Δευτέρα 16/11/20   7º μάθημα

ΘΔΟ $SS_{reg} = \hat{\beta}^T X^T Y - n\bar{Y}^2$.

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum (\hat{Y}_i^2 - 2\hat{Y}_i\bar{Y} + \bar{Y}^2) = \sum \hat{Y}_i^2 - 2\bar{Y}\sum \hat{Y}_i + n\bar{Y}^2$$

$$= \hat{Y}^T\hat{Y} + n\bar{Y}^2 - 2\bar{Y}\sum\hat{Y}_i \xrightarrow{\hat{Y}=X\hat{\beta}} \hat{\beta}^T X^T X \beta + n\bar{Y}^2 - 2\bar{Y}\sum\hat{Y}_i$$

Αλλά $\hat{\underset{\sim}{\beta}}^T (X^T X)^{-1} Y = \hat{\underset{\sim}{\beta}}^T (X^T X)^{9} (X^T X)^{-1} X^T \underset{\sim}{Y}) = \hat{\underset{\sim}{\beta}}^T X^T \underset{\sim}{Y}$.

Άρα $SS_{reg} = \hat{\underset{\sim}{\beta}}^T X^T \underset{\sim}{Y} + n\bar{Y}^2 - 2\bar{Y}\sum\hat{Y}_i$ (1) Αναλύουμε το $\sum\hat{Y}_i$

Το $\underset{\sim}{\beta}$ ικανοποιεί τις κανονικές εξισώσεις $X^T X \beta = X^T \underset{\sim}{Y}$

$\Rightarrow X^T(X\underset{\sim}{\hat{\beta}} - \underset{\sim}{Y}) = \underset{\sim}{0} \Rightarrow [X^T(X\underset{\sim}{\beta} - \underset{\sim}{Y})]^T = \underset{\sim}{0}^T$

$\Rightarrow (X\underset{\sim}{\tilde{\beta}} - \underset{\sim}{Y})^T X = 0 \Rightarrow (\underset{\sim}{\hat{Y}} - \underset{\sim}{Y})^T X = \underset{\sim}{0}$

$$\Rightarrow \underbrace{(\hat{Y}_1 - Y_1, \hat{Y}_2 - Y_2, \dots \hat{Y}_n - Y_n)}_{1\times n} \underbrace{\begin{pmatrix} 1 & X_{11} & X_{12} & \dots & X_{1n} \\ 1 & X_{21} & X_{22} & & X_{2n} \\ & & & \vdots & \\ 1 & X_{n1} & X_{n2} & & X_{nn} \end{pmatrix}}_{n\times n+1} = \underset{\sim}{0}^T = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}^T$$

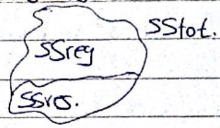
$\Rightarrow \sum_{i=1}^{n} (\hat{Y}_i - Y_i) = 0 \Rightarrow \sum\hat{Y}_i = \sum Y_i \Rightarrow \sum\hat{Y}_i = n\bar{Y}$

Τότε (1) $\Rightarrow SS_{reg} = \underset{\sim}{\hat{\beta}} X^T \underset{\sim}{Y} + n\bar{Y}^2 - 2\bar{Y}n\bar{Y} =$

$$\Rightarrow \boxed{SS_{reg} = \underset{\sim}{\hat{\beta}} X^T \underset{\sim}{Y} - n\bar{Y}^2}$$

Επομένως και στο πολλαπλό γραμμικό μοντέλο ισχύει ότι

$$\boxed{SStot = SSreg + SSres}$$



Πίνακας Ανάλυσης

Πηγή Μεταβλ.	SS	β.ε	MS.	F-πηλίκο
μοντέλο π.γ.π	SSreg.	p	$MSreg = SSreg/p$.	$F = \frac{MSreg}{MSres}$
Υπόλοιπα.	SSres	n-1-p.	$MSres = SSres/(n-1-p)$	
Ολική μεταβλητότητα	SStot.	n-1		

Οι παρατηρήσεις είναι παρόμοιες με το απλό γραμμικό μοντέλο

Αν $SSreg >> SSres$ ⟶ Υποσχόμενο μοντέλο (δεν το υιοθετώ, ελέγχω

Αν $MSreg >> MSres$ και άλλα πράγματα αλλά έχω μια ιδέα ότι

Αν η τιμή του F >> πολύ μεγάλη μάλλον θα είναι αυτό που θέλω)

Πρόταση: Υπό τις υποθέσεις (ι1, ι2) για τα σφάλματα, το MSres είναι αμερόληπτος εκτιμητής της σ^2, δηλ. $E(MSres) = \sigma^2$

Συντελεστής Προσδιορισμού Προσαρμοστικότητας

Ίδια ερμηνεία με τον συντελεστή προσδιορισμού στο α.γ.μ

$$R^2 = \frac{SSreg}{SStot} = 1 - \frac{SSres}{SStot}.$$

- R^2 είναι καθαρός αριθμός
- $R^2 \in [0,1]$.

