Δευτέρα 23/11/20

Αξίωμα Εξτρα Αθροίσματος Τετραγώνων

Γενική Γραμμική Υπόθεση:

Θεωρούμε το μοντέλο ΠΓΠ.

$\underset{\sim}{Y} = X\cdot\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}$ και $\underset{\sim}{\varepsilon} \sim N_n(0, \sigma^2 I_n)$.

$\Leftrightarrow$ $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i$, $i=1,\dots,n$.

Θεωρούμε τη γενική γραμμική υπόθεση (γγυ)

$H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$, όπου A ένας $q\times(p+1)$ πίνακας, $\operatorname{rank}(A)=q$.

και $\underset{\sim}{c}$ ένα q-διάστατο διάνυσμα.

Η Ho έχει ενδιαφέρον γιατί για κατάλληλη επιλογή του A και του c δημιουργούνται πλήθος άλλες ενδιαφέρουσες υποθέσεις στο μοντέλο.

Π.χ.

(1) Αν $A=I_{p+1}$ και $\underset{\sim}{c}=\underset{\sim}{0}$ τότε $A\underset{\sim}{\beta}=\underset{\sim}{c} \Leftrightarrow H_0: \beta_i=0\ \forall i$ και αν την αποδεχτούμε έχουμε ότι δεν υφίσταται μοντέλο παλινδρόμησης.

(2) $A=I_{p+1}$ και $\underset{\sim}{c}\neq\underset{\sim}{0}$ τότε. $H_0: A\underset{\sim}{\beta}=c \Leftrightarrow H_0: \beta_i=\beta_i^*$, όπου β_i^* το i-στοιχείο $\underset{\sim}{c}$

(3) Αν $A=\begin{pmatrix} 0 & & & \\ & 1 & & 0 \\ & & 1 & \\ 0 & & 0 & \\ & & & \ddots \, 0 \end{pmatrix}$, $\underset{\sim}{c}=\underset{\sim}{0}$ τότε. $\underset{\sim}{\beta}_{(m)}=\underset{\sim}{0}$

όπου $\beta_{(m)}$ ένα διάνυσμα που περιέχει κάποιες από τις συνιστώσες του $\underset{\sim}{\beta}$

Για A και. $A\underset{\sim}{\beta}=\underset{\sim}{c} \Leftrightarrow \begin{pmatrix} 0 & & \\ & 1 & 0 \\ & & 1 \\ 0 & & \ddots\, 0 \end{pmatrix}\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_p\end{pmatrix} = \underset{\sim}{0} \Leftrightarrow$

$\begin{pmatrix}0\\ \beta_1\\ \beta_2\\ 0\\ \vdots\\ 0\end{pmatrix} = \begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix} \rightarrow \beta_1=0$ και $\beta_2=0$ οπότε αν $\beta_{(2)}=(\beta_1,\beta_2)^T$

τότε $H_0: A\underset{\sim}{\beta}=\underset{\sim}{c} \Leftrightarrow H_0: \beta_1=\beta_2=0$

Αν την αποδεχτώ, δεν υπάρχει γραμμική σχέση της Y με την X_1, X_2

Περιγραφή Τεστ

- Θεωρώ το πλήρες μοντέλο (Π.Μ)

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \dots + b_p X_{ip} + \varepsilon_i, \quad \forall i = 1, 2, \dots n.$$

και θεωρώ επίσης το Ho μοντέλο, το οποίο προκύπτει από το πλήρες μοντέλο, αν πάρω υπόψη μου την Ho: $A\underset{\sim}{B} = \underset{\sim}{c}$

π.χ) αν $A = \begin{pmatrix} 0 & & & \\ & 1 & & 0 \\ 0 & & 1 & \\ & & & \ddots \\ & & & 0 \end{pmatrix}$ τότε λαμβάνοντας υπόψη τη γ.γ.υ τότε Ho: $b_1 = b_2 = 0$

