Διάλεξη 6η 12/11/2019

Παλινδρόμηση & Ανάλυση Διακύμανσης

Γενική γραμμική υπόθεση.

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

Έστω ένας πίνακας σταθερών A διαστάσεων $q \times (p+1)$

Γενική Γραμμική Υπόθεση: $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$

$q\times(p+1)$ $(p+1)\times 1$ $q\times 1$

$q\times 1$

Η H_0 περικλείει ως ειδικές περιπτώσεις τις ακόλουθες ενδιαφέρουσες:

(1) Αν $A \equiv I_{p+1}$ και $\underset{\sim}{c} = \underset{\sim}{0}_{p+1}$ τότε $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c} \rightsquigarrow H_0: \beta_0 = \beta_1 \dots = \beta_p = 0$

Αν δεν μπορεί να απορριφθεί $\rightsquigarrow$ ✗ μοντέλο πχπ

(2) Αν $A = I_{p+1}$ και $\underset{\sim}{c} = \underset{(p+1)\times 1}{\underset{\sim}{\beta}^*}$, με β^* γνωστό τότε $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$

$\rightsquigarrow H_0: \beta_0 = \beta_0^*, \dots \beta_p = \beta_p^*$

Αν δεν μπορεί να απορριφθεί τότε η μεταβλητή της Y για μοναδιαία μεταβολή της X_i όταν οι υπόλοιπες X θεωρηθούν σταθερές είναι ίση με β_i^*

(3) Αν $A = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & & \\ \vdots & & \ddots & \\ 0 & & & 1 \end{pmatrix}$ και $\underset{\sim}{c} = 0$ τότε $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c} \sim H_0: \underset{\sim}{\beta}_m = 0$

όπου $\underset{\sim}{\beta}_m$ διάνυσμα που αποτελείται από κάποιες μηδενικές συνιστώσες του $\underset{\sim}{\beta}$

Πχ αν $A = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & & & & 0 \end{pmatrix}$ τότε $A\underset{\sim}{\beta} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & & & & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} 0 \\ \beta_1 \\ \beta_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

Άρα $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c} = \underset{\sim}{0} \rightarrow H_0: \beta_1 = \beta_2 = 0$ (1)

- Είναι σημαντική η $H_0: \beta_1 = \beta_2 = 0$;

ΝΑΙ πάρα πολύ γιατί αν δεν μπορώ να την απορρίψω τότε οι όροι που περιέχουν την X_1 και X_2 πρέπει να διαγραφούν και το μοντέλο θα γίνει: $Y_i = \beta_0 + \beta_3 X_{i3} + \beta_4 X_{i4} + \dots + \beta_p X_{ip} + \varepsilon_i$

Περιγραφή της κατασκευής στατιστικού τεστ για τον έλεγχο της $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$.

Έστω ΠΜ το πλήρες μοντέλο: $Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip} + \varepsilon_i,\ i = 1, \dots, n$

Έστω H_0Μ το H_0-μοντέλο που διαμορφώνεται από το ΠΜ αν λάβω υπόψη μου τη $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$.

πχ Έστω η $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$ είναι:

$$H_0: \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$\Updownarrow$

$H_0: \beta_1 = 0$ και $\beta_2 = 0$

Άρα το H_0Μ είναι $Y_i = \beta_0 + \beta_3 X_{i3} + \dots + \beta_p X_{ip} + \varepsilon_i,\ i = 1, \dots, n$

Έστω $\hat{\underset{\sim}{\beta}}$ και $\hat{\underset{\sim}{\beta}}_{H_0}$ οι ΕΕΤ της παραμέτρου $\underset{\sim}{\beta}$ υπό το ΠΜ και H_0Μ, αντίστοιχα.

