Παλινδρόμηση του 12/12 

Ανάλυση Διακύμανσης

## Μέθοδος γραμμικών αντιθέσεων ή Μέθοδος Scheffe

### ΟΡΙΣΜΟΣ

Κάθε γραμμικός συνδυασμός των κύριων επιδράσεων $a_i, i=1..I$ των επιπέδων του παράγοντα στην εξαρτημένη μεταβλητή της μορφής
$L = \sum_{i=1}^{I} c_i a_i$ με $\sum_{i=1}^{I} c_i = 0$ λέγεται γραμμική αντίθεση

Η γραμμική αντίθεση επιτρέπει την διατύπωση συγκρίσεων περισσοτέρων των 2 $a_i$, $i=1,..,I$

π.χ. α) Μας ενδιαφέρει $a_i = a_{i'}$ (ΕΣΔ)

$$\left.\begin{array}{l} \text{Αν } c_k = 0 \quad k \neq i,j \\ \text{Αν } c_i = 1, \quad c_{i'} = 1 \end{array}\right\} \Rightarrow L = \sum_{i=1}^{I} c_i a_i = 0 \Leftrightarrow a_i = a_{i'}$$

β) $c_k = 0 \quad \forall k \neq i, i', l$

$c_i = 1, \quad c_{i'} = -\frac{1}{2} = c_l$

$L = \sum_{i=1}^{I} c_i a_i = a_i - \frac{1}{2} a_{i'} - \frac{1}{2} a_l$

$L = 0 \Leftrightarrow a_i = \frac{1}{2} a_{i'} + \frac{1}{2} a_l$

Η επίδραση του i-επιπέδου είναι ίση με το ημιάθροισμα των επιδράσεων των επιπέδων i και l.

Έχει ενδιαφέρον η κατασκευή στατιστικού τεστ για έλεγχο $H_0: L=0$.

Θεωρώ ότι οι υποθέσεις για τα σφάλματα ικανοποιούνται

δηλ: $\varepsilon_{ij} \sim N(0, \sigma^2)$ $i=1,\dots,I$ , $j=1,\dots,J_i$ και είναι ανεξάρτητα

λόγω των υποθέσεων για τα σφάλματα $\bar{Y}_{i.} \sim N(\mu+a_i, \frac{\sigma^2}{J_i})$ $i=1,\dots,I$

Με την κατά Wald προσέγγιση ένα τεστ για έλεγχο $H_0: L=0$

θα στηρίζεται σε ένα εκτιμητή της $L$

Θεωρώ τον εξής εκτιμητή της $L$: $\hat{L} = \sum_{i=1}^{I} c_i \hat{a}_i$ , $\hat{a}_i = \bar{Y}_{i.} - \bar{Y}_{..}$ $i=1,\dots,I$

$$\hat{L} = \sum_{i=1}^{I} c_i \hat{a}_i = \sum_{i=1}^{I} c_i (\bar{Y}_{i.} - \bar{Y}_{..}) = \sum_{i=1}^{I} c_i \bar{Y}_{i.} - \bar{Y}_{..} \sum_{i=1}^{I} c_i \Rightarrow \hat{L} = \sum_{i=1}^{I} c_i \bar{Y}_{i.}$$

Το $\hat{L}$ είναι γραμμικός συνδυασμός των ανεξάρτητων

$\bar{Y}_{i.} \sim N(\mu+a_i, \frac{\sigma^2}{J_i})$ , $i=1,\dots,I$ Άρα $\hat{L} \sim$ Normal

- $E(\hat{L}) = E(\sum_{i=1}^{I} c_i \bar{Y}_{i.}) = \sum_{i=1}^{I} c_i E(\bar{Y}_{i.}) = \sum_{i=1}^{I} c_i (\mu + a_i) = \mu \sum c_i + \sum_{i=1}^{I} c_i a_i \Rightarrow$

$\Rightarrow E(\hat{L}) = \sum_{i=1}^{I} c_i a_i = L$

- $Var(\hat{L}) = Var(\sum_{i=1}^{I} c_i \bar{Y}_{i.}) \xlongequal{\bar{Y}_{i.} \text{ ανεξ}} \sum_{i=1}^{I} c_i^2 Var\, \bar{Y}_{i.} = \sum_{i=1}^{I} c_i^2 \frac{\sigma^2}{J_i} \Rightarrow$

$\Rightarrow Var(\hat{L}) = \sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{J_i}$ $i=1,\dots,I$

Άρα $\hat{L} \sim N(L, \sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{J_i})$ $i=1,\dots,I$

Υπό την $H_0: L=0$ το $\hat{L} \sim N(0, \sigma^2 \sum_{i=1}^{I} \frac{c_i}{J_i})$

