν οι υποθέσεις δεν τηρούνται.

Στην περίπτωση που έχουμε 2 γρουπ όλα οι έλεγχοι είναι ισοδύναμοι με τον έλεγχο του Hotelling.

Θα πρέπει να τονιστεί πως, όπως και στην απλή ANOVA χρειαζόμαστε την υπόθεση ότι οι πίνακες διακυμάνσεων είναι ίσοι για να είναι αξιόπιστα τα αποτελέσματα. Χωρίς αυτή την υπόθεση η μεθοδολογία δεν ισχύει πια. Για να ελέγξουμε την υπόθεση ότι όλοι οι πίνακες διακύμανσης είναι ίσοι μπορούμε να χρησιμοποιήσουμε τον έλεγχο Box-M που περιγράψαμε στο κεφάλαιο 5.

Σε περίπτωση που απορρίψουμε την υπόθεση των ίσων πινάκων διακυμάνσεων, όπως και στην μονομεταβλητή περίπτωση, χρειάζεται να κάνουμε μετασχηματισμούς ώστε να μετασχηματίσουμε τα δεδομένα να έχουν ίσους πίνακες διακυμάνσεων. Αυτό σημαίνει πως πρέπει κανείς πρώτα να εξετάσει ποιες διακυμάνσεις ή συνδιακυμάνσεις του πίνακα είναι αυτές που δημιουργούν το πρόβλημα. Επειδή γενικά πολυμεταβλητοί μετασχηματισμοί δεν είναι καθόλου απλοί προτιμάται ο ερευνητής να αρχίσει από απλούς μετασχηματισμούς κάποιων από τις μεταβλητές που έχουν διαφορετικές διακυμάνσεις και στη συνέχεια να προσπαθήσει να αντιμετωπίσει το πρόβλημα με ταυτόχρονους μετασχηματισμούς σε όλον τον πίνακα διακυμάνσεων.

Πάντως ο έλεγχος αυτός είναι πολύ ευαίσθητος σε αποκλίσεις από την κανονικότητα. Δηλαδή όταν τα δεδομένα δεν προέρχονται από κανονικούς πληθυσμούς η ισχύς του ελέγχου είναι πολύ μικρή.

Ολοκληρώνοντας αυτή τη σύντομη παρουσίαση της μεθόδου MANOVA πρέπει να τονίσουμε πως μπορεί να γενικευτεί παρόμοια με την περίπτωση της απλής ANOVA. Δηλαδή μπορούμε να έχουμε δύο παράγοντες ή και περισσότερους, μπορούμε να ορίσουμε αλληλεπιδράσεις, μπορούμε να κάνουμε πολλαπλούς ελέγχους για να βρούμε ποιες ομάδες διαφέρουν μεταξύ τους κλπ. Επίσης μπορούμε να χρησιμοποιήσουμε πειραματικούς σχεδιασμούς σε συνδυασμό με MANOVA.

6.4 Πίνακας Ανάλυσης Διακύμανσης

Από όσα αναφέρθηκαν πιο πάνω η ιδέα της MANOVA είναι παρόμοια με αυτή της μονομεταβλητής περίπτωσης. Δηλαδή τη συνολική μεταβλητότητα την σπάμε σε δύο κομμάτια. Στην απλή ANOVA κυρίαρχο ρόλο την πρακτική εφαρμογή της μεθόδου έπαιξε ο πίνακας ανάλυσης διακύμανσης. Και στη MANOVA μπορούμε να κατασκευάσουμε τον αντίστοιχο πίνακα ανάλυσης διακύμανσης που έχει ως εξής

Πηγή	SSP πίνακας	Be	Ttest
Μεταξύ γρουπ (between)	$B = \sum_{j=1}^k n_j (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})'$	$k-1$	$\Lambda = \frac{ W }{ W+B }$
Μεσα στα γρουπ (within)	$W = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{\cdot j})(x_{ij} - \bar{x}_{\cdot j})'$	$n-k$	
Συνολική	$T = \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_{\cdot\cdot})(x_{ij} - \bar{x}_{\cdot\cdot})'$	$n-1$	

Παρατηρείστε πως οι διαφορές είναι ελάχιστες. Για παράδειγμα αντί να έχουμε αθροίσματα τετραγώνων τώρα πια έχουμε πίνακες με αθροίσματα τετραγώνων. Συνήθως αυτοί οι πίνακες ονομάζονται sum of squared cross products (ή SSCP). Αν κανείς ελέγξει τους τύπους με τους οποίους προκύπτουν οι πίνακες αυτοί θα δει πως είναι ακριβώς αντίστοιχοι με τα αθροίσματα τετραγώνων της ANOVA. Μάλιστα αποδεικνύεται πολύ εύκολα πως αν $p=1$, δηλαδή έχω μόνο μια μεταβλητή, τότε ο πίνακας ταυτίζεται με τον πίνακα της απλής ANOVA. Ακόμα και η ελεγχουσυνάρτηση Λ του Wilks σε αυτή την περίπτωση είναι ανάλογη με την τιμή της ελεγχουσυνάρτησης F για τον απλό έλεγχο.

