4ο μάθημα:

Παλινδρόμηση
26/10/20

①

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, 2, \ldots n$

ΕΕΤ $\hat{\beta}_0, \hat{\beta}_1$

Υποθέσεις για τα σφάλματα

Γενική: Τα σφάλματα είναι ασυσχέτιστα ανά δύο και $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots n$.

Οι υποθέσεις έχουν κάποιες συνέπειες στην $Y \Rightarrow Y_i$ ασυσχέτιστες ανά δύο

$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ $i = 1, \ldots n$.

Θεώρημα:

Αν οι υποθέσεις για τα σφάλματα ικανοποιούνται τότε $\frac{SS_{res}}{\sigma^2} \sim X^2_{n-1}$

Απόδειξη

$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ $i = 1, \ldots n$.

$$\frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma} \sim N(0,1) \Rightarrow \frac{\left[Y_i - (\beta_0 + \beta_1 X_i)\right]^2}{\sigma^2} \sim N^2(0,1) \equiv X^2_1$$

$$\xRightarrow{\text{Υιοθ}} \sum_{i=1}^{n} \frac{(Y_i - (\beta_0 + \beta_1 X_i))^2}{\sigma^2} \sim X^2_{\sum_{i=1}^{n} 1} = X^2_n.$$

Θ.ν.δ.ο
$$\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)}{\sigma^2} = \frac{\sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2}{\sigma^2} \sim X^2_{n-2}$$

Αν οι β_0 και β_1 αντικατασταθούν με τους ΕΕΤ $\hat{\beta}_0$ και $\hat{\beta}_1$ τότε

$$\sum_{i=1}^{n} \frac{(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2}{\sigma^2} \sim X^2_{n-2}$$

όπου οι αρχικοί n βαθμοί ελευθερίας της X^2 μειώνονται κατά 2 λόγω των $\underline{2}$ δεσμεύσεων που οι εκφράσεις των ΕΕΤ επιβάλλουν στο άθροισμα

(ο τύπος του $\hat{\beta}_0$ που σχετίζεται με το Y_i
>> $\hat{\beta}_1$ >>)

Έλεγχος Υποθέσεων και Διαστήματα Εμπιστοσύνης για τις παραμέτρους β_0 και β_1

τ.μ. $X \longrightarrow$ κατανομή $f(x;\theta)$. θ: άγνωστη παράμετρος

Πρόβλημα: Εκτίμηση (προσέγγιση) της άγνωστης θ.

(i) Εκτιμητής σε σημείο θ

(ii) Εκτίμηση σε διάστημα εμπιστοσύνης για την θ

(iii) Έλεγχοι υποθέσεων.

Υπάρξη τ.δ. $X_1, \ldots X_n$

Κανονική Κατανομή $N(\mu, \sigma^2)$

① Εκτιμητής της μ είναι $\bar{X}$

② Δ.Ε της μ $\left(\bar{X} - Z_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\frac{a}{2}} \frac{\sigma}{\sqrt{n}}\right)$, σ^2 γνωστό
αν είναι άγνωστο όπου $\sigma^2 = S^2$

③ Z-τέστ για έλεγχο $H_0: \mu = \mu_0$ έναντι H_a: $\mu > \mu_0$, $\mu < \mu_0$, $\mu \neq \mu_0$

η ΣΣ. $Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$ υπό την H_0

$\sigma/\sqrt{n}$ ← Τυπική απόκλιση του εκτιμητή

κρίσιμη περιοχή $\begin{cases} Z \geq Z_a \\ Z \leq -Z_a \\ |Z| \geq Z_{a/2} \end{cases}$

Τέστ σημαντικότητας / Τέστ κατά Wald.

Έστω τ.δ $X_1, \ldots, X_n$ από $f(x, \theta)$, $\theta \in \Theta$

Έστω έλεγχος $H_0: \theta = \theta_0$ έναντι H_a: $\theta > \theta_0$, $\theta < \theta_0$, $\theta \neq \theta_0$

α) η ΣΣΤ. $= \frac{\text{Εκτιμητής της } \theta - \theta_0}{\text{Τυπική απόκλιση του εκτιμητή}}$ ← Αν ο εκτιμητής είναι ο ΕΜΠ έχει $N(0,1)$ υπό την H_0

β) Η κ.π. διαμορφώνεται από τη ΣΣΤ. και την H_0, H_a.