Υπό την $H_0: L=0$ το $\dfrac{\hat{L}}{\left(\sigma \sum_{i=1}^{I} \dfrac{c_i^2}{J_i}\right)^{1/2}} \sim N(0,1) \Rightarrow$

$$\Rightarrow \frac{\hat{L}^2}{\sigma^2 \sum_{i=1}^{I} \frac{c_i^2}{J_i}} \sim \chi_1^2 \text{ υπό την } H_0: L=0$$

Θυμίζεται ότι $\dfrac{SSres}{\sigma^2} \sim \chi^2_{N-I}$

Συμβολίζω $MS_L = \dfrac{\hat{L}^2}{\sum_{i=1}^{I} \dfrac{c_i^2}{J_i}}$ και θεωρώ το $F_L \overset{\text{ορ}}{=} \dfrac{MS_L}{MSres} =$

$$= \frac{\dfrac{\hat{L}^2}{\sum_{i=1}^{I} \frac{c_i^2}{J_i}}}{SSres/(N-I)} = \frac{\dfrac{\hat{L}^2}{\sigma^2 \sum c_i/J_i}}{\dfrac{SSres}{\sigma^2(N-I)}} \sim \frac{\chi_1^2}{\chi^2_{N-I}/(N-I)} \overset{\text{Ανεξ } MSres,\ MS_L}{=\!=\!=} F_{1,N-I}$$

υπο την $H_0$

Μορφή της κ.π. Μεγάλες τιμές του $F_L$ οδηγούν σε μεγάλες τιμές του $\hat{L}$ και επειδή $\hat{L}$ εκτιμητής του $L$, μεγάλες τιμές του $F_L$ οδηγούν σε απόρριψη της $H_0: L=0$

κ.π $F_L \geq c$

Προσδιορισμός $c$ : $a = P(\text{Απορ } H_0 \mid H_0 \text{ αληθ}) = P(F_L \geq c \mid F_L \sim F_{1,N-I}) =$

$$= P(F_{1,N-I} \geq c) \Rightarrow c = F_{1,N-I,a}$$

Συγκεντρωτικά για τον έλεγχο της $H_0: L=0$ χρησιμοποιείται η Σ.Σ.Τ

$F_L = \dfrac{MS_L}{MSres}$, $MS_L = \dfrac{\hat{L}^2}{\sum_{i=1}^{I} \dfrac{c_i^2}{J_i}}$ με κατανομή την $F_{1,N-I}$ υπό $H_0: L=0$

και κ.π μεγέθους $a$ την $F_L \geq F_{1,N-I,a}$

# Ανάλυση Υπολοίπων

Υπόλοιπα : $e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - (\hat{\mu} + \hat{a}_i) = Y_{ij} - (\bar{Y}_{..} + \bar{Y}_{i.} - \bar{Y}_{..}) \Rightarrow$

$\rightarrow e_{ij} = Y_{ij} - \bar{Y}_{i.} \quad i = 1, \dots, I \quad j = 1, \dots, J_i$

Τα $e_{ij}$ χρησιμοποιούνται όπως και στα μοντέλα παλινδρόμησης για τον έλεγχο ικανοποίησης των υποθέσεων για τα σφάλματα

Η ανάλυση διακύμανσης κατά ένα παράγοντα ως ειδική περίπτωση του γενικού γραμμικού μοντέλου $\underset{\sim}{Y} = X \underset{\sim}{b} + \varepsilon$

Μοντέλο Ανάλυσης Διακύμανσης κατά ένα παράγοντα

$$\underbrace{Y_{ij} = \mu_i + \varepsilon_{ij}}_{\mu_i \to b_i} \Longleftrightarrow Y_{ij} = \underbrace{\mu + a_i}_{b_i} + \varepsilon_{ij}, \quad i = 1, \dots, I, \; j = 1, \dots, J_i$$

$$\underset{\sim}{Y} = \underset{N \times 1}{\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1J_1} \\ Y_{21} \\ Y_{22} \\ \vdots \\ Y_{2J_2} \\ \vdots \\ Y_{I1} \\ Y_{I2} \\ \vdots \\ Y_{IJ_I} \end{pmatrix}} \qquad X = \underset{N \times I}{\begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}} \qquad \underset{\sim}{b} = \underset{I \times 1}{\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_I \end{pmatrix}}$$

③

$$\underset{\sim}{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1J_1} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2J_2} \\ \vdots \\ \varepsilon_{I1} \\ \varepsilon_{I2} \\ \vdots \\ \varepsilon_{IJ_I} \end{pmatrix}_{N\times 1}
\qquad \underset{\sim}{Y} = \underset{\sim}{b}X + \underset{\sim}{\varepsilon} \iff \begin{cases} Y_{ij} = \mu_i + \varepsilon_{ij}, & \text{αν } b_i = \mu_i \\ Y_{ij} = \mu + a_i + \varepsilon_{ij}, & \text{αν } b_i = \mu + a_i \end{cases}$$

