15/10/2019

Διάλεξη 2η

Παλινδρόμηση και Ανάλυση Διακύμανσης.

Σεμινάρια Στατιστικής
Παρασκευή 18/10/2019
ώρα 12:00 Αιθ. 307α

Απλή Γραμμική Παλινδρόμηση (αγπ)

Τμ. $\begin{array}{c|c} X & X_1, X_2 \dots , X_n \\ \hline Y & Y_1, Y_2, \dots , Y_n \end{array}$ αγπ δηλ $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i = 1, 2, \dots, n$

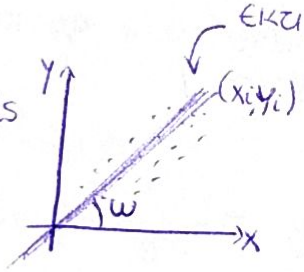
(Y_i: εξαρτημένη μτβ; ε_i: σφάλματα; β_0, β_1: παράμετροι; X_i: ανεξάρτητη)

ΕΕΤ των β_0, β_1.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Εκτιμώμενο Μοντέλο: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

Γραφικά: Διάγραμμα Διασποράς



$\hat{\beta}_1 = \varepsilon\varphi\omega$

Ερμηνεία των $\hat{\beta}_0, \hat{\beta}_1$

Το $\hat{\beta}_0$ εκφράζει την τιμή της Y όταν $X = 0$

Το $\hat{\beta}_1$ εκφράζει την μεταβολή της Y σε μοναδιαία μεταβολή της X.

— o — o — o — o — o — o —

Υπόλοιπα: Εκφράζουν την απόκλιση του μοντέλου από την πραγματικότητα.

$$e_i = Y_i - \hat{Y}_i$$

(Y_i: πραγματικότητα; $\hat{Y}_i$: μοντέλο)

Ιδιότητα Υπολοίπων: $\sum_{i=1}^{n} e_i = 0$

Αποδ: $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum_{i=1}^{n} (Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_1 X_i)$

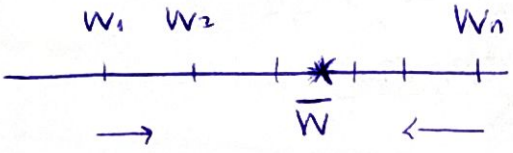
$= \sum_{i=1}^{n} (Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{X})$

$= \sum_{i=1}^{n} Y_i - n\bar{Y} = n\bar{Y} - n\bar{Y} = 0$ (όμοιως) $\Big\} = 0$ Άρα $\sum_{i=1}^{n} e_i = 0$.

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ στο μοντέλο αχπ.

Αν $W_1, .. W_n$ είναι τυχαίο δείγμα, η δειγματική διακύμανση

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (W_i - \bar{W})^2.$$



S^2 ← μέτρο μεταβλητότητας των παρατηρήσεων $W_1, .., W_n$ σε σχέση με $\bar{W}$.

Αναλόγα η μεταβλητότητα των $Y_1, .. Y_n$ είναι $\underbrace{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}_{SStot}$ =

αν $\pm \hat{Y}_i$ τότε :

$$\underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{SSreg} + \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{SSres} \quad \left(\begin{matrix}\text{Αποδ}\\ \text{Άσκηση}\end{matrix}\right)$$

Ολική Μεταβλητότητα = Αθρ. τετραγ στο μοντέλο αχπ + Αθρ. Τετραγ Υπόλοιπα

$$\boxed{SStot = SSreg + SSres}$$

- Αν $SSreg >> SSres$ τότε το μεγαλύτερο μέρος της ολικής μεταβλητότητας περιγράφεται από το μοντέλο της αχπ.
  Άρα το μοντέλο της αχπ είναι υποσχόμενο
- Αν $SSreg << SSres$ τότε το μοντέλο της αχπ δεν φαίνεται να είναι υποσχόμενο

Πίνακας ΑΝΑΔΙΑ μοντέλου αχπ

Πηγη μεταβλητοτητας	SS	β.ε	MS	F πηλίκο
Μοντελο αχπ	$SSreg = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$	1	$MSreg = \frac{SSreg}{1}$	$F = \frac{MSreg}{MSres}$
Υπολοιπα	$SSres = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	(n-2)	$MSres = \frac{SSres}{n-2}$	
Ολικη	$SStot = \sum_{i=1}^{n}(y_i - \bar{y})^2$	n-1		

β.ε : πλήθος ανεξάρτητων πληροφοριών τις οποίες πρέπει να διαθέτουμε ώστε να μπορούμε να υπολογίσουμε το αντίστοιχο άθροισμα τετραγώνων.

