Γραμμική Συσχέτιση και Γραμμική Παλινδρόμηση

Ένας βουλευτής σκέφτηκε ότι η δημοτικότητά του έχει σχέση με τον αριθμό των τηλεοπτικών του εμφανίσεων. Αποφάσισε λοιπόν να συγκεντρώσει δεδομένα για να διαπιστώσει την αλήθεια ή όχι του συλλογισμού του αλλά και να μετρήσει το μέγεθος της σχέσης αυτής. Μετά από 18 μήνες κατάφερε να συγκεντρώσει τα παρακάτω στοιχεία:

Δημοτικότητα	Τηλεοπτικές εμφανίσεις
36,7	12
37,2	15
36,4	14
38,4	18
37,9	17
37,3	14
39,2	19
40,2	18
39,9	16
40,6	20
40,2	21
41,5	23
41,7	22
41,2	20
42,3	23
42,1	25
42,5	24
42,3	26

Φτάνοντας στην προεκλογική περίοδο ο βουλευτής πιστεύει ότι ήρθε η ώρα να αξιολογήσει τα στοιχεία του έτσι ώστε να διαμορφώσει την στρατηγική του όσον αφορά τις τηλεοπτικές του εμφανίσεις. Αν με Y συμβολίσουμε την μεταβλητή της δημοτικότητας και X την μεταβλητή των τηλεοπτικών εμφανίσεων, τότε εξετάζοντας μεμονωμένα την κάθε μία από τις παραπάνω μεταβλητές μπορούμε να δώσουμε τα παρακάτω στοιχεία:

