Παλινδρόμηση και Ανάλυση Διακύμανσης

Υποθέσεις για τα σφάλματα $\underset{\sim}{\varepsilon}$

1) $E(\underset{\sim}{\varepsilon}) = \underline{0}$		1) $E(\underset{\sim}{Y}) = X\underset{\sim}{\beta}$
2) $Var(\underset{\sim}{\varepsilon}) = \sigma^2 I_n$	$\xrightarrow{\text{συνέπειες}}$	2) $Var(\underset{\sim}{Y}) = \sigma^2 I_n$
3) $\underset{\sim}{\varepsilon} \sim N_n(0, \sigma^2 I_n)$		3) $\underset{\sim}{Y} \sim N_n(X\underset{\sim}{\beta}, \sigma^2 I_n)$

Πολυδιάστατη κανονική κατανομή

Η τ.μ W έχει μονοδιάστατη κανονική κατανομή αν

$$f_W(w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(w-\mu)^2}$$

Το τ.δ $\underset{\sim}{W}$ έχει την n-διάστατη κανονική κατανομή και θα συμβολίζω

$$\underset{\sim}{W} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \sim N_n\left(\mu = E(W) = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \Sigma = Var(W) = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{pmatrix}\right)$$

αν η σ.π.π (ή από κοινού κατανομή των $W_1, \dots, W_n$) είναι

$$f_{\underset{\sim}{W}}(w) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{w}-\underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{w}-\underset{\sim}{\mu})}$$

όπου $\sigma_{ii} = Var(W_i) = \sigma_i^2$, $\sigma_{ij} = Cov(W_i, W_j)$, $\underset{\sim}{w} \in R^n$

Για να ορίζεται πρέπει $\Sigma > 0$ (θετικά ορισμένος)

Ιδιότητες της $N_n(\underset{\sim}{\mu}, \Sigma)$

1) Αν $\underset{\sim}{W} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \sim N_n(\underset{\sim}{\mu}, \Sigma)$ τότε $w_i \sim N(\mu_i, \sigma_{ii} = \sigma_i^2)$ $i = 1, \dots, n$

(Το αντίστροφο δεν ισχύει πάντα. Ισχύει μόνο αν οι w_i είναι ασυσχέτιστες δηλαδή $Cov(w_i, w_j) = 0$)

2) Αναλλοίωτο της $N_n(\underset{\sim}{\mu}, \Sigma)$

Αν $\underset{\sim}{W} \sim N_n(\underset{\sim}{\mu}, \Sigma)$ και A ένας $q \times n$ πίνακας τότε

$A\underset{\sim}{W} \sim N_q(A\underset{\sim}{\mu}, A\Sigma A')$ (ή $A\underset{\sim}{W} + \underset{\sim}{b} \sim N_q(A\underset{\sim}{\mu} + \underset{\sim}{b}, A\Sigma A')$ $\underset{\sim}{b} \in \mathbb{R}^q$

Ιδιότητες των ΕΕΤ $\hat{\underset{\sim}{\beta}}$

1) Οι Ε.Ε.Τ. $\hat{\underset{\sim}{\beta}} \sim N_{p+1}(\underset{\sim}{\beta}, \sigma^2(X^T X)^{-1})$

$\boxed{\text{Απόδειξη}}$

Δείξαμε ότι $E(\hat{\underset{\sim}{\beta}}) = \beta$ και $Var(\hat{\underset{\sim}{\beta}}) = \sigma^2 (X^T X)^{-1}$

Αρκεί νδο $\hat{\underset{\sim}{\beta}} \sim N_{p+1}$

Πράγματι $\hat{\beta} = \underbrace{(X^T X)^{-1} X^T}_{A \to (p+1) \times n} Y$

Άρα το $\hat{\underset{\sim}{\beta}}$ έχει $(p+1)$-διάστατη κατανομή ως γραμμικός μετασχηματισμός του $\underset{\sim}{Y}$ και επειδή $\underset{\sim}{Y} \sim N_n$ από την 3η υπόθεση για τα σφάλματα.

