9ο online μάθημα
23/4/2020

2ος τρόπος: Γραμμικός Συντελεστής Συσχέτισης

Έστω X, Y δύο συνεχείς τ.μ. Εισάγεται τότε η έννοια της συνδιακύμανσης των X, Y η οποία συμβολίζεται $Cov(X,Y)$ (↓ Covariance) και ορίζεται:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

όπου $\mu_X = EX$, $\mu_Y = EY$

Εύκολα προκύπτει

$$Cov(X,Y) = E(XY) - (EX)(EY)$$

Παρατήρηση 1η: Η συνδιακύμανση είναι ένα μέτρο της συμμεταβλητότητας των X, Y. Έχει ως μονάδες μέτρησης το γινόμενο των μονάδων μέτρησης των X, Y. Με στόχο να απαλλαγούμε από τις μονάδες οδηγούμαστε στο συντελεστή συσχέτισης

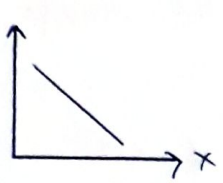
$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var\,X}\sqrt{Var\,Y}}$$

Θεωρητικός - Πληθυσμιακός γραμμικός συντελεστής συσχέτισης

- Είναι καθαρός αριθμός
- Αποδεικνύεται (μάθημα 531) ότι:

α) $-1 \leq \rho(x,y) \leq 1$

β) $\rho(x,y) = -1$ έχουμε τέλεια αρνητική γραμμική συσχέτιση



γ) $\rho(x,y) = 1$ έχουμε τέλεια θετική γραμμική συσχέτιση

δ) $\rho(x,y) = 0$ τότε δεν υπάρχει γραμμική σχέση

Επομένως για να δούμε αν υπάρχει ή όχι γραμμική σχέση μεταξύ των X, Y αρκεί να ελέγξω: την:

$H_0: \rho(x,y) = 0$

$H_{\alpha}: \rho(x,y) \neq 0$

- Αν αυτός ο έλεγχος απορρίπτεται συμαίνει ότι υπάρχει γραμμική σχέση.
- Αν δεν απορρίπτεται συμαίνει ότι δεν συσχέτονται γραμμικά.

Ο έλεγχος "βρέπει" να στηριχθεί σ' ένα δειγματικό αναλογο του $\rho(x,y)$. Είναι:

$$r = r_{x,y} = r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

ή ισοδύναμα

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Παρατηρήσεις: • $r(x,y) = r(y,x)$

• Ορίζεται για ποσοτικές μεταβλητές

• $r(x,y) > 0$ υποδηλώνει θετική συσχέτιση.

ενώ $r(x,y) < 0$ αρνητική.

• Τιμές του $r(x,y)$ "κοντά" στο 0 δηλώνουν μη υπαρξη γραμμικής σχέσης.

• ΠΡΟΒΟΛΗ ΕΝΔΕΙΚΤΙΚΩΝ ΓΡΑΦΗΜΑΤΩΝ

• Η μη υπαρξη γραμμικής σχέσης δεν συμαίνει την μη υπαρξη σχέσης.

Έλεγχος για τον $\rho(x,y)$

$H_0: \rho(x,y) = 0$

έναντι μιας εκ των

$H_1: \rho(x,y) > 0$ (θετική γραμμική συσχέτιση) ή $H_1: \rho(x,y) < 0$ (αρνητική γραμμική συσχέτιση) ή $H_1: \rho(x,y) \neq 0$ (γραμμική συσχέτιση)

Υπό την προϋπόθεση ότι (X,Y) έχουν δισδιάστατη κανονική και δεν υπάρχουν ακέραιες τιμές.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \overset{H_0}{\sim} t_{n-2}$$

(r: στατιστικός συντελεστής συσχέτισης; n: πλήθος (X_i, Y_i))

Άρα περιοχή απόρριψης

$t > t_{n-2,\alpha}$ ή $t < -t_{n-2,\alpha}$ ή $|t| \geq t_{n-2,\alpha/2}$ αντίστοιχα.