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

- Βρίσκω τους ΕΕΤ του ΠΜ και HoM και έστω $\underset{\sim}{\hat{b}}$ και $\underset{\sim}{\hat{b}}_{Ho}$ αντίστοιχα.
- Βρίσκω τα εκτιμώμενα μοντέλα $\underset{\sim}{\hat{Y}} = X\underset{\sim}{\hat{b}}$ και $\underset{\sim}{\hat{Y}}_{Ho} = X\underset{\sim}{\hat{b}}_{Ho}$

Έχοντας βρει τα $\underset{\sim}{\hat{Y}}, \hat{Y}_{Ho}$ μπορούμε να βρούμε τα SSres

- $SSres(ΠΜ) := (\underset{\sim}{Y} - \underset{\sim}{\hat{Y}})^T (Y - \hat{Y})$

$SSres(HoM) := (\underset{\sim}{Y} - \hat{Y}_{HoM})^T (\underset{\sim}{Y} - \hat{Y}_{Ho})$

Θεωρούμε τη διαφορά $SSres(HoM) - SSres(ΠΜ)$.

⇒ Αν η διαφορά είναι μικρή τότε $SSres(HoM) \simeq SSres(ΠΠ)$ τότε το HoM συμπεριφέρεται ανάλογα με το πλήρες μοντέλο και δεν μπορώ να απορρίψω την Ho

⇒ Αν $SSres(HoM) >> SSres(ΠΜ)$ τότε το HoM ερμηνεύει πολύ μικρότερο κομμάτι ολικής μεταβλητότητας σε σχέση με το ΠΜ. και οδηγούμαστε σε απόρριψη της Ho

Αποδεικνύεται ότι

① $SSres(HoM) - SS(ΠΜ) = (A\underset{\sim}{\hat{b}} - \underset{\sim}{c})^T [A(X^T X)^{-1} A^T]^{-1} (A\underset{\sim}{\hat{b}} - \underset{\sim}{c})$

$$\frac{SSres(HoM) - SS(ΠΜ)}{\sigma^2} \sim X_q^2 \quad \text{υπό την Ho: } A\underset{\sim}{b} = \underset{\sim}{c}$$

Εφόσον $\frac{SSres(ΠΜ)}{\sigma^2} \sim \chi^2_{n-p-1}$ και SSres(ΠΜ) ανεξάρτητο των $\hat{\beta}$

τότε SSres(ΠΜ) ανεξάρτητο SSres(H₀M) − SSres(ΠΜ).

~~SSres~~

Έτσι υπό την $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$

$$F = \frac{(SSres(H_0M) - SSres(ΠΜ))/q}{SSres(ΠΜ)/(n-p-1)} \sim F_{q,n-p-1} \text{, υπό την } H_0$$

$$= \frac{(n-p-1)(A\hat{\underset{\sim}{\beta}} - \underset{\sim}{c})^T [A(X^TX)^{-1}A^T]^{-1} (A\hat{\underset{\sim}{\beta}} - \underset{\sim}{c})}{q\, SSres(ΠΜ)} \sim F_{q,n-p-1} \text{, υπό την } H_0$$

Επομένως για τον έλεγχο της γενικής γραμμικής υπόθεσης $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$ η στατιστική συνάρτηση του τεστ είναι

$$F = \frac{(n-p-1)(A\hat{\underset{\sim}{\beta}} - \underset{\sim}{c})^T [A(X^TX)^{-1}A^T]^{-1} (A\hat{\underset{\sim}{\beta}} - \underset{\sim}{c})}{q\, SSres(ΠΜ)} \sim F_{q,n-p-1}.$$

υπό την H_0 και κρίσιμη περιοχή επιπέδου σημαντικότητας a την $F \geq F_{q,n-p-1,a}$

Ανάλυση Υπολοίπων:

Υπόλοιπα: βοηθούν στο να ελεγχθεί η ικανοποίηση των υποθέσεων για τα σφάλματα του μοντέλου

Μοντέλο: $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \Rightarrow \underset{\sim}{\varepsilon} = \underset{\sim}{y} - X\underset{\sim}{\beta}$

Υπόλοιπα: $\underset{\sim}{e} = \underset{\sim}{y} - \hat{\underset{\sim}{y}} = \underset{\sim}{y} - X\hat{\underset{\sim}{\beta}}$

$\Rightarrow$ Τα υπόλοιπα είναι εκτιμητές των σφαλμάτων.