σε πολλά φαινόμενα του πραγματικού κόσμου και σε στατιστικά προβλήματα συνυπάρχουν μεταβλητές οι οποίες μπορεί να σχετίζονται μεταξύ τους. Είναι επίσης συχνά επιθυμητή η προβλεψη της τιμής μιας μεταβλητής που εξετάζεται, αν είναι γνωστές οι τιμές άλλων μεταβλητών που πιθανολογείται ότι μπορούν να επιδράουν σε αυτήν.

Παραδείγματα πιθανής ύπαρξης συσχέτισης μεταξύ μεταβλητών υπάρχουν πάρα πολλά. Μερικά τέτοια είναι:

- Η επίδραση της ζήτησης ή και του κόστους της διαφημιστικής εκστρατείας μιας υπηρεσίας ή ενός προϊόντος στην τιμή παροχής ή πώλησης αντίστοιχα
- Οι επιδράσεις του χρόνου μελέτης και παρακολούθησης ενός μαθήματος από τους φοιτητές στην επιδοσή τους στο τελικό βαθμό
- Η επίδραση του εισοδήματος μιας οικογένειας ή και του διαθέσιμου χρόνου άδειας των γονέων στον αριθμό των ημερών που αφιερώνονται ανά έτος για διακοπές καδώς και πολλά άλλα.

(6ε)
130
Η στατιστική ανάλυση που εξετάζει την σχέση
εξάρτησης μεταξύ δύο ή περισσότερων μεταβλητών
και στοχεύει στην πρόβλεψη της τιμής μιας μεταβ-
λητής που επηρεάζεται από τις υπόλοιπες
ονομάζεται παλινδρόμηση

Η μεταβλητή τις οποίας οι τιμές επιδρούν στην
τιμή μιας άλλης μεταβλητής ονομάζεται
ανεξάρτητη ενώ η άλλη εξαρτημένη.

Οι ανεξάρτητες μεταβλητές μπορεί να είναι
μία ή και περισσότερες. Αν έχουμε μόνο μια
ανεξάρτητη μεταβλητή τότε έχουμε απλή παλιν-
δρόμηση ενώ όταν έχουμε δύο ή περισσότερες
τότε έχουμε πολλαπλή παλινδρόμηση

Απλή γραμμική παλινδρόμηση
Έχουμε μόνο μια ανεξάρτητη μεταβλητή X
και μια εξαρτημένη Y
Το μαθηματικό μοντέλο είναι:

$$Y = a + b \cdot X + \varepsilon$$

a : σταθερά της γραμμής παλινδρόμησης

b : η κλίση της -//- -//-

ε : το τυχαίο σφάλμα της παλινδρόμησης

- Ο συντελεστής α αποτελεί το σημείο τομής της γραμμής παλινδρόμησης με τον άξονα της εξαρτημένης μεταβλητής Y και εκφράζει την προβεπόμενη τιμή της Y όταν $x=0$
- Ο συντελεστής β αποτελεί την κλίση της γραμμής και είναι η μέση μεταβολή του Y που αντιστοιχεί σε μεταβολή της X κατά μια μονάδα.
Αν $b > 0$ υπάρχει θετική γραμμική συσχέτιση μεταξύ X και Y
Αν $b < 0$ έχουμε αρνητική γραμμική συσχέτιση μεταξύ των X και Y
- Το ϵ εφάλμα παλινδρόμησης αποτελεί την διαφορά μεταξύ της πραγματικής παρατήρησης της μεταβλητής Y από την προβεπόμενη που προκύπτει από το μοντέλο.
Θεωρείται ότι το εφάλμα έχει μέση τιμή μηδέν και σταθερή μεταβλητότητα σ^2 για όλες τις τιμές του X
- Οι παραδοχές για την δημιουργία και εφαρμογή του μοντέλου παλινδρόμησης είναι

