Διάλεξη 3η

Παλινδρόμηση και Ανάλυση Διακύμανσης 22/10/2019

Συμπερασματολογία (τεστ και διαστήματα εμπιστοσύνης) ως παραμέτρους β_0 και β_1 της α.γ.π $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ $i = 1, 2, \dots n$

Έχει νοημα να ενδιαφερθούμε για έλεγχο υποδέσεων ως παραμέτρους β_0 και β_1 ;

πχ Έχει νόημα να ενδιαφερθώ για έλεγχο της $H_0: \beta_1 = \beta_1^*$, β_1^* χωρίς ?
ΝΑΙ

Έχει νόημα να κατασκευάσω τεστ για έλεγχο $H_0: \beta_1 = \beta_1^*$ γιατί μου δίνει τη δυνατότητα να ελέγξω συγκεκριμένη μεταβλητή της Υ. για μοναδιαία μεταβολή της Χ.
Ακόμη, αν $\beta_1^* = 0$ τότε η αποδοχή της $H_0: \beta_1 = 0$ σημαίνει ότι δεν υφίσταται γραμμική σχέση μεταξύ των Χ και Υ.

Πώς θα κατασκευάσω στατιστικό τεστ για έλεγχο της $H_0: \beta_1 = \beta_1^*$ ή $H_0: \beta_0 = \beta_0^*$?

Τεχνική κατασκευής τεστ του Wald
ή κατασκευάζουμε τεστ σημαντικότητας.

Έστω τυχαίο δείγμα $W_1, .. W_n$ από $N(\mu, \sigma^2)$
Έστω ο έλεγχος $H_0: \mu = \mu_0$ έναντι $H_a: \mu \neq \mu_0$.

(I) σ^2 γνωστή (Z-τεστ)

Το στατιστικό τεστ για τον έλεγχο της $H_0: \mu = \mu_0$ έχει ΣΣΤ την $Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ με κατανομή $N(0,1)$ υπό την H_0 και κ.π $|z| \geq z_{\alpha/2}$

(II) σ^2 άγνωστο (t-τεστ)

Το στατιστικό τεστ για έλεγχο την $H_0: \mu = \mu_0$ έχει ΣΣΤ την $t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$ με κατανομή t_{n-1} υπό την H_0 και κπ $|t| \geq t_{n-1, \alpha/2}$.

* Παρατηρώντας την Z και t η ΣΣΤ έχει τη μορφή:

$$\text{ΣΣΤ} = \frac{\text{Εκτιμητής της παραμέτρου που εμφανίζεται στην } H_0 \; - \; \text{Η παράμετρος που εμφανίζεται στην } H_0}{\text{Τυπική Απόκλιση του Εκτιμητή}}$$

- Χρειάζομαι την κατανομή της ΣΣΤ υπό την H_0
- Χρειάζεται η μορφή της κρίσιμης περιοχής (περιοχή απόρριψης της H_0)
- Υπολογισμός του κρίσιμου σημείου $a = P(\text{Απορρ } H_0 / H_0 \text{ αληθής})$.

Υποθέτω ότι οι υποθέσεις για τα σφάλματα ει ισχύαν

Κατασκευή τεστ για έλεγχο της $H_0: \beta_1 = \beta_1^*$ (β_1^* γνωστό)

έναντι $H_a: \beta_1 \neq \beta_1^*$.

Ξεκινώ από εκτιμητή της β_1: $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right)$

υπό την $H_0: \beta_1 = \beta_1^*$, $\hat{\beta}_1 \sim N\left(\beta_1^*, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right)$

$$\frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\sigma^2/\sum(x_i-\bar{x})^2}} \sim N(0,1) \text{ υπό την } H_0$$

ΠΡΟΒΛΗΜΑ

Το σ^2 είναι άγνωστο

Ισχύει ότι $\frac{SSres}{\sigma^2} \sim \chi^2_{n-2}$

Θεωρώ το:

επειδή $\hat{\beta}_1$ ανεξ SSres.

$$t_{n-2} \equiv \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}} \underset{H_0}{\overset{\text{υπό}}{\sim}} t = \frac{\frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\sigma^2/\sum(x_i-\bar{x})^2}}}{\sqrt{\frac{SSres}{\sigma^2}/n-2}}$$

$$\Rightarrow t = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\frac{SSres/(n-2)}{\sum(x_i-\bar{x})^2}}} = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\frac{MSres}{\sum(x_i-\bar{x})^2}}}$$

ή

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{n-2} \text{ υπο } H_0$$

Η ΣΣΤ είναι $t = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{Var}(\hat{\beta}_1)}} \sim t_{n-2}$ υπο Ho

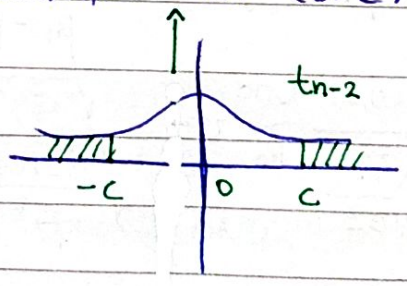
Και $\hat{Var}(\hat{\beta}_1) = \frac{MSres}{\Sigma(x_i - \bar{x})^2}$

Μικρές τιμες του t $\Rightarrow$ δεχόμαι την Ho
Για να απορρ την Ho θα παίρω μεγάλες τιμές του t.

Μορφή κρισιμης περιοχής. Μεγάλες τιμές της t.
Επειδή το t παίρνει και αρνητικές τιμές
μορφή της Κ.π είναι $|t| \geq c$ $\left(\Leftrightarrow \begin{matrix} t \geq c \\ \text{ή} \\ t \leq -c \end{matrix}\right)$

Υπολογισμός κρισιμου σημείου c

Πάντα $\alpha = P(\text{απορ Ho} / \text{Ho αληθής}) = P(|t| \geq c / \text{Ho αληθής})$
$= P(t \geq c \text{ ή } t \leq -c / \text{Ho αληθής}) = 2P(t \geq c / \text{Ho αληθής})$



$= 2P(t \geq c / t \sim t_{n-2}) = 2P(t_{n-2} \geq c)$
$\Rightarrow P(t_{n-2} \geq c) = \alpha/2$

άρα $c = t_{n-2, \alpha/2}$

Για τον έλεγχο της Ho: $\beta_1 = \beta_1^*$ v Ha: $\beta_1 \neq \beta_1^*$ η ΣΣΤ είναι:
$t = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{Var}(\hat{\beta}_1)}}$ με κατανομή t_{n-2} υπο την Ho

Και κη $|t| \geq t_{n-2, \alpha/2}$

όπου $\hat{Var}(\hat{\beta_1}) = \frac{MSres}{\sum(x_i - \bar{x})^2}$

Αντίστοιχα για την β_0.

⋮

όμως $\hat{Var}(\hat{\beta_0}) = \frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2} \cdot MSres$ (Η μόνη διαφορά σε σχέση με το β_1)

Κατασκευή διαστήματος εμπιστοσύνης (Δ.Ε) με βαθμό εμπιστοσύνης (β.ε) $100(1-\alpha)\%$ για την β_1.

Διάστημα εμπιστοσύνης

Έστω Τ.Δ $W_1, \dots W_n$ από πληθυσμό με κατανομή $f(x,\theta)$ (πχ εκθ $f(x,\theta) = \theta e^{-\theta x}, x>0$)

(*) Ορισμός: Ονομάζουμε δ.ε για την παράμετρο θ με β.ε $100(1-\alpha)\%$ κάθε διάστημα (L, U) όπου $L = L(W_1, \dots, W_n)$ και $U = U(W_1, \dots, W_n)$ τ.ω $P(L < \theta < U) = 1-\alpha$. ΜΕΘΟΔΟΣ ΑΝΤΙΣΤΡΕΠΤΗΣ ΠΟΣΟΤΗΤΑΣ.

πχ δ.ε για το μ $N(\mu, \sigma^2)$ → $\left(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ σ^2 γνωστό

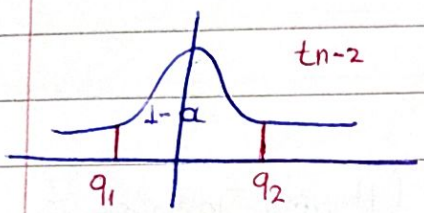
→ $\left(\bar{X} - t_{n-1, \frac{\alpha}{2}}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1, \frac{\alpha}{2}}\cdot\frac{S}{\sqrt{n}}\right)$ σ^2 άγνωστο

Βασίζομαι στον $\hat{\beta}_1$ και ειδικότερα στην $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} \sim t_{n-2}$

είναι αντιστρεπτή ποσότητα.

Αφού η t έχει t_{n-2}, υπάρχουν q_1, q_2 με $q_1 < q_2$ τ.ω

$$1-\alpha = P(q_1 < t < q_2) = P\left(q_1 < \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} < q_2\right)$$



λύνω προς β_1 προσεκτικά.

...