Τα στατιστικά πακέτα χρησιμοποιούν διαφοροποιημένους τύπους:

Μια βελτιωμένη μορφή για το R^2 είναι ο $1 - \frac{MSres}{MStot}$.

Περαιτέρω Μελέτη του μοντέλου π.γ.π.
Έστω $\underset{\sim}{\varepsilon} \sim N_n(\underset{\sim}{0}, \sigma^2 I_n)$ ικανοποιείται

Β1) Θεώρημα Gauss-Markov

Οι ΕΕΤ έχουν την μικρότερη διακύμανση μεταξύ όλων των άλλων αμερόληπτων εκτιμητών της $\underset{\sim}{\beta}$ που είναι γραμμικές συναρτήσεις των εξαρτημένων μεταβλητών $Y_1, Y_2, \dots, Y_n$.

Β2) Αν $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{0}, \sigma^2 I_n)$ τότε οι ΕΕΤ $\hat{\underset{\sim}{\beta}} \sim N_{p+1}(\underset{\sim}{\beta}, \sigma^2 (X^T X)^{-1})$

Απόδ.

Ισχύει Αν $\underset{\sim}{W} \sim N_n(\underline{\mu}, \Sigma)$ τότε $\underset{a \times n}{A} \underset{n \times 1}{W} \sim N_a(A\underline{\mu}, A \Sigma A^{-1})$

$$\left.\begin{aligned} \hat{\underset{\sim}{\beta}} &= (X^T X)^{-1} X^T Y \\ \underset{\sim}{Y} &\sim N_n(X\underset{\sim}{\beta}, \sigma^2 I_n) \end{aligned}\right\} \Rightarrow \hat{\underset{\sim}{\beta}} \sim N_{p+1}(\underset{\sim}{\beta}, \sigma^2 (X^T X)^{-1}$$

Β3)

Αν $\varepsilon \sim N_n(0, \sigma^2 I_n)$ τότε ΕΕΤ της παραμέτρου $\underset{\sim}{\beta}$ ταυτίζεται με τον ΕΜΠ της $\underset{\sim}{\beta}$

Οι ΕΕΤ προκύπτουν αν ελαχιστοποιήσω ως προς $\underset{\sim}{\beta}$ την ποσότητα
$$S^2 = \sum \varepsilon_i^2 = \underset{\sim}{\varepsilon}^T \underset{\sim}{\varepsilon} = (\underset{\sim}{Y} - X\underset{\sim}{\beta})^T (\underset{\sim}{Y} - X\underset{\sim}{\beta}).$$

Πιθανοφάνεια = Από κοινού κατανομή των $Y_1, Y_2, \dots, Y_n$ $= f_{\underline{Y}}(\underline{y}) \overset{Y \sim N_n(X\beta, \sigma^2 I_n)}{=}$

$$= \frac{1}{(2\pi)^{n/2} (\sigma^2)^{n/2}} e^{-\frac{1}{2}(\underline{Y} - X\underline{\beta})^T (\sigma^2 I_n)^{-1} (\underline{Y} - X\underline{\beta})} =$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}(\underline{Y} - X\underline{\beta})^T (\underline{Y} - X\underline{\beta})}$$

Οι ΕΜΠ των $\underset{\sim}{\beta}$ προκύπτουν από την μεγιστοποίηση ως προς $\underset{\sim}{\beta}$ της L ή του: $\log L$

$$\log \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} - \frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta).$$

Για την εύρεση των ΕΜΠ του $\underset{\sim}{\beta}$, μεγιστοποιώ ως προς $\underset{\sim}{\beta}$ το ~~[illegible]~~ $-\frac{1}{\sigma^2}(\underline{Y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta})$ ή ελαχιστοποιώ την ποσότητα $(\underline{Y}-X\underline{\beta})^T(\underline{Y}-X\underline{\beta})$

Και αφού προσδιορίζονται από την ελαχιστοποίηση της ίδιας ποσότητας τότε οι ΕΕΤ και ΕΜΠ ταυτίζονται ∎

Παρατήρηση

Ο ΕΜΠ της σ^2 προκύπτει από την μεγιστοποίηση ως προς σ^2 της $L(\hat{\underset{\sim}{\beta}}|\sigma^2)$ και είναι $\hat{\sigma}^2 = \frac{SS_{res}}{n}$ (άσκηση) είναι αμερόληπτος?

B4) Αν $\underset{\sim}{\varepsilon} \sim N_n(0, \sigma^2 I_n)$ τότε $\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-p-1}$

Η απόδειξη απαιτεί γνώση θεωρίας κατανομών τετραγωνικών μορφών

B5) οι ΕΕΤ $\hat{\underset{\sim}{\beta}}$ είναι ανεξάρτητοι του MS_{res} ή SS_{res}

B6) F-πηλίκο του Ανάλυσης:

Για τον έλεγχο της $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$, δηλαδή τον έλεγχο ύπαρξης του μοντέλου της πολλαπλής γρ. παλινδρόμησης χρησιμοποιούμε ως στατιστική συνάρτηση του τεστ. $F = \frac{MS_{reg}}{MS_{res}} \overset{H_0}{\sim} F_{p,n-p-1}$ και

κρίσιμη περιοχή μεγέθους a: $F \geq F_{p,n-p-1,a}$.

