ΘΕΜΑΤΙΚΗ ΕΝΟΤΗΤΑ ΔΕΟ 13

Eclass4U

The best Choice for you

ΣΗΜΕΙΩΣΕΙΣ

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

ΗΜΕΡΟΜΗΝΙΑ ΑΝΑΡΤΗΣΗΣ:15-01-22

ΣΥΝΤΑΚΤΗΣ:ΣΠΥΡΟΣ ΒΛΑΧΟΠΟΥΛΟΣ

Eclass4U
The best Choice for you

ΘΕΡΜΟΠΥΛΩΝ 17
ΠΕΡΙΣΤΕΡΙ

100Μ ΑΠΟ ΤΗ ΣΤΑΣΗ
ΜΕΤΡΟ «ΠΕΡΙΣΤΕΡΙ»

ΤΗΛΕΦΩΝΟ: 210-5711484
ΚΙΝΗΤΟ: 6970401981
EMAIL:grammateia.eclass4u@gmail.com
ΤΟΠΟΘΕΣΙΑ WEB : www.eclass4u.gr
SOCIAL MEDIA:

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Μέχρι στιγμής, στην στατιστική, ασχοληθήκαμε με τη μελέτη ενός πληθυσμού (δείγματος) ως προς ένα και μόνο χαρακτηριστικό (εισόδημα, ηλικία, βαθμοί σε μια εξέταση κλπ). Σε πολλές άλλες περιπτώσεις έχει μεγάλο ενδιαφέρον **η ταυτόχρονη μελέτη δυο μεταβλητών** ώστε να προσδιοριστεί ο τρόπος με τον οποίο αυτές σχετίζονται.

Για παράδειγμα, είναι προφανές ότι το ύψος και το βάρος ενός παιδιού σχετίζονται (όσο ψηλότερο είναι ένα παιδί, τόσο μεγαλύτερο βάρος αναμένεται να έχει). Το οικογενειακό εισόδημα και τα συνολικά οικογενειακά έξοδα διαβίωσης, επίσης σχετίζονται (όσο μεγαλύτερο το συνολικό εισόδημα, τόσο περισσότερα χρήματα αναμένεται να ξοδεύει η οικογένεια για τις καταναλωτικές της ανάγκες). Υπάρχει, όμως, και αρνητική συσχέτιση ανάμεσα σε δυο μεταβλητές (δηλαδή όταν αυξάνεται η μία, μειώνεται αντίστοιχα η άλλη). Παράδειγμα, οι πωλήσεις σε κιλά σε ένα ιχθυοπωλείο και η τιμή των ψαριών ανά κιλό, σχετίζονται αρνητικά (όσο μεγαλύτερη η τιμή τόσο λιγότερα κιλά ψαριών αναμένεται να πουληθούν).

Σε τέτοιους διμεταβλητούς πληθυσμούς, η μια μεταβλητή είναι η ανεξάρτητη (Χ) και η δεύτερη μεταβλητή είναι η εξαρτημένη (Υ). Στο παράδειγμα με το βάρος ενός παιδιού και το ύψος, η ανεξάρτητη μεταβλητή Χ είναι το ύψος (καθώς αυτό είναι ανεξάρτητο του βάρους) ενώ το βάρος είναι η εξαρτημένη μεταβλητή (αφού εξαρτάται από το ύψος). Στο παράδειγμα του ιχθυοπωλείου η ανεξάρτητη μεταβλητή Χ είναι η τιμή σε κιλά, ενώ η εξαρτημένη μεταβλητή Υ είναι οι πωλήσεις σε κιλά (αφού οι πωλήσεις εξαρτώνται από την τιμή του προϊόντος).