Οι εκτιμώμενες τιμές του $\underset{\sim}{Y}$ είναι $\hat{\underset{\sim}{Y}} = X\hat{\underset{\sim}{\beta}}$ για το ΠΜ

$\hat{\underset{\sim}{Y}}_{H_0} = X\hat{\underset{\sim}{\beta}}_{H_0}$ για το H_0Μ

Επομένως τα αντίστοιχα αθροίσματα τετραγώνων των υπολοίπων θα είναι:

$SSres(\Pi M) \stackrel{op}{=} (\underset{\sim}{Y}-\hat{\underset{\sim}{Y}})'(\underset{\sim}{Y}-\hat{\underset{\sim}{Y}})$

$SSres(H_0M) \stackrel{op}{=} (\underset{\sim}{Y}-\hat{\underset{\sim}{Y}}_{H_0})'(\underset{\sim}{Y}-\hat{\underset{\sim}{Y}}_{H_0})$

Αν $SSres(H_0M) >> SSres(\Pi M)$ τότε το μοντέλο κάτω από την $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$ ερμηνεύει πολύ μικρότερο κομμάτι από την ολική μεταβλητότητα και επομένως είναι το H_0M λιγότερο προτιμητέο από το ΠΜ.
Άρα το H_0M πρέπει να απορριφθεί
Άρα το τεστ για τον έλεγχο θα πρέπει να βασιστεί στη σύγκριση των $SSres(H_0M)$ με το $SSres(\Pi M)$.

Αποδεικνύεται (αλγεβρικά):

$$SSres(H_0M) - SSres(\Pi M) = (A\hat{\underset{\sim}{\beta}} - \underset{\sim}{c})'[A(X'X)^{-1}A']^{-1}(A\hat{\underset{\sim}{\beta}} - \underset{\sim}{c})$$

Επίσης αποδεικνύεται (πιθανοθεωρητικά με κατανομές τετραγωνικών μορφών).

$$\frac{SSres(H_0M) - SSres(\Pi M)}{\sigma^2} \sim \chi^2_q \quad \text{υπό την } H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}.$$

Επίσης $\frac{SSres(\Pi M)}{\sigma^2} \sim \chi^2_{n-p-1}$.

Επίσης $\hat{\underset{\sim}{\beta}}$ ανεξ του $SSres(\Pi M)$
Άρα και $SSres(H_0M) - SSres(\Pi M)$ ανεξ του $SSres(\Pi M)$

Έτσι υπό $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$ η στατιστική συνάρτηση:

$$F = \frac{SSres(H_0M) - SSres(\Pi M)/\sigma^2 \cdot q}{SSres(\Pi M)/\sigma^2(n-p-1)} \sim F_{q,n-p-1} \text{ υπό την } H_0$$

Συχνεντρωτικά για τον έλεγχο της $H_0: A\underset{\sim}{\beta} = \underset{\sim}{c}$
Χρησιμοποιείται η στατιστική συνάρτηση:

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

Και κ.π μεγάλες τιμές του F ($F \geq C$, το C υπολογ από $\alpha = P(\text{Απορρ } H_0 / H_0 \text{ αληθεύει})$)

$$F \geq F_{q, n-p-1, \alpha}$$

Ανάλυση Υπολοίπων
- Διερεύνηση των Ιδιοτήτων των υπολοίπων *
- Αξιοποίηση των υπολοίπων στον έλεγχο των υποθέσεων για τα σφάλματα. (αφού τα υπόλοιπα εκτιμούν τα σφάλματα του μοντέλου) **

* Μοντέλο $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

Υπόλοιπα: $\underset{\sim}{e} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}} = \underset{\sim}{Y} - X\hat{\underset{\sim}{\beta}}$, $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{Y}$

Ιδιοτητες των Υπολοιπων.

