28/11/16 Παλινδρ.

Ανάλυση υπολοίπων

ο Υπόλοιπα: Βοηθούν να ελεχθεί η ικανοποίηση των υποθέσεων δια
των σφαλμάτων μοντέλου.

Θεωρητικές ιδιότητες υπολοίπου

Μοντέλο $\quad Y = X\beta + \varepsilon \Rightarrow \varepsilon = Y - X\beta$

Υπόλοιπα $\quad e = Y - \hat{Y} \overset{\hat{Y} = X\hat{\beta}}{\Longrightarrow} e = Y - X\hat{\beta}$

τα υπόλοιπα e είναι εκτιμητές των σφαλμάτων ε

Ιδιότητα 1. Τα υπόλοιπα είναι γραμμικές συναρτήσεις των παρατηρήσεων Y και των σφαλμάτων ε

πράγματι $\quad e = Y - \hat{Y} = Y - X\hat{\beta} = Y - X(X'X)^{-1}X'Y \Rightarrow$

$$e = (I_n - X(X'X)^{-1}X')Y \Rightarrow e = (I_n - P)$$

όπου $P = X(X'X)^{-1}X'$

Παρατήρηση ο P είναι: 1. συμμετρικός $P' = P$

2. Ταυτοδύναμος $P^2 = PP = P$

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

2. $P^2 = (X(X'X)^{-1}X')(X(X'X)^{-1}X')$
$= X(X'X)^{-1}X'X(X'X)^{-1}X' = X(X'X)^{-1}X' = P$

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

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

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

<u>Ιδιότητα 2</u> Αν οι υποθέσεις δια τα σφάλματα ικανοποιούνται ($\varepsilon \sim N(\underset{\sim}{0}, \sigma^2)$)

τότε το διάνυσμα των υπολοίπων $\underset{\sim}{e} \sim N_n(\underset{\sim}{0}, (I_n - P)\sigma^2)$

όπου $P = X(X'X)^{-1}X'$

<u>Επιπλέον</u> $e_i \sim N(0, \sigma^2(1 - P_{ii}))$ όπου P_{ij} το (i,j) στοιχείο του P

<u>Γνωρίζουμε</u> ότι αν $\underset{\sim}{W} \sim N(\mu, \Sigma)$ και A πίνακας $m \times n$ τότε

$$\underbrace{\underset{m\times n}{A}\underset{n\times 1}{\underset{\sim}{W}}}_{m\times 1} \sim N_m(A\mu, A\Sigma A')$$

Επειδή το $\underset{\sim}{e} = (I_n - P)\underset{\sim}{\varepsilon}$ είναι γραμμική συνάρτηση των $\underset{\sim}{\varepsilon} \sim N_n(\underset{\sim}{0}, \sigma^2 I_n)$

τα $\underset{\sim}{e}$ έχουν n-διάστατη κανονική κατανομή

Αρκεί λοιπόν να βρω τα $\underline{E(\underset{\sim}{e})}$ και $\underline{Var(\underset{\sim}{e})}$

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

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

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

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

$$= \sigma^2(I_n - P) = \sigma^2 \begin{pmatrix} 1-P_{11} & -P_{12} & \cdots & -P_{1n} \\ & 1-P_{22} & & \\ & & \ddots & \\ -P_{n1} & & & 1-P_{nn} \end{pmatrix}$$

Επιπλέον αφού $\underset{\sim}{e} \sim N_n(\underset{\sim}{0}, \sigma^2(I_n - P))$ καθε $e_i \sim N(0,$ το i-διαγώνιο στοιχείο του $\sigma^2(I_n-P)$

$\equiv N(0, \sigma^2(1-P_{ii}))$

$\forall i = 1, \ldots, n$

<u>Ιδιότητα 3</u> Τα $e_i, \hat{Y}_i$ ασυσχέτιστα ή ισοδύναμα $Cov(\underset{\sim}{e}, \hat{\underset{\sim}{Y}}) = 0$

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

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

$$* \; Cov(A\underset{\sim}{w}, B\underset{\sim}{w}) = A\,Var(w)B'$$

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

Ιδιότητα 4 : Κανονικοποιημένα Υπόλοιπα κ Studentized υπόλοιπα

$t_i \sim t_{n-p-1}, \ \forall i=1,\dots,n$

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-p_{ii}}} = \frac{e_i}{MSres\sqrt{1-p_{ii}}}, \quad i=1,\dots,n$$

αποδείξαμε $e_i \sim N(0, \sigma^2(1-p_{ii}))$

$$= \frac{e_i}{\sigma\sqrt{1-p_{ii}}} \sim N(0,1) \quad \forall i=\cancel{0},\dots,n$$ και εδώ θέλω να καταλήξω σε t-κατανομή