Ο κλάδος της στατιστικής που εξετάζει τη σχέση δυο τέτοιων (ή και περισσότερων) μεταβλητών ονομάζεται **ανάλυση παλινδρόμησης** (regression analysis). Η απλούστερη περίπτωση παλινδρόμησης είναι η απλή **γραμμική παλινδρόμηση** κατά την οποία η συσχέτιση ανάμεσα στην ανεξάρτητη μεταβλητή Χ και την εξαρτημένη μεταβλητή Υ προσεγγίζεται ικανοποιητικά από μια ευθεία (ευθεία παλινδρόμησης). Μια ευθεία, δηλαδή, της μορφής $\Upsilon = \alpha \cdot \Chi + \beta$

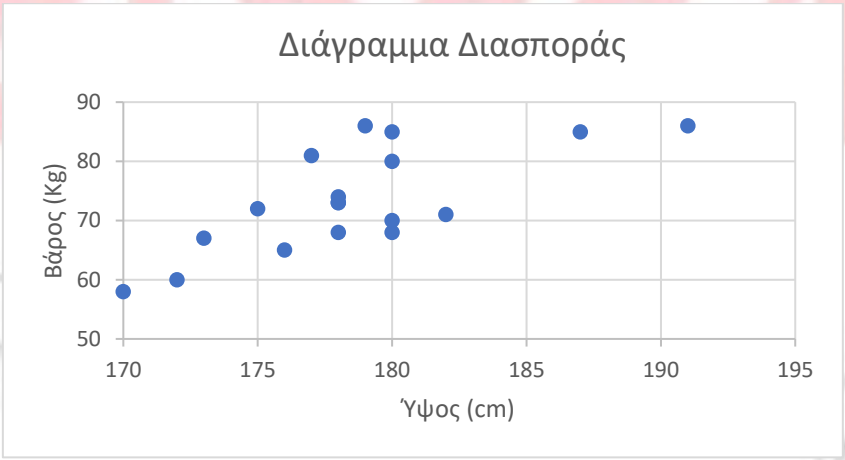
SITE: www.Eclass4U.gr
Τ. 210-5711484 ΚΙΝ. 6970401981
E-MAIL:grammateia.eclass4u@gmail.com

Ας δούμε το πρόβλημα εύρεσης της ευθείας παλινδρόμησης με ένα παράδειγμα.

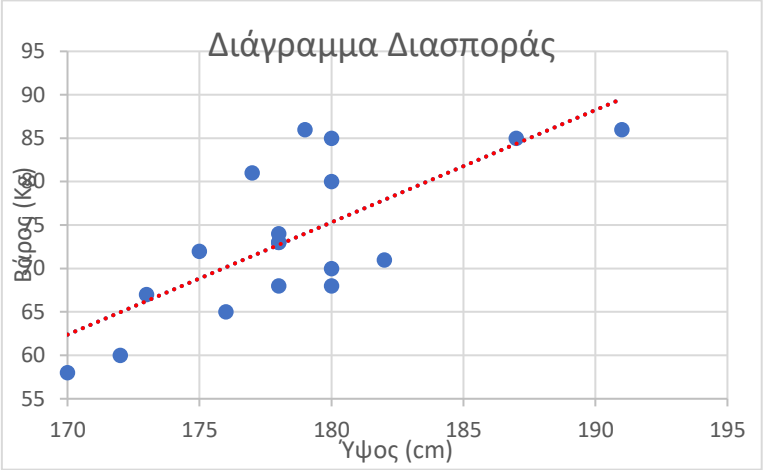
Ο παρακάτω πίνακας δίνει τα ύψη Χ (σε cm) και τα βάρη Υ (σε kg) 18 μαθητών.