## Άσκηση

Θεωρώ το μοντέλο ανάλυσης διακύμανσης κατά ένα παράγοντα.
$Y_{ij} = \mu + a_i + \varepsilon_{ij}$ $i = 1,\dots,I$, $j = 1,\dots,J_i$ με τα $\varepsilon_{ij} \sim N(0, \sigma^2)$
και ανεξάρτητα $i = 1,\dots,I$, $j = 1,\dots,J_i$

a) νδο $E(\hat{a}_i) = a_i$

b) νδο οι $EMΠ(\mu, a_i) = EET(\mu, a_i)$

## Λύση

a. $\hat{a}_i = \bar{Y}_{i.} - \bar{Y}_{..}$

$$E(\hat{a}_i) = E(\bar{Y}_{i.} - \bar{Y}_{..}) = E(\bar{Y}_{i.}) - E(\bar{Y}_{..}) = E\left(\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij}\right) - E\left(\frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} Y_{ij}\right) =$$

$$= \frac{1}{J_i}\sum_{j=1}^{J_i} E(Y_{ij}) - \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} E(Y_{ij}) = \frac{1}{J_i}\sum_{j=1}^{J_i} E(\mu + a_i + \varepsilon_{ij}) - \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i} E(\mu + a_i + \varepsilon_{ij})$$

$\underline{E(\varepsilon_{ij}) = 0}$ $$\frac{1}{J_i}\sum_{j=1}^{J_i}(\mu + a_i) - \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(\mu + a_i) = \frac{1}{J_i}(J_i\mu + J_i a_i) - \frac{1}{N}\left(N\mu + \underbrace{\sum_{i=1}^{I} J_i a_i}_{0}\right) =$$

λόγω περιορισμού

$= \mu + a_i - \mu = a_i$

β. Οι ΕΕΤ προκύπτουν από την ελαχιστοποίηση του $S = \sum_{i=1}^{I}\sum_{j=1}^{J_i} e_{ij}^2 =$

$$= \sum_{i=1}^{I}\sum_{j=1}^{J_i} (Y_{ij} - \mu - a_i)^2 \text{ ως προς } \mu \text{ και } a_i, \quad i = 1,\dots,I \qquad (1)$$

Οι ΕΜΠ προκύπτουν από τη μεγιστοποίηση της πιθανοφάνειας:

Πιθανοφάνεια $\equiv$ Από κοινού κατανομή των δεδομένων $Y_{ij}$ $= \prod_{i=1}^{I}\prod_{j=1}^{J_i} f_{Y_{ij}}(y_{ij}) \xrightarrow[Y_{ij} \sim N(\mu + a_i, \sigma^2)]{\varepsilon_{ij} \sim N(0, \sigma^2)}$

Άρα η πιθανοφάνεια $L = \frac{1}{(\sigma\sqrt{2\pi})^N} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \mu - a_i)^2}$

Οι ΕΜΠ προκύπτουν από μεγιστοποίηση της $L$ ως προς $\mu$, $a_i$

ή μεγιστοποίηση του $\log L$

ή μεγιστοποίηση του $-N\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \mu - a_i)^2$

ή μεγιστοποίηση του $-\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \mu - a_i)^2$

ή ελαχιστοποίηση ως προς $\mu$ και $a_i$ $i = 1,\dots,I$ του $\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij} - \mu - a_i)^2$ (2)

Από (1), (2) οι ΕΕΤ$(\mu, a_i) \equiv$ ΕΜΠ$(\mu, a_i)$

## <u>Μοντέλο Ανάλυσης Διακύμανσης κατά δύο παράγοντες</u>

Θεωρώ μια ποσοτική μεταβλητή $Y$ 2 παραγ.:

- $A$: $1, 2, \dots, I$ επίπεδα
- $B$: $1, 2, \dots, J$ επίπεδα

Η ανάλυση διακύμανσης κατά 2 παράγοντες διερευνά την σχέση ανάμεσα σε μια ποσοτική μεταβλητή Y και σε 2 παράγοντες (ποιοτικές μεταβλητές) A, B που ο A εμφανίζεται σε I-επίπεδα και ο B εμφανίζεται σε J-επίπεδα.