Εμπειρικά στα Γραμμικά Μοντέλα

β.ε SStot = μέγεθος δείγματος − 1

β.ε SSμοντέλου = πλήθος ανεξάρτητων μεταβλ

β.ε SSυπολοίπων = αφαίρεση

Αν MSreg >> MSres τοτε Υποσχόμενο Μοντέλο

Αν MSreg << MSres τότε Μη Υποσχόμενο

Οι β.Ε του SStot είναι n-1.

$$SStot = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \Rightarrow y_n + \sum_{i=1}^{n-1} y_i = n\bar{y} \Rightarrow y_n = n\bar{y} - \sum_{i=1}^{n-1} y_i$$

$$y_n - \bar{y} = n\bar{y} - \sum_{i=1}^{n-1} y_i - \bar{y} = (n-1)\bar{y} - \sum_{i=1}^{n-1} y_i \Rightarrow y_n - \bar{y} = -\sum_{i=1}^{n-1} (y_i - \bar{y})$$

Άρα $(y_n - \bar{y})^2 = \left[\sum_{i=1}^{n-1} (y_i - \bar{y}) \right]^2$

$$SStot = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n-1} (y_i - \bar{y})^2 + (y_n - \bar{y})^2$$

$$= \sum_{i=1}^{n-1} (y_i - \bar{y})^2 + \left[\sum_{i=1}^{n-1} (y_i - \bar{y}) \right]^2$$

← απαιτούνται n-1 ανεξ. πληροφ. για να υπολογιστεί.
οι $y_1 - \bar{y}, y_2 - \bar{y}, \ldots y_{n-1} - \bar{y}$.

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

⇓

$$SSreg = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \underline{\underline{\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i}} \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2$$

$\underline{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$ $\sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$

Συντελεστής Προσδιορισμού ή Συντελεστής Προσαρμοστικότητας

$$R^2 = \frac{SSreg}{SStot}, \quad R^2 = 1 - \frac{SSres}{SStot}$$

Ιδιότητες R^2

① Ο R^2 είναι καθαρός αριθμός (απαλλαγμένος από μονάδες μέτρησης)

② $0 \le R^2 \le 1$

Λόγω της (2) ο R^2 είναι ποσοστό και εκφράζει το ποσοστό της ολικής μεταβλητότητας της εξαρτημένης μεταβλητής που ερμηνεύεται από το μοντέλο της αγπ.

Τιμές του R^2 κοντά στο 1 σημαίνουν υποσχόμενο μοντέλο

Μικρές τιμές του R^2 σημαίνουν όχι υποσχόμενο μοντέλο.

Για την περαιτέρω ανάπτυξη του μοντέλου της αγπ $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, $i=1,..,n$ απαιτούνται κάποιες πιθανοθεωρητικές υποθέσεις που διατυπώνονται πάνω στα σφάλματα ε_i, $i=1,..,n$ του μοντέλου.

- ΥΠΟΘΕΣΕΙΣ ΓΙΑ ΤΑ ΣΦΑΛΜΑΤΑ.

Τα σφάλματα είναι τ.μ.

① $E(\varepsilon_i) = 0$, $\forall i=1,..,n$

② Η διακύμανση των σφαλμάτων είναι σταθερή

$Var(\varepsilon_i) = \sigma^2$, $\forall i=1,..,n$

③ Τα σφάλματα ε_i είναι ασυσχέτιστα

$Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$, $i,j=1,2,..,n$

④ $\varepsilon_i \sim N(0, \sigma^2)$, $i=1,..,n$

ΥΠΕΝΘΥΜΙΣΗ

$Cov(Z,W) = E[(Z-E(Z))(W-E(W))]$

$= E(ZW) - (EZ)(EW)$

Αν $Z=W$

$Cov(Z,W) = Var(Z)$

Οι υποθέσεις για τα σφάλματα αντανακλούν και στα Y_i

① $E(Y_i) = \beta_0 + \beta_1 X_i$ $i=1,..,n$

② $Var(Y_i) = \sigma^2$, $i=1,..,n$

③ $Cov(Y_i, Y_j) = 0$, $i \neq j$ $i,j=1,..,n$ τα Y_i ασυσχέτιστα

④ $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$

ΘΕΩΡΗΜΑ Αν οι υποθέσεις για τα σφάλματα ικανοποιούνται τότε:

α) $\hat{\beta}_0 \sim N(\beta_0, Var(\hat{\beta}_0) = \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right])$

↑ αμερόληπτος

β) $\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2})$

Απόδ

β) $\hat{\beta}_1 = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \frac{\sum(x_i-\bar{x})y_i - \bar{y}\sum(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$ (0)

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

ο $\hat{\beta}_1$ είναι γραμμικός συνδιασμός των y_i, $i=1,2..n$

Ισχύει ότι αν $W_i \sim Normal$ τότε κάθε γραμμικός συνδιασμός των $W_i \sim Normal$

Άρα και ο $\hat{\beta}_1 \sim Normal$ ως γραμμικός συνδιασμός των $y_i \sim Normal$ λόγω της (4).

$$E(\hat{\beta}_1) = E\left[\sum \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2} y_i\right] = \sum \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2} E(y_i) = \sum \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2}$$

$$= \sum \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2}(\beta_0+\beta_1 x_i)$$

$$= \beta_0 \sum \frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2} + \beta_1 \sum \frac{x_i^2-\bar{x}x_i}{\sum(x_i^2-2\bar{x}x_i+\bar{x}^2)} = \beta_0 \frac{\sum(x_i-\bar{x})}{\sum(x_i-\bar{x})^2} + \beta_1 \frac{\sum x_i^2 - \bar{x}\sum x_i}{\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2}$$

$$= \beta_1 \frac{\sum x_i^2 - \bar{x}n\bar{x}}{\sum x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2} = \beta_1 \frac{\sum x_i^2 - n\bar{x}^2}{\sum x_i^2 - n\bar{x}^2} = \beta_1$$

Από ΘΠΣ ισχύει:

Αν W_i $i=1,\dots,n$ είναι τ.μ. τότε $Var\left(\sum_{i=1}^{n} a_i W_i\right) = \sum_{i=1}^{n} a_i^2 Var(W_i) + \sum_{i=1}^{n}\sum_{j=1} a_i a_j Cov(W$

Ειδική περίπτωση: Αν W_i ασυσχέτιστες ($Cov(W_i, W_j) = 0$) τότε

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

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} \frac{X_i - \bar{X}}{\sum (x_i - \bar{x})^2} Y_i\right) \overset{\substack{\text{Υποθέσεις} \\ Y_i \text{ ασυσχ}}}{=\!=} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}\right)^2 Var(Y_i)$$

$$\overset{\text{Υποθέσεις}}{=\!=} \sigma^2 \sum \frac{(X_i - \bar{X})^2}{\left[\sum (x_i - \bar{x})^2\right]^2} = \sigma^2 \frac{\sum (x_i - \bar{x})^2}{\left[\sum (x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$Var(Y_i) = \sigma^2$

Παρατηρώ ότι η σ^2 ($= Var(\varepsilon_i) = Var(Y_i)$) είναι <u>άγνωστη</u>

Άρα η σ^2 πρέπει να εκτιμηθεί.

ΘΕΩΡΗΜΑ Υπό τις υποθέσεις για τα σφάλματα του μοντέλου της αγπ το MSres είναι αμερόληπτος εκτιμητής της σ^2.

δηλ $E(MSres) = \sigma^2$.

<u>Αποδ</u> σε φυλλάδιο που έδωσε.

Απομερόληψη αλόβιτη:

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

Είναι $MS_{res} = \frac{SS_{res}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$

Ισχύει $SS_{res} = SS_{tot} - SS_{reg}$

$$E(SS_{tot}) = E\left\{\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right\} = \sum_{i=1}^{n} E(Y_i-\bar{Y})^2 = \sum_{i=1}^{n} E\left\{b_0+b_1x_i+\varepsilon_i - \right.$$

$$\left. -\frac{1}{n}\sum_{i=1}^{n}(b_0+b_1x_i+\varepsilon_i)\right\}^2 = \sum_{i=1}^{n} E\left\{\cancel{b_0}+b_1x_i+\varepsilon_i - \cancel{b_0} - b_1\bar{x} - \bar{\varepsilon}\right\}^2.$$

Μετά τις πράξεις και λαμβάνοντας υπόψη ότι:

$$E(\varepsilon_i - \bar{\varepsilon}) = E(\varepsilon_i) - \frac{1}{n}\sum_{i=1}^{n} E(\varepsilon_i) = 0 - 0 = 0$$