Μαθητής	**Ύψος Χ**	**Βάρος Υ**	Μαθητής	**Ύψος Χ**	**Βάρος Υ**
1	**170**	**58**	10	**178**	**68**
2	**172**	**60**	11	**179**	**86**
3	**173**	**67**	12	**180**	**68**
4	**175**	**72**	13	**180**	**80**
5	**176**	**65**	14	**180**	**70**
6	**177**	**81**	15	**180**	**85**
7	**178**	**73**	16	**182**	**71**
8	**178**	**74**	17	**187**	**85**
9	**178**	**73**	18	**191**	**86**

Αν παραστήσουμε τα ζεύγη των παρατηρήσεων (Χ,Υ) σε ένα σύστημα ορθογωνίων αξόνων, παίρνουμε αυτό που αποκαλείται **διάγραμμα διασποράς** των παρατηρήσεων.



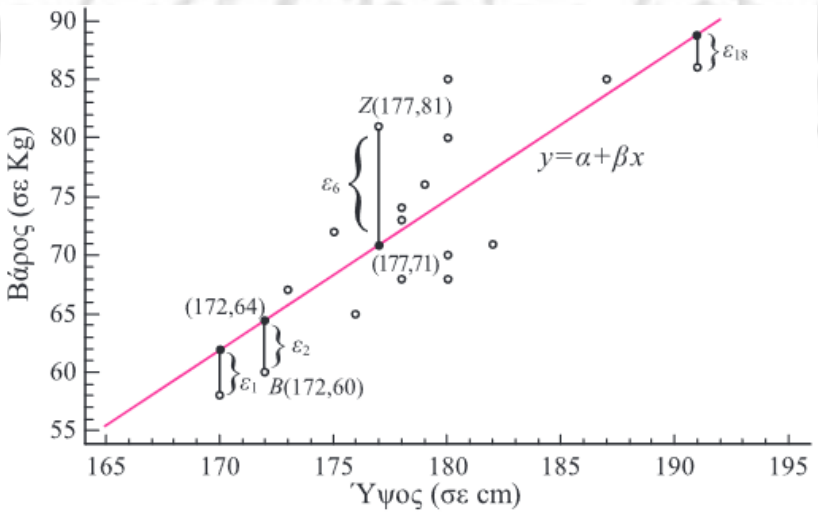
Από το παραπάνω διάγραμμα παρατηρούμε ότι τα σημεία (Χ,Υ) είναι συγκεντρωμένα γύρω από μια ευθεία η οποία προσεγγιστικά «με το μάτι» θα ήταν περίπου η παρακάτω.



Διαπιστώνουμε, δηλαδή, ότι παρά τη «διασπορά» των σημείων, υπάρχει μια ευθεία γύρω από την οποία είναι «συγκεντρωμένα». Συνεπώς υπάρχει μια ευθεία της μορφής $\Upsilon = \alpha \cdot \Chi + \beta$ που καλούμαστε να εντοπίσουμε πολύ πιο ικανοποιητικά από την προσέγγιση «με το μάτι» που έχουμε κάνει μέχρι στιγμής. Αυτό σημαίνει ότι πρέπει να εκτιμήσουμε τις παραμέτρους α και β της εξίσωσης της ευθείας με κάποιο πολύ πιο ακριβή κι αντικειμενικό τρόπο.

Μια μέθοδος που χρησιμοποιείται για την εκτίμηση των παραμέτρων α και β, άρα και την εύρεση της «καλύτερης» ευθείας που προσαρμόζεται στα δεδομένα είναι η «**μέθοδος ελαχίστων τετραγώνων**».

Το κάθε σημείο του διαγράμματος απέχει από την ευθεία που φέραμε μια κατακόρυφη απόσταση ε_1, ε_2,…ε_{18}. (όπως φαίνεται στο ακόλουθο διάγραμμα διασποράς)



Η μέθοδος ελαχίστων τετραγώνων στηρίζεται στην ελαχιστοποίηση του αθροίσματος των **τετραγώνων** των σφαλμάτων.

