Διάλεξη 8η 26/11/2019

Παλινδρόμηση και Ανάλυση Διακύμανσης

Ανάλυση Διακύμανσης κατά 1 παράγοντα $i=1,\dots,I$

$Y_{ij}=\mu+a_i+\varepsilon_{ij}$ $j=1,\dots,J_i$

ΕΕΤ: $\hat{\mu}=\overline{Y}_{..}$, $\hat{a}_i=\overline{Y}_{i.}-\overline{Y}_{..}$

Εκτιμώμενο μοντέλο: $\hat{Y}_{ij}=\hat{\mu}+\hat{a}_i=\overline{Y}_{i.}$

Πίνακας ΑΝΑΔΙΑ : $SS_{tot}=SS_{tr}+SS_{res}$

Υποθέσεις για τα Σφάλματα:

- (i) $E(\varepsilon_{ij})=0 \rightarrow E(Y_{ij})=\mu+a_i$
- (ii) $Var(\varepsilon_{ij})=\sigma^2 \rightarrow Var(Y_{ij})=\sigma^2$
- (iii) $Cov(\varepsilon_{ij},\varepsilon_{kl})=0 \rightarrow Cov(Y_{ij},Y_{kl})=0$
- (iv) $\varepsilon_{ij}\sim N(0,\sigma^2) \rightarrow Y_{ij}\sim N(\mu+a_i,\sigma^2)$

Ιδιότητες (Αν οι υποθ για τα σφάλματα ισχύουν)

(1) Οι ΕΕΤ $\hat{\mu}$ και $\hat{a}_i$, $i=1,\dots,I$ είναι αμερόληπτοι εκτιμητές των μ και a_i. Δηλ $E(\hat{\mu})=\mu$, $E(\hat{a}_i)=a_i$ $i=1,\dots,n$ (Αποδ Άσκηση)

(2) Οι ΕΕΤ των μ και a_i συμπίπτουν με τους ΕΜΠ. (Αποδ Άσκηση)

(3). ΘΕΩΡΗΜΑ: Αν οι υποθ. για τα σφάλματα ικανοποιούνται, τότε υπό την πλευρική συνθήκη $\sum_{i=1}^{I} J_i a_i=0$, ισχύουν: α) $E(MS_{res})=\sigma^2$

β) $E(MS_{tr})=\sigma^2+\frac{1}{I-1}\sum_{i=1}^{I}J_i a_i^2$.-

Αποδ α) (Ισχύει αν $W_1,\dots,W_n$ τδ από πληθυσμό με διακύμανση σ^2 τότε $E(S_W^2)=\sigma^2$ όπου $S_W^2=\frac{1}{n-1}\sum_{i=1}^{n}(W_i-\overline{W})^2$.)

Άρα αφού $Y_{i1}, Y_{i2}, \dots, Y_{iJ_i}$ είναι τυχ. δείγμα από πληθυσμό με διακύμανση σ^2.

$$E\left(\frac{1}{J_i-1}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i.})^2\right)=\sigma^2. \quad ⊛$$

$$MS_{res}=\frac{SS_{res}}{N-I}=\frac{1}{N-I}\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i.})^2$$

$$E(MS_{res})=\frac{1}{N-I}\sum_{i=1}^{I}\left[E\left(\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i.})^2\right)\right]$$

$$\overset{⊛}{=}\frac{1}{N-I}\sum_{i=1}^{I}\left[(J_i-1)\sigma^2\right]$$

$$=\frac{\sigma^2}{N-I}\sum_{i=1}^{I}(J_i-1)$$

$$=\frac{\sigma^2}{N-I}\left(\sum_{i=1}^{I}J_i-I\right)$$

$$=\frac{\sigma^2}{N-I}(N-I)$$

$$=\sigma^2.$$

Απδ

β) Από το μοντέλο $Y_{ij}=\mu+a_i+\varepsilon_{ij}$ και την πλευρική συνθήκη $\sum_{i=1}^{I}J_i a_i=0$

$$Y_{i.}=J_i\cdot\mu+J_i a_i+\varepsilon_{i.} \quad \text{και} \quad Y_{..}=N\mu+\varepsilon_{..}$$

$$E(Y_{i.})=J_i(\mu+a_i), \quad Var(Y_{i.})=J_i\sigma^2$$
$$E(Y_{..})=N\mu, \quad Var(Y_{..})=N\sigma^2.$$

$$E(SS_{tr})=E\left(\sum_{i=1}^{I}J_i(\bar{Y}_{i.}-\bar{Y}_{..})^2\right)=E\left(\sum_{i=1}^{I}J_i\left(\frac{Y_{i.}}{J_i}-\frac{Y_{..}}{N}\right)^2\right)$$

