Μάθημα 2ο 16/10/17

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΣΤΟ ΜΟΝΤΕΛΟ ΤΗΣ Α.Γ.Π.

Ανάλυση της ολικής μεταβλητότητας στο μοντέλο της α.γ.π.

Ολική μεταβλητότητα $\approx$ Δειγματική διακύμανση των $Y_1, \dots, Y_n$. ($Var(Y)$)

Δειγματική διακύμανση των $Y_1, \dots, Y_n \longrightarrow S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

Ολική μεταβλητότητα στα $Y_1, \dots, Y_n \simeq \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

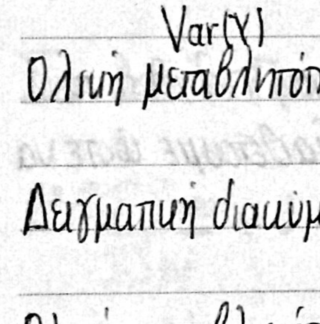
ολική μεταβλητότητα

μεταβλητότητα που οφείλεται στο μοντέλο της α.γ.π.

μεταβλητότητα που οφείλεται στα υπόλοιπα

προσθαφαιρώ $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$ (ΑΣΚΗΣΗ)

$$\underset{total}{SS_{tot}} = \underset{regration}{SS_{reg}} + SS_{res}$$



($MS_{reg} >> MS_{res}$) το ίδιο

- Αν $SS_{reg} >> SS_{res}$ τότε το μοντέλο της α.γ.π. φαίνεται να είναι υποσχόμενο (καλό)
- Αν $SS_{reg} << SS_{res}$ τότε το μοντέλο δεν φαίνεται να είναι υποσχόμενο

- Άλλη μορφή για το άθροισμα τετραγ. που οφείλεται στην παλινδρόμηση SS_{reg}

$$SS_{reg} = \Sigma(\hat{Y}_i - \bar{Y})^2 = \Sigma(\hat{b}_0 + \hat{b}_1 X - \bar{Y})^2 = \Sigma(\bar{Y} - \hat{b}_1 \bar{X} + \hat{b}_1 X_i - \bar{Y})^2 = \hat{b}_1^2 \Sigma(X_i - \bar{X})^2$$

Άρα λοιπόν προσπαθούμε να αναδείξουμε αυτή τη σχέση κατασκευάζοντας έναν πίνακα:

Πίνακας ανάλυσης διακύμανσης του μοντέλου της α.γ.π.

ΠΙΝΑΚΑΣ ΑΝΑΔΙΑ

πηγή μεταβλητότητας	Αθρ. τετρ. SS	β.ε	MS μέσο τετραγ.	F πηλίκο
μοντέλο α.γ.π.	SS_{reg}	1	$MS_{reg} = SS_{reg}/1$	$F = \frac{MS_{reg}}{MS_{res}}$
υπόλοιπο	SS_{res}	$n-2$	$MS_{res} = SS_{reg}/(n-2)$	
ολική μεταβλητότητα	SS_{tot}	$n-1$		

Για να αξιοποιήσω αυτό τον πίνακα χρησιμοποιώ το 2ο κριτήριο που προαναφέρθηκε

ΟΡΙΣΜΟΣ: Βαθμός ελευθερίας ενός αθροίσματος τετραγώνου είναι το πλήθος των ανεξάρτητων πληροφοριών στο $Y_1, .., Y_n$ τις οποίες πρέπει να διαθέτουμε ώστε να μπορούμε να υπολογίσουμε το αντίστοιχο άθροισμα.

Ας πούμε για το SS_{tot} ξέρω $SS_{tot} = \Sigma(Y_i - \bar{Y})^2$ δηλαδή αρκεί να έχω n-πληροφορίες για να το υπολογίσω: $Y_1 - \bar{Y}$, $Y_2 - \bar{Y}$, $Y_n - \bar{Y}$ όμως!!! $\bar{Y} = 1/n \Sigma Y_i$ δηλ. μπορώ να εκφράσω το Y_n συναρτήσει των άλλων άρα έχω n-1 β.ε

ΑΠΟΔΕΙΞΗ ότι β.ε. του SS_{tot} είναι n-1

- $SS_{tot} = \Sigma(Y_i - \bar{Y})^2$

όπου $\bar{Y} = \frac{1}{n} \Sigma Y_i \Rightarrow \Sigma Y_i = n\bar{Y} \Rightarrow Y_n + \sum_{i=1}^{n-1} Y_i = n\bar{Y} \Rightarrow Y_n = n\bar{Y} - \sum_{i=1}^{n-1} Y_i$

- $Y_n - \bar{Y} = n\bar{Y} - \sum_{i=1}^{n-1} Y_i - \bar{Y} = (n-1)\bar{Y} - \sum_{i=1}^{n-1} Y_i \Rightarrow Y_n - \bar{Y} = -\sum_{i=1}^{n-1}(Y_i - \bar{Y}) \Rightarrow$

$\Rightarrow (Y_n - \bar{Y})^2 = \left[\sum_{i=1}^{n-1}(Y_i - \bar{Y})\right]^2$

δηλαδή $SS_{tot} = \sum(Y_i - \bar{Y})^2 = \sum_{i=1}^{n-1}(Y_i - \bar{Y})^2 + (Y_n - \bar{Y})^2 =$
$= \sum_{i=1}^{n-1}(Y_i - \bar{Y})^2 + \left[\sum_{i=1}^{n}(Y_i - \bar{Y})\right]^2$

Άρα, χρειάζεται να ξέρω $Y_1 - \bar{Y}, \dots, Y_{n-1} - \bar{Y}$ για να υπολογίσω το SS_{tot}.
Άρα, οι β.ε. είναι $n-1$.

Εμπειρικός κανόνας για β.ε.

Το SS_{tot} έχει β.ε.: πλήθος παρατηρήσεων -1
(μέγεθος δείγματος) -1

Το SS_{reg} έχει β.ε. όσο πλήθος ανεξάρτητων μεταβλητών (για την περίπτωσή μας μόνο η X).

Το SS_{res} έχει β.ε. την διαφορά

Συντελεστής Προσδιορισμού ή προσαρμοστικότητας

$R^2 \stackrel{\text{ορ}}{=} \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$ από τη σχέση $SS_{reg} + SS_{res} = SS_{tot}$.

και οι 2 ιδιότητες είναι οι εξής:
1) ο R^2 είναι καθαρός αριθμός
2) $0 \le R^2 \le 1$ και άρα εκφράζεται ως ποσοστό!
Ενδιαφέρουν παρουσιάζουν οι τιμές 0 και 1.

- Τιμή του R^2 κοντά στο 1: δηλαδή θα πρέπει $SS_{reg} \simeq SS_{tot}$ δηλ. $SS_{reg} >> SS_{res}$ υποσχόμενο μοντέλο
- Τιμή του R^2 κοντά στο 0: δηλαδή θα πρέπει $SS_{reg} << SS_{res}$ μη υποσχόμενο (κακό) μοντέλο

<u>Μια ερμηνεία για το R^2</u>: εκφράζει το ποσοστό της ολικής μεταβλητότητας των $Y_1, \dots, Y_n$ που ερμηνεύονται από το μοντέλο μας.

Επανάληψη

α.γ.π. $Y_i = \underbrace{\beta_0 + \beta_1}_{\text{παράμετροι}} X_i + \varepsilon_i \quad i=1,\dots,n$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \ i=1,\dots,n, \quad e_i = Y_i - \hat{Y}_i$$

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

Αν $SS_{res} << SS_{reg} \rightsquigarrow$ καταλληλότητα μοντέλου.

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

$0 \leq R^2 \leq 1$

↓ φτωχό μοντέλο ↓ κατάλληλο μοντέλο.

ΥΠΟΘΕΣΕΙΣ ΣΤΑ ΣΦΑΛΜΑΤΑ ΤΟΥ ΜΟΝΤΕΛΟΥ Α.Γ.Π.