1. Ανεξαρτησία: Οι επιμέρους παρατηρήσεις της μεταβλητής Y θεωρείται ότι είναι τυχαίες και ανεξάρτητες μεταξύ τους
 2. Κανονικότητα: Η στατιστική κατανομή της εξαρτημένης μεταβλητής είναι κανονική
 3. Γραμμικότητα Ο μέσος της κάθε κανονικής κατανομής της μεταβλητής Y_i ισούται με $y_i = \alpha_i + \beta_i x_i$
 4. Για κάθε συγκεκριμένη τιμή της ανεξάρτητης μεταβλητής αντιστοιχούν πολλές τιμές της εξαρτημένης μεταβλητής που κατανομούνται κανονικά. Στο δείγμα των παρατηρήσεων περιλαμβάνεται μόνο μια από αυτές
- Το πρόβλημα της απλής γραμμικής παλινδρόμησης διατυπώνεται ως εξής:
τα δεδομένα του προβλήματος αποτελούνται από n δειγματικά ζεύγη παρατηρήσεων $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$. Αναζητείται ευθεία που αντιπροσωπεύει όσο το δυνατό καλύτερα τα δεδομένα ζεύγη παρατηρήσεων. Το κάθε ζεύγος (x_i, y_i) δημιουργεί την βιναρτησιακή σχέση $y_i = \alpha_i + \beta_i x_i + \epsilon_i$

Η ανάλυση που χρησιμοποιείται για τον προσδιορισμό της καλύτερης δυνατής συνάρτησης που συνδέει τις μεταβλητές X και Y ονομάζεται μέθοδος ελαχίστων τετραγώνων. Η μέθοδος

αυτή ελαχιστοποιεί το άθροισμα των τετραγώνων των εφαλμάτων $\epsilon_i = Y_i - \hat{Y}_i$

$$SSE = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

όπου Y_i = παρατήρηση i και $\hat{Y}_i = a + b \cdot X_i$ η τιμή που εκτιμάται από την ευθεία παλινδρόμησης

Για τον προσδιορισμό των τιμών των a και b που ελαχιστοποιούν το άθροισμα των τετραγώνων των εφαλμάτων λύνουμε το σύστημα

$$\frac{\partial SSE}{\partial a} = 0 \quad \text{και} \quad \frac{\partial SSE}{\partial b} = 0$$

Από την επίλυση του παραπάνω συστήματος προσδιορίζονται οι τιμές των εκτιμητριών $\hat{a}$ και $\hat{b}$ των παραμέτρων a και b της γραμμής παλινδρόμησης

$$\hat{a} = \bar{Y} - \hat{b} \cdot \bar{X}$$

$$\hat{b} = \frac{n \sum_{i=1}^n X_i \cdot Y_i - \sum_{i=1}^n Y_i \cdot \sum_{i=1}^n X_i}{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2}$$

Παρατηρήσεις

(σελ 134)

- α) Για δεδομένη τιμή x_i της μεταβλητής X η εκτίμηση $\hat{y}_i = \hat{a} + \hat{b} \cdot x_i$ αναφέρεται στη μέση τιμή της μεταβλητής Y και όχι στην πραγματική της τιμή
- β) Οι προβλέψεις των τιμών της μεταβλητής Y από τις τιμές της μεταβλητής X μέσω της ευθείας ελαχίστων τετραγώνων πρέπει να γίνονται μόνο για τιμές της X που βρίσκονται μέσα ή έστω πολύ κοντά στο διάστημα τιμών που έχει πραγματοποιηθεί η συλλογή των δεδομένων
- γ) Η εξίσωση παλινδρόμησης δεν επιτρέπει την αντίστροφη πρόβλεψη, δηλαδή τον προσδιορισμό του X αν γνωρίζουμε τιμές του Y . Για το σκοπό αυτό απαιτείται ο εφ' αρχής προσδιορισμός της ευθείας ελαχίστων τετραγώνων της X ($X = \gamma + \delta \cdot Y + \epsilon$)

Παράδειγμα 11.1. Μία εμπορική εταιρεία επιθυμεί να εξετάσει τη διακριτή συνεισφορά που έχουν στα μηνιαία καθαρά έσοδα (σε €) ενός σημαντικού προϊόντος, οι α) δαπάνες διαφήμισης (σε €) και β) ο αριθμός των διαθέσιμων πωλητών. Για το σκοπό αυτό συγκεντρώθηκαν τα αντίστοιχα δεδομένα των τελευταίων 16 μηνών που καταγράφονται στον ακόλουθο πίνακα.