Καταλήγουμε, έτσι, στην **ευθεία ελαχίστων τετραγώνων** ή **ευθεία παλινδρόμησης** της Υ πάνω στην Χ. Στο τυπολόγιο δίνεται ως εξής

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

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

$$\hat{\beta}_0 = \frac{\sum Y_i}{n} - \hat{\beta}_1 \frac{\sum X_i}{n} = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Οι $\hat{\beta}_0$ και $\hat{\beta}_1$ λέγονται εκτιμήτριες. Στις ασκήσεις μας δεν χρειάζεται να τις γράφουμε με το «καπελάκι». Παρατηρούμε ότι η εκτιμήτρια $\hat{\beta}_1$ έχει δυο τύπους ο αριστερά τύπος είναι ο τύπος «με απόκλιση από τους μέσους»

$$\hat{\beta}_1 = \frac{\sum[(\mathrm{Y_i} - \bar{\mathrm{Y}}) \cdot (\mathrm{X_i} - \bar{\mathrm{X}})]}{\sum[(\mathrm{X_i} - \bar{\mathrm{X}})^2]}$$

Αφού περιέχει τις διαφορές $\mathrm{Y_i} - \bar{\mathrm{Y}}$ και $\mathrm{X_i} - \bar{\mathrm{X}}$ (δηλαδή πόσο «αποκλίνει» η κάθε παρατήρηση $\mathrm{Y_i}$ και $\mathrm{X_i}$ από τον αντίστοιχο αριθμητικό μέσο).

Ο δεύτερος τύπος (που προτιμάται επειδή είναι πιο «γρήγορος») είναι ο

$$\hat{\beta}_1 = \frac{\sum \mathrm{X_i} \cdot \mathrm{Y_i} - \frac{\sum \mathrm{X_i} \cdot \sum \mathrm{Y_i}}{\mathrm{n}}}{\sum \mathrm{X_i^2} - \frac{(\Sigma \mathrm{X_i})^2}{\mathrm{n}}}$$

Αν ζητηθεί από την άσκηση, θα χρησιμοποιείται ο πρώτος τύπος, διαφορετικά θα επιλέγουμε τον δεύτερο.

Για την εκτιμήτρια $\hat{\beta}_0$ ο τύπος είναι ο ακόλουθος, και θα τον χρησιμοποιούμε αφού έχουμε υπολογίσει την τιμή της $\hat{\beta}_1$

$$\hat{\beta}_0 = \frac{\sum Y_i}{n} - \hat{\beta}_1 \cdot \frac{\sum X_i}{n} = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

Να σημειωθεί ότι n= το πλήθος των ζευγών (X, Y)

Ερμηνείες των $\hat{\beta}_0$ και $\hat{\beta}_1$ της ευθείας παλινδρόμησης $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$

(μπορεί να ζητηθεί και ως «να ερμηνευθούν οι συντελεστές του γραμμικού υποδείγματος)

Για την $\hat{\beta}_0$: Αν η τιμή της μεταβλητής X γίνει 0, τότε η τιμή της μεταβλητής Y είναι ίση με $\hat{\beta}_0$

Για την $\hat{\beta}_1$: Αν η τιμή της μεταβλητής X αυξηθεί κατά μία μονάδα, τότε η μεταβλητή Y θα μεταβληθεί κατά $\hat{\beta}_1$

Επιστρέφοντας στο αρχικό παράδειγμα των 18 μαθητών και τον πίνακα με τα ύψη και τα αντίστοιχα βάρη τους, η ευθεία παλινδρόμησης που προκύπτει με χρήση των τύπων είναι η

$$Y = -156,9 + 1,29 \cdot X$$

Ερμηνεία των β_0 και β_1 :

Για την $\hat{\beta}_1 = 1,29$: (Αν η τιμή της μεταβλητής X αυξηθεί κατά μία μονάδα, τότε η μεταβλητή Y θα μεταβληθεί κατά $\hat{\beta}_1$)

Με όρους της άσκησης : Αν το ύψος αυξηθεί κατά 1 cm τότε το βάρος θα αυξηθεί κατά 1,29 kg