2) Οι ΕΕΤ $\hat{\underset{\sim}{\beta}}$ είναι ΑΟΕΑ

$\boxed{\text{ΘΕΩΡΗΜΑ}}$ Gauss - Markov

Οι ΕΕΤ $\hat{\underset{\sim}{\beta}}$ έχουν τη μικρότερη διακύμανση μεταξύ όλων των αμερόληπτων εκτιμητών της παραμέτρου $\underset{\sim}{\beta}$ που είναι γραμμικές συναρτήσεις των εξαρτημένων μεταβλητών $Y_1, \dots, Y_n$.

(Απόδειξη Απαιτητική και Παραλείπεται.)

3) Οι Ε.Ε.Τ $\hat{\underset{\sim}{\beta}}$ ταυτίζονται με τον ΕΜΠ της παραμέτρου $\underset{\sim}{\beta}$

Περαιτέρω ο ΕΜΠ της σ^2 είναι $\hat{\sigma}^2 = \frac{SSres}{n} = \frac{n-p-1}{n} MSres$

ΕΕΤ $\to$ ελαχιστοποίηση του $S = \sum_i^n \varepsilon_i^2 = \underset{\sim}{\varepsilon}^T \underset{\sim}{\varepsilon} = (\underset{\sim}{Y} - X\underset{\sim}{\beta})^T (\underset{\sim}{Y} - X\underset{\sim}{\beta})^{-1}$ (1)

$L =$ Από κοινού κατανομή δεδομένων $=$ Από κοινού κατανομή $\underset{\sim}{Y} =$

$$= f_{\underset{\sim}{Y}}(\underset{\sim}{y}) \overset{Y \sim N_n(X\underset{\sim}{\beta}, \sigma^2 I_n)}{=\!=\!=} \frac{1}{(2\pi)^{n/2} |\sigma^2 I_n|^{1/2}} e^{-\frac{1}{2}(\underset{\sim}{y} - X\underset{\sim}{\beta})^T (\sigma^2 I_n)^{-1} (\underset{\sim}{y} - X\underset{\sim}{\beta})} =$$

$$\overset{|aA| = a^n |A|}{=\!=\!=} \frac{1}{(2\pi)^{1/2} (\sigma^2)^{n/2} |I_n|} e^{-\frac{1}{2}(\underset{\sim}{y} - X\underset{\sim}{\beta})' \frac{1}{\sigma^2} I_n^{-1} (\underset{\sim}{y} - X\underset{\sim}{\beta})} \Rightarrow$$

$$\Rightarrow L = L(\underset{\sim}{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(\underset{\sim}{y} - X\underset{\sim}{\beta})^T (\underset{\sim}{y} - X\underset{\sim}{\beta})}$$

$$\log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(\underset{\sim}{y} - X\underset{\sim}{\beta})^T (\underset{\sim}{y} - X\underset{\sim}{\beta})$$

• Για την εύρεση των ~~[illegible]~~ Ε.Μ.Π του $\underset{\sim}{b}$ μεγιστοποιώ ως προς $\underset{\sim}{b}$ το $\log L$ ή μεγιστοποιώ ως προς $\underset{\sim}{b}$ το

$-\frac{1}{2\sigma^2}(y-Xb)^T(y-X\underset{\sim}{b})$ ή ελαχιστοποιώ ως προς $\hat{\underset{\sim}{b}}$ το

$(\underset{\sim}{y}-X\underset{\sim}{b})^T(\underset{\sim}{y}-X\underset{\sim}{b})$ (2)