Ιδιότητες:

1] Τα υπόλοιπα είναι γραμμικές συναρτήσεις των παρατηρήσεων στην εξαρτημένη μεταβλητή $\underset{\sim}{Y}$ και των σφαλμάτων $\underset{\sim}{\varepsilon}$

Αποδ:

Πράγματι, $\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}} = \underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}} = \underset{\sim}{Y} - X(X^T X)^{-1} X^T Y \Rightarrow$

$$\Rightarrow \underset{\sim}{e} = (I_n - \underbrace{X(X^T X)^{-1} X^T}_{P}) Y = (I_n - P)\underset{\sim}{Y}, \text{ όπου } P = X(X^T X)^{-1} X^T$$

Ιδιότητες του P: i) P συμμετρικός $\Rightarrow P^T = P$

$$P^T = (X(X^T X)^{-1} X^T)^T = X(X^T X)^{-1} X^T = P$$

ii) P ταυτοδύναμος $\Rightarrow P^2 = P \Leftrightarrow P^2 = (X(X^T X)^{-1} X^T)^2 = (X(X^T X)^{-1} X^T)(X(X^T X)^{-1} X^T)$

$$= X(X^T X)^{-1} X^T = P$$

Μένει νδο $\underset{\sim}{e}$ είναι γραμμ. συναρτήσεις των $\underset{\sim}{\varepsilon}$

$\underset{\sim}{e} = (I_n - P)\underset{\sim}{Y} = (I_n - P)(X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}) = (I_n - P)\underset{\sim}{\varepsilon} + (I_n - P)X\underset{\sim}{\beta}$.

Ανδο $(I_n - P)X\underset{\sim}{\beta} = 0$

$(I_n - P)X\underset{\sim}{\beta} = I_n X\underset{\sim}{\beta} - PX\underset{\sim}{\beta} = X\underset{\sim}{\beta} - X(X^T X)^{-1} X^T X\underset{\sim}{\beta} = X\underset{\sim}{\beta} - X\underset{\sim}{\beta} = 0$

Άρα $\underset{\sim}{e} = (I_n - P)\underset{\sim}{\varepsilon}$.

2] Αν οι υποθέσεις για τα σφάλματα ικανοποιούνται, δηλ. $\underset{\sim}{\varepsilon} \sim N(0, \sigma^2 I_n)$ τότε το διάνυσμα $\underset{\sim}{e} \sim N_n(\underset{\sim}{0}, (I_n - P)\sigma^2)$.

Αποδ] Γνωρίζουμε ότι αν $\underset{\sim}{W} \sim N_n(\underset{\sim}{\mu}, \Sigma)$ και αν A ένας $m \times n$ πίνακας τότε $AW = N_q(A\mu, A\Sigma A^T)$

Επειδή τα $\underset{\sim}{e} = (I_n - P)\underset{\sim}{\varepsilon}$ είναι γραμμική συνάρτηση των σφαλμάτων $\underset{\sim}{\varepsilon} \sim N_n(0, \sigma^2 I_n)$ τότε τα υπόλοιπα $\underset{\sim}{e}$ έχουν n-διάστατη κανονική κατανομή.

$E(\underset{\sim}{e}) = E((I_n - P)\underset{\sim}{\varepsilon}) = (I_n - P)\cdot 0 = 0$.

$Var(\underset{\sim}{e}) = Var((I_n - P)\underset{\sim}{\varepsilon}) = (I_n - P)\ Var(\underset{\sim}{\varepsilon})(I_n - P)^T$

$= (I_n - P)\sigma^2 I_n (I_n - P^T)^T = (I_n - P)\sigma^2 I_n (I_n^T - P^T) = \sigma^2 (I_n - P)(I_n - P)$
$= \sigma^2 (I_n - P)^2 = \sigma^2 (I_n - P - P + P^2) = \sigma^2 (I_n - P)$

Άρα $\underset{\sim}{e} \sim N_n(\underset{\sim}{0}, (I_n - P)\sigma^2)$

Συνέπεια:
Επειδή το $\underset{\sim}{e}$ έχει πολυδιάστατη κανονική τότε κάθε συνιστώσα του έχει μονοδιάστατη κανονική με μέση τιμή 0 και διακύμανση το i-διαγώνιο στοιχείο του $\sigma^2(I_n - P)$
δηλαδή

$e_i \sim N(0, \sigma^2(1 - P_{ii}))$, P_{ii} το (i,j) στοιχείο του πίνακα P.

