Παλινδρόμηση 10/10/16 ①

▶ Απλή Γραμμική Παλινδρόμηση

Δύο ή περισσότερες τυχαίες μεταβλητές π.χ X, Y

και έστω τυχαίο δείγμα $X_1 \ldots X_n$ από τη X

και έστω τ.δ Y_1, Y_n από τη Y

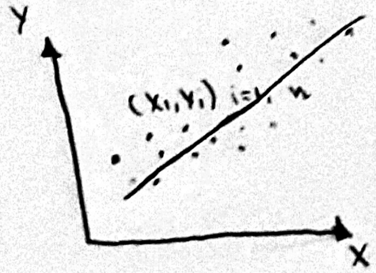
οπότε έχω δεδομένα

X	x_1	x_2	x_n
Y	y_1	y_2	y_n

Ερώτημα: Υπάρχει κάποια σχέση (γραμμική;) που να συνδέει τις X, Y;

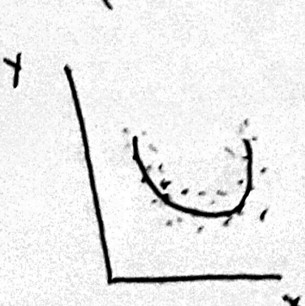
1ο Βήμα για να απαντήσω στο ερώτημα

Διάγραμμα διασποράς



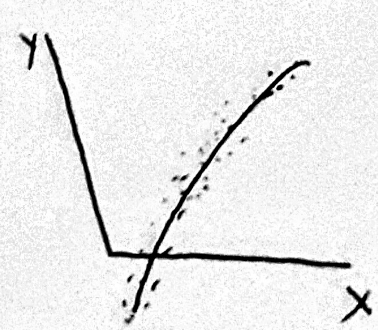
οι μεταβλητές X, Y τείνουν να προσαρμόζονται από μία ευθεία
$Y = B_0 + B_1 X + \varepsilon$

Στην περίπτωση που είχα τέτοιο διάγραμμα δεν θα μπορούσα να το προσεγγίσω από μια ευθεία



$Y = B_0 + B_1 X + B_2 X^2 + \varepsilon$

ή



$Y = B_0 + B_1 \log x + \varepsilon$

Οι σχέσεις αυτές δεν είναι απόλυτες, αλλά περιμένω τα σημεία να παλατζάρουν - παλινδρομούν γύρω από τη καμπύλη που τείνει να τα περιγράψει

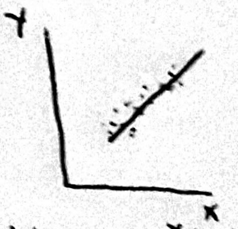
Αυτό ακριβώς περιγράφει το -ε-

Περιοριζόμαστε σε γραμμικές σχέσεις

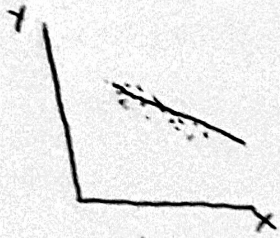
Το διάγραμμα διασποράς εκφράζει

1. Τη μορφή της σχέσης X και Y π.χ γραμμική σχέση $Y=B_0+B_1x+\varepsilon$

2. Κατεύθυνση της γραμμικής σχέσης



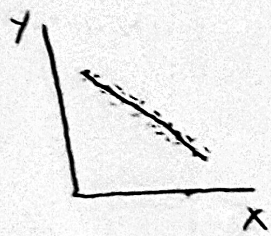
Μεγαλώνουν η X και η Y
άρα θετική κατεθ.

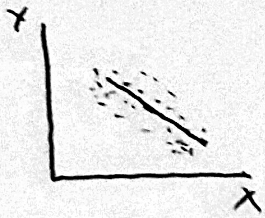


Μεγαλώνει η X μικραίνει η Y
άρα αρνητική κατευθ

3. Πόσο έντονη είναι η γραμμική σχέση

δηλαδή πόσο παλινδρομούν απο το γραμμικό μοντέλο που τείνει να τα περιγράψει





Δηλαδή η $Y=B_0+B_1X+\varepsilon$ είναι το <u>μοντέλο απλής γραμμικής παλινδρόμησης</u>

