4º μάθημα

25/10/21

Τυπική Απόκλιση της πρόβλεψης $\hat{Y_k}$.

α.γ.π. : $Y_i = b_0 + b_1 \cdot X_i + \varepsilon_i$, με $i = 1, \ldots, n$

Εκτιμώμενο Μοντέλο: $\hat{Y} = \hat{b_0} + \hat{b_1} \cdot X$

Πρόβλεψη της Y για την τιμή X_k της X :

$\hat{Y_k} = \hat{b_0} + \hat{b_1} \cdot X$

Πόσο καλή είναι η πρόβλεψη $\hat{Y_k}$?

Αρκεί να μπορώ να βρω την $Var(\hat{Y_k}) = ?$

$$\left.\begin{aligned} \hat{Y_k} &= \hat{b_0} + \hat{b_1} \cdot X_k \\ \hat{b_0} &= \bar{Y} - \hat{b_1} \cdot \bar{X} \end{aligned}\right\} \hat{Y_k} = \bar{Y} + \hat{b_1} \cdot (X_k - \bar{X}).$$

Πάω να βρω την $Var(\hat{Y_k}) = ?$

$$\left.\begin{aligned} &Var(\hat{Y_k}) = Var(\bar{Y} + \hat{b_1} \cdot (X_k - \bar{X})) \\ &Var\left(\sum_{i=1}^{n} a_i W_i\right) = \sum_{i=1}^{n} a_i^2 Var W_i + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{m} a_i a_j Cov(W_i, W_j) \end{aligned}\right\}$$

- Αν οι W_i ανεξ/τες: $Var\left(\sum_{i=1}^{n} a_i W_i\right) = \sum_{i=1}^{n} a_i^2 Var(W_i)$

$$\Rightarrow Var(\hat{Y_k}) = Var(\bar{Y}) + (X_k - \bar{X})^2 Var(\hat{b_1}) + (X_k - \bar{X}) \cdot Cov(\bar{Y}, \hat{b_1}) \quad (*)$$

Αλλά $Cov(\bar{Y}, b_1) = Cov\left(\frac{1}{n} \cdot \frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2} \cdot \underbrace{Var(Y_i)}_{\sigma^2}\right) =$

$$= \sum_{i=1}^{n} \frac{1}{n} \cdot \frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2} \cdot Var(Y_i) = \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2} = 0 \quad (**)$$

Καθώς από Θ.Π.Σ. ισχύει ότι:

$$Cov\left(\sum_{i=1}^{n} a_i W_i, \sum_{i=1}^{n} b_i W_i\right) = \sum_{i=1}^{n} a_i b_i Var(W_i)$$

Υπενθ.: $\sum_{i=1}^{n} \frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2} = \frac{1}{\Sigma(X_i - \bar{X})^2} \cdot \Sigma(X_i - \bar{X}) =$

$$= \frac{1}{\Sigma(X_i - \bar{X})^2} (\Sigma X_i - n\bar{X}) = \frac{1}{\Sigma(X_i - \bar{X})^2} \cdot (\cancel{n\bar{X} - n\bar{X}}^{\,0}) = 0.$$

Από (*),(**) έχω: $\boxed{Var(\hat{Y}_k) = Var(\bar{Y}) + (X_k - \bar{x})^2 \cdot Var(\hat{b}_1)}$

μόνο αυτό μπορεί να αλλάξει όταν X_k κοντά στο $\bar{x}$

Άρα προβλέπω τιμές της X_k κοντά στο $\bar{x}$. Συνεπώς αν έχω μικρή διακύμανση → θα έχω καλή πρόβλεψη.

Συμπερασματικά: Η πρόβλεψη είναι καλή αν η $Var(\hat{Y}_k)$ είναι μικρή, δηλ. αν το $(X_k - \bar{x})^2$ είναι μικρό. Άρα πρέπει να χρησιμοποιώ το μοντέλο της α.γ.π. για πρόβλεψη για τιμές X_k της X, όχι μακριά από την $\bar{x}$.

