13/4/2016

Απλή Γραμμική Παλινδρόμηση

- $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad , i=1,2,\dots,n$

$\varepsilon_i \sim N(0,\sigma^2)$ τυχαία σφάλματα ανεξάρτητες τ.μ.

$E(Y_i) = \beta_0 + \beta_1 x_i$: άγνωστη συνάρτηση παλινδρόμησης

$\varepsilon_i = Y_i - E(Y_i)$

$Y_i \sim N(E(Y_i), \sigma^2)$ $\qquad s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ $\qquad \frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}$

$(\hat{\varepsilon}_i =)\ e_i = Y_i - \hat{Y}_i$, $\quad \sum_{i=1}^{n} e_i = 0 \quad , \quad \sum_{i=1}^{n} e_i x_i = 0$

$$\min \sum_{i=1}^{n} e_i^2 \longrightarrow \begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \end{cases}$$

$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\right)$, $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$

- $(1-a)100\%$ Δ.Ε για β_1 : $\hat{\beta}_1 \pm t_{\frac{a}{2}, n-2} \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$

$H_0: \beta_1 = 0 \quad , H_a: \beta_1 \neq 0$

$$B_1 = \frac{\hat{\beta}_1 - \beta_1}{s\Big/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \sim t_{n-2} \text{ όταν } H_0 \text{ αληθής}$$

κρίσιμη περιοχή $|B_1| \geq t_{a/2, n-2}$

Στο παράδειγμα της ΕΒΟ

$B_1 = 42.6$ και $t_{a/2, n-2} = 2.306$ για $a = 0.05$ και $n = 10$

- $\hat{\beta}_1 = 2$, $\hat{\beta}_0 = 10$, $\hat{Y}_0 = 10 + 2x_0$

Η Ανάλυση της Διακύμανσης

$$Y_i - \bar{Y} = \underbrace{Y_i - \hat{Y}_i}_{=e_i} + \underbrace{\hat{Y}_i - \bar{Y}}_{\text{παλινδρόμηση}} \Rightarrow$$

$$\underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{\substack{\text{ολικό} \\ \text{άθροισμα τετραγώνων} \\ \text{total sum of squares} \\ (SS_{tot})}} = \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{\substack{\text{άθροισμα} \\ \text{τετραγώνων} \\ \text{υπολοίπων} \\ (SS_{res})}} + \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{\substack{\text{άθροισμα} \\ \text{τετραγώνων} \\ \text{παλινδρόμησης} \\ (SS_{reg})}} + 2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

όμως:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(\hat{\beta}_0 + \hat{\beta}_1 X_i - \bar{Y}) =$$

$$= \sum_{i=1}^{n} (Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i)(\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i - \bar{Y}) =$$

$$= \sum_{i=1}^{n} \left[(Y_i - \bar{Y}) - \hat{\beta}_1 (X_i - \bar{X}) \right] \hat{\beta}_1 (X_i - \bar{X}) =$$

$$= \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) - (\hat{\beta}_1)^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 =$$

$$= \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) - \hat{\beta}_1 \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 = 0$$

Άρα

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}) \Rightarrow$$

$$SS_{tot} = SS_{res} + SS_{reg}$$

$$\hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = (\hat{\beta}_1)^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$R^2 = \frac{SS_{reg}}{SS_{tot}}$$: συντελεστής προσδιορισμού

(Θέλουμε $R^2 \to 1$)

Για το SS_{tot} έχουμε $n-1$ βαθμούς ελευθερίας γιατί έχουμε n όρους και $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

Για το SS_{res} έχουμε $n-2$ γιατί έχουμε n όρους και $\sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} e_i x_i = 0$

Για το SS_{reg} έχουμε 1 βαθμό ελευθερίας γιατί

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{\beta}_0 - \hat{\beta}_1 x_i - \bar{Y})^2 = \sum_{i=1}^{n}(\bar{Y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{Y})^2 =$$

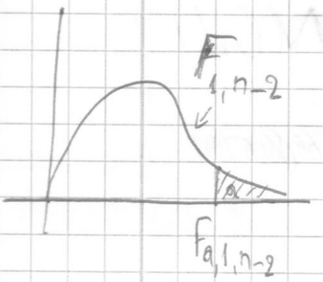
$$= (\hat{\beta}_1)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$$ (x_i όχι τυχαίες μεταβλητές)