B7) Αν $\varepsilon \sim N_n(0, \sigma^2 I_n)$ τότε για τον έλεγχο της $H_0\ \beta_i = \beta_i^*$ (β_i^* γνωστό) χρησιμοποιείται η ΣΣΤ $t_i = \frac{\hat{\beta}_i - \beta_i^*}{\sqrt{MS_{res}}\sqrt{C_{i+1,i+1}}}$ με κατανομή

t_{n-p-1} υπό την H_0 και κ.π. $|t| \geq t_{n-p-1,\frac{a}{2}}$ όπου $C_{i+1,i+1}$ είναι το στοιχείο της $(i+1,i+1)$ θέσης του $(X^T X)^{-1}$

Κατασκευή του τεστ

Επειδή $\hat{\underset{\sim}{\beta}} \sim N_{p+1}(\underset{\sim}{\beta}, \sigma^2 (X^T X)^{-1}) \Rightarrow \hat{\beta_i} \sim N(\beta_i, \sigma^2 c_{i+1,i+1})$

και προφανώς υπό την $H_0: \beta_i = \beta_i^*$,

$\hat{\beta_i} \sim N(\beta_i^*, \sigma^2 c_{i+1,i+1}) \Rightarrow$

$$\Rightarrow \boxed{\frac{\hat{\beta_i} - \beta_i^*}{\sigma \sqrt{c_{i+1,i+1}}} \overset{H_0}{\sim} N(0,1).}$$

Επίσης $\frac{SSres}{\sigma^2} \sim X^2_{n-p-1}$ και $\hat{\beta}$ ανεξάρτητο του $MSres$ άρα $\hat{\beta_i}$ ανεξ. του $SSres$

Έτσι $t_i = \frac{\hat{\beta_i} - \beta_i^*}{\sqrt{MSres}\sqrt{c_{i+1,i+1}}} = \frac{(\hat{\beta_i} - \beta_i^*)/\sqrt{c_{i+1,i+1}}}{\sqrt{\frac{SSres}{n-p-1}}}$ διαιρώ με σ^2

$$= \frac{(\hat{\beta_i} - \beta_i^*)/\sqrt{\sigma c_{i+1,i+1}}}{\sqrt{\frac{SSres}{\sigma^2(n-p-1)}}} \sim \frac{N(0,1)}{\sqrt{\frac{X^2_{n-p-1}}{n-p-1}}} \Rightarrow$$

$\Rightarrow t_i \sim t_{n-p-1}$ υπό την H_0

δηλ. $\hat{\beta_i} \neq \beta_i^* \neq \beta_i$ υποδηλώνεται απορρ. της H_0

Η κρίσιμη περιοχή είναι μεγάλες τιμές του t_i, εντωτής $|t_i| > c$

$a = P(\text{Απορ } H_0 | H_0 \text{ αληθής}) = P(|t_i| > c \mid t_i \sim t_{n-p-1}) \Rightarrow \dots \Rightarrow$

$\Rightarrow c = t_{n-p-1, \frac{a}{2}}$.

Παρατήρηση: Πρακτική αξιοποίηση της B7

① Η B7 δίνει την δυνατότητα ελέγχου ύπαρξης γραμμικής σχέσης ατομικά, Αν δηλ η κάθε X_i συνδέεται γραμμικά με την Y εφαρμόζοντας το τεστ B7 για $\beta_i^* = 0$

(2) Βασίζεται στην ερμηνεία του εκτιμητή $\hat{\beta}_i$, $i=1,..p$. σύμφωνα με την οποία ο $\hat{\beta}_i$ εκφράζει την μεταβολή της Y σε μοναδιαία μεταβολή της X_i, (όταν οι άλλες μεταβλητές θεωρηθούν σταθερές)

Άρα η υπόθεση $H_0: \beta_i = \beta_i^*$ δίνει την δυνατότητα να ελέγξω αν η μοναδιαία μεταβολή της Y σε μοναδιαία μεταβολή της X_i είναι μεγέθους β_i^*

Άσκηση

Για το μοντέλο της α.γ.π $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i=1...n$ ν.δ.ο

i) $r^2_{XY} = R^2$, όπου r_{XY} ο συντελεστής συσχέτισης του Pearson και R^2 ο συντελεστής προσδιορισμού

ii) Αν $r_{Y,\hat{Y}}$ είναι ο συντελεστής συσχέτισης του Pearson μεταξύ της παρατηρούμενης Y και της εκτιμώμενης $\hat{Y}$ τότε

$r^2_{Y\hat{Y}} = R^2 = r^2_{XY}$