Μήνας	Έσοδα πωλήσεων	Διαφημιστική δαπάνη	Αριθμός πωλητών
1 ^{ος}	1050	160	32
2 ^{ος}	1260	285	47
3 ^{ος}	1470	540	23
4 ^{ος}	2160	261	68
5 ^{ος}	1950	360	32
6 ^{ος}	2400	690	17
7 ^{ος}	2370	495	58
8 ^{ος}	3150	948	75
9 ^{ος}	3570	720	98
10 ^{ος}	4410	1140	43
11 ^{ος}	4500	1395	76
12 ^{ος}	5610	1560	89
13 ^{ος}	5190	1380	108
14 ^{ος}	5670	1260	76
15 ^{ος}	5160	1710	65
16 ^{ος}	6840	1860	93

Αναλύστε ο προσδιορισμός και η ερμηνεία των δύο επιθετικών απλής γραμμικής παλινδρόμησης μεταξύ εσόδων από τις πωλήσεις και διαφημιστικής δαπάνης, πωλήσεων και αριθμού πωλητών αντίστοιχα.

Αναλυτική λύση

Για την ανάλυση της απλής γραμμικής παλινδρόμησης (μία εξαρτημένη και μια ανεξάρτητη μεταβλητή) χρησιμοποιείται η μέθοδος των ελαχίστων τετραγώνων, εφαρμόζοντας τη σχέση 11.3. Για τη διευκόλυνση των απαιτούμενων υπολογισμών με 'χαρτί και μολύβι' καταρτίζεται ο ακόλουθος πίνακας.

α) Πωλήσεις - Διαφημιστική δαπάνη

Μήνας	Πωλήσεις (Y)	Διαφήμιση (X)	Y · X	Y ²	X ²
1 ^{ος}	1050	162	170100	1102500	26244
2 ^{ος}	1260	285	359100	1587600	81225
3 ^{ος}	1470	540	793800	2160900	291600
4 ^{ος}	2160	261	563760	4665600	68121
5 ^{ος}	1950	360	702000	3802500	129600
6 ^{ος}	2400	690	1656000	5760000	476100
7 ^{ος}	2370	495	1173150	5616900	245025
8 ^{ος}	3150	948	2986200	9922500	898704
9 ^{ος}	3570	720	2570400	12744900	518400
10 ^{ος}	4410	1140	5027400	1944900	1299600
11 ^{ος}	4500	1395	6277500	20250000	1946025
12 ^{ος}	5610	1560	8751600	31472100	2433600
13 ^{ος}	5190	1380	7162200	26936100	1904400
14 ^{ος}	5670	1260	7144200	32148900	1587600
15 ^{ος}	5160	1710	8823600	26625600	2924100
16 ^{ος}	6840	1860	12722400	46785600	3459600
Σύνολα	56760	14766	66883410	251029800	11289944

Για τον υπολογισμό των εκτιμητριών $\hat{\beta}$ και $\hat{\alpha}$ των παραμέτρων β και α της γραμμής παλινδρόμησης από την εφαρμογή των σχέσεων 11.6 και 11.5 προκύπτει το ακόλουθο αποτέλεσμα.

$$\begin{aligned}\hat{\beta} &= \frac{n \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n Y_i \sum_{i=1}^n X_i}{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} = \\ &= \frac{16 \times 66883410 - 56760 \times 14766}{16 \times 18289944 - 14766^2} = 3,11 \\ \hat{\alpha} &= \bar{Y} - \hat{\beta} \bar{X} = \frac{56760}{16} - 3,11 \frac{14766}{16} = 677,4\end{aligned}$$

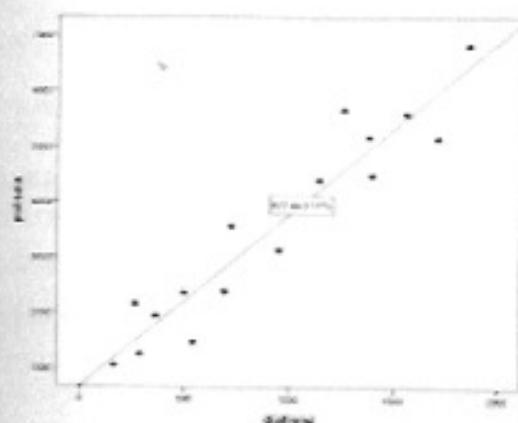
Επομένως η ευθεία ελαχίστων τετραγώνων του μοντέλου γραμμικής παλινδρόμησης των ετήσιων πωλήσεων και της διαφημιστικής δαπάνης είναι:

$$\hat{Y} = 677,4 + 3,11X$$

Από την παραπάνω εξίσωση προκύπτει ότι τα έσοδα των ετήσιων πωλήσεων εκτιμώνται σε 677,4 € αν η διαφημιστική δαπάνη είναι μηδενική, ενώ για κάθε

... που δαπανάται για διαφήμιση οι πωλήσεις θα αυξηθούν κατά 3,11 €. Συνεπώς, ότι η διαφημιστική δαπάνη είναι ιδιαίτερα αποδοτική (με απόδοση επένδυσης του 300%) και συνιστάται η αύξησή της.

