Μάθημα 3ο 23/10/17

Στατιστική συμπερασματολογία (τεστ, δ.ε.) για τις b_0, b_1, παραμέτρους του μοντέλου της α.γ.π.
Θεωρώ το μοντέλο της α.γ.π. $Y_i = b_0 + b_1 X_i + \varepsilon_i$, $i=1,\dots,n$ και υποθέτω ότι οι υποθέσεις για τα σφάλματα ικανοποιούνται.

Αναγκαιότητα τεστ για b_0, b_1

Έχει νόημα $H_0: b_1 = b_1^*$ (b_1^* γνωστό) ;;;
$H_0: b_1 = 0$

ΝΑΙ γιατί $H_0: b_1 = b_1^*$ ελέγχει αν η μεταβλητότητα της Y για μοναδιαία μεταβολή της X είναι ίση με b_1^*
ή γιατί αν ισχύει η $H_0: b_1 = 0$ τότε $\nexists$ σχέση μεταξύ Y και X.

Τεστ για έλεγχο $H_0: b_1 = b_1^*$ (b_1^* γνωστό) έναντι $H_a: b_1 \neq b_1^*$

Τεστ κατά Wald: Έστω τ.δ. $W_1, \dots, W_n$ από $N(\mu, \sigma^2)$, $H_0: \mu = \mu_0$ (μ_0 γνωστό)
$H_a: \mu \neq \mu_0$

σ^2 γνωστό $\longrightarrow$ Z τεστ, $Z = \dfrac{\overline{W} - \mu_0}{\sigma/\sqrt{n}}$

σ^2 άγνωστο $\longrightarrow$ t-test, $t = \dfrac{\overline{W} - \mu_0}{S/\sqrt{n}}$, $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (W_i - \overline{W})^2$

της παραμέτρου που με ενδιαφέρει

$$\Sigma.\Sigma.T. = \frac{\text{Εκτιμητής} - E(\text{Εκτιμητή})}{\text{Τυπική απόκλιση (εκτιμητή)}}$$ (τουλάχιστον για μεγάλα δείγματα)

(i) Κατανομή της Σ.Σ.Τ. υπό την H_0

(ii) Κρίσιμο σημείο $\leftarrow$ $a = P(\text{απορ. } H_0 \mid H_0 \text{ αληθεύει}) = P(\text{σφάλμα τύπου I})$.

Αφού θέλω τεστ για b_1 στηρίζομαι σε ένα εκτιμητή της b_1, την $\hat{b}_1 \sim N\left(b_1, \dfrac{\sigma^2}{\sum (X_i - \overline{X})^2}\right)$

Κατασκευή στατιστικού τεστ:

Υπό την $H_0: \beta_1=\beta_1^*$, $\hat{\beta}_1 \sim N\left(\beta_1^*, \dfrac{\sigma^2}{\sum(X_i-\bar{X})^2}\right)$ έναντι $H_a: \beta_1 \neq \beta_1^*$ (β_1^* γνωστό)

(Ανάλογα για την $H_0: \beta_0=\beta_0^*$)

$$\frac{\hat{\beta}_1-\beta_1^*}{\sqrt{\dfrac{\sigma^2}{\sum(X_i-\bar{X})^2}}} \sim N(0,1) \text{ υπό την } H_0: \beta_1=\beta_1^*$$

$t_v \simeq \dfrac{N(0,1)}{\sqrt{\chi^2_v/v}}$ οπότε, εφόσον $\dfrac{SS_{res}}{\sigma^2} \sim \chi^2_{n-2}$ θα έχουμε:

$$t = \frac{\dfrac{\hat{\beta}_1-\beta_1^*}{\sqrt{\dfrac{\cancel{\sigma^2}}{\sum(X_i-\bar{X})^2}}}}{\sqrt{\dfrac{\dfrac{SS_{res}}{\cancel{\sigma^2}}}{(n-2)}}} = \frac{\hat{\beta}_1-\beta_1^*}{\dfrac{1}{\sqrt{\sum(X_i-\bar{X})^2}}\sqrt{\dfrac{SS_{res}}{n-2}}} = \frac{\hat{\beta}_1-\beta_1^*}{\dfrac{1}{\sqrt{\sum(X_i-\bar{X})^2}}\sqrt{MS_{res}}} \Rightarrow$$

$$\Rightarrow t = \frac{\hat{\beta}_1-\beta_1^*}{\sqrt{\dfrac{MS_{res}}{\sum(X_i-\bar{X})^2}}}$$

Το $t = \dfrac{N(0,1)}{\sqrt{\dfrac{\dfrac{SS_{res}}{\sigma^2}}{(n-2)}}} = \dfrac{N(0,1)}{\sqrt{\chi^2_{n-2}/(n-2)}} \overset{\text{ανεξ.}}{\sim} t_{n-2}$ $\hat{\beta}_1$ ανεξάρτητο από SS_{res}.