$$= E\left(\sum_{i=1}^{I} \frac{Y_{i.}^2}{J_i} - \frac{Y_{..}^2}{N}\right) = \sum_{i=1}^{I} \frac{1}{J_i} E(Y_{i.}^2) - \frac{1}{N} E(Y_{..}^2)$$

$$= \sum_{i=1}^{I} \frac{1}{J_i}\left(Var(Y_{i.}) + (EY_{i.})^2\right) - \frac{1}{N}\left(Var(Y_{..}) + (EY_{..})^2\right)$$

$$= \sum_{i=1}^{I} \frac{1}{J_i}\left(J_i\sigma^2 + J_i^2(\mu+a_i)^2\right) - \frac{1}{N}\left(N\sigma^2 + N^2\mu^2\right)$$

πράξεις

$$(I-1)\sigma^2 + \sum_{i=1}^{I} J_i a_i^2$$

Άρα $E(MS_{tr}) = E\left(\frac{1}{I-1} SS_{tr}\right)$

$$= \frac{1}{I-1} E(SS_{tr})$$

$$= \frac{1}{I-1}\left[(I-1)\sigma^2 + \sum_{i=1}^{I} J_i a_i^2\right]$$

$$= \sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} J_i a_i^2$$

Έλεγχος της $H_0: a_1 = a_2 = \dots = a_I$

Πρακτικά σημαίνει ότι όλα τα επίπεδα του παράγοντα ισοεπιδρούν στην εξαρτημένη μεταβλητή Y.

Για την κατασκευή τεστ για έλεγχο της Ho βολεύει να θεωρήσω την ισοδύναμη (διότι H_0 $a_i = a$, $i=1,\dots I \rightarrow Y_{ij} = \mu + a + \varepsilon_{ij} = \mu^* + \varepsilon_{ij}$ $H_0: a_i = 0$ $i=1,\dots,I \rightarrow Y_{ij} = \mu + \varepsilon_{ij}$)

$H_0: a_1 = a_2 = \dots = a_I = 0.$

$\Downarrow$

στηριζόμαστε στο γεγονός ότι $E(MS_{res}) = \sigma^2$ κ $E(MS_{tr}) = \sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} J_i a_i^2$

Προφανώς αν Ho αληθής τότε $E(MS_{res}) = E(MS_{tr})$

ή αν Ho αληθής τότε $MS_{res} \simeq MS_{tr}$

Μαθηματική Λογική: Αν $A \Rightarrow B$ τότε $\sim B \Rightarrow \sim A$.

Αν MSres πολύ διαφορετικό MStr τότε η Ho δεν είναι αληθής

Άρα, Αν MSres -//- MStr -//- Ho απορρ.

Άρα ένα τεστ θα βασιστεί στη σύγκριση των MSres κ MStr.

ΘΕΩΡΗΜΑ: (Διερευνά τις κατανομές SSres, SStr)

Αν οι υποθέσεις για τα σφάλματα ικανοποιούνται τότε υπό την Ho $\alpha_1=\alpha_2=\dots=\alpha_I=0$ έχουμε:

α) $\frac{SSres}{\sigma^2} \sim X^2_{N-I}$ β) $\frac{SStr}{\sigma^2} \sim X^2_{I-1}$.

Αποδ

Ισχύει ότι $Y_{ij} \sim N(\mu+\alpha_i, \sigma^2)$ $i=1,\dots,I$

Επομένως υπο Ho : $\alpha_1=\dots=\alpha_I=0$ τότε $j=1,\dots,J_i$

το $Y_{ij} \sim N(\mu, \sigma^2)$.

$$SSres = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (Y_{ij}-\bar{Y}_{i\cdot})^2$$

$$SStr = \sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2$$

α)

(Γνωρίζω Αν $W_1,\dots,W_n$ τδ από $N(\mu,\sigma^2)$ τότε $\frac{(n-1)S^2}{\sigma^2} \sim X^2_{n-1}$, $S^2=\frac{1}{n-1}\sum(W_2-\bar{W})^2$)

Επειδή $Y_{i1},\dots,Y_{iJ_i}$ είναι τδ από $N(\mu,\sigma^2)$ $\forall i$

$$\frac{\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i\cdot})^2}{\sigma^2} \sim X^2_{J_i-1} \quad \forall i=1,\dots,I$$

Y_{ij} ανεξαρτ. $\Rightarrow \sum_{i=1}^{I}\frac{\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i\cdot})^2}{\sigma^2} \sim X^2_{\sum_{i=1}^{I}(J_i-1)}$

$$\Rightarrow \frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i\cdot})^2}{\sigma^2} \sim X^2_{\sum^I J_i - I} \equiv X^2_{N-I}$$