1. Τα υπόλοιπα $\underset{\sim}{e}$ είναι γραμμική συνάρτηση των $\underset{\sim}{\varepsilon}$ και $\underset{\sim}{Y}$

Πράγματι,

$$\underset{\sim}{e} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}} = \underset{\sim}{Y} - X\hat{\underset{\sim}{\beta}} = \underset{\sim}{Y} - X(X'X)^{-1}X'Y = (I_n - \underbrace{X(X'X)^{-1}X'}_{P})\underset{\sim}{Y}$$

$$\Rightarrow \boxed{\underset{\sim}{e} = (I_n - P)\underset{\sim}{Y}} \text{ όπου } P = X(X'X)^{-1}X'$$

Επίσης,

$$\underset{\sim}{e} = (I_n - P)\underset{\sim}{Y} = (I_n - P)(X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}) = (I_n - P)X\underset{\sim}{\beta} + (I_n - P)\underset{\sim}{\varepsilon}.$$
$$= X\underset{\sim}{\beta} - PX\underset{\sim}{\beta} + (I_n - P)\underset{\sim}{\varepsilon} = X\underset{\sim}{\beta} - X(X'X)^{-1}X'X\underset{\sim}{\beta} + (I_n - P)\underset{\sim}{\varepsilon}.$$
$$= X\underset{\sim}{\beta} - X\underset{\sim}{\beta} + (I_n - P)\underset{\sim}{\varepsilon} \Rightarrow \boxed{\underset{\sim}{e} = (I_n - P)\underset{\sim}{\varepsilon}}$$

ΠΑΡΑΤΗΡΗΣΗ: Για τον $P = X(X'X)^{-1}X'$

i) Ο P είναι συμμετρικός $P = P'$

ii) Ο P είναι ταυτοδύναμος, $P^2 = P$.

Αποδ

i) $P' = [X(X'X)^{-1}X']' \overset{(AB)' = B'A'}{=\!=\!=} X(X'X)^{-1}X' = P$

$(A^{-1})' = (A')^{-1}$

ii) $P^2 = [X(X'X)^{-1}X'][X(X'X)^{-1}X'] = X(X'X)^{-1}\underbrace{X'X}_{I_n}(X'X)^{-1}X'$

$= X(X'X)^{-1}X' = P.$

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

Επειδή $\underset{\sim}{e} = (I_n - P)\underset{\sim}{\varepsilon}$ και $\underset{\sim}{\varepsilon} \sim N_n \Rightarrow \underset{\sim}{e} \sim N_n$

από υποθ σφαλμάτων

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

Ισχύει $Var(A\underset{\sim}{w}) = A \cdot Var(\underset{\sim}{w})\, A'$

$Var(\underset{\sim}{e}) = Var[(I_n - P)\underset{\sim}{\varepsilon}] \Rightarrow$

$Var(\underset{\sim}{e}) = (I_n - P)\, Var(\underset{\sim}{\varepsilon})\, (I_n - P)'$

$= (I_n - P)\sigma^2(I_n - P)$ $\quad P = P'$

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

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

$= \sigma^2(I_n - \underbrace{2P}_{\text{''}} I_n + P) = \sigma^2(I_n - P)$

$2P$

ΠΑΡΑΤΗΡΗΣΗ: Αφού $Var(\underset{\sim}{e}) = \sigma^2(I_n - P)$

$Var(e_i)$ = i-διαγώνιο στοιχείο του $\sigma^2(I_n - P)$

$= \sigma^2(1 - p_{ii})$ όπου p_{ii} το i-διαγώνιο στοιχείο του P

$Cov(e_i, e_j) = -\sigma^2 p_{ij}$, όπου p_{ij} το (ij) στοιχείο του P

3. Τα e_i και $\hat{Y}_i$ ασυσχέτιστα

δηλ $Cov(e_i, \hat{Y}_i) = 0$.