Υποθέσεις σφάλματα: (Μοντέλο α.γ.π.: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i=1,\dots,n$)

1) $E(\varepsilon_i) = 0$, $i=1,\dots,n$
2) $Var(\varepsilon_i) = \sigma^2$, $i=1,\dots,n$, τα σφάλματα έχουν κοινή διακύμανση.
3) Τα ε_i ασυσχέτιστα ανά δύο: $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i,j = 1,\dots,n, \ i \neq j$
4) $\varepsilon_i \sim N(0, \sigma^2)$, $i=1,\dots,n$

Συνέπειες στα Y_i ($Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$):

1) $E(Y_i) = \beta_0 + \beta_1 X_i$, $i=1,\dots,n$
2) $Var(Y_i) = \sigma^2$, $i=1,\dots,n$
3) $Cov(Y_i, Y_j) = 0$, $i,j=1,\dots,n$, $i \neq j$ (γραμμικές συναρτήσεις ασυσχέτιστων είναι ασυσχέτιστες)
4) $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, $i=1,\dots,n$.

ΘΕΩΡΗΜΑ: Αν οι υποθέσεις για τα σφάλματα ικανοποιούνται τότε ισχύουν τα εξής:

α) $\hat{b}_1 \sim N\left(b_1, \underbrace{\frac{\sigma^2}{\sum(x_i-\bar{x})^2}}_{Var(\hat{b}_1)}\right)$

β) $\hat{b}_0 \sim N\left(b_0, \underbrace{\frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma^2}_{Var(\hat{b}_0)} = \sigma^2\left\{\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right\}\right)$

ΠΑΡΑΤΗΡΗΣΗ: $\hat{b}_0, \hat{b}_1$ είναι αμερόληπτοι των b_0, b_1, δηλ. των παραμέτρων που εκτιμούν.

Ισχύει: Αν $W_1, W_2, \dots, W_m$ είναι κανονικές τ.μ. τότε $\sum_{i=1}^{m} a_i W_i \sim N\left(\sum_{i=1}^{m} a_i E(W_i),\right.$ $Var(\sum a_i W_i) = \sum a_i^2 Var(W_i) + \sum_{\substack{i=1 \\ i\neq j}}^{m}\sum_{j=1}^{m} a_i a_j Cov(W_i, W_j)\Big)$

ΑΠΟΔΕΙΞΗ:

α) $\hat{b}_1 = \frac{\sum(x_i-\bar{x})(Y_i-\bar{Y})}{\sum(x_i-\bar{x})^2} = \sum_{i=1}^{n} \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2} Y_i$

Το $\hat{b}_1$ είναι γραμμικός συνδυασμός των ανεξάρτητων Y_i

Y_i ασυσχέτιστα $\quad Y_i \sim N$

Άρα, $\hat{b}_1 \sim$ Normal

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

$= \sum_{i=1}^{n} \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2} E(b_0+b_1x_i+\varepsilon_i) = \sum_{i=1}^{n} \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2} (b_0+b_1x_i+\cancelto{0}{E(\varepsilon_i)}) =$

$= b_0 \cancelto{0}{\sum_{i=1}^{n} \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})}} + b_1 \sum_{i=1}^{n} \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2} \cdot x_i =$

$= b_1 \frac{\sum(x_i^2-\bar{x}x_i)}{\sum(x_i^2-2\bar{x}x_i+\bar{x}^2)} = b_1 \frac{\sum x_i^2-\bar{x}\sum x_i}{\sum x_i^2-2\bar{x}\sum x_i+n\bar{x}^2} = b_1 \frac{\sum x_i^2-n\bar{x}^2}{\sum x_i^2-2n\bar{x}^2+n\bar{x}^2} =$

$= b_1 \frac{\sum x_i^2-n\bar{x}^2}{\sum x_i^2-n\bar{x}^2} = b_1.$

ΠΟΡΙΣΜΑ: Αν W_i ασυσχέτιστες $(Cov(W_i,W_j)=0,\ i\neq j)$ τότε:

$\sum_{i=1}^{m} a_i W_i \sim N\left(\sum_{i=1}^{m} a_i E(W_i),\ Var\left(\sum_{i=1}^{m} a_i W_i\right) = \sum_{i=1}^{m} a_i^2 Var(W_i)\right)$

$$Var(\hat{\beta}_1) = Var\left[\sum_{i=1}^{n} \frac{X_i - \bar{X}}{\sum(X_i-\bar{X})^2} Y_i\right]$$

Αν $W_1, \dots, W_n$ τ.μ. τότε: $Var(\sum_{i=1}^n a_i W_i) = \sum_{i=1} a_i^2 Var(W_i) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ i \neq j}}^{n} a_i a_j Cov(W_i, W_j)$

Αν $W_1, \dots, W_n$ ασυσχέτιστες τότε: $Var\left(\sum_{i=1} a_i W_i\right) = \sum_{i=1} a_i^2 Var W_i$

Επομένως: $$Var(\hat{\beta}_1) = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sum(X_i-\bar{X})^2}\right)^2 Var(Y_i) = \sigma^2 \cdot \sum_{i=1}^{n} \frac{(X_i-\bar{X})^2}{[\sum(X_i-\bar{X})^2]^2} =$$

$$= \sigma^2 \frac{\sum(X_i-\bar{X})^2}{[\sum(X_i-\bar{X})^2]^2} = \frac{\sigma^2}{\sum(X_i-\bar{X})^2}$$

ΠΑΡΑΤΗΡΗΣΗ: Το σ^2 άγνωστο. Άρα, για να αξιοποιηθούν οι κατανομές αυτές το σ^2 πρέπει να εκτιμηθεί.

ΘΕΩΡΗΜΑ: Από τις υποθέσεις για τα σφάλματα $E(MS_{res}) = \sigma^2$, δηλ. ο αμερόληπτος εκτιμητής της σ^2 είναι το MS_{res}. (χωρίς απόδειξη)

ΠΑΡΑΤΗΡΗΣΗ: Το $\hat{\sigma}^2 = MS_{res}$ ένας αμερόληπτος εκτιμητής της σ^2

ΘΕΩΡΗΜΑ: Υπό τις υποθέσεις για τα σφάλματα, $\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-2}$.

ΑΠΟΔΕΙΞΗ:

Αν $W_1, \dots, W_n \sim N(\mu, \sigma^2)$ τότε: $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (W_i - \bar{W})^2$

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

$$\Rightarrow \frac{Y_i - (\beta_0 + \beta_1 X_i)}{\sigma} \sim N(0,1)$$

$$\Rightarrow \frac{[Y_i - (\beta_0 + \beta_1 X_i)]^2}{\sigma^2} \sim \chi^2_1$$

(Y_i κανονικές και ασυσχέτιστες, οι χ_1^2 ανεξάρτητες)

$$\Rightarrow \frac{\sum_{i=1}^{n}[Y_i - (\beta_0 + \beta_1 X_i)]^2}{\sigma^2} \sim \sum_{i=1}^{n} \chi_1^2 \equiv \chi^2_{\sum_{i=1}^{n} 1}.$$

Άρα, $\frac{\sum(Y_i - (\beta_0 + \beta_1 X_i))^2}{\sigma^2} \sim \chi^2_n$ (A)

$$SS_{res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \quad (B)$$

Από τις (A), (B) $\Rightarrow \dfrac{\sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2}{\sigma^2} \sim \chi^2_{n-2}$ (?)

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

αντικαθιστώντας τα β_0, β_1 με $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι σαν να βάζω 2 δεσμούς άρα από n πάω σε $n-2$

ΘΕΩΡΗΜΑ: Οι $\underbrace{\hat{\beta}_0 \text{ και } \hat{\beta}_1}_{ΕΕΤ}$ είναι ανεξάρτητοι του MS_{res} και άρα ανεξάρτητοι του SS_{res}.

(χωρίς απόδειξη)