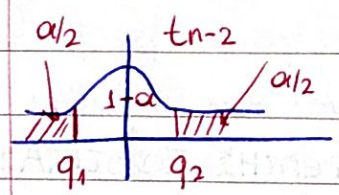
...

$$= P\left(\hat{\beta}_1 - q_2\sqrt{\hat{Var}(\hat{\beta}_1)} < \beta_1 < \hat{\beta}_1 - q_1\sqrt{\hat{Var}(\hat{\beta}_1)}\right)$$

Από τον ορισμό του Δ.Ε (*) ένα Δ.Ε για την β_1 με β.Ε $100(1-\alpha)\%$ είναι το $\left(\hat{\beta}_1 - q_2\sqrt{\hat{Var}(\hat{\beta}_1)}, \hat{\beta}_1 - q_1\sqrt{\hat{Var}(\hat{\beta}_1)}\right)$

Επιλέγω τα q_1 και q_2 ώστε να κόβουν ίσες ουρές από την t_{n-2} κατανομή

δηλ $P(t_{n-2} > q_2) = P(t_{n-2} < q_1) = \frac{\alpha}{2}$.



Από την $P(t_{n-2} > q_2) = \frac{\alpha}{2} \Rightarrow q_2 = t_{n-2, \alpha/2}$

Από την $P(t_{n-2} < q_1) = \frac{\alpha}{2}$

$\Rightarrow 1 - P(t_{n-2} > q_1) = 1 - \frac{\alpha}{2}$

$\Rightarrow q_1 = t_{n-1, 1-\alpha/2}$

Συμπερασματικά το $100(1-\alpha)\%$ Δ.Ε για β_1 είναι

$$\left(\hat{\beta}_1 - t_{n-2, \frac{\alpha}{2}}\sqrt{\hat{Var}(\hat{\beta}_1)}, \hat{\beta}_1 - t_{n-2, 1-\frac{\alpha}{2}}\sqrt{\hat{Var}(\hat{\beta}_1)}\right)$$

με $\hat{Var}(\hat{\beta}_1) = MSres / \sum(x_i - \bar{x})^2$

Το $100(1-\alpha)\%$ Δε για το β_0 είναι:

$$\left(\hat{\beta_0} - t_{n-2,\frac{\alpha}{2}}\sqrt{\hat{Var}(\hat{\beta_0})},\ \hat{\beta_0} - t_{n-2,1-\frac{\alpha}{2}}\sqrt{\hat{Var}(\hat{\beta_0})}\right)$$

με $\hat{Var}(\hat{\beta_0}) = \dfrac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\, MSres$

εκτίμηση - πρόσθιση του Y_K.

Έστω η πρόβλεψη $\hat{Y}_K = \hat{\beta_0} + \hat{\beta_1} X_K$.

Ποια η $Var(\hat{Y}_K)$?

$$\hat{Y}_K = \hat{\beta_0} + \hat{\beta_1} X_K$$

$$= \bar{Y} - \hat{\beta_1}\bar{X} + \hat{\beta_1} X_K = \bar{Y} + \hat{\beta_1}(X_K - \bar{X})$$

$$Var(\hat{Y}_K) = Var\left[\bar{Y} + \hat{\beta_1}\underbrace{(X_K - \bar{X})}_{\text{σταθερή ποσότητα}}\right]$$

Γνωρίζω ότι:

$$Var\left(\sum_{i=1}^{n} a_i W_i\right) = \sum_{i=1}^{n} a_i^2\, Var(W_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ i\neq j}}^{n} a_i a_j\, Cov(W_i, W_j)$$

Αν W_i ασυσχέτιστη τότε

$$Var\left(\sum_{i=1}^{n} a_i W_i\right) = \sum_{i=1}^{n} a_i^2 Var(W_i)$$

Άρα $Var(\hat{Y}_K) = Var(\bar{Y}) + (X_K - \bar{X})^2 Var(\hat{\beta_1}) + (X_K - \bar{X})\, Cov(\bar{Y}, \hat{\beta_1})$ (1)

όπου $Cov(\bar{Y}, \beta_1) = Cov\left(\frac{1}{n}\sum Y_i,\ \sum \frac{x_i - \bar{x}}{\sum(x_i-\bar{x})^2} Y_i\right)$

Αν $W_1, \dots W_n$ ασυσχ. τότε $Cov(\sum a_i W_i, \sum \beta_i W_i)$

$$= \sum_{i=1}^{n} a_i \beta_i\, Var(W_i)$$

$$\text{Cov}(\bar{Y}, \hat{\beta}_1) = \sum_{i=1}^{n} \frac{1}{n} \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})} \text{Var}(y_i)$$

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

$$= \frac{\sigma^2}{n} \frac{1}{\sum (x_i - \bar{x})^2} \sum_{i=1}^{n} (x_i - \bar{x}) \quad (\to 0)$$

$$= 0 \qquad (2)$$

Από (1) κ (2)

$$\boxed{\text{Var}(\hat{y}_k) = \text{Var}(\bar{y}) + (x_k - \bar{x})^2 \text{Var}(\hat{\beta}_1)}$$

- Η διακύμανση μικραίνει όταν x_k πολύ κοντά στο $\bar{x}$.