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

Θεωρώ οπότε

$$\frac{\dfrac{e_i}{\sigma\sqrt{1-p_{ii}}}}{\sqrt{\left(\dfrac{SSres}{\sigma^2}\right)\Big/(n-p-1)}} = \frac{\dfrac{e_i}{\sigma\sqrt{1-p_{ii}}}}{\sqrt{\dfrac{MSres}{\sigma^2}}} = \frac{\dfrac{e_i}{\cancel{\sigma}\sqrt{1-p_{ii}}}}{\dfrac{\sqrt{MSres}}{\cancel{\sigma}}} = \frac{e_i}{\sqrt{MSres}\sqrt{1-p_{ii}}} \overset{op.}{=} t_i$$

Αρα

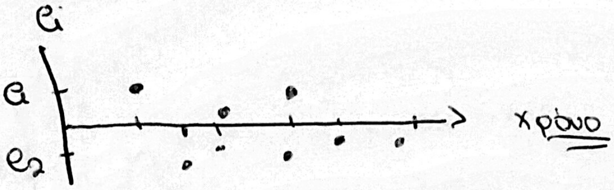
$$t_i = \frac{\dfrac{e_i}{\sigma\sqrt{1-p_{ii}}}}{\sqrt{\dfrac{SSres}{\sigma^2}\Big/(n-p-1)}} = \frac{N(0,1)}{\sqrt{\chi^2_{n-p-1}/(n-p-1)}} \sim t_{n-p-1}$$

e_i ανεξ SSres διότι τα $\hat{\underset{\sim}{\beta}}$ ανεξ SSres και $\underset{\sim}{e} = \underset{\sim}{y} - \hat{\underset{\sim}{y}} = \underset{\sim}{y} - X\hat{\underset{\sim}{\beta}}$

[αυτή η κατανομή χρησιμοποιείται για τρόπους ανίχνευσης ακραίων καταστάσεων]

(I) Έλεγχος Ασυσχέτιστου των σφαλμάτων
($Cov(\varepsilon_i, \varepsilon_j) = 0$ $i,j=1,\dots n$ $i \neq j$)

a. Γραφική παράσ. υπολοίπων ως προς τον χρόνο



Αν τα σημεία είναι τυχαία κατανεμημένα σε μία ζώνη

γύρω από το 0 (αφού $E(\hat{e}_i)=0$) τότε

έχουμε ισχυρή ένδειξη ότι τα σφάλματα είναι

ασυσχέτιστα

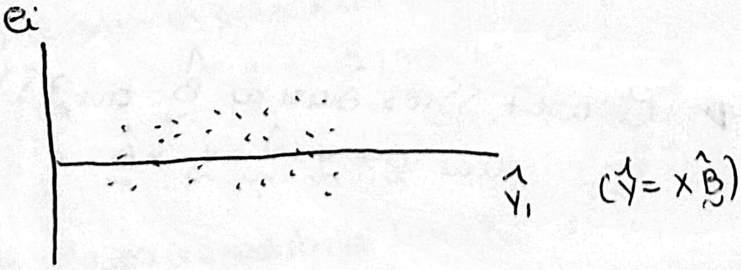
(δηλ δεν υπάρχει κάποιος τρόπος που το ένα επηρεάζει το

άλλο)

β. Μη παραμετρικό test των ροών

γ. Test Durbin-Watson

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

a. 

Αν τα σημεία του διαγράμματος κατανέμονται τυχαία

σε μία ζώνη γύρω από το 0 τότε υπάρχει ισχυρή

ένδειξη ότι η υπόθεση της σταθερής διακύμανσης

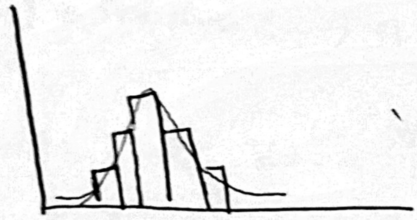
ικανοποιείται. Και αυτό γιατί αποδείξαμε ότι

$Cov(\underset{\sim}{e}, \underset{\sim}{\hat{Y}}) = 0$

III Έλεγχος της κανονικότητας των σφαλμάτων ($\epsilon_i \sim N(0,\sigma^2)$)

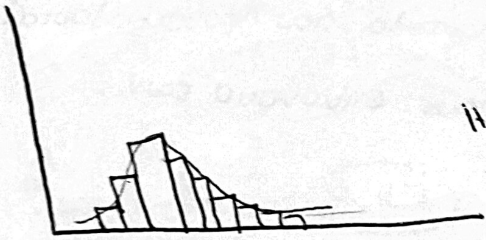
Αρκεί να ελέγξω την κανονικότητα των $e_1, \dots, e_n$