F-τεστ για την παλινδρόμηση ή F-τεστ για τον έλεγχο της $H_0: b_1 = 0$ (διότι αν $b_1 = 0$ ∄ μοντέλο α.γ.π)

Αποδείξαμε ότι: $E(MSres) = \sigma^2$

Θδο: $E(MSreg) = \sigma^2 + b_1^2 \cdot \Sigma(X_i - \bar{x})^2$

Απόδειξη: $MSreg = \frac{SSreg}{1}$, άρα $E(MSreg) = E(SSreg) =$

$$= E\left(\hat{b}_1^2 \cdot \sum_{i=1}^{n} (X_i - \bar{x})^2\right) = \sum_{i=1}^{n} (X_i - \bar{x})^2 \cdot E(\hat{b}_1^2) =$$

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

$$= \sum_{i=1}^{n} (X_i - \bar{x})^2 \cdot \left[\frac{\sigma^2}{\Sigma(X_i - \bar{x})^2} + b_1^2 \right] = \sigma^2 + b_1^2 \cdot \Sigma(X_i - \bar{x})^2$$

- Αν η $H_0: b_1 = 0$ είναι αληθής, τότε $E(MSreg) = E(MSres)$
ή αν $H_0: b_1 = 0$ είναι αληθής, τότε $MSreg \approx MSres$

} ⇒ Ισχύει: Αν η πρόταση $A \Rightarrow B$, τότε $\sim B \Rightarrow \sim A$

⇒ Αν MSreg πολύ διαφορετικό από το MSres, τότε $b_1 \neq 0$ και επομένως η $H_0: b_1 = 0$ πρέπει να απορριφθεί.

Άρα ένα τεστ για τον έλεγχο της $H_0: b_1 = 0$ μπορεί να στηριχτεί στη σύγκριση των MSreg και MSres

Θα στηριχτώ για τη σύγκριση των MSreg και MSres στο πηλίκο $F = \frac{MSreg}{MSres}$

Χρειαζόμαστε κατανομή του F κάτω από την Ho.

Γνωρίζω: $\frac{SSres}{\sigma^2} \sim X^2_{n-2}$

Διερεύνηση κατανομής του SSreg

$$SSreg = \hat{b}_1^2 \cdot \Sigma(X_i - \bar{X})^2, \quad \hat{b}_1 \sim N\left(b_1, \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}\right)$$

$$\frac{\hat{b}_1 - b_1}{\sigma / \sqrt{\Sigma(X_i - \bar{X})^2}} \sim N(0,1) \text{ και υπό την } H_0: b_1 = 0$$

$$\frac{\hat{b}_1}{\sigma / \sqrt{\Sigma(X_i - \bar{X})^2}} \sim N(0,1) \Rightarrow \frac{\hat{b}_1^2}{\sigma^2 / \Sigma(X_i - \bar{X})^2} \sim X^2_1 \Rightarrow$$

$$\frac{\Sigma(X_i - \bar{X})^2 \cdot \hat{b}_1^2}{\sigma^2} \sim X^2_1 \Rightarrow \frac{SSreg}{\sigma^2} \sim X^2_1 \text{ υπό } H_0: b_1 = 0$$

Επομένως: $F = \frac{MSreg}{MSres} = \frac{SSreg/1}{SSres/(n-2)} = \frac{SSreg/\sigma^2 \cdot 1}{SSres/\sigma^2(n-2)} =$

$$= \frac{SSreg/\sigma^2 \cdot 1}{SSres/\sigma^2 \cdot (n-2)} \sim \frac{X_1^2/1}{X_{n-2}^2/(n-2)} \equiv F_{1,n-2}$$

με την προϋπόθεση αριθμητής και παρονομαστής να είναι ανεξάρτητοι μεταξύ τους.

<u>Μορφή της κ.π. (Κρίσιμης Περιοχής)</u> :

Πολύ διαφορετικές τιμές των MSreg, MSres ή πολύ μεγάλη τιμή του $F = \frac{MSreg}{MSres}$ ή $\boxed{F \geq c}$

<u>Προσδιορισμός του κ.σ. c</u> :

$a = P(\text{απορ. } H_0 \mid H_0 \text{ αληθής}) = P(F \geq c \mid F \sim F_{1,n-2}) =$
$= P(F_{1,n-2} \geq c) \Rightarrow \boxed{c = F_{1,n-2,a}}$

Για τον έλεγχο ύπαρξης παλινδρόμησης, δηλ. $H_0: \beta_1 = 0$ η $\underbrace{\text{ΣΣΤ}}_{\text{Στατ. Συν. Τεστ.}}$ είναι $F = \frac{MSreg}{MSres} \sim F_{1,n-2}$ υπό την H_0.