ΠΑΡΑΤΗΡΗΣΗ.

Όταν χρησιμοποιούμε το μοντέλο της αγη για προβλέψεις πρέπει να το χρησιμοποιούμε για τιμές της ανεξάρτητης μεταβλητής X κοντά στο $\bar{x}$, γιατί τότε $\text{Var}(\hat{y}_k)$ γίνεται μικρή και άρα η πρόβλεψη είναι αξιόπιστη

. Το F-τεστ για την Παλινδρόμηση ή Το F-τεστ για τον έλεγχο της $H_0: \beta_1 = 0$.

Γνωρίζουμε ότι $E(MSres) = \sigma^2$

Να βρούμε $E(MSreg) = ?$

$$MSreg = \frac{SSreg}{1} = SSreg = \hat{\beta}_1^2 \sum (x_i - \bar{x})^2$$

$$E(MSreg) = E(\hat{\beta}_1^2 \sum (x_i - \bar{x})^2) = \sum (x_i - \bar{x})^2 E(\hat{\beta}_1^2)$$

$\text{Var}(w) = E(w^2) - (E(w))^2$

$$\sum (x_i - \bar{x})^2 \left[\text{Var}(\hat{\beta}_1) + (E\hat{\beta}_1)^2 \right]$$

$$= \Sigma(x_i-\bar{x})^2\left[\frac{\sigma^2}{\Sigma(x_i-\bar{x})^2}+\beta_1^2\right] = \sigma^2+\beta_1^2\,\Sigma(x_i-\bar{x})^2$$

Συμπερασματικά: $E(MS_{res}) = \sigma^2$

$E(MS_{reg}) = \sigma^2+\beta_1^2\,\Sigma(x_i-\bar{x})^2$

Παρατηρώ ότι υπό την $H_0: \beta_1=0$

$E(MS_{res}) = E(MS_{reg}) = \sigma^2$.

Άρα αν $\beta_1=0$ τότε $MS_{res} \approx MS_{reg}$

(Αν $A \Rightarrow B$ τότε $\sim B \Rightarrow \sim A$ ($\sim$: άρνηση))

Αν MS_{res} είναι πολύ διαφορετικό από το MS_{reg} τότε $\beta_1 \neq 0$, δηλ τότε Απορ $H_0: \beta_1=0$.

ΑΡΑ ένα στατιστικό τεστ για τον έλεγχο της $H_0: \beta_1=0$ μπορεί να βασιστεί στη σύγκριση του MS_{res} με το MS_{reg}.

Ειδικότερα: Αν MS_{res} πολύ διαφορετικό από MS_{reg} τότε απορρίπτω την $H_0: \beta_1=0$.

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

Μήπως και $\frac{SS_{reg}}{\sigma^2} \sim \chi^2$??

Είναι $SS_{reg} = \hat{\beta}_1^2\,\Sigma(x_i-\bar{x})^2$

Ισχύει $\hat{\beta_1} \sim N\left(\beta_1, \frac{\sigma^2}{\Sigma(x_i-\bar{x})^2}\right)$

και υπό την $H_0: \beta_1=0$, $\hat{\beta_1} \sim N\left(0, \frac{\sigma^2}{\Sigma(x_i-\bar{x})^2}\right)$

υπο την $H_0: \beta_1=0$

$$\frac{\hat{\beta_1}}{\frac{\sigma}{\sqrt{\Sigma(x_i-\bar{x})^2}}} \sim N(0,1)$$

$\Rightarrow$ υπό την $H_0: \beta_1=0$

$$\eta \ \frac{\sqrt{\Sigma(x_i-\bar{x})^2}\hat{\beta_1}}{\sigma} \sim N(0,1) \Rightarrow \frac{\Sigma(x_i-\bar{x})^2\hat{\beta_1}}{\sigma^2} \sim N(0,1)^2 \equiv \chi^2_1.$$

$\Rightarrow$ υπό την $H_0: \beta_1=0$ το $\frac{SSreg}{\sigma^2} \sim \chi^2_1$

Θεωρώ της ποσότητα

$$F = \frac{MSreg}{MSres} = \frac{SSreg/1}{SSres/(n-2)}$$ διαιρώ με σ^2 αριθμ και παρανομ.