Αποδ

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

$$= Cov\left((I_n - P)\underset{\sim}{Y}, X(X'X)^{-1}X'\underset{\sim}{Y}\right)$$

$$= Cov\left((I_n - P)\underset{\sim}{Y}, P\underset{\sim}{Y}\right)$$

Όπως Ισχύει $Cov(A\underset{\sim}{Y}, B\underset{\sim}{Y}) = A\,Var(\underset{\sim}{Y})\,B'$

$$= Cov(\underset{\sim}{e}, \underset{\sim}{\hat{Y}}) = (I_n - P)\,Var(\underset{\sim}{Y})\,P'$$

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

$$= \sigma^2 (I_n - P)\,P$$

$$= \sigma^2 [I_n P - P^2]$$

$$= \sigma^2 (P - P)$$

$$= \sigma^2 \cdot 0 \quad \text{← ταυτοδυναμικός πίνακας}$$

$$= 0$$

4. Μαθηματικοποιημένα Υπόλοιπα.

Ορίζονται: $t_i = \dfrac{e_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}$ $\hat{\sigma} = \sqrt{MSres}$, p_{ii} το i-διαγώνιο στοιχείο του $P = X(X'X)^{-1}X'$

$$= \frac{e_i}{\sqrt{\hat{Var}(e_i)}}$$

Ισχύει $t_i \sim t_{n-p-1}$.

Αποδ $t_i \sim t_{n-p-1}$

Επειδή $\underset{\sim}{e} \sim N_n(\underset{\sim}{0}, \sigma^2(I_n - P))$ $\longrightarrow$ το $e_i \sim N(0, \sigma^2(1 - p_{ii}))$

$$\Rightarrow \frac{e_i - 0}{\sigma\sqrt{1 - p_{ii}}} \sim N(0,1) \quad \text{ή} \quad \frac{e_i}{\sigma\sqrt{1 - p_{ii}}} \sim N(0,1)$$

Επιπλέον $\dfrac{SSres}{\sigma^2} \sim \chi^2_{n-p-1}$

$$ti \sim \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}} = \frac{e_i/\sqrt{1-p_{ii}}}{\sqrt{MSres}} = \frac{e_i/\sigma\sqrt{1-p_{ii}}}{\sqrt{\frac{SSres}{(n-p-1)\sigma^2}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-p-1}}{(n-p-1)}}} \sim t_{n-p-1}$$

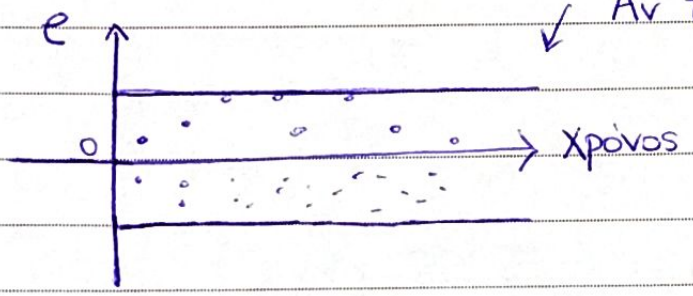
γιατί e_i και SSres ανεξάρτητα αφού

$\underset{\sim}{e} = \underset{\sim}{Y} - \underset{\sim}{\hat{Y}} = \underset{\sim}{Y} - X\underset{\sim}{\hat{\beta}}$ και $\underset{\sim}{\hat{\beta}}$ ανεξ SSres.

**

(I) Έλεγχος του ασυσχέτιστου των σφαλμάτων: $Cov(\varepsilon_i, \varepsilon_j) = 0$.

α) Γραφική Μέθοδος



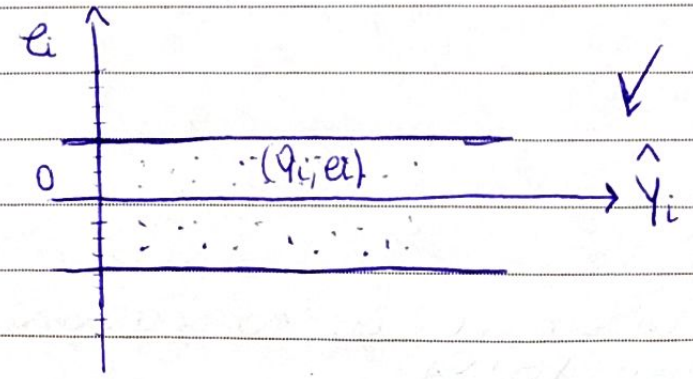
Αν τα υπόλοιπα e_i κατανέμονται τυχαία σε μια μόνη ζώνη από το 0, τότε έχω ένδειξη του ασυσχέτιστου