• Απλή = μεταξύ μόνο 2 μεταβλητών

ή αλλιώς $Y_i=B_0+B_1X_i+\varepsilon_i$, $i=1,n$

Η μεταβλητή X ονομάζεται ανεξάρτητη ή επεξηγηματική

Η Y ονομάζεται εξαρτημένη ή απόκριση

Τα ε_i ονομάζονται σφάλματα του μοντέλου, (η απόσταση του σημείου απο το μοντέλο)

πως ;

Το B_i ονομάζονται παράμετροι του μοντέλου

2º Βήμα Κατασκευή του Μοντέλου α.γ.π

Εύρεση εκτιμητών των παραμέτρων B_0, B_1

① Εκτιμήτες Ελαχίστων Τετραγώνων (Ε.Ε.Τ) των B_0, B_1 (Pearson, ~1900)

Η ιδέα της μεθόδου εκτίμησης: Θέλω να βρώ την 'καλύτερη' ευθεία που θα συμπεριλαμβάνει τα σημεία

Το κριτήριο είναι τα σημεία να συμπεριλαμβάνονται όσο πιο κοντά γίνεται στη γραμμή.
δηλαδή έτσι που τα ε_i ελαχιστοποιούνται

Άρα λοιπόν αρκεί να βρώ τα B_0, B_1 που ελαχιστοποιούν το

$$S=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(Y_i-B_0-B_1X_i)^2$$

(το τετράγωνο είναι για να μην έχω αρνητικά)

• $\frac{\partial S}{\partial B_0}=0$ και $\frac{\partial S}{\partial B_1}=0$

δηλαδή $\begin{cases} nB_0+B_1\sum X_i=\sum Y_i \\ B_0\sum X_i+B_1\sum X_i^2=\sum X_iY_i \end{cases}$ κανονικές εξισώσεις

Οπότε, η επίλυση των κανονικών εξισώσεων οδηγεί στους ΕΕΤ των B_0 και B_1 που είναι

▶ $$\hat{B}_1=\frac{\sum X_iY_i-(\sum X_i)(\sum Y_i)/n}{\sum X_i^2-\frac{(\sum X_i)^2}{n}}=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$$

▶ $$\hat{B}_0=\bar{Y}-\hat{B}_1\bar{X}$$

Τελικά η καλύτερη ευθεία που θα μπορούσα να έχω, δηλαδή με τα μικρότερα σφάλματα προκύπτει από την αντικατάσταση των παραμέτρων από τους εκτιμήτες τους: $\hat{Y}=\hat{B}_0+\hat{B}_1X$

Εκτιμώμενο μοντέλο α.γ.π

άποψη ερμηνείας ▶ το $\hat{B}_0$ παριστά την τιμή της Y για $X=0$

▶ το $\hat{B}_1$ παριστά την μεταβολή της Y σε μοναδιαία μεταβολή της X

3ο Βήμα

Διάγνωση της ορθότητας του μοντέλου

▶ Υπόλοιπα: Ορίζονται ως οι αποκλίσεις του μοντέλου από την πραγματικότητα που το μοντέλο επιδιώκει να περιγράψει

Ορισμός: $e_i = Y_i - \hat{Y}_i \quad i=1,n$

Ιδιότητα των υπολοίπων $\sum_{i=1}^{n} e_i = 0$

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - \hat{Y}_i = \sum_{i=1}^{n} (Y_i - \hat{B}_0 - \hat{B}_1 X_i)$$

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

$$= \sum_{i=1}^{n} \overset{\to 0}{(Y_i - \bar{Y})} + B_1 \sum \overset{0}{(\bar{X} - X_i)} = 0$$

άρα $\sum (Y_i - \bar{Y}) = \sum Y_i - \sum \bar{Y} = n\bar{Y} - n\bar{Y} = 0$

και $\sum (\bar{X} - X_i) = 0$ το ίδιο

Μια πρώτη ένδειξη είναι τα υπόλοιπα e_i να είναι κοντά στο 0 δηλ να μην υπάρχει διαφορά μεγάλη από την πραγματικότητα στο μοντέλο της α.δ.π

▶ Ανάλυση Ολικής Μεταβλητότητας

Ολική μεταβλητότητα $\simeq$ Δειγματική διακύμανση

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

o Ολική μεταβλητότητα $= \sum (Y_i - \bar{Y})^2$

o Ολική μεταβλητότητα διέπεται

Εδώ θέλω να δω πως η ολική μεταβλητότητα εξηγείται από το μοντέλο, άρα πρέπει να το εξηγήσω

Το πιο απλό είναι να το προσθαφαιρέσω

③

Δηλαδή $\pm \hat{y}_i$

οπότε μετά από πράξεις (απόδειξη σε σημ.)