$\Rightarrow$

$$F = \frac{SSreg/\sigma^2}{SSres/\sigma^2(n-2)} = \frac{\chi_1^2/1}{\chi^2_{n-2}/(n-2)} \equiv F_{1,n-2}$$

Συμπερασματικά: Υπό την $H_0: \beta_1=0$ το F-πηλίκο

$$F = \frac{MSreg}{MSres} \sim F_{1,n-2}.$$

- Μορφή κπ Μεγάλες τιμές του F (γιατί τότε το MSreg πολύ διαφορετικό MSres).
  δηλ $F \geq c$.

• Υπολογισμός Κ.σ. C:

$\alpha = P(\text{απφ } H_0 / H_0 \text{ αληθής})$
$= P(F \geq c / F \sim F_{1,n-2}) = P(F_{1,n-2} \geq c)$

ορισμό ⟹ Εκατοστιαίων σημείων της $F_{1,n-2}$

$$\boxed{c = F_{1,n-2,\alpha}}$$

Για τον έλεγχο της $H_0: \beta_1 = 0$
η ΣΣΤ είναι:

$$F = \frac{MSreg}{MSres} \quad \text{με κατανομή } F_{1,n-2}$$

Υπό την $H_0: \beta_1 = 0$ και Κ.Π. $F \geq F_{1,n-2,\alpha}$.

ΠΑΡΑΤΗΡΗΣΗ.

Το F τεστ για τον έλεγχο της $H_0: \beta_1 = 0$ είναι ισοδύναμο με το t-τεστ διότι:

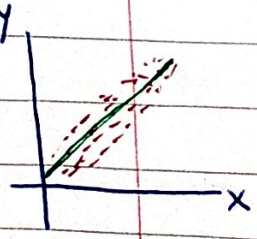
i) $F = t^2$

ii) $F_{1,n-2,\alpha} = t^2_{n-1,\alpha/2}$

Συντελεστής:

Συσχέτισης - Pearson.

Διάγραμμα Διασποράς: (δ.δ) το Μαθηματικό ανάλογο ⟹ του δ.δ



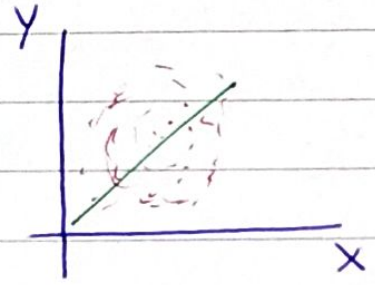
$$r_{x,y} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

Ιδιότητες του $r_{x,y}$.

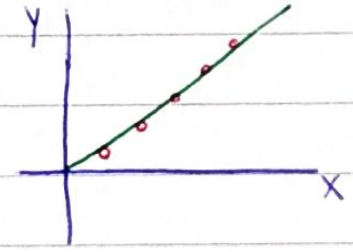
① $r_{x,y}$ καθαρός αριθμός

② $-1 \leq r_{x,y} \leq 1$

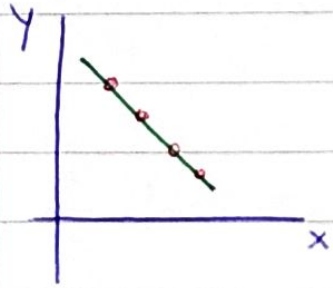
Ιδιαίτερο ενδιαφέρον παρουσιάζουν οι τιμές $0, -1, 1$.



$r_{x,y} = 0$ ← αντιστοιχεί στην μη ύπαρξη γραμμικής σχέσης



$r_{x,y} = 1$ ← πλήρης θετική γραμμική σχέση.



$r_{x,y} = -1$ ← απόλυτη αρνητική γραμμική σχέση

ΠΑΡΑΤΗΡΗΣΗ

Ο Συντελεστής Pearson εκφράζει <u>μόνο</u> γραμμική σχέση.
Δηλ το $r_{x,y} = 0$ σημαίνει ότι $\nexists$ γραμμική σχέση
Μπορεί όμως να $\exists$ άλλου είδους σχέση.

πχ:

X	-3	-1	0	1	3
Y	9.54	9.95	10	9.95	9.54

$r_{x,y} = 0$ Άρα $\nexists$ γραμμική σχέση μεταξύ X,Y
όμως $X^2 + Y^2 = 100$