α. Ιστόγραμμα συχνοτήτων



Ένδειξη κανονικότητας

<u>Ενώ</u>



Η καμπύλη προσομοιάζει περισσότερο την χ^2 κατανομή

β. Kolmogorov-Smirnov

γ. Shapro-Wilk

Ακραίες και Επιρεάζουσες Παρατηρήσεις

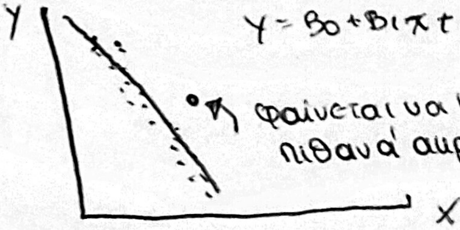
Γενικά στη στατιστική ακραίες παρατηρήσεις είναι εκείνες που διαφοροποιούνται σημαντικά από τις υπόλοιπες.

π.χ $2, 3, 7, 4, 0, -12, \underline{20}$

↳ πιθανή ακραία παρατήρηση

Ειδικότερα στα γραμμικά μοντέλα ακραίες παρατηρήσεις είναι εκείνες που απέχουν σημαντικά από τις υπόλοιπες <u>και</u> είναι εκείνες που δεν προσαρμόζονται σε κάποιο γραμμ. μοντέλο στο οποίο προσαρμόζονται οι υπόλοιπες

$Y = \beta_0 + \beta_1 x + \varepsilon$



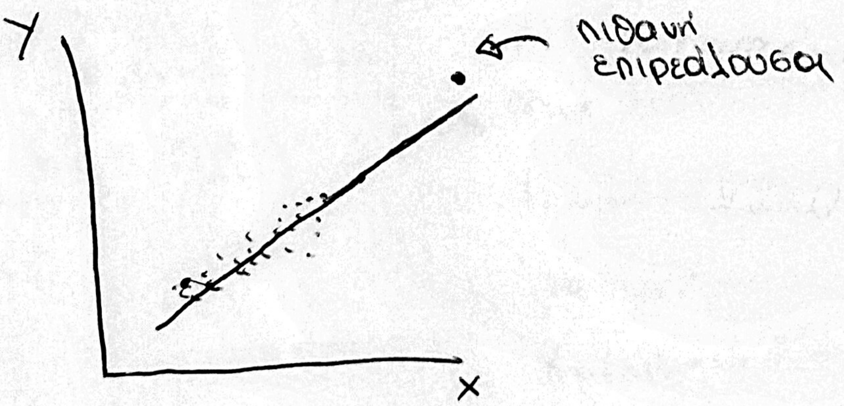
φαίνεται να μην προσαρμόζεται στο μοντέλο, πιθανά ακραία

Έχουν αναπτυχθεί τεχνικές για να ελέγχεται αν μια παρατήρηση είναι ακραία.

• Στα γραμμικά μοντέλα: αν $|\theta|$ είναι μεγάλη τότε η αντίστοιχη παρατήρηση είναι πιθανά ακραία

Επηρεάζουσες παρατηρήσεις

Είναι παρατηρήσεις που απέχουν από τις άλλες που μπορεί να προσαρμόζονται στο μοντέλο που προσαρμόζονται οι υπόλοιπες αλλά επηρεάζουν σημαντικά τους εκτιμητές.



Ανίχνευση επηρεαζουσών παρατηρήσεων

Test που βασίζεται στην απόσταση Cook

Ασκ.3

Θερμ	-5	-4	1	2	5
Αποδοση	1	5	9	13	18

α.

X = Θερμοκρασια ← Ανεξαρτητη

Y = Απόδοση ← Εξαρτημένη

Μοντέλο → $Y = B_0 + B_1 X + \varepsilon$

ή $Y_i = B_0 + B_1 X_i + \varepsilon_i \quad i=1,\ldots,n$

Ε.Ε.Τ.

μετά από πράξεις

$$\hat{B}_1 = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = 1{,}44$$

$$\hat{B}_0 = \bar{Y} - \hat{B}_1 \bar{X} = 9{,}27$$

Ερμηνεία ΕΕΤ.

→ η αναμενόμενη απόδοση σε μοναδιαία μεταβολή της θερμοκρασίας είναι 1,44

→ η αναμενόμενη απόδοση για θερμοκρασία 0° βαθ. είναι 9,27

και άρα το εκτιμώμενο μοντέλο είναι

$\hat{Y} = \hat{B}_0 + \hat{B}_1 X$ δηλ

$\boxed{\hat{Y} = 9{,}27 + 1{,}44x}$ → Αξία εκτιμώμενου μοντέλου:

Μπορώ να δώσω τιμές στο X και

να προβλέπω το Y

Χρησιμοποιώ το εκτιμώμενο μοντέλο

για προβλέψεις για τιμές του X κοντά

στο $\bar{X}$

Θεωρώ ότι οι υποθέσεις για τα σφάλματα ικανοποιούνται

β. Ανάλυση, αρχικό test $H_0: B_1=0$

Πηγή μεταβλ.	ΑΝΑΛΥΣΗ A.T.	B.E	M.S	F-πηλίκο
παλινδρόμηση (Μοντέλο)	226,95	1	226,95	96,17
Υπόλοιπα	21,23	9	2,36	
ολική	248,18	10		