Πίνακας ΑΝΑλυσης ΔΙΑκύμανσης (ΑΝΑΔΙΑ)

Προέλευση Μεταβλητότητας	Άθροισμα Τετραγώνων	Βαθμοί Ελευθερίας	Μέσα Τετράγωνα	f-test
Παλινδρόμηση	$SS_{reg} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = (\hat{\beta}_1)^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$	1	$MS_{reg} = \frac{SS_{reg}}{1} = SS_{reg}$ (mins square)	$F = \frac{MS_{reg}}{MS_{res}}$ απόρριψη: $F > F_{a,1,n-2}$
Ολική Μεταβλητότητα	$SS_{tot} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$n-1$	($MS_{tot} = \frac{SS_{tot}}{n-1}$)	
Υπόλοιπο μεταβλητότητας (απομένει από την παλινδρόμηση)	$SS_{res} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n-2$	$MS_{res} = \frac{SS_{res}}{n-2} = S^2$	

$$F = \frac{MS_{reg}}{MS_{res}} = \frac{\dfrac{SS_{reg}}{1}}{\dfrac{SS_{reg}}{n-2}} \equiv \frac{\dfrac{SS_{reg}/1}{\sigma^2}}{\dfrac{SS_{reg}/(n-2)}{\sigma^2}} \equiv \frac{X^2_{1/1}}{X^2_{n-2/n-2}} \equiv F_{1,n-2}$$ όταν $\beta_1 = 0$

Κρίσιμη περιοχή είναι $F \geq F_{a,1,n-2}$ για επίπεδο σημαντικότητας a



$$\frac{SS_{reg}}{\sigma^2} = \frac{(\hat{\beta}_1)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma^2} \sim X^2_1 \quad \text{για } \beta_1 = 0$$

αφού: $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2\right) \Longrightarrow$

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \sim N(0,1) \Longrightarrow \frac{(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma^2} \overset{\beta_1 = 0}{\sim} X^2_1$$

$$E\left((\hat{\beta}_1)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2\right) \underset{= aE(X)}{\overset{E(aX) =}{=}} \sum_{i=1}^{n} (x_i - \bar{x})^2 \, E\left((\hat{\beta}_1)^2\right) \qquad \underline{\underline{\operatorname{Var} X = E(X^2) - (E(X))^2}}$$

(η απόδειξη είναι δοσ)

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 \left[\operatorname{Var}(\hat{\beta}_1) + \left(E(\hat{\beta}_1)\right)^2 \right] =$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 \left(\frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \beta_1^2 \right) =$$

$$= \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Έλεγχος Καταλληλότητας του Μοντέλου

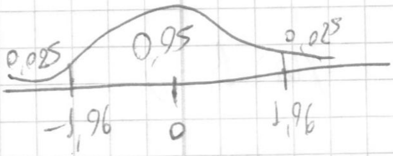
(επιτυχίας)

Ανάλυση Υπολοίπων

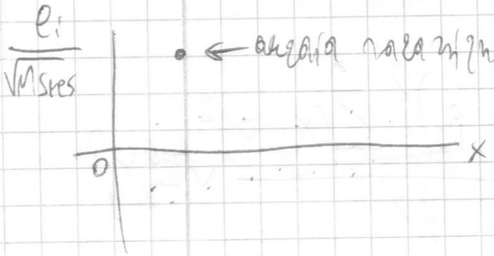
τυχαία σφάλματα $\varepsilon_i = Y_i - E(Y_i)$, $\varepsilon_i \sim N(0, \sigma^2)$, $\frac{\varepsilon_i}{\sigma} \sim N(0,1)$

$e_i = Y_i - \hat{Y}_i$, $\frac{e_i}{s} = \frac{e_i}{\sqrt{MS_{res}}} \overset{*}{\sim} N(0,1)$