3] Τα e_i, $\hat{Y}_i$ είναι ασυσχέτιστα $\Leftrightarrow Cov(\underset{\sim}{e}, \underset{\sim}{\hat{Y}}) = 0$

$Cov(\underset{\sim}{e}, \underset{\sim}{\hat{Y}}) = cov((I_n - P)\underset{\sim}{Y}, \underset{\sim}{\hat{Y}}) = Cov((I_n - P)\underset{\sim}{Y}, X\underset{\sim}{\hat{\beta}}) =$

$= Cov((I_n - P)\underset{\sim}{Y}, X(X^TX)^{-1}X^T\underset{\sim}{Y}) = Cov((I_n - P)\underset{\sim}{Y}, P\underset{\sim}{Y})$

Γνωρίζουμε ότι $Cov(AW, BW) = A\,var(\underset{\sim}{W}) \cdot B^T$ επομένως έχω

$(I_n - P) \cdot \underbrace{Var(\underset{\sim}{Y})}_{=\sigma^2 I_n} \cdot P^T = (I_n - P) \cdot \sigma^2 I_n \cdot P = \sigma^2 (I_n - P) P =$

$= \sigma^2 (I_n \cdot P - P^2) = \sigma^2 (P - P) = 0$

Τα $\underset{\sim}{e}$ και $\underset{\sim}{\hat{Y}}$ είναι ανεξάρτητα· όντας ασυσχέτιστες και ισόνομες

4] Μαθητικοποιημένα Υπόλοιπα (Studentized Υπόλοιπα).

$t_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1 - P_{ii}}} = \dfrac{e_i}{\sqrt{MSres \cdot (1 - P_{ii})}}$, $i = 1, 2, \ldots n$ όπου

$t_i \sim t_{n-p-1}$, $\forall i = 1, 2, \ldots n$.

Αποδ:

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-P_{ii}}} = \frac{e_i}{\sqrt{MS_{res}}\sqrt{1-P_{ii}}}, \quad i=1,\dots,n$$

$$e_i \sim N(0, \sigma^2(1-P_{ii})) \rightarrow \frac{e_i}{\sigma\sqrt{1-P_{ii}}} \sim N(0,1) \quad \forall i = 1,2,\dots,n$$

$$\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-1-p}$$

(*) Τα e_i είναι ανεξάρτητα με τα SS_{res} γιατί $\hat{\beta}$ ανεξ. SS_{res} και $\tilde{e} = \tilde{y} - \hat{\tilde{y}} = \tilde{y} - X\hat{\tilde{\beta}}$.

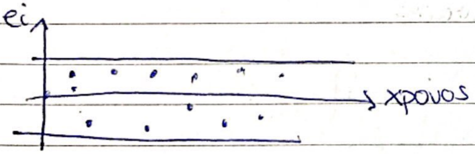
Θεωρώ το πηλίκο $\dfrac{\frac{e_i}{\sigma\sqrt{1-P_{ii}}}}{\sqrt{\frac{SS_{res}}{\sigma^2}/(n-p-1)}} = \dfrac{e_i}{\sqrt{MS_{res}}\cdot\sqrt{1-P_{ii}}} = \dfrac{e_i}{\hat{\sigma}\sqrt{1-P_{ii}}} \overset{*}{\sim} t_{n-p-1} \quad \forall i=1,2,\dots,n.$

Άρα $t_i \sim t_{n-p-1}$

Έλεγχος υποθέσεων για τα σφάλματα

(I) Έλεγχος ασυσχέτιστου των σφαλμάτων:

α) Γραφική παράσταση των υπολοίπων ως προς τον χρόνο



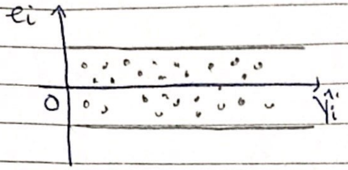
Εάν τα υπόλοιπα είναι τυχαία κατανεμημένα σε μια ζώνη γύρω από το 0 αφού ($E(e_i) = 0$) τότε έχουμε ένδειξη ότι τα σφάλματα είναι ασυσχέτιστα.