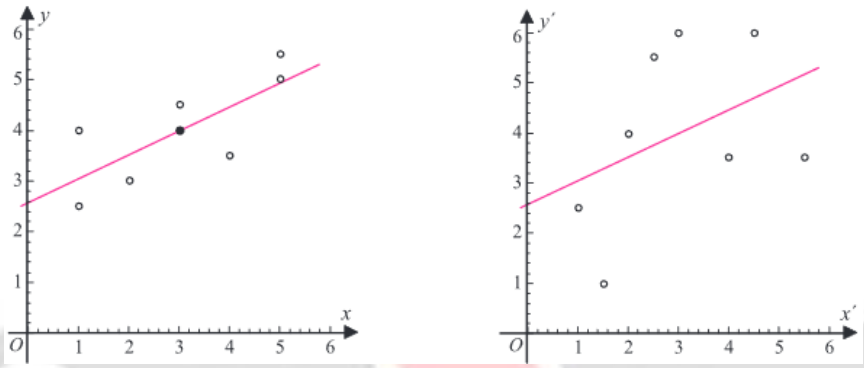
Για την $\hat{\beta}_0 = -156,9$: (Αν η τιμή της μεταβλητής X γίνει 0, τότε η τιμή της μεταβλητής Y είναι ίση με $\hat{\beta}_0$)

Με όρους της άσκησης : Αν το ύψος γίνει ίσο με μηδέν, τότε το βάρος θα γίνει ίσο με -156,9 kg

(προφανώς τέτοια περίπτωση δεν είναι ρεαλιστική, αλλά στο στάδιο της ερμηνείας των εκτιμητριών τέτοια μη ρεαλιστικά συμπεράσματα ενδέχεται να προκύπτουν)

ΓΡΑΜΜΙΚΗ ΣΥΣΧΕΤΙΣΗ

Στα παρακάτω διαγράμματα διασποράς παρουσιάζονται δυο διαφορετικά σμήνη σημείων, τα οποία, όμως, καταλήγουν στην ίδια ακριβώς ευθεία παλινδρόμησης $(y = 2{,}58 + 0{,}47 \cdot x)$



Είναι προφανές ότι τα σημεία του πρώτου σμήνους είναι πολύ πιο συγκεντρωμένα γύρω από την ευθεία, σε αντίθεση με το δεύτερο του οποίου τα σημεία είναι πιο διασκορπισμένα γύρω από την ίδια ευθεία.

Στην πρώτη περίπτωση η **γραμμική σχέση** των δυο μεταβλητών είναι πιο ισχυρή από την δεύτερη (δηλαδή το κατά πόσο τα σημεία «συγκλίνουν» σε μια ευθεία)

Το μέτρο που μας δίνει το βαθμό συγκέντρωσης των σημείων του διαγράμματος διασποράς γύρω από την ευθεία παλινδρόμησης, είναι ο συντελεστής γραμμικής συσχέτισης (ή συντελεστής συσχέτισης) r .

Στο τυπολόγιο, και πάλι, έχουμε δυο διαθέσιμους τύπους, έναν με απόκλιση από τους μέσους και έναν χωρίς απόκλιση από τους μέσους.

$$r = \frac{\sum[(Y_i - \overline{Y}) \cdot (X_i - \overline{X})]}{\sqrt{\sum(Y_i - \overline{Y})^2 \cdot \sum(X_i - \overline{X})^2}} = \frac{\sum X_i Y_i - \frac{\sum X_i \cdot \sum Y_i}{n}}{\sqrt{\left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right] \cdot \left[\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}\right]}}$$