Από (1), (2) προκύπτει ότι ΕΕΤ ≡ ΕΜΠ

• Ο Ε.Μ.Π της σ^2 προκύπτει από μεγιστοποίηση ως προς σ^2 του $\log L$ ή του $-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}(\underset{\sim}{y}-X\hat{\underset{\sim}{b}})^T(\underset{\sim}{y}-X\hat{\underset{\sim}{b}})$ ή του $-\frac{n}{2}\log\sigma^2-\frac{1}{2\sigma^2}SSres$ το οποίο μεγιστοποιείται για $\boxed{\hat{\sigma}^2=\frac{SSres}{n}}$

4) $\frac{SSres}{\sigma^2}\sim\chi^2_{n-p-1}$

Η απόδειξη απαιτεί γνώση θεωρίας κατανομών τετραγωνικών μορφών

5) Οι ΕΕΤ $\hat{b}$ είναι ανεξάρτητοι του MSres ή του SSres

Η απόδειξη ομοίως με πριν.

6) Για τον έλεγχο της $b_i=b_i^*$ (b_i^* γνωστό) $i=1,\dots,p$

Χρησιμοποιείται η ΣΣΤ $t_i=\frac{\hat{b}_i-b_i^*}{\sqrt{MSres}\sqrt{C_{i+1,i+1}}}$

με $t_i \sim t_{n-p-1}$ υπό την H_0 και κ.π. $|t_i| \geq t_{n-p-1,\frac{a}{2}}$ $i=1,..,p$

όπου $c_{i+1,i+1}$ το $(i+1)$-διαγώνιο στοιχείο του πίνακα $(X^TX)^{-1}$

Παρατήρηση

Έχει ενδιαφέρον πρακτικό η περίπτωση $b_i^*=0$ γιατί αν δεν μπορώ να απορρίψω την $H_0: b_i^*=0$ σημαίνει ότι η αντίστοιχη ανεξάρτητη μεταβλητή X_i, $i=1,..,p$ δεν συνδέεται γραμμικά με την Y.

Επειδή $\underset{\sim}{\hat{b}} \sim N_{p+1}\left(\underset{\sim}{b}, \sigma^2 (X^TX)^{-1}\right)$ $\quad \underset{\sim}{\hat{b}} = \begin{pmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \hat{b}_p \end{pmatrix}$

κάθε συνιστώσα έχει μονοδιάστατη κανονική Άρα

$\hat{b}_i \sim N(b_i, \sigma^2 c_{i+1,i+1})$, $i=1,..,p$ και υπό την $H_0: b_i = b_i^*$

$\hat{b}_i \sim N(b_i^*, \sigma^2 c_{i+1,i+1})$, $i=1,..,p$ $\Rightarrow$

Υπό την $H_0: b_i = b_i^*$ $\quad \dfrac{\hat{b}_i - b_i^*}{\sigma\sqrt{c_{i+1,i+1}}} \sim N(0,1)$

$$\text{Θεωρώ} \quad \frac{\dfrac{\hat{b}_i - b_i^*}{\sigma\sqrt{c_{i+1,i+1}}}}{\sqrt{\dfrac{SSres}{\sigma^2}/(n-p-1)}} \sim \frac{N(0,1)}{\sqrt{\chi^2_{n-p-1}/(n+p-1)}} = t_{n-p-1}$$

$$\Rightarrow \frac{\hat{b}_i - b_i^*}{\sqrt{MSres}\sqrt{c_{i+1,i+1}}} = t_i$$

Άρα καταλήγει ότι υπό $H_0: b_i = b_i^*$ $\quad t_i = \dfrac{\hat{b}_i - b_i^*}{\sqrt{MS_{res}}\sqrt{c_{i+1,i+1}}} \sim t_{n-p-1}$

Η μορφή της κ.π είναι μεγάλες τιμές του t_i δηλ $|t_i| > c$

$a = P(\text{Απορ } H_0 \,/\, H_0 \text{ αληθ.}) \Rightarrow \ldots \Rightarrow c = t_{n-p-1, \frac{a}{2}}$