Εντοπίζεται ενδιαφέρον στην ανίχνευση των επιπέδων των δύο παραγόντων που ασκούν σημαντικότερη επίδραση στην εξαρτημένη μεταβλητή Y.

| Παράγοντας A | Παράγοντας B: 1 | 2 | ... | J | Σύνολα | |
|---|---|---|---|---|---|---|
| 1 | $Y_{11}$ | $Y_{12}$ | ... | $Y_{1J}$ | $Y_{1.}$ | $\bar{Y}_{1.}$ |
| 2 | $Y_{21}$ | $Y_{22}$ | ... | $Y_{2J}$ | $Y_{2.}$ | $\bar{Y}_{2.}$ |
| ⋮ | | | $Y_{ij}$ | ⋮ | | ⋮ |
| I | $Y_{I1}$ | $Y_{I2}$ | ... | $Y_{IJ}$ | $Y_{I.}$ | $\bar{Y}_{I.}$ |
| Σύνολα | $Y_{.1}$ | $Y_{.2}$ | ... | $Y_{.J}$ | $Y_{..}$ | |
| | $\bar{Y}_{.1}$ | $\bar{Y}_{.2}$ | ... | $\bar{Y}_{.J}$ | | $\bar{Y}_{..}$ |

Παριστά μια μέτρηση της Y υπό το επίπεδο i του A και j του B $\forall i=1,\dots,I,\ j=1,\dots,J$

Θεωρώ μια παρατήρηση ανά κελί δηλαδή μία παρατήρηση ανά δοκιμασία. Περισσότερες παρατηρήσεις ανά δοκιμασία οδηγούν σε μοντέλα με αλληλεπίδραση.

$$Y_{i.} = \sum_{j=1}^{J} Y_{ij}, \quad \bar{Y}_{i.} = \frac{1}{I}\sum_{i=1}^{J} Y_{ij}$$

$$Y_{.j} = \sum_{i=1}^{I} Y_{ij}, \quad \bar{Y}_{.j} = \frac{1}{I}\sum_{i=1}^{I} Y_{ij}$$

$$Y_{..} = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij} = \sum_{i=1}^{I} Y_{i.} = \sum_{j=1}^{J} {}_{.j}$$

$$\bar{Y}_{..} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij} = \frac{1}{I}\sum_{i=1}^{I} Y_{i.} = \frac{1}{J}\sum_{j=1}^{J} Y_{.j}$$

Γενικός Δειγματικός Μέσος

Μοντέλο

$$Y_{ij} = \mu + a_i + b_j + \varepsilon_{ij} \qquad \begin{matrix} i=1,\dots,I \\ j=1,\dots,J \end{matrix}$$

σφάλματα μοντέλου

παρατήρηση της εξαρτ. μετ. Y

κοινή επίδραση στην Y όλων των επιπέδων των 2 παραγόντων

Ατομική επίδραση στην Y του i-επιπέδου του παράγοντα A

Ατομική επίδραση στην Y του j-επιπέδου του B

Εκτιμητές Ελαχίστων τετραγώνων του μοντέλου

Προέρχονται από την ελαχιστοποίηση $S = \sum_{i=1}^{I}\sum_{j=1}^{J} \varepsilon_{ij}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\mu-a_i-b_j)^2$

$\frac{\partial S}{\partial \mu}=0$ , $\frac{\partial S}{\partial a_i}=0 \quad i=1,\dots,I$ $\quad \frac{\partial S}{\partial b_j}=0 \quad j=1,\dots,J$

Σύστημα κανονικών εξισώσεων

$$\begin{cases} IJ\mu + J\sum_{i=1}^{I} a_i + I\sum_{j=1}^{J} b_j = Y_{..} \\ J\mu + Ja_i + \sum_{j=1}^{J} b_j = Y_{i.}\ , \ i=1,\dots,I \\ I\mu + \sum_{i=1}^{I} a_i + Ib_j = Y_{.j}\ , \ j=1,\dots,J \end{cases}$$

∄ μοναδική λύση. Για να την πετύχω καταφεύγω στην υιοθέτηση πλευρικών συνθηκών

Διαισθητικά ένας εκτιμητής της μ πρέπει να είναι $\bar{Y}_{..}$

Πρέπει $\sum_{i=1}^{I} a_i = \sum_{j=1}^{J} b_j = 0$

Αν υιοθετήσω ως πλευρικές συνθήκες τις $\sum_{i=1}^{I} a_i = \sum_{j=1}^{J} b_j = 0$

Τότε οι μοναδικές λύσεις του συστήματος των κανονικών εξισώσεων

είναι $\hat{\mu} = \bar{Y}_{..}$ , $\hat{a}_i = \bar{Y}_{i.} - \bar{Y}_{..} \quad i=1,\dots,I$

$\hat{b}_j = \bar{Y}_{.j} - \bar{Y}_{..} \quad j=1,\dots,J$