<u>Παρατήρηση</u>: Για τον έλεγχο της $H_0: \beta_1 = 0$ ∃ διαθέσιμο και το t-τεστ. Το t-τεστ και το F-τεστ είναι ισοδύναμα γιατί $F = t^2$ και (το αντίστροφο εκατοστιαίο σημείο) $\boxed{F_{1,n-2,a} = t^2_{n-2,\frac{a}{2}}}$

(στην πράξη χρησιμοποιείται το F γιατί είναι στον πίνακα Anova)

<u>Συντελεστής Συσχέτισης Pearson</u> :

X	$X_1, \ldots, X_n$
Y	$Y_1, \ldots, Y_n$

Ο δειγματικός συντελεστής συσχέτισης Pearson

$$r_{x,y} = r = \frac{\Sigma(X_i-\bar{X})\cdot(Y_i-\bar{Y})}{\sqrt{\Sigma(X_i-\bar{X})^2\cdot\Sigma(Y_i-\bar{Y})^2}}$$

Ιδιότητες:

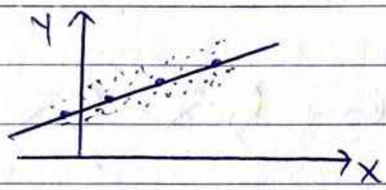
① Ο $r_{x,y}$ είναι απαλλαγμένος από μονάδες μέτρησης

② Είναι συμμετρικός, δηλ. $r_{x,y} = r_{y,x}$

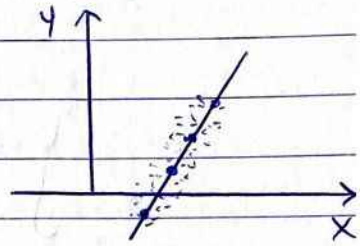
③ Παίρνει τιμές: $-1 \le r_{x,y} \le 1$

(απόδειξη μέσω ανισότητας Cauchy-Schwartz).

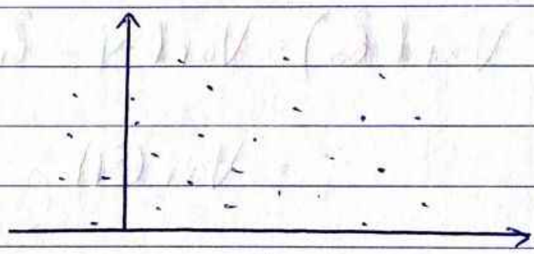
• $r_{x,y} = 1$



• $r_{x,y} = -1$



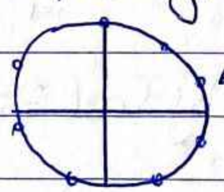
• $r_{x,y} = 0$



π.χ.

X	-3	-1	0	1	3
Y	9,54	9,95	10	9,95	9,54

Αν κάνω πράξεις, βρίσκω $r_{x,y} = 0$. Άρα δεν μπορώ να υποθετήσω γραμμική σχέση. Δηλ. έχω $X^2+Y^2=100$, άρα



όλα τα σημεία θα βρίσκονται πάνω στον κύκλο.

Νδο: $\hat{b_0} \sim N\left(b_0, \sigma^2\left\{\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(X_i-\bar{X})^2}\right\} = \frac{\sigma^2 \cdot \Sigma X_i^2}{n\Sigma(X_i-\bar{X})^2}\right)$

Λύση: $\hat{b}_0 = \bar{Y} - \hat{b}_1 . \bar{x}$

Το $\hat{b}_0 \sim$ Normal ως γρ. συνδυασμός των $\bar{Y}$ και $\hat{b}_1$ που έχουν κανονική κατανομή. Άρα:

- $E(\hat{b}_0) = E(\bar{Y} - \hat{b}_1 . \bar{x}) = E(\bar{Y}) - \bar{x} . E(\hat{b}_1) =$
$= E(\bar{Y}) - \bar{x} . b_1$

- $E(\bar{Y}) = E\left(\frac{1}{n} . \Sigma Y_i\right) = \frac{1}{n} . \Sigma E(Y_i) = \frac{1}{n} . \Sigma (b_0 + b_1 . x_i) =$
$= b_0 + b_1 . \bar{x}$

$\Rightarrow E(\hat{b}_0) = b_0 + \cancel{b_1 . \bar{x}} - \cancel{b_1 . \bar{x}} = b_0.$

$$Var(\hat{b}_0) = Var(\bar{Y} - \hat{b}_1 . \bar{x}) = Var(\bar{Y}) + \bar{x}^2 . Var(\hat{b}_1) - \bar{x} . \cancel{Cov(\bar{Y}, \hat{b}_1)}^{\,0}$$

$$= Var(\bar{Y}) + \bar{x}^2 . \frac{\sigma^2}{\Sigma (x_i - \bar{x})^2} = \frac{\sigma^2}{n} + \frac{\sigma^2 . \bar{x}^2}{\Sigma (x_i - \bar{x})^2}$$