7) Για τον έλεγχο της $H_0: b_1 = b_2 = \ldots = b_p = 0$ (δηλαδή του ελέγχου ύπαρξης του μοντέλου της Π.Γ.Π) χρησιμοποιούμε ως Σ.Σ.Τ. την $F = \dfrac{MS_{reg}}{MS_{res}} \sim F_{p, n-p-1}$ υπό την H_0

και κ.π. μεγέθους a $\quad F \geq F_{p, n-p-1, a}$

Γενική Γραμμική Υπόθεση

Θεωρούμε το μοντέλο Π.Γ.Π $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ $\quad \underset{\sim}{\varepsilon} \sim N_n(0, \sigma^2 I_n)$

με τον όρο γενική γραμμική υπόθεση εννοούμε $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$ όπου A ένας $q \times (p+1)$ πίνακας γνωστός και $\underset{\sim}{c}$ ένα q-διάστατο διάνυσμα γνωστό

Ερώτημα: Γιατί μια τόσο γενική H_0?

Γιατί μπορεί να εφαρμοστεί για κατάλληλα A και $\underset{\sim}{c}$ ώστε να εκφράζει χρήσιμες υποθέσεις

Παράδειγμα

1. Έστω $A = I_{p+1}$ και $\underset{\sim}{c} = \underset{\sim}{0}$ τότε $H_0 : A\underset{\sim}{b} = \underset{\sim}{c} \iff H_0 : b_i = 0 \ \forall i$
2. Έστω $A = I_{p+1}$ και $\underset{\sim}{c} \neq \underset{\sim}{0}$ τότε $H_0 : A\underset{\sim}{b} = \underset{\sim}{c} \iff H_0 : b_i = b_i^*$

   όπου b_i^* το i-στοιχείο του $\underset{\sim}{c}$
3. $H_0 : A\underset{\sim}{b} = \underset{\sim}{c} \iff H_0 \ \underset{\sim}{b_\mu} = \underset{\sim}{0}$ όπου $\underset{\sim}{b_\mu}$ ένα υπό-διάνυσμα από συνιστώσες του $\underset{\sim}{b}$ π.χ $\underset{\sim}{b_\mu} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$

   Τότε $A = \begin{pmatrix} 0 & & & \\ & 1 & & 0 \\ & & 1 & \\ & 0 & & 0 \ddots 0 \end{pmatrix}$ $\underset{\sim}{c} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

   $H_0 : A\underset{\sim}{b} = \underset{\sim}{c} \iff H_0 : b_1 = b_2 = 0$ (πρακτική αξία: Αν την δεχτώ οι X_1, X_2 δεν συνδέονται γραμμικά με την y)

Περιγραφή Τεστ.

Για την κατασκευή του τεστ θεωρώ το πλήρες μοντέλο (Π.Μ)

$$y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_p X_{ip} + \varepsilon_i, \ i = 1, \dots, n$$

και Ho μοντέλο (Ho.Μ). Δηλαδή το μοντέλο το οποίο προκύπτει από το Π.Μ υπό την $H_0 : A\underset{\sim}{b} = \underset{\sim}{c}$

π.χ $A = \begin{pmatrix} 0 & & \\ & 1 & 0 \\ & & 1 \ddots \\ 0 & & 0 \end{pmatrix}$, $\underset{\sim}{c} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

Τότε $H_0 : A\underset{\sim}{b} = \underset{\sim}{c} \iff \begin{pmatrix} 0 & & \\ & 1 & 0 \\ & & 1 \\ 0 & & \ddots 0 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \Rightarrow b_1 = b_2 = 0$

Άρα το ΗοΜ : $Y_i = b_0 + b_3 X_{i3} + b_4 X_{i4} + \dots + b_p X_{ip} + \varepsilon_i \quad , \; i=1,\dots,n$

Έστω $\hat{\underset{\sim}{b}}$ οι ΕΕΤ του ΠΜ και έστω $\hat{\underset{\sim}{b}}_{HoM}$ οι ΕΕΤ του ΗοΜ.