$H_a: \rho > 0$ $H_a: \rho < 0$ $H_a: \rho \neq 0$

Παρατήρηση: Για τον γενικότερο έλεγχο

$H_0: \rho(x,y) = \rho_0$ $(\rho_0 \in [-1,1])$, $H_1: \rho > \rho_0$ ή $H_1: \rho < \rho_0$ ή $H_1: \rho \neq \rho_0$

Χρησιμοποιείται το στατιστικό

$$W = \frac{Z - E(Z)}{\sqrt{\text{Var } Z}} \overset{H_0}{\sim} N(0,1)$$

όπου

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}$$

$$EZ = \frac{1}{2} \ln \frac{1+r_0}{1-r_0}$$

$$\text{Var } Z = \frac{1}{n-3} \quad]$$

Εκτός

Μόνο για κανονικοποιημένους πληθυσμούς

Περιοχές απόρριψης:

$$W \geq Z_\alpha \text{ ή } W \leq -Z_\alpha \text{ ή } |W| \geq Z_{\alpha/2}$$

Αφού διαπιστωθεί η ύπαρξη γραμμικής σχέσης, στόχος είναι ο προσδιορισμός της με απώτερο σκοπό την αξιοποίηση της για την πρόβλεψη της μιας μεταβλητής όταν η άλλη θεωρείται δεδομένη. Η γραμμική παλινδρόμησης της Y πάνω στην X εκφράζεται από τη σχέση:

$$Y = b_0 + b_1 X + \varepsilon$$

με b_0, b_1 σταθερές αλλά άγνωστες ποσότητες και ε το τυχαίο σφάλμα που παριστάνει την απόκλιση της Y από την γραμμική παλινδρόμησης στο σημείο X.

Η σχέση προφανώς θα ικανοποιείται από τα διαθέσιμα δεδομένα

για i

Επομένως $Y_i = b_0 + b_1 X_i + \varepsilon_i \quad i = 1, \dots, n$

1ο Βήμα: Εκτίμηση των παραμέτρων b_0, b_1

Πώς; Προσδιορισμός των b_0, b_1 έτσι ώστε να ελαχιστοποιείται

το αθροίσμα των τετραγώνων των σφαλμάτων. Η μέθοδος αυτή εκτίμησης είναι γνωστή ως μέθοδος ελαχίστων τετραγώνων θέλουμε να ελαχιστοποιήσουμε την

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \quad \text{ως προς } \beta_0, \beta_1$$

Είναι

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

Οι εκτιμητές, έστω $\hat{\beta}_0, \hat{\beta}_1$ προκύπτουν από την επίλυση του συστήματος

$$\begin{cases} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \end{cases}$$

ΓΝΩΣΤΟ ΩΣ ΣΥΣΤΗΜΑ ΚΑΝΟΝΙΚΩΝ ΕΞΙΣΩΣΕΩΝ

ή ισοδύναμα

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i \\ \sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \end{cases}$$

Από την πρώτη προκύπτει ότι

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

Ενώ έπειτα έχουμε

$$\sum_{i=1}^{n} X_i Y_i = \bar{Y} \sum_{i=1}^{n} X_i - \hat{\beta}_1 \bar{X} \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \Rightarrow$$

$$\hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \right) = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i}{n}$$

$$\Rightarrow \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

ή

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

Παρατήρηση: Κανονικά θα έπρεπε για να είμαστε σίγουροι ότι είναι σημεία ελαχίστου, να ελέγξουμε τις δεύτερες παραγώγους και εδιανό πίνακα.

Εκτιμώμενη τιμή $\hat{Y}$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

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

$\hat{\beta}_0$: τιμή της $\hat{Y}$ για $x=0$

$\hat{\beta}_1$: μεταβολή της $\hat{Y}$ για μοναδιαία μεταβολή της X

Γιατί;

Έστω (x_1, Y) τότε

$$\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

ενώ για $x_2 = x_1 + 1$ είναι

$$\hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1 (x_1 + 1) \Rightarrow \hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_1 \Rightarrow \hat{Y}_2 = \hat{Y}_1 + \hat{\beta}_1$$

Υπόλοιπο: $e_i = y_i - \hat{y}_i \quad i = 1, \dots, n$

Παρατηρήσεις: ① Το σημείο $(\bar{x}, \bar{y})$ ανήκει στην ευθεία παλινδρόμησης.

Όντως για $x = \bar{x}$ είναι

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \xrightarrow{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}} \hat{y} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$$

② Τα υπόλοιπα αθροίζουν στο 0

Όντως

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

$$= 0$$

Η πιο απλά ισχύει από το σύστημα κανονικών εξισώσεων ότι

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

③ Ισχύει ότι $\sum_{i=1}^{n} x_i e_i = 0$. Όντως είναι η 2η εξίσωση του συστή-

ματος κανονικών εξισώσεων.