Συνοψίζοντας μέχρι εδώ η ΣΣΤ:

Υπό την $H_0: \beta_1=\beta_1^*$ η $t = \dfrac{\hat{\beta}_1-\beta_1^*}{\sqrt{\dfrac{MS_{res}}{\sum(X_i-\bar{X})^2}}} \sim t_{n-2}$ ή $t = \dfrac{\hat{\beta}_1-\beta_1^*}{\sqrt{\widehat{Var}(\hat{\beta}_1)}}$ καθώς $\hat{\sigma}^2 = MS_{res}$

Κρίσιμη περιοχή:

Αν το t έχει μεγάλες τιμές τότε $\hat{\beta}_1$ πολύ διαφορετικό από β_1^*, τότε και το β_1 πολύ διαφορετικό από το β_1^*, άρα απορ. H_0.

Τελικά μεγάλες τιμές του t οδηγούν σε απόρριψη της $H_0: \beta_1=\beta_1^*$

Άρα η κρίσιμη περιοχή αποτελείται από μεγάλες τιμές του t, δηλ. $|t| \geq c$

Άρα, μορφή κ.π. $|t| \geq c$.

- Αρκεί να προσδιορίσω το κρίσιμο σημείο c.

Πάντα από $P(\text{σφάλμα τύπου I}) = P(\text{απορ. } H_0 \mid H_0 \text{ αληθεύει}) = a = P(|t| \geq c \mid t \sim t_{n-2})$

$= P(|t_{n-2}| \geq c) =$

$= P(t_{n-2} \geq c \text{ ή } t_{n-2} \leq -c)$

$= 2P(t_{n-2} \geq c)$

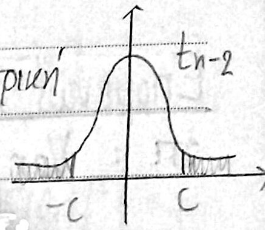
$\Rightarrow c = t_{n-2,\frac{a}{2}}$

Παρατηρώντας ότι $\sqrt{\dfrac{MS_{res}}{\Sigma(x_i-\bar{x})^2}} = \sqrt{\dfrac{\hat{\sigma}^2}{\Sigma(x_i-\bar{x})^2}} = \sqrt{\hat{Var}(\hat{\beta}_1)}$ ↳ ένας εκτιμητής

Οπότε: για τον έλεγχο της $H_0: \beta_1 = \beta_1^*$ έναντι $H_a: \beta_1 \neq \beta_1^*$ η ΣΣΤ είναι

$t = \dfrac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{Var}(\hat{\beta}_1)}}$ με κατανομή t_{n-2} υπό H_0 και κ.π. μεγέθους a, την $|t| \geq t_{n-2,\frac{a}{2}}$

όπου $\hat{Var}(\hat{\beta}_1) = \dfrac{MS_{res}}{\Sigma(x_i-\bar{x})^2}$



Κατασκευή δ.ε. για την παράμετρο β_1

Επειδή $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}} \sim t_{n-2}$ είναι αντίστρεπτη, δηλ. $\exists\, q_1, q_2$ ώστε:

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

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

Αντίστρεπτη ποσότητα
1. Συνάρτηση της παραμέτρου
2. Συνάρτηση των δεδομένων
3. Η κατανομή της είναι ανεξάρτητη από την παράμετρο.

Άρα, $100(1-a)\%$ δ.ε. για την β_1 είναι $\left(\hat{\beta}_1 - q_2\sqrt{\hat{Var}(\hat{\beta}_1)},\ \hat{\beta}_1 - q_1\sqrt{\hat{Var}(\hat{\beta}_1)}\right)$

Το δ.ε. ίσων ουρών προκύπτει για $q_2 = -q_1$ και $q_1 = -t_{n-2,a/2}$, $q_2 = t_{n-2,a/2}$ } λόγω συμμετρίας

Οπότε Το $100(1-a)\%$ δ.ε. για β_1 είναι $\left(\hat{\beta}_1 - t_{n-2,\frac{a}{2}}\sqrt{\hat{Var}(\hat{\beta}_1)},\ \hat{\beta}_1 + t_{n-2,\frac{a}{2}}\sqrt{\hat{Var}(\hat{\beta}_1)}\right)$