Τότε $\hat{\underset{\sim}{Y}} = X\hat{\underset{\sim}{b}}$ και $\hat{Y}_{HoM} = X\hat{\underset{\sim}{b}}_{HoM}$ αντίστοιχα.

Μπορώ να βρω το $SS_{res}(ΠΜ) \overset{ορ}{=} (\underset{\sim}{Y} - \hat{\underset{\sim}{Y}})^T (\underset{\sim}{Y} - \hat{\underset{\sim}{Y}})$

— " — $SS_{res}(HoM) \overset{ορ}{=} (\underset{\sim}{Y} - \hat{\underset{\sim}{Y}}_{HoM})^T (\underset{\sim}{Y} - \hat{\underset{\sim}{Y}}_{HoM})$

Θεωρώ την διαφορά $SS_{res}(HoM) - SS_{res}(ΠΜ)$

→ Αν η διαφορά είναι μικρή $\Rightarrow SS_{res}(HoM) \simeq SS_{res}(Π.Μ) \Rightarrow$

$\Rightarrow$ Το HoM συμπεριφέρεται ανάλογα με το πλήρες μοντέλο $\Rightarrow$

$\Rightarrow$ Δεν μπορώ να απορρίψω $H_0 : A\underset{\sim}{b} = \underset{\sim}{c}$

→ Άρα μεγάλη τιμή της διαφοράς οδηγεί σε απόρριψη της H_0

<u>Αλγεβρικά αποδεικνύεται ότι:</u>

$$\boxed{SS_{res}(HoM) - SS_{res}(ΠΜ) = (A\hat{\underset{\sim}{b}} - \underset{\sim}{c})^T \left[A(X^TX)^{-1}A^T\right]^{-1} (A\hat{\underset{\sim}{b}} - \underset{\sim}{c})}$$

<u>Στατιστικά αποδεικνύεται</u>

$$\boxed{\frac{SS_{res}(HoM) - SS_{res}(ΠΜ)}{\sigma^2} \sim \chi^2_q \quad \text{υπό την } H_0 : A\underset{\sim}{b} = \underset{\sim}{c}}$$

Η $\dfrac{SS_{res}(ΠΜ)}{\sigma^2} \sim \chi^2_{n-p-1}$

$\Rightarrow$

Θεωρώ $F = \dfrac{\left(\dfrac{SSres(Ho.Μ) - SSres(Π.Μ)}{\sigma^2}\right)/q}{\left(\dfrac{SS(Π.Μ)}{\sigma^2}\right)/(n-p-1)} =$

$$= \frac{(n-p-1)\left(SSres(Ho.Μ) - SSres(Π.Μ)\right)}{q\, SSres(Π.Μ)} \Rightarrow$$

$$\Rightarrow F = \frac{(n-p-1)(A\hat{\underset{\sim}{b}} - \underset{\sim}{c})^T \left[A(X^TX)^{-1}A^T\right]^{-1} (A\hat{\underset{\sim}{b}} - \underset{\sim}{c})}{q\, SSres}$$ όπου $SSres = SSres(Π.Μ)$

Προφανώς $F \sim F_{q,n-p-1}$ υπό την Ho

Συνοπτικά

Για τον έλεγχο της γενικής γραμμικής υπόθεσης $Ho: A\underset{\sim}{b} = \underset{\sim}{c}$

η Σ.Σ.Τ. είναι $F = \dfrac{(n-p-1)(A\hat{\underset{\sim}{b}} - \underset{\sim}{c})^T \left[A(X^TX)^{-1}A^T\right]^{-1} (A\hat{\underset{\sim}{b}} - \underset{\sim}{c})}{q\, SSres}$

με κατανομή την $F_{q,n-p-1}$ υπό Ho και κ.π με ε.σ. a

την $F \geq F_{q,n-p-1,a}$.