γ) Το κρίσιμο σημείο προσδιορίζεται από $a = P(\text{Απορ } H_0 | H_0 \text{ αληθ})$

Έλεγχος υποθέσεων και διαστήματα εμπιστοσύνης για τις παραμέτρους β_0 και β_1

Ποια η πρακτική αξία στατιστικού τέστ, π.χ. για την β_1?

① H_0: $\beta_1 = \beta_1^*$ (β_1^* γνωστό) έναντι H_a: $\beta_1 \neq \beta_1^*$)

X = έξοδα διαφήμισης

Y = έσοδα από την διαφήμιση

Το $\hat{\beta}_1$ εκφράζει τη μεταβολή της Y στη μοναδιαία μεταβολή της X

② Αν $\beta_1^* = 0$, τότε η H_0 γίνεται $H_0: \beta_1 = 0$.

Αν δεν μπορώ να απορρίψω την $H_0: \beta_1 = 0$, τότε $\nexists$ γραμμική σχέση μεταξύ X και Y $\Rightarrow$ $\nexists$ παλινδρόμηση

Πολυδιάσταση ②

26/10/2020

Στατιστικό τεστ για έλεγχο της

$H_0: b_0 = b_0^*$ (b_0^* γνωστό) έναντι $H_a: b_0 \neq b_0^*$

Ακολουθούμε την τεχνική Wald.

Αξιοποιώ τον εκτιμητή $\hat{b}_0$ της b_0 για τον οποίο έχουμε αποδείξει ότι

$\hat{b}_0 \sim N\left(b_0^*, \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma^2\right)$ υπό την $H_0: b_0 = b_0^*$

Άρα $\frac{\hat{b}_0 - b_0^*}{\sqrt{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma^2}} \sim N(0,1)$

Η συνάρτηση αυτή θα μπορούσε να είναι στατιστική συνάρτηση του τεστ. περιέχει όμως το σ^2 που είναι άγνωστο. Άρα πρέπει το σ^2 με κάποιον τρόπο να απαλοιφεί.

Θεωρώ την σ.σ. t, $t = \frac{\frac{\hat{b}_0 - b_0^*}{\sqrt{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\cancel{\sigma^2}}}}{\sqrt{\frac{SSres}{\cancel{\sigma^2}}/(n-2)}} =$ ~~$\frac{\hat{b}_0 - b_0^*}{\sqrt{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}}}$~~

$= \frac{\hat{b}_0 - b_0^*}{\sqrt{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}MSres}} \overset{\text{υπό την } H_0: b_0 = b_0^*}{\sim} \frac{N(0,1)}{\sqrt{\frac{X^2_{n-2}}{n-2}}} \sim t_{n-2}$. (όταν η $N(0,1)$ είναι ανεξάρτητη του $\sqrt{\frac{X^2_{n-2}}{n-2}}$)

Ισχύει γιατί $\hat{b}_0$ ανεξάρτητο του SSres

Άρα η στατιστική συνάρτηση του τεστ είναι

$$\boxed{t = \frac{\hat{b}_0 - b_0^*}{\sqrt{\frac{\sum x_i^2}{\sum(x_i-\bar{x})^2}MSres}}}$$ με κατανομή t_{n-2}, υπό την $H_0: b_0 = b_0^*$

Αλλά $t = \frac{\hat{b}_0 - b_0^*}{\sqrt{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}MSres}} = \frac{\hat{b}_0 - b_0^*}{\sqrt{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\hat{\sigma}^2}}$ γιατί $E(MSres) = \sigma^2$

$= \frac{\hat{b}_0 - b_0^*}{\sqrt{\hat{Var}(\hat{b}_0)}}$ όπου $\hat{Var}(\hat{b}_0) = \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}MSres$

Άρα $t = \dfrac{\hat{b}_0 - b_0^*}{\sqrt{\hat{Var}(\hat{b}_0)}}$ με κατανομή t_{n-2} υπό την H_0: $b_0 = b_0^*$

όπου $\hat{Var}(\hat{b}_0) = \dfrac{\sum x_i^2}{n\sum(x_i - \bar{x})^2} MSres$.