④ Ισχύει ότι $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$

Όντως

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \Rightarrow \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i$$

Όπως από την 1η από κανονικές εξισώσεις:

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad \text{Ισχύει}$$

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta} \sum_{i=1}^{n} x_i$$

Ιδιότητες των εκτιμητών ελαχίστων τετραγώνων

Έχουμε δείξει ότι

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

$$\sum(x_i-\bar{x})(y_i-\bar{y}) = \sum(x_i-\bar{x})y_i - \sum(x_i-\bar{x})\bar{y}$$

ή

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

ή καθώς $\sum_{i=1}^{n}(x_i-\bar{x})\bar{y} = \bar{y}\sum_{i=1}^{n}(x_i-\bar{x}) = 0$

$$\sum(x_i-\bar{x}) = \sum x_i - n\bar{x} = \sum x_i - n\frac{\sum x_i}{n}$$

Επίσης $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

Θ.Σ.Ο. $E\hat{\beta}_0 = \beta_0$

$E\hat{\beta}_1 = \beta_1$

υπό την υπόθεση ότι $E(\varepsilon_i)=0$

Είναι

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

$$E\left[\sum_{i=1}^{n}(x_i-\bar{x})y_i\right]$$

$$= \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}E[(x_i-\bar{x})y_i] = \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{i=1}^{n}(x_i-\bar{x})E\,y_i$$

Όμως $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

άρα $Ey_i = \beta_0 + \beta_1 x_i$

Επομένως

$$E\hat{\beta}_1 = \frac{1}{\Sigma(x_i-\bar{x})^2} \sum_{i=1}^{n} (x_i-\bar{x})(\beta_0+\beta_1 x_i) =$$

$$= \frac{1}{\Sigma(x_i-\bar{x})^2} \left[\beta_0 \sum_{i=1}^{n} (x_i-\bar{x}) + \beta_1 \sum_{i=1}^{n} (x_i-\bar{x}) x_i \right] =$$

$$= \frac{1}{\Sigma(x_i-\bar{x})^2} \beta_1 \Sigma (x_i-\bar{x})^2 = \beta_1$$

Χρησιμοποιήσαμε ότι

$$\sum_{i=1}^{n} (x_i-\bar{x}) x_i = \sum_{i=1}^{n} (x_i-\bar{x})(x_i-\bar{x})$$

Τώρα είναι:

$$E\hat{\beta}_0 = E\bar{Y} - \bar{x}\, E\hat{\beta}_1 \Rightarrow E\hat{\beta}_0 = E\frac{\Sigma Y_i}{n} - \bar{x}\beta_1 \Rightarrow$$

$$E\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} (\beta_0+\beta_1 x_i) - \bar{x}\beta_1 \Rightarrow E\hat{\beta}_0 = \beta_0$$

Πρόταση: Υπό την υπόθεση ότι $E(\varepsilon_i)=0$, $\mathrm{Var}(\varepsilon_i)=\sigma^2$, υ.δ.ο.

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

$$\mathrm{Var}\,\hat{\beta}_0 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i-\bar{x})^2} \right)$$

$$\mathrm{Var}\,\hat{\beta}_1 = \mathrm{Var}\left[\frac{\sum_{i=1}^{n} (x_i-\bar{x}) Y_i}{\Sigma(x_i-\bar{x})^2} \right] \Rightarrow$$

$$\text{Var}\,\hat{\beta}_1 = \left\{ \frac{1}{\sum (x_i - \bar{x})^2} \right\}^2 \text{Var}\left(\sum_{i=1}^{n} (x_i - \bar{x}) Y_i \right)$$

$$\Rightarrow \text{Var}\,\hat{\beta}_1 = \left\{ \frac{1}{\sum (x_i - \bar{x})^2} \right\}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \,\text{Var}\, Y_i$$

$$\rightarrow \text{Var}\,\hat{\beta}_1 = \left\{ \frac{1}{\sum (x_i - \bar{x})^2} \right\}^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma^2$$

και αποδείχθηκε το ζητούμενο.

$$\text{Var}\,\hat{\beta}_0 = \text{Var}\,(\bar{Y} - \hat{\beta}_1 \bar{x})$$

Προσοχή:

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

άρα δεν είναι ανεξάρτητες τ.μ. για να μπορέσω να πω ότι

π.χ. $\text{Var}\,(a_1 X_1 + a_2 X_2) = a_1^2 \,\text{Var}\, X_1 + a_2^2 \,\text{Var}\, X_2$

Είναι όμως

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

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

Επομένως:

$$\text{Var}\,\hat{\beta}_0 = \sum_{i=1}^{n} \left\{ \frac{1}{n} - \frac{\bar{x}\,(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right\}^2 \text{Var}\, Y_i$$

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

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

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

Παρατήρηση: Πιο όμορφες αποδείξεις στο μάθημα του 3ου εξαμήνου!!!