▶ Να υπολογισθεί η $Cov(\hat{b}_0, \hat{b}_1) = ?$

Λύση: $Cov(\hat{b}_0, \hat{b}_1) = Cov(\bar{Y} - \hat{b}_1 . \bar{x}, \hat{b}_1)$

ισχύει ότι $Cov\left(\sum_{i=1}^{n} a_i z_i, \sum_{i=1}^{n} b_i w_i\right) =$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j \, Cov(z_i, w_j)$$

Άρα: $Cov(\hat{b}_0, \hat{b}_1) = Cov(\bar{Y}, \hat{b}_1) - \bar{x} . Cov(\hat{b}_1, \hat{b}_1) =$

$$= 0 - \bar{x} . Var(\hat{b}_1) =$$

$$= -\bar{x} . \frac{\sigma^2}{\Sigma (x_i - \bar{x})^2}$$

<u>Άσκηση 1η:</u> Έστω το μοντέλο α.γ.π.
$Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i \,,\ i=1,\dots,n$

Υποθέτουμε $\varepsilon_i \sim N(0,\sigma^2)$ με $i=1,\dots,n$ ασυσχέτιστα

α) Νδο Ε.Ε.Τ. των β_0,β_1 ταυτίζονται με Ε.Μ.Π. των β_0,β_1

β) Ο Ε.Μ.Π. της $\underset{\text{κοινή διακ/ση}}{\sigma^2}$ είναι: $\hat{\sigma}^2 = \dfrac{SS_{res}}{n}$

Έστω τ.δ. μεγέθους n: $X_1,\dots,X_n$ από $f(x,\theta)$ με $\theta \in \Theta$, θ άγνωστο.

<u>Πιθανοφάνειας:</u> $L(\theta) \overset{\text{ορ}}{=}$ η από κοινού κατανομή του δείγματος $X_1,\dots,X_n$ =
$= f_{X_1,\dots,X_n} \underset{\text{ανεξ.}}{\overset{X_1,\dots,X_n}{=}} \prod_{i=1}^{n} f_{X_i}(x_i,\theta)$

Ε.Μ.Π: $\hat{\theta} \overset{\text{ορ}}{=} \arg\max_{\theta\in\Theta} L(\theta)$

Ο ΕΜΠ $\hat{\theta}$ προκύπτει από τη μεγιστοποίηση ως προς θ της $L(\theta)$ ή του $\log L(\theta)$.

<u>Λύση:</u> α) ΕΕΤ ελαχιστοποιούν το $S = \sum_{i=1}^{n} \varepsilon_i^2$

$-\text{"}-$ ως προς β_0,β_1 $S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i)^2$

Πιθανοφάνεια στο μοντέλο της α.γ.π. είναι η από κοινού κατανομή των $Y_1,\dots,Y_n$

Δηλ. $L = L(\beta_0,\beta_1) = f_{Y_1,\dots,Y_n}(y_1,\dots,y_n) \underset{\text{ανεξ.}}{\overset{Y_1,\dots,Y_n}{=}} \prod_{i=1}^{n} f_{Y_i}(y_i)$ και επειδή

Αλλά: $Y_i \sim N(\beta_0 + \beta_1 \cdot X_i, \sigma^2)$, $i = 1, \dots, n$

Άρα: $f_{Y_i}(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2}$

Επομένως: $L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-1}{2\sigma^2} \cdot (Y_i - \beta_0 - \beta_1 X_i)^2} \Rightarrow$

$$L = \frac{1}{(\sigma\sqrt{2\pi})^n} \cdot e^{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i)^2}$$

- Οι ΕΜΠ των β_0, β_1 προκύπτουν από τη μεγιστοποίηση ως προς β_0, β_1 της L
- ή από μεγιστοποίηση ως προς β_0, β_1 της $\log L$
- ή —װ— του $-n \cdot \log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2} \Sigma (Y_i - \beta_0 - \beta_1 X_i)^2$
- ή από μεγ. ως προς β_0, β_1 του $-\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i)^2$
- ή ελαχ. του $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 \cdot X_i)^2$

Παρατηρώ ότι οι Ε.Ε.Τ. και οι Ε.Μ.Π. προκύπτουν από ελαχιστοποίηση της ίδιας ποσότητας, άρα θα ταυτίζονται.

β) → Σπίτι.