με $r \in [-1,1]$

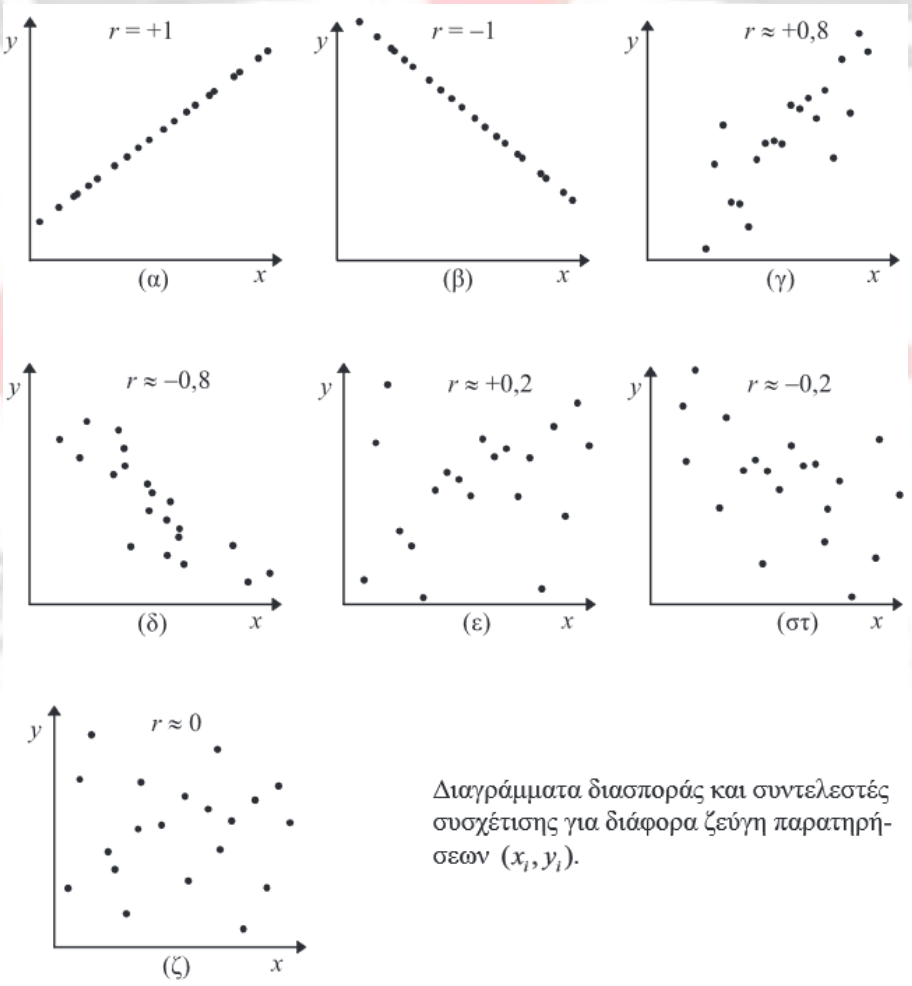
Αν $\mathbf{0 < r < 1}$ τότε έχουμε **θετική γραμμική συσχέτιση** (η ευθεία παλινδρόμησης είναι αύξουσα) (σχήμα 9 και ε)

Αν $\mathbf{-1 < r < 0}$ τότε έχουμε **αρνητική γραμμική συσχέτιση** (η ευθεία παλινδρόμησης είναι φθίνουσα) (σχήμα δ και στ)

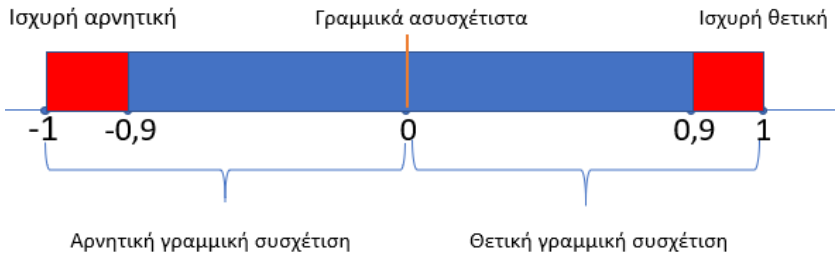
Αν $\mathbf{r = +1}$ τότε έχουμε **τέλεια θετική γραμμική συσχέτιση** (όλα τα σημεία του σμήνους βρίσκονται πάνω σε μια αύξουσα ευθεία) (σχήμα α)