(*) προσεγγιστικά

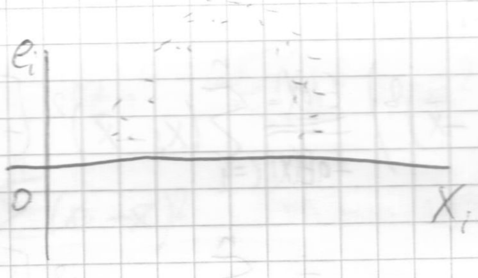


95% του $\frac{e_i}{\sqrt{MS_{res}}}$ στο $[-2, 2]$

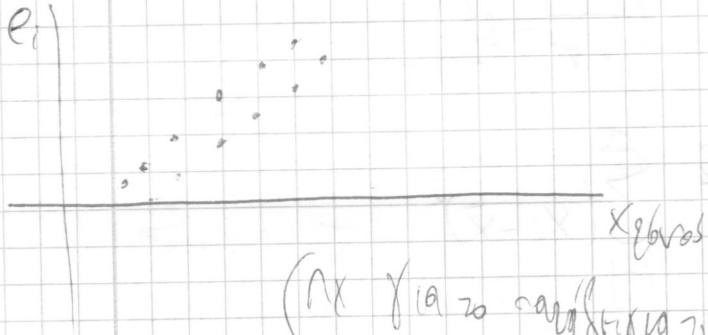


ομοιομορφία και σταθερότητα της διακύμανσης ✓

ένδειξη μη σταθερής διακύμανσης (αυξάνεται σχηματίζοντας μια χοάνη προς τα δεξιά)



ένδειξη καμπυλόγραμμης σχέσης



(π.χ. για το παράδειγμα της ΕΒΟ)

Χρειάζεται να προσθέσουμε κι άλλες παραμέτρους για να ληφθεί υπόψη και ο χρόνος

(εκτός ύλης)

Παράδειγμα 1 Λογαριθμικός Μετασχηματισμός

$$Y = \gamma_0 \gamma_1^X + \varepsilon \quad , E(\varepsilon) = 1 \quad , Var(\varepsilon) = \sigma^2$$

$$\log_{10} Y = \log_{10} \gamma_0 + X\log_{10}\gamma_1 + \varepsilon'$$

$$Y' = \beta_0 + X\beta_1 + \varepsilon'$$

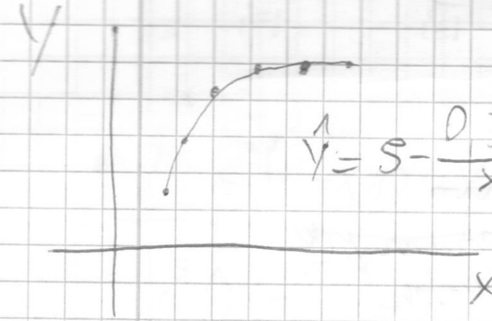
$$(X,Y) \longrightarrow (X, \log_{10} Y = Y')$$

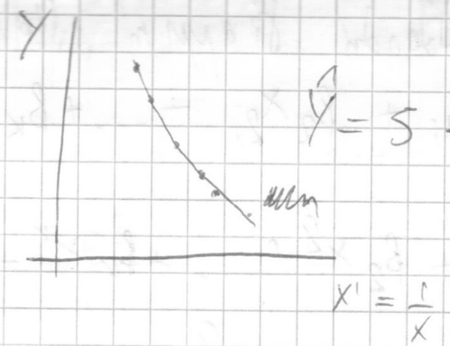
Παράδειγμα 2 Αντίστροφος Μετασχηματισμός

$$Y = \beta_0 + \frac{\beta_1}{X} + \varepsilon \quad , \quad (X,Y) \longrightarrow \left(X' = \frac{1}{X}, Y\right)$$

$$Y = \beta_0 + \beta_1 X' + \varepsilon$$

πχ





14/4/2016

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad E(Y_i) = \beta_0 + X_i \beta_1 \qquad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$H_0 : \beta_1 = 0 \quad V \quad H_a : \beta_1 \neq 0$$

$$t = \frac{\hat{\beta}_1 - \underbrace{\beta_1}_{\to 0}}{s / \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}} \quad (\sim t_{n-2} \text{ όταν } H_0 \text{ αληθής}) \qquad |t| \geq t_{\frac{a}{2}, n-2}$$