<u>Διαμόρφωση κρίσιμης περιοχής</u>

Η κ.π θα διαμορφωθεί για μεγάλες τιμές της ΣΣΤ t. είτε θετικές είτε αρνητικές. Γιατί? Γιατί μεγάλες τιμές του t σημαίνει $\hat{b}_0$ πολύ διαφορετικό του b_0^* ή b_0 πολύ διαφορετικό του b_0^* (αφού το $\hat{b}_0$ είναι εκτιμητής της b_0). λόγος που σημαίνει ότι η H_0 πρέπει να απορριφθεί.

Άρα η κ.π έχει τη μορφή $t >>$ ή $t <<$. Δηλ. $|t| \geq c$

Πώς θα προσδιορίσω το κρίσιμο σημείο c?

Το κ.σ πάντα υπολογίζεται από $a = P(\text{Απορ } H_0 | H_0 \text{ αληθ})$

$$a = P(\text{Απορ } H_0 | H_0 \text{ αληθής}) = P(|t| \geq c | H_0 \text{ αληθ}) = P(|t_{n-2}| \geq c) =$$
$$= P(t_{n-2} \geq c \text{ ή } t_{n-2} \leq -c) \xrightarrow{\text{τσυμ.}} 2P(t_{n-2} \geq c).$$

$$\Rightarrow P(t_{n-2} \geq c) = \frac{a}{2} \Rightarrow c = t_{n-2, \frac{a}{2}}.$$

Συγκεντρωτικά Για τον έλεγχο της H_0: $b_0 = b_0^*$ έναντι H_a: $b_0 \neq b_0^*$

η ΣΣΤ είναι $t = \dfrac{\hat{b}_0 - b_0^*}{\sqrt{\hat{Var}(\hat{b}_0)}} \sim t_{n-2}$ υπό H_0.

, $\hat{Var}(b_0) = \dfrac{\sum x_i^2}{n\sum(x_i - \bar{x})^2} MSres$

και κ.π $|t| \geq t_{n-2, \frac{a}{2}}$

Εντελώς ανάλογα (Άσκηση) Για τον έλεγχο H_0: $b_1 = b_1^*$ έναντι της H_a: $b_1 \neq b_1^*$

η ΣΣΤ είναι $t = \dfrac{\hat{b}_1 - b_1^*}{\sqrt{\hat{Var}(\hat{b}_1)}} \sim t_{n-2}$ υπό H_0, $\hat{Var}(b_1) = \dfrac{1}{\sum(x_i - \bar{x})^2} MSres$

και η κ.π $|t| \geq t_{n-2, \frac{a}{2}}$

Παρατήρηση:

Ο έλεγχος έχει αξία γιατί; $H_0: \beta_1 = \beta_1^*$

i) ελέγχει υπόθεση ότι η μεταβολή της Y είναι ίση με β_1^* για μοναδιαία μεταβολή της X.

ii) Αν $\beta_1^* = 0$ και η $H_0: \beta_1 = 0$ δεν μπορεί να απορριφθεί τότε $\nexists$ μοντέλο α.γ.π.

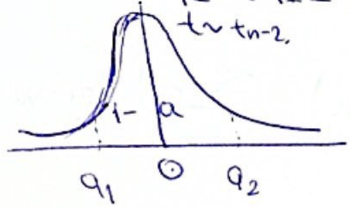
Διάστημα εμπιστοσύνης για την β_0:

Αντιστρεπτή ποσότητα είναι η $t = \dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \sim t_{n-2}$,

$$\widehat{Var}(\hat{\beta}_0) = \frac{\sum X_i^2}{n\sum(X_i-\bar{X})^2} MS_{res}.$$

Αφού η t είναι αντιστρεπτή με κατανομή t_{n-2}, $\exists\, q_1, q_2$ $(q_1 \le q_2)$ τ.ω.

$$1-a = P(q_1 \le t \le q_2)$$



$$= P\left(q_1 \le \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \le q_2\right)$$

$$= P\left(\hat{\beta}_0 - q_2\sqrt{\widehat{Var}(\hat{\beta}_0)} \le \beta_0 \le \hat{\beta}_0 - q_1\sqrt{\widehat{Var}(\hat{\beta}_0)}\right)$$