Αν $\mathbf{r = -1}$ τότε έχουμε **τέλεια αρνητική γραμμική συσχέτιση** (όλα τα σημεία του σμήνους βρίσκονται πάνω σε μια φθίνουσα ευθεία) (σχήμα β)

Αν $\mathbf{r = 0}$ τότε **δεν υπάρχει γραμμική συσχέτιση**. Οι μεταβλητές Χ και Υ είναι γραμμικά ασυσχέτιστες (σχήμα ζ). Αυτό δε σημαίνει, όμως, ότι δεν συσχετίζονται με άλλο τρόπο. Είναι δυνατό να έχουν άλλη μορφή εξάρτησης (όχι, όμως, σε ευθεία) .



Διαγράμματα διασποράς και συντελεστές συσχέτισης για διάφορα ζεύγη παρατηρήσεων (x_i, y_i).



Αν δίνονται οι τυπικές αποκλίσεις S_X και S_Y των X και Y, τότε ο συντελεστής συσχέτισης είναι ίσος με

$$\mathbf{r = \frac{S_X}{S_Y} \cdot \beta_1}$$

Τέλος, πρέπει να σημειωθεί ότι ο συντελεστής συσχέτισης r είναι πάντα ομόσημος του συντελεστή β_1 (του συντελεστή της μεταβλητής X)

SITE: www.Eclass4U.gr
T. 210-5711484 ΚΙΝ. 6970401981
E-MAIL:grammateia.eclass4u@gmail.com

ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ

Ο Συντελεστής Προσδιορισμού R^2 εκφράζει το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής Y που ερμηνεύεται από την ανεξάρτητη μεταβλητή X και παίρνει τιμές μεταξύ 0 και 1 $(0 \le R^2 \le 1)$

Ισχύει $R^2 = r^2$ συνεπώς αν έχει προηγηθεί ο υπολογισμός του συντελεστή συσχέτισης r, απλώς τον υψώνουμε στο τετράγωνο για να βρεθεί ο συντελεστής προσδιορισμού R^2, αντί να αντικαταστήσουμε σε κάποιον από τους τύπους που δίνονται.

Οι τύποι είναι και πάλι με απόκλιση από τους μέσους και χωρίς απόκλιση από τους μέσους :

$$R^2 = \frac{(\sum[(Y_i - \overline{Y}) \cdot (X_i - \overline{X})])^2}{\sum(Y_i - \overline{Y})^2 \cdot \sum(X_i - \overline{X})^2} = \frac{\left(\sum X_i Y_i - \frac{\sum X_i \cdot \sum Y_i}{n}\right)^2}{\left[\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right] \cdot \left[\sum Y_i^2 - \frac{(\sum Y_i)^2}{n}\right]}$$

Όσο μεγαλύτερη είναι η τιμή του R^2, τόσο καλύτερη είναι η προσαρμογή των δεδομένων στο γραμμικό μοντέλο ή αλλιώς τόσο καλύτερα το γραμμικό μοντέλο εκφράζει τα δεδομένα.

Παράδειγμα :

Έστω οι μεταβλητές X : ετήσιο οικογενειακό εισόδημα και Y : ετήσιες καταναλωτικές δαπάνες. Αν υπολογιστεί $R^2 = 0{,}95$ τότε η ερμηνεία του συντελεστή προσδιορισμού είναι «το 95% της μεταβλητότητας των ετήσιων οικογενειακών δαπανών ερμηνεύεται από το ετήσιο οικογενειακό εισόδημα»

SITE: www.Eclass4U.gr
T. 210-5711484 ΚΙΝ. 6970401981
E-MAIL:grammateia.eclass4u@gmail.com