β) Αν $W_1, \dots, W_n$ τδ από $N(\mu, \sigma^2)$

το $\overline{W} \sim N(\mu, \sigma^2/n)$

Επειδή $Y_{i1}, Y_{i2}, \dots, Y_{iJ_i}$ είναι τ.δ. από $N(\mu, \sigma^2)$

οπότε $\overline{Y}_{i\cdot} \sim N\left(\mu, \frac{\sigma^2}{J_i}\right) \quad \forall\, i=1,\dots,I$

$$\Rightarrow \frac{\overline{Y}_{i\cdot} - \mu}{\sqrt{\frac{\sigma^2}{J_i}}} \sim N(0,1) \quad \forall\, i=1,\dots,I$$

$$\frac{(\overline{Y}_{i\cdot} - \mu)\sqrt{J_i}}{\sigma} \sim N(0,1) \xrightarrow[\text{στο τετράγωνο}]{\text{υψώνω}} \frac{(\overline{Y}_{i\cdot} - \mu)^2 J_i}{\sigma^2} \sim N(0,1)^2 \equiv \chi^2_1$$

Y_{ij} ανεξ

$$\sum_{i=1}^{I} \frac{J_i (\overline{Y}_{i\cdot} - \mu)^2}{\sigma^2} \sim \chi^2_{\sum_{i=1}^{I} 1}$$

$$\sum_{i=1}^{I} \frac{J_i (\overline{Y}_{i\cdot} - \mu)^2}{\sigma^2} \sim \chi^2_I$$

$$\sum_{i=1}^{I} \frac{J_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2}{\sigma^2} \sim \chi^2_{I-1}$$

$$\frac{SStr}{\sigma^2} \sim \chi^2_{I-1}$$

Για τον έλεγχο της $H_0: \alpha_1 = \dots = \alpha_I = 0$

θεωρώ το $\boxed{F = \frac{MStr}{MSres}}$

SStr ανεξ SSres

$$= \frac{SStr/(I-1)}{SSres/(N-I)} = \frac{SStr/\sigma^2(I-1)}{SSres/\sigma^2(N-I)} \sim \frac{\chi^2_{I-1}/(I-1)}{\chi^2_{N-I}/(N-I)}$$

$\equiv F_{I-1, N-I}$

υπό την H_0.

και κρίσιμη περιοχή μεγάλες τιμές του F δηλ $F \geq c$.

Υπολογισμός του κρίσιμου σημείου c.

$$\alpha = P(\text{απορ } H_0 / H_0 \text{ αληθ}) = P(F \geq c \,/\, F \sim F_{I-1,N-I})$$
$$= P(F_{I-1,N-I} \geq c)$$
$$\Rightarrow c = f_{I-1,N-I}$$

Συγκεντρωτικά : Για τον έλεγχο της H_0 $\alpha_1 = \alpha_2 = \dots = \alpha_I = 0$
$(\Leftrightarrow H_0 : \alpha_1 = \dots = \alpha_I)$

χρησιμοποιείται η ΣΣΤ $F = \dfrac{MStr}{MSres}$ με κατανομή $F_{I-1,N-I}$

υπό H_0 και κ.π $F \geq F_{I-1,N-I,\alpha}$.

Στην πράξη αναμένουμε η H_0 να απορρ.
Αυτό σημαίνει ότι τα επίπεδα του παράγοντα δεν ασκούν όλα την ίδια επίδραση στην Y.
Αυτό σημαίνει ότι υπάρχουν επίπεδα που ασκούν σημαντικότερη επίδραση στην Y.
Άρα το ενδιαφέρον επικεντρώνεται στον εντοπισμό των επιπέδων που ασκούν τη σημαντικότερη επίδραση στην Y ή ακόμη και στην κατάταξη των επιπέδων από το σημαντικότερο προς το λιγότερο σημαντικό, για την Y.
Αυτό επιτυγχάνεται με την τεχνική των πολλαπλών συγκρίσεων.

ΠΟΛΛΑΠΛΕΣ ΣΥΓΚΡΙΣΕΙΣ

Ελαχίστης Σημαντικής (*) Διαφοράς

Tukey

Γραμμικών Αντιθέσεων (*)

Ελαχίστης Σημαντικής Διαφοράς (ΕΣΔ)

Πολλαπλά τεστ που ελέγχουν ανά δύο την ισότητα ή τη διαφορετικότητα των α_i $i=1,\dots,I$.