Επομένως $SSreg >> SSres$ ή

$\underbrace{MSreg >> MSres}$ Απορρ H_0

δηλαδή το μοντέλο δείχνει να 'στέκει' άρα δεν μπορώ να δεχτώ ότι το $B_1=0$

γ. Στατιστικός Test έλεγχο $H_0: B_1=0$

t-test

F-test: το οποίο και έχω από τον πίνακα ανάλυσης αφού F-πηλίκο $= 96,17 > F_{1,9,0.05} = 5,12$

$\Rightarrow$ Απορρ H_0

δ. Το $(1-\alpha)100\%$ δ.ε για την B_1

- $\left(\hat{B}_1 - t_{9,0.025}\sqrt{\hat{Var}(B_1)},\ \hat{B}_1 + t_{9,0.025}\sqrt{\hat{Var}(B_1)}\right) = (1,11,\ 1,77)$

Αναμενόμενη απόδ. μοναδιαίας μεταβολ. θερμοκρ.

- $\hat{Var}(B_1) = \dfrac{\hat{\sigma}^2}{\Sigma(x_i-\bar{x})^2} = \dfrac{MSres}{\Sigma(x_i-\bar{x})^2}$

(9)

Ασκ. Η $\quad Y_i = B_0 + B_1 X_i + \varepsilon_i \quad i=1,\dots,n$

α. $r^2_{X,Y} = R^2 \qquad$ β. $r^2_{Y,\hat{Y}} = R^2 = r^2_{X,Y}$

α.

$$R^2 = \frac{SS_{res}}{SS_{tot}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(\hat{B}_0 + \hat{B}_1 X_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

$$= \frac{\sum(\bar{Y} - \hat{B}_1 \bar{X} + \hat{B}_1 X_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{\hat{B}_1^2 \sum(X_i - \bar{X})^2}{\sum(Y_i - \bar{Y})^2} \quad (1)$$

Αλλά

$$\hat{B}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \quad \text{και} \quad r_{X,Y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})}\sqrt{\sum(Y_i - \bar{Y})}}$$

δηλαδή $\hat{B}_1 = \left[\frac{\sum(Y_i - \bar{Y})^2}{\sum(X_i - \bar{X})^2}\right]^{1/2} r_{X,Y} \quad (2)$

οπότε άμεσα πλέον βγαίνει το συμπέρασμα

β. $r_{Y,\hat{Y}} \overset{\text{ορισ.}}{=} \frac{\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}})}{\sqrt{\sum(Y_i - \bar{Y})^2}\sqrt{\sum(\hat{Y}_i - \bar{\hat{Y}})^2}} \quad (1)$

- $\bar{\hat{Y}} = \frac{1}{n}\sum \hat{Y}_i = \frac{1}{n}\sum(\hat{B}_0 + \hat{B}_1 X_i)$

$$= \frac{1}{n}\hat{B}_0 + \hat{B}_1 \bar{X} = \bar{Y} - \hat{B}_1 \bar{X} + \hat{B}_1 \bar{X}$$

$$= \bar{Y} \quad (2)$$

- $r^2_{Y,\hat{Y}} = \frac{[\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})]^2}{\sum(Y_i - \bar{Y})^2 \sum(\hat{Y}_i - \bar{Y})^2}$

$$= \frac{[\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})]^2}{\sum(Y_i - \bar{Y})^2 \hat{B}_1^2 \sum(X_i - \bar{X})^2} \quad (3)$$

$$[\sum(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})]^2 = [\sum(Y_i - \bar{Y})(\hat{B}_0 + \hat{B}_1 X_i - \bar{Y})]^2$$

$$= [\sum(Y_i - \bar{Y})(\cancel{\bar{Y}} - \hat{B}_1 \bar{X} + \hat{B}_1 X_i - \cancel{\bar{Y}})]^2 \quad (4)$$

$$= \hat{B}_1^2 [\sum(Y_i - \bar{Y})(X_i - \bar{X})]^2$$

Τελικά

$$r^2_{Y,\hat{Y}} = \frac{\cancel{\hat{B}_1^2}[\sum(Y_i - \bar{Y})(X_i - \bar{X})]^2}{\sum(Y_i - \bar{Y})^2 \cancel{\hat{B}_1^2} \sum(X_i - \bar{X})^2}$$

$$= \frac{\sum(Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum(Y_i - \bar{Y})^2}\sqrt{\sum(X_i - \bar{X})^2}} \overset{\text{α.}}{=} R^2$$