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

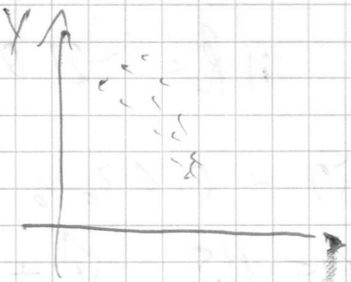
Δειγματικός Συντελεστής Συσχέτισης (του Pearson)

$$r = r(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad , \quad -1 \le r \le 1$$

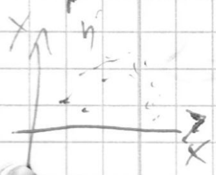
για τη δειγμα: (X_i, Y_i), $i=1,2,\dots,n$

$r \simeq 0$

$r \simeq 1$



$r \simeq -1$



$$r^2 = \frac{\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \cdot \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} =$$

$$= \frac{\hat{\beta}_1 \sum_{i=1}^{n}(x_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = \frac{SS_{reg}}{SS_{tot}} = R^2$$

Αν επιπλέον τα δεδομένα ακολουθούν την κανονική κατανομή τότε:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \quad \left(\sim t_{n-2} \text{ όταν } \rho=0\right)$$

Έτσι για τον έλεγχο της ~~...~~ $H_0: \rho=0 \; v \; H_a: \rho \neq 0$

κρίσιμη περιοχή $|t| \ge t_{\frac{a}{2}, n-2}$

Για το παράδειγμα της ΕΒΟ: $r = \frac{6800}{\sqrt{3400 \cdot 13660}} = 0,9978$

$$t = \frac{0,9978\sqrt{10-2}}{1-0,9978^2} = 42,6 > 2,306 = t_{0,025,8}$$ απορρίπτεται η $H_0: \rho=0$

Άσκηση 5.7 σελ 192

X	130	148	154	162	158	152	144	138	167
Y	31	42	45	46	49	43	46	36	50

$\hat{\beta}_1 = 0{,}436$ i) $\hat{\beta}_0 = -21{,}939$

$\hat{Y}_0 = -21{,}939 + 0{,}436 x_0$

ii) για $x_0 = 115$, $\hat{Y}_0 \pm t_{\alpha/2, n-2}\, s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \Rightarrow$

95% διάστημα εμπιστοσύνης $[21{,}976,\ 34{,}360]$

Προέλευση Μεταβλητότητας	Άθροισμα Τετραγώνων	Βαθμοί ελευθερίας	Μέσο τετράγωνο	F - τεστ
Παλινδρόμηση	195,634	1	195,634	$F = 34{,}049 > F_{0{,}05,1,8} = 5{,}32$ απορρίπτεται η H_0
Υπόλοιπα	45,966	8	5,746	
Ολική Μεταβλητότητα	241,600	9		

5.2 Άσκηση 191 σελίδα

$Y_i = \beta_0 + \beta_1 (x_i - \bar{x}) + \varepsilon_i$, $i = 1, 2, \ldots, n$ $\hat{\beta}_0, \hat{\beta}_1 = ;$

$\min \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 (x_i - \bar{x}))^2 = Q(\beta_0, \beta_1) \Rightarrow$

$\frac{dQ}{d\beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1(x_i - \bar{x})) = 0 \rightarrow \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \underbrace{\cancel{\sum_{i=1}^{n}(x_i - \bar{x})}}_{0} = n\hat{\beta}_0$

$\Rightarrow \hat{\beta}_0 = \bar{Y}$

$\frac{dQ}{d\beta_1} = -2 \sum_{i=1}^{n} (\beta_0 - \beta_1 (x_i - \bar{x}))(x_i - \bar{x}) = 0 \rightarrow \sum_{i=1}^{n} (x_i - \bar{x}) Y_i = \hat{\beta}_0 \underbrace{\cancel{\sum_{i=1}^{n}(x_i - \bar{x})}}_{0} + \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i-\bar{x})y_i}{\sum_{i=1}^{n} (x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n} (x_i-\bar{x})}$$