Ανάλογα Για τον έλεγχο $H_0: \beta_0 = \beta_0^*$ (β_0^* γνωστό) έναντι $H_a: \beta_0 \neq \beta_0^*$ η ΣΣΤ είναι

$t = \dfrac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\hat{Var}(\hat{\beta}_0)}}$, με κατανομή t_{n-2}, υπό H_0 και κ.π. μεγέθους a την

$|t| \geq t_{n-2,\frac{a}{2}}$ όπου $\hat{Var}(\hat{\beta}_0) = \sqrt{\dfrac{\Sigma(x_i)^2}{n\,\Sigma(x_i-\bar{x})^2}}\, MS_{res}$ και $100(1-a)\%$

είναι το $\left(\hat{\beta}_0 - t_{n-2,\frac{a}{2}}\sqrt{\hat{Var}(\hat{\beta}_0)},\ \hat{\beta}_0 + t_{n-2,\frac{a}{2}}\sqrt{\hat{Var}(\hat{\beta}_0)}\right)$

Διακύμανση της πρόβλεψης

Το μοντέλο της α.γ.π. χρησιμοποιείται για προβλέψεις.
Ειδικότερα το εκτιμώμενο μοντέλο χρησιμοποιείται για προβλέψεις.
Έτσι η πρόβλεψη της Y όταν $X = X_k$ είναι $\hat{Y}_k = \hat{b}_0 + \hat{b}_1 X_k$

ΕΡΩΤΗΜΑ:

Όμως το θέμα είναι, πόσο καλή είναι η πρόβλεψη;
Η πρόβλεψη $\hat{Y}_k$ είναι καλή όταν $\hat{Var}(\hat{Y}_k)$ είναι μικρή.

- Υπολογισμός $Var(\hat{Y}_k)$

$$\hat{Y}_k = \hat{b}_0 + \hat{b}_1 X_k = \bar{Y} - \hat{b}_1 \bar{X} + \hat{b}_1 X_k = \bar{Y} - \hat{b}_1(\bar{X} - X_k)$$

Επομένως, λοιπόν: $Var(\hat{Y}_k) = Var(\bar{Y} - \hat{b}(\bar{X} - X_k))$ για να το υπολογίσω θυμάμαι

ότι: $Var\left(\sum_{i=1}^{n} a_i W_i\right) = \sum a_i^2 Var(W_i) + \sum_i \sum_j a_i a_j Cov(W_i, W_j)$ για $i,j = 1,\dots,n$, $i \neq j$

οπότε: $Var(\hat{Y}_k) = Var(\bar{Y}) + (\bar{X} - X_k)^2 Var(\hat{b}_1) - (\bar{X} - X_k) Cov(\bar{Y}, \hat{b}_1)$ (1)

Ξέρω ότι η συνδιακύμανση είναι: $Cov(\bar{Y}, \hat{b}_1) = Cov\left(\frac{1}{n}\sum Y_i, \sum \frac{(X_i - \bar{X})}{\sum(X_i - \bar{X})^2} Y_i\right) =$

$Cov\left(\sum_{i=1}^{n} a_i W_i, \sum_{j=1}^{m} b_j Z_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(W_i, Z_j)$

$$= Cov\left(\sum \frac{1}{n} Y_i, \sum \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2} Y_i\right) \quad (2)$$

Άρα, τελικά, από (2), (3) η συνδιακύμανση είναι: ($W_i = Z_i = Y_i$)
($Cov(Y_i, Y_j) = 0$, Y_i ανεξάρτητα $i \neq j$)

$$Cov(\bar{Y}, \hat{b}_1) = Cov\left(\sum_{i=1}^{n} \underbrace{\frac{1}{n}}_{a_i} Y_i, \sum_{i=1}^{n} \underbrace{\frac{(X_i - \bar{X})}{\sum(X_i - \bar{X})^2}}_{b_i} Y_i\right) = \sum_{i=1}^{n} \frac{1}{n} \frac{X_i - \bar{X}}{\sum(X_i - \bar{X})^2} Var(Y_i) =$$

$$= \sum_{i=1}^{n} \frac{1}{n} \frac{(X_i - \bar{X})}{\sum(X_i - \bar{X})^2} \sigma^2 = \frac{\sigma^2}{n} \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{\sum(X_i - \bar{X})^2} = \frac{\sigma^2}{n} \frac{\overbrace{\sum(X_i - \bar{X})}^{0}}{\left(\sum(X_i - \bar{X})\right)^2} \quad (4)$$