β) Μη παραμετρικό τεστ των ροών

γ) Τεστ Durbin-Watson

II) Έλεγχος σταθερής διακύμανσης ($Var(\varepsilon_i)=\sigma^2$) $i=1,\dots n$.

α) Γραφική παράσταση των e_i και $\hat{Y}_i$



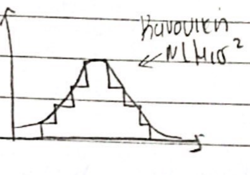
Αν τα $(\hat{Y}_i, e_i)$, $i=1,\dots n$ είναι κατανεμημένα τυχαία σε μια ζώνη γύρω από το 0 τότε υπάρχει ισχυρή ένδειξη ότι η υπόθεση της σταθερής διακύμανσης ισχυροποιείται (Γιατί?)

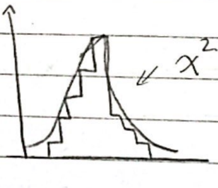
↳ Αποδείξαμε ότι $Var(\varepsilon_i)=\sigma^2$ τότε $Cov(e_i, \hat{Y}_i)=0$, $i=1,\dots n$.

β) Τεστ Levene, Τεστ White.

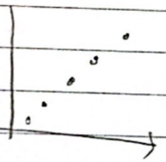
III) Έλεγχος υπόθεσης της κανονικότητας των σφαλμάτων: ($\varepsilon_i \sim N(0,\sigma^2)$, $i=1,\dots n$)

α) Ιστόγραμμα συχνοτήτων (Ανασφαλές





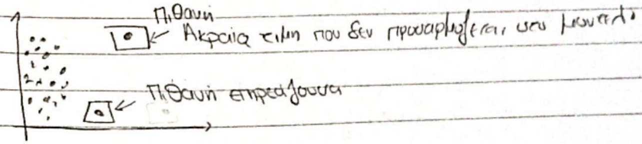
β) Γραφική παράσταση των e_i σε χαρτί πιθανότητας $(\frac{i}{n+1}, e_i)$



γ) Σ.Τ ελέγχου κανονικότητας Kolmogorov-Smirnov, Shapiro-Wilk

Ακραίες και Επηρεάζουσες Παρατηρήσεις:

Γενικά στην στατιστική ακραίες παρατηρήσεις λέμε εκείνες που διαφοροποιούνται σημαντικά από τις υπόλοιπες.



Ειδικότερα στα Γραμμικά μοντέλα ακραίες παρατηρήσεις είναι εκείνες που απέχουν σημαντικά από τις υπόλοιπες και είναι εκείνες που δεν προσαρμόζονται σε κάποιο γραμμικό μοντέλο στο οποίο προσαρμόζονται οι υπόλοιπες.

Πως ανιχνεύουμε ακραίες - επηρεάζουσες παρατηρήσεις

Ακραίες: • Αν $|e_k|$ είναι μεγάλο τότε η αντίστοιχη παρατήρηση είναι πιθανά ακραία
• Χρησιμοποιούνται τα t_i σε γραφικές παραστάσεις.

Καλούμε επηρεάζουσες εκείνες τις παρατηρήσεις που απέχουν από τις άλλες αλλά τείνουν να προσαρμόζονται στο μοντέλο όπως προσαρμόζονται οι υπόλοιπες αλλά επηρεάζονται σημαντικά από τους εκτιμητές.

Έχουν αναπτυχθεί τεστ για τον έλεγχο μιας παρατήρησης αν είναι επηρεάζουσα ή όχι όπως το τεστ που βασίζεται στην απόσταση Cook.