β) Τεστ ροών

γ) Τεστ Durbin-Watson

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

α) Γραφική Μέθοδος (Γραφική παράσταση των $e_i, \hat{Y}_i$ $i=1,\dots,n$)



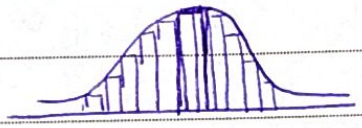
Αν τα $(\hat{Y}_i, e_i)$ $i=1,\dots,n$ κατανέμονται τυχαία σε μια μόνη ζώνη από το 0 έχω ισχυρή ένδειξη ότι η διακύμανση των σφαλμάτων είναι κοινή και σταθερή

Θεωρητική δικαιολόγηση
Αποδ. ότι $Var(\varepsilon_i)=\sigma^2$ τότε $Cov(e_i, \hat{Y}_i)=0$.

β) Τεστ του Levene.

(III) Κανονικότητα των σφαλμάτων ε_i $i=1,..,n$
δηλ ότι τα $\varepsilon_i \sim N$

α) Ιστόγραμμα $e_1,... e_n$



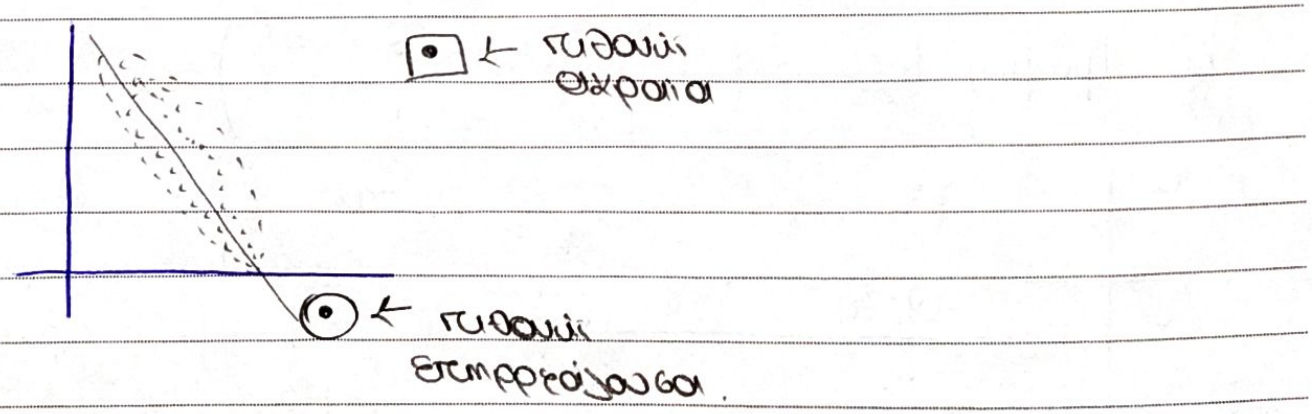
β) Normal probab. Plots ← SPSS κλπ.

γ) Kolmogorov-Smirnov

δ) Shapiro - Wilks

— o — o —

▸ Ακραίες - Επηρεάζουσες παρατηρήσεις
Γενικά παρατηρήσεις που διαφοροποιούνται αισθητά από τις υπόλοιπες



• Μια παρατήρηση Y_k είναι ακραία αν το κ-υπόλοιπο e_k που αντιστοιχεί σ'αυτή είναι $|e_k|>3$.
• Μια παρατήρηση Y_k είναι επηρεάζουσα
→ τέτ βαθμίζεται στην απόσταση Cook.