$$\left(\sum(X_i - \bar{X}) = \sum X_i - n\bar{X} = 0\right)$$

δηλ. από (1), (4)

$$\mathbf{Var(\hat{Y}_k) = Var(\bar{Y}) + (\bar{X} - X_k)^2 \, Var(\hat{b}_1)}$$

Για να είναι η πρόβλεψη Y_k όσο πιο αξιόπιστη μπορεί, πρέπει $Var(\hat{Y}_k)$ όσο δυνατόν ελάχιστη.
Άρα, τι πρέπει να συμβαίνει για να είναι η $Var(\hat{Y}_k)$ όσο πιο μικρή γίνεται,
δηλ. αρκεί το X_k να είναι κοντά στο $\bar{X}$ ή το $(X_k - \bar{X})^2$ όσο το δυνατόν ελάχιστο

Συμπέρασμα: Όταν το εκτιμώμενο μοντέλο $\hat{Y} = \hat{b}_0 + \hat{b}_1 X$ χρησιμοποιείται για προβλέψεις θα πρέπει η πρόβλεψη να γίνεται για τιμή του X κοντά στο σύνολο των δεδομένων, για την X_k κοντά στο $\bar{X}$

Το F-τεστ για έλεγχο παλινδρόμισης ή

το F-τεστ για έλεγχο της $H_0: b_1=0$ έναντι $H_a: b_1 \neq 0$.

Για τον έλεγχο αυτό κατασκευάσαμε t-τεστ ($H_0: b_1 = b_1^*$, b_1 γνωστό)

Ιδέα: Γνωρίζουμε ότι $E(MSres)=\sigma^2$

Ισχύει: $E(MSreg) = \sigma^2 + b_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2$

Πράγματι, $E(MSreg) = E\left[\frac{SSreg}{1}\right] = E(SSreg) = E\left[\hat{b}_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2\right] =$

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

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

$$\Rightarrow \boxed{E[MSreg] = \sigma^2 + b_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Υπό την $H_0: b_1=0$ τότε $E(MSres) = E(MSreg)$ ή $MSres \approx MSreg$.

Προτασιακός Λογισμός: Αν $A \Rightarrow B$ τότε $\sim B \Rightarrow \sim A$

$\Rightarrow$

Αν MSres πολύ διαφορετικό από MSreg, τότε η $H_0: b_1=0$ δεν ισχύει, άρα απορρίπτεται.

Συμπερασματικά:

- Άρα, ένα τεστ για τον έλεγχο της $H_0: b_1=0$ μπορεί να στηριχτεί στη σύγκριση των MSreg με το MSres.

Γνωρίζοντας ότι $\frac{SSres}{\sigma^2} \sim \chi^2_{n+2}$, συγκρίνω τα MSres και MSreg με βάση το

πηλίκο τους: $\frac{MSreg}{MSres} = \frac{SSreg/1}{SSres/(n-2)}$ ή μπορώ να ελέγξω την διαφορά τους αν είναι κοντά στο μηδέν.

▸ Απομένει να βρω κατανομή για SSreg.

{Το $\frac{SSreg}{\sigma^2} \overset{\text{υπό } H_0: \beta_1=0}{\sim} \chi_1^2$ αν οι υποθέσεις για τα σφάλματα ικανοποιούνται}

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

Υπό τις υποθέσεις για σφάλματα: $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right)$ υπό την $H_0: \beta_1 = 0$.

$$\hat{\beta}_1 \sim N\left(0, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right) \Rightarrow \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{\sigma^2}{\sum(x_i-\bar{x})^2}}} \sim N(0,1) \Rightarrow$$

$$\Rightarrow \frac{\hat{\beta}_1 \sqrt{\sum(x_i-\bar{x})^2}}{\sigma} \sim N(0,1) \Rightarrow \frac{\hat{\beta}_1^2 \sum(x_i-\bar{x})^2}{\sigma^2} \sim \chi_1^2 \Rightarrow$$

$$\Rightarrow \frac{SSreg}{\sigma^2} \sim \chi_1^2 \quad \text{υπό την } H_0: \beta_1 = 0.$$

Θεωρώ τη στατιστική συνάρτηση: $F = \frac{MSreg}{MSres} = \frac{SSreg/1}{SSres/(n-2)} = \frac{(SSreg/\sigma^2)/1}{(SSres/\sigma^2)/(n-2)} =$

$$F \equiv \frac{\chi_1^2/1}{\chi_{n-2}^2/n-2} \sim F_{1,n-2} \text{ υπό } H_0: \beta_1 = 0.$$

χ_1^2 και χ_{n-2}^2 ανεξάρτητες
↓
αν SSreg ανεξάρτητο SSres. (SSreg εξαρτάται από $\hat{\beta}_1$ το οποίο είναι ανεξάρτητο SSres)
↓
Ισχύει αλλά η απόδειξη ξεφεύγει από το μάθημα.