έχω $\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

μεταβλητότητα που οφείλεται στο μοντέλο

ολική μεταβλητ.

μεταβλητότητα που οφείλεται στα υπόλοιπα

Δηλαδή $\boxed{SS_{tot} = SS_{reg} + SS_{res}}$

Αν το μεγαλύτερο μέρος από την ολική μεταβλητότητα περιέχει SSres τότε αυτό είναι κακό.

Δηλαδή ένα δεύτερο κριτήριο είναι Αν $SS_{reg} > SS_{res}$ τότε το μοντέλο της α.γ.π φαίνεται να είναι υποσχόμενο

αλλιώς συμβαίνει το αντίθετο

• Άλλη μορφή για το άθροισμα τετραγ. που οφείλεται στην παλινδρ. SS_{reg}

$$SS_{reg} = \sum (\hat{y}_i - \bar{y})^2 = \sum (\hat{\beta}_0 + \hat{\beta}_1 x - \bar{y})^2$$
$$= \sum (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2$$
$$= \hat{\beta}_1^2 \sum (x_i - \bar{x})^2$$

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

<u>Πίνακας Ανάλυσης Διακύμανσης του μοντέλου της α.δ.π</u>

<u>ΠΙΝΑΚΑ</u> <u>ΑΝΑΔΙΑ</u>

Πηγή μεταβλητότητας	Αθρ. Τετρ. SS	Β.Ε	M.S μέσο τετραγ.	F πηλ.
μοντέλο α.δ.π	SS_{reg}	1	$MS_{reg} = \frac{SS_{reg}}{1}$	$F = \frac{MS_{reg}}{MS_{res}}$
υπόλοιπο	SS_{res}	$n-2$	$MS_{res} = \frac{SS_{reg}}{n-2}$	
ολική μεταβλητότητα	SS_{tot}	$n-1$		

Για να αξιοποιήσω αυτόν τον πίνακα χρησιμοποιώ το 2º κριτήριο που προαναφέραμε

- 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

~~Παραδ~~
Ερμηνεία

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!! $\bar{Y} = \frac{1}{n}\Sigma Y_i$ δηλαδή μπορώ να εκφράσω το Y_n συναρτήσει των άλλων άρα έχω $n-1$ Β.Ε

④

Απόδειξη ότι Β.Ε του SS_{tot} είναι $n-1$

- $SS_{tot} = \sum (y_i - \bar{y})^2$

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

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

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

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

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

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 $y_1 - \bar{y} \ldots \quad y_{n-1} - \bar{y}$

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
(διότι την περίπτωσή μας είναι μόνο η Χ)

Το SS_{res} έχει Β.Ε την διαφορά

Συντελεστής προσδιορισμού ή προσαρμοστικότητας

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

απο τη σχέση $SS_{reg} + SS_{res} = SS_{tot}$

και 2 ιδιότητες είναι οι εξής:

1. ο R^2 είναι καθαρός αριθμός

2. $0 \leq R^2 \leq 1$ και άρα εκφράζεται και ως ποσοστό!

Ενδιαφέρον παρουσιάζουν οι τιμές 0 και 1

▶ Τιμή του R^2 κοντά στο 1: δηλαδή θα πρέπει $SS_{reg} \simeq SS_{tot}$

δηλ $SS_{reg} \gg SS_{res}$

υποσχόμενο μοντέλο

▶ Τιμή του R^2 κοντά στο 0: δηλαδή θα πρέπει

$SS_{reg} \ll SS_{res}$ μη υποσχόμενο μοντέλο

Μια ερμηνεία για το R^2: Εκφράζει το ποσοστό της ολικής μεταβλητότητας των Y_1, Y_n που ερμηνεύεται απο το μοντέλο μας