Για τη καλύτερη αναπαράσταση της κατάστασης μπορεί να κατασκευαστεί γραφικά το ακόλουθο διάγραμμα διασποράς με την ευθεία παλινδρόμησης μεταξύ εσόδων πωλήσεων και διαφημιστικής δαπάνης.



Π.1: Διάγραμμα διασποράς εσόδων πωλήσεων και διαφημιστικής δαπάνης.

... η ανάλυση πραγματοποιείται και για την κατασκευή της ευθείας παλινδρόμησης των πωλήσεων και αριθμού πωλητών.

Πωλήσεις - Αριθμός πωλητών

Πωλήσεις (Y)	Πωλητές (Z)	Y · Z	Y ²	Z ²
1050	32	33600	1102500	1024
1260	47	59220	1587600	2209
1470	23	33810	2160900	529
2160	68	146880	4665600	4624
1950	32	62400	3802500	1024
2400	17	40800	5760000	289
2370	58	137460	5616900	3364
3150	75	236250	9922500	5625
3570	98	349860	12744900	9604
4410	43	189630	19448100	1849
4500	76	342000	20250000	5776
5610	89	499290	31472100	7921
5190	108	560520	26936100	11664
5670	76	430920	32148900	5776
5160	65	335400	26625600	4225
6840	93	636120	46785600	8649
56760	1000	4094160	251029800	74152

Από την εφαρμογή των αντίστοιχων σχέσεων (11.6 και 11.5) αντιστοίχως, εκτιμήτριας των παραμέτρων α και β αυτής της γραμμής παλινδρόμησης, θεωτώντας ως Z τη μεταβλητή του αριθμού των πωλητών – προκύπτουν ότι:

$$\hat{\beta} = \frac{n \sum_{i=1}^n Z_i Y_i - \sum_{i=1}^n Y_i \sum_{i=1}^n Z_i}{n \sum_{i=1}^n Z_i^2 - (\sum_{i=1}^n Z_i)^2} = \frac{16 \cdot 4094160 - 56760 \cdot 1000}{16 \cdot 74152 - 1000^2} = 46,92$$

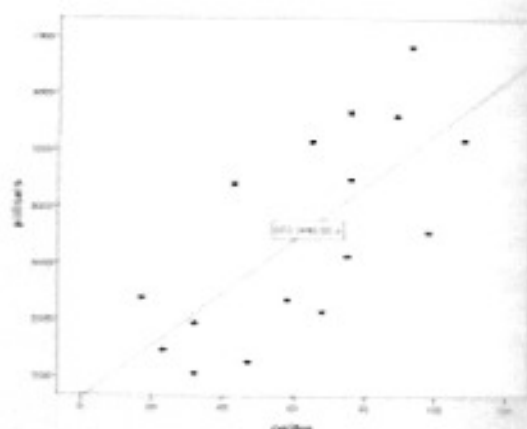
$$\hat{\alpha} = Y - \beta Z = \frac{56760}{16} - 46,92 \cdot \frac{1000}{16} = 615,3$$

οπότε η ευθεία ελαχίστων τετραγώνων του μοντέλου είναι η

$$\hat{Y} = 615,3 + 46,92 \cdot Z$$

Ερμηνεύοντας τις παραμέτρους της γραμμής παλινδρόμησης συνάγεται ότι, αν δεν χρησιμοποιηθεί κανείς πωλητής τα έσοδα από τις ετήσιες πωλήσεις κυμαίνονται σε ~615 €, ενώ η χρησιμοποίηση καθενός πωλητή θα τα αυξήσει κατά ~47 €. Μένει πλέον να συγκριθεί αυτή η συνεισφορά με το συνολικό μηνιαίο κόστος χρήσης ενός πωλητή για το συγκεκριμένο προϊόν προκειμένου να ληφθεί μια ορθολογική απόφαση.

Μπορεί τέλος να σχεδιαστεί το διάγραμμα διασποράς μεταξύ εσόδων πωλήσεων και αριθμού πωλητών (βλ. σχήμα 11.2).



Σχήμα 11.2: Διάγραμμα διασποράς εσόδων πωλήσεων και αριθμού πωλητών