Άρα ένα $100(1-a)\%$ δ.ε για την β_0 είναι

$$\left(\hat{\beta}_0 - q_2\sqrt{\widehat{Var}(\hat{\beta}_0)}\ ,\ \hat{\beta}_0 - q_1\sqrt{\widehat{Var}(\hat{\beta}_0)}\right)$$

Το δ.ε ελαχίστου μήκους για την β_0 προκύπτει για $q_1 = -t_{n-2,\frac{a}{2}}$, $q_2 = t_{n-2,\frac{a}{2}}$

Συγκεντρωτικά: Το ΔΕ ελαχίστου μήκους για την β_0 με β.ε $100(1-a)\%$ είναι

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

όπου $\widehat{Var}(\hat{\beta}_0) = \dfrac{\sum X_i^2}{n\sum(X_i-\bar{X})^2} MS_{res}$

Το δ.ε ελαχίστου μήκους για την β_1 με δ.ε $100(1-a)\%$ είναι

$$\left(\hat{\beta_1} - t_{n-2,\frac{a}{2}}\sqrt{\hat{Var}(\hat{\beta_1})}\ ,\ \hat{\beta_0} + t_{n-2,\frac{a}{2}}\sqrt{\hat{Var}(\hat{\beta_1})}\right)$$

όπου $\hat{Var}(\hat{\beta_1}) = \dfrac{1}{\Sigma(X_i-\bar{X})^2} MSres$

<u>Τυπική Απόκλιση της πρόβλεψης $\hat{Y_k} = \hat{\beta_0} + \hat{\beta_1} X_k$</u>

Σκοπός είναι η εύρεση της $Var(\hat{Y_k})$, αν είναι μικρή η πρόβλεψη Y_k είναι αξιόπιστη

$$\hat{Y_k} = \hat{\beta_0} + \hat{\beta_1} X_k = \bar{Y} - \hat{\beta_1}\bar{X} + \hat{\beta_1} X_k.$$

$$\Rightarrow \hat{Y_k} = \bar{Y} + (X_k - \bar{X})\beta_1$$

$$\left.\begin{array}{l} Var(\hat{Y_k}) = Var(\bar{Y} + (X_k - \bar{X})\hat{\beta_1}) \\ Var(\Sigma a_i W_i) = \Sigma a_i^2 Var(W_i) + \sum_{\substack{i=1 \\ i \neq j}}^{n}\sum_{j=1}^{n} a_i a_j Cov(W_i, W_j) \end{array}\right\} \Rightarrow$$

$$Var(\hat{Y_k}) = Var(\bar{Y}) + (X_k - \bar{X})^2 Var(\hat{\beta_1}) + (X_k - \bar{X}) Cov(\bar{Y}, \hat{\beta_1})$$

$$\left.\begin{array}{l} Cov(\bar{Y}, \hat{\beta_1}) = Cov\left(\Sigma\frac{1}{n} Y_i ,\ \Sigma \frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2} Y_i\right) \\ Cov\left(\sum_{i=1}^{n} a_i W_i , \sum_{i=1}^{n} b_i W_i\right) = \sum_{i=1}^{n} a_i b_i Var(W_i) \end{array}\right\} \Rightarrow$$

$$Cov(\bar{Y}, \hat{\beta_1}) = \sum_{i=1}^{n} \frac{1}{n}\frac{X_i - \bar{X}}{\Sigma(X_i - \bar{X})^2} Var(Y_i) = \frac{1}{n}\sigma^2 \cancel{\Sigma\frac{(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}}^{=0}$$

Άρα $\boxed{Var(\hat{Y_k}) = Var(\bar{Y}) + (X_k - \bar{X})^2 Var(\hat{\beta_1})}$

<u>Παρατήρηση:</u> Η διακύμανση $\hat{Y_k}$ ελαχιστοποιείται για $X_k = \bar{X}$

Άρα στην πράξη το μοντέλο της α.γ.π πρέπει να χρησιμοποιείται για την πρόβλεψη της Y για τιμές της X κοντά στο $\bar{X}$, και τα δηλ στη μάζα των μετρήσεων της X.