Δηλ. προχωρά σε ελέγχους
$H_0: \alpha_i = \alpha_{i'}$ έναντι $H_\alpha: \alpha_i \neq \alpha_{i'}$ για $i, i' = 1, \dots, I$, $i \neq i'$

Για τον έλεγχο αυτό και σύμφωνα με την τεχνική Wald στηριζόμαστε σε εκτιμητές των παραμέτρων που εμφανίζονται στην H_0 και ειδικά στους $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}$ και $\hat{\alpha}_{i'} = \bar{Y}_{i'\cdot} - \bar{Y}_{\cdot\cdot}$ και στη διαφορά τους

$$\hat{\alpha}_i - \hat{\alpha}_{i'} = \bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}$$

Προφανώς αν η $\hat{\alpha}_i - \hat{\alpha}_{i'}$ είναι μεγάλη το α_i διαφέρει από το $\alpha_{i'}$ και η $H_0: \alpha_i = \alpha_{i'}$ πρέπει να απορριφθεί.
Επομένως το τεστ θα στηριχτεί στην $\hat{\alpha}_i - \hat{\alpha}_{i'} = \bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}$ $i, i' = 1, \dots, I$ $i \neq i'$

Επειδή $Y_{i1}, \dots, Y_{iJ_i}$ τδ από $N(\mu + \alpha_i, \sigma^2)$

$$\left.\begin{array}{l} \text{το } \bar{Y}_{i\cdot} \sim N(\mu + \alpha_i, \sigma^2 / J_i) \\ \text{Όμοια } \bar{Y}_{i'\cdot} \sim N(\mu + \alpha_{i'}, \sigma^2 / J_{i'}) \end{array}\right\} \text{ Επειδή } \bar{Y}_{i\cdot} \text{ ανεξ } \bar{Y}_{i'\cdot}$$

$$\text{ή } \bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} \sim N\left(\alpha_i - \alpha_{i'}, \sigma^2\left(\frac{1}{J_i} + \frac{1}{J_{i'}}\right)\right)$$

$$\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot} \sim N\left(0, \sigma^2\left(\frac{1}{J_i} + \frac{1}{J_{i'}}\right)\right) \text{ υπό } H_0: \alpha_i = \alpha_{i'}$$

Λόγω του ότι $\dfrac{\bar{Y}_{i.} - \bar{Y}_{i'.}}{\sigma\sqrt{\frac{1}{J_i}+\frac{1}{J_{i'}}}} \sim N(0,1)$ και $\dfrac{SS_{res}}{\sigma^2} \sim X^2_{N-I}$.

$$\frac{\bar{Y}_{i.} - \bar{Y}_{i'.}}{\sqrt{MS_{res}}\sqrt{\frac{1}{J_i}+\frac{1}{J_{i'}}}} \sim t_{N-I} \text{ υπό την } H_0 : \alpha_i = \alpha_{i'}$$

Επομένως για τον έλεγχο της $H_0 : \alpha_i = \alpha_{i'}$ χρησιμοποιείται η

$$\frac{\bar{Y}_{i.} - \bar{Y}_{i'.}}{\sqrt{MS_{res}\left(\frac{1}{J_i}+\frac{1}{J_{i'}}\right)}} \sim t_{N-I} \text{ υπό } H_0$$

και κ.π. $\left|\dfrac{\bar{Y}_{i.} - \bar{Y}_{i'.}}{\sqrt{MS_{res}\left(\frac{1}{J_i}+\frac{1}{J_{i'}}\right)}}\right| \geq t_{N-I,\alpha/2}$

ή $|\bar{Y}_{i.} - \bar{Y}_{i'.}| \geq \sqrt{MS_{res}\left(\frac{1}{J_i}+\frac{1}{J_{i'}}\right)}\, t_{N-I,\alpha/2}$

ΣΥΝΟΨΗ της μεθόδου ΕΣΔ.

$$ΕΣΔ = t_{N-I,\alpha/2}\sqrt{MS_{res}\left(\frac{1}{J_i}+\frac{1}{J_{i'}}\right)}$$

Αν $|\bar{Y}_{i.} - \bar{Y}_{i'.}| \geq ΕΣΤ$ τότε η $H_0 : \alpha_i = \alpha_{i'}$ απορρ.

Αν απορρίψω την $H_0 : \alpha_i = \alpha_{i'}$ κάποιο από τα δύο ασκεί σημαντικότερη επίδραση στην Y.

Ποιο?

Αν $\bar{Y}_{i.} > \bar{Y}_{i'.}$ τότε το i-επίπεδο ασκεί σημαντικότερη επίδραση στην Y από το i.

Αν $\overline{Y}_{i\cdot} < \overline{Y}_{i'\cdot}$ τότε το i'-επίπεδο ασκεί σημαντικότερη επίδραση στην Y από το i.