Συγκεντρωτικά:

• Για τον έλεγχο της $H_0: \beta_1 = 0$ η σ.σ. του τεστ είναι $F = \frac{MSreg}{MSres}$ με κατανομή $F_{1,n-2}$ υπό H_0 και κ.π. της μορφής: μεγάλες τιμές του F, δηλ. $F \geq c$ → κρίσιμο σημείο
$\Rightarrow F \geq F_{1,n-2,a}$

Υπολογισμός κ.σ. c:

Μορφή της κρίσιμης περιοχής: Μεγάλες τιμές του F, δηλαδή $F \geq c$.
$a = P(\text{Απορ. } H_0 \mid H_0 \text{ αληθ.}) = P(F \geq c \mid F \sim F_{1,n-2}) = P(F_{1,n-2} \geq c) \Rightarrow c = F_{1,n-2,a}$.

ΠΑΡΑΤΗΡΗΣΗ: Το F τεστ ισοδύναμο με το t-τεστ γιατί $F = \frac{MSreg}{MSres} = \frac{SSreg}{SSres} =$ →
$\left(t = \frac{\hat{\beta}_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{n-2}\right)$

$$= \frac{\hat{\beta}_1^2 \sum (x_i-\bar{x})^2}{MSres} = \left(\frac{\hat{\beta}_1}{\sqrt{\frac{MSres}{\sum (x_i-\bar{x})^2}}} \right)^2 = \left(\frac{\hat{\beta}_1 - 0}{\sqrt{\hat{Var}(\hat{\beta}_1)}} \right)^2 = t^2$$

Συντελεστής Συσχέτισης Pearson

Πληθυσμιακή μορφή: Αν X, Y είναι τ.μ. $\longrightarrow \rho(X,Y) = \dfrac{Cov(X,Y)}{\sqrt{Var X}\sqrt{Var Y}}$

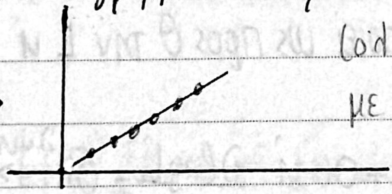
Δειγματική μορφή: Αν έχω τ.δ. $x_1, \ldots, x_n$, $y_1, \ldots, y_n$ $\longrightarrow r_{X,Y} = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$

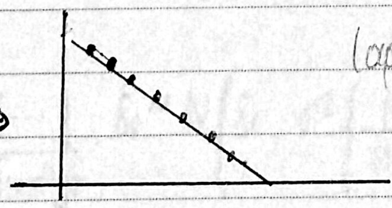
Ο $\overset{r_{X,Y}}{\text{συντε}}$ εκτιμά τον πληθυσμιακό συντελεστή συσχέτισης του Pearson και χρησιμοποιείται για ελέγχους της μορφής $H_0: \rho=0$ ή $H_0: \rho=\rho_0$.

ΙΔΙΟΤΗΤΕΣ:

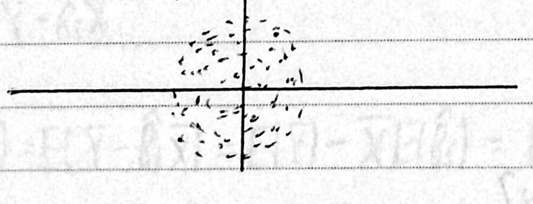
1) Είναι καθαρός αριθμός (απαλλαγμένος, δηλ. από μονάδες μέτρησης)
2) Συμμετρικός $r_{X,Y} = r_{Y,X}$
3) $-1 \leq r_{X,Y} \leq 1$ (από ανισότητα Cauchy-Schwarz)

Το $r_{X,Y}$ σχετίζεται: Διάγραμμα διασποράς.

- $r_{X,Y} = 1 \longleftrightarrow$  (όλα τα σημεία απολύτως πάνω σε μια ευθεία με θετική κλίση)

- $r_{X,Y} = -1 \longleftrightarrow$  (αρνητική κλίση)

- $r_{X,Y} = 0 \Leftrightarrow \nexists$ γραμμική σχέση μεταξύ των X και Y



Αν $r_{X,Y} = 0$ τότε $\nexists$ γραμμική σχέση, **μπορεί** όμως να $\exists$ άλλου είδους σχέση.