παίρνουμε:

$$E(SS_{tot}) = b_1^2 \sum_{i=1}^{n}(x_i-\bar{x})^2 + \sum_{i=1}^{n} E(\varepsilon_i - \bar{\varepsilon})^2. \qquad (1)$$

Αλλά:

$$\sum_{i=1}^{n} E(\varepsilon_i-\bar{\varepsilon})^2 = \sum_{i=1}^{n}\left(E(\varepsilon_i^2) - 2E(\varepsilon_i\bar{\varepsilon}) + E(\bar{\varepsilon}^2)\right) = \sum_{i=1}^{n} E(\varepsilon_i^2) - 2E\left(\sum_{i=1}^{n}\varepsilon_i\bar{\varepsilon}\right) +$$

$$+ n\,E(\bar{\varepsilon}^2) = \sum_{i=1}^{n} E(\varepsilon_i^2) - 2E\left(n\bar{\varepsilon}\cdot\bar{\varepsilon}\right) + n\,E(\bar{\varepsilon}^2) = \sum_{i=1}^{n} E(\varepsilon_i^2) - nE(\bar{\varepsilon}^2).$$

$$= \sum_{i=1}^{n}\left[Var(\varepsilon_i) + \cancel{[E(\varepsilon_i)]^2}^{0}\right] - n\left[Var(\bar{\varepsilon}) + \cancel{[E(\bar{\varepsilon})]^2}^{0}\right]$$

$$= \sum_{i=1}^{n}\sigma^2 - n\,Var\left(\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\right) = n\sigma^2 - n\frac{1}{n^2}\sum_{i=1}^{n} Var(\varepsilon_i) = n\sigma^2 - \frac{1}{n}\sum_{i=1}^{n}\sigma^2$$

και τελικά: $\sum_{i=1}^{n} E(\varepsilon_i-\bar{\varepsilon})^2 = (n-1)\sigma^2 \qquad (2)$

Από τις (1) και (2):

$$E(SS_{tot}) = b_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2 + (n-1)\sigma^2 \qquad (3)$$

Επίσης $SS_{reg} = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \hat{b}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$

και $E(SS_{reg}) = \left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\} E(\hat{b}_1^2)$

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

Λαμβάνοντας υπόψη ότι

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{και} \quad E(\hat{\beta}_1) = \beta_1 ,$$

μετά από πράξεις προκύπτει:

$$E(SS_{reg}) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (4)$$

Έτσι από τις (3) και (4) έχουμε:

$$E(MS_{res}) = \frac{1}{n-2} E(SS_{res}) = \frac{1}{n-2} \left[E(SS_{tot}) - E(SS_{reg}) \right]$$

$$= \frac{1}{n-2} \left\{ \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + (n-1)\sigma^2 - \sigma^2 - \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\}$$

$$= \frac{1}{n-2} (n-2)\sigma^2$$

$$= \sigma^2 \qquad \blacktriangle$$

ΘΕΩΡΗΜΑ Οι ΕΕΤ $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι ανεξάρτητα των SS_{res}, MS_{res}.

ΘΕΩΡΗΜΑ Υπό τις υποθέσεις για τα σφάλματα

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

Απόδ

Λόγω των υποθέσεων για τα σφάλματα $\forall i=1,2,\dots,n$ το $Y_i \sim N(\beta_0+\beta_1 X_i, \sigma^2)$

άρα (τυποποίηση) $\frac{Y_i-(\beta_0+\beta_1 X_i)}{\sigma} \sim N(0,1) \quad \forall i=1,\dots,n$

$$\Rightarrow \frac{[Y_i-(\beta_0+\beta_1 X_i)]^2}{\sigma^2} \sim N^2(0,1) \equiv \chi^2_1 \quad \forall i=1,\dots,n$$

Y_i ανεξ

$$\Rightarrow \sum \frac{(Y_i-(\beta_0+\beta_1 X_i))^2}{\sigma^2} \sim \chi^2_{\sum_{i=1}^n 1} \equiv \chi^2_n \quad (*)$$

λόγω του ότι
τα Y_i ασυσχέτιστα
Και Κανονικά

$$SS_{res} = \sum_{i=1}^n (Y_i-\hat{Y}_i)^2$$

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

$$\Rightarrow \sum_{i=1}^n \frac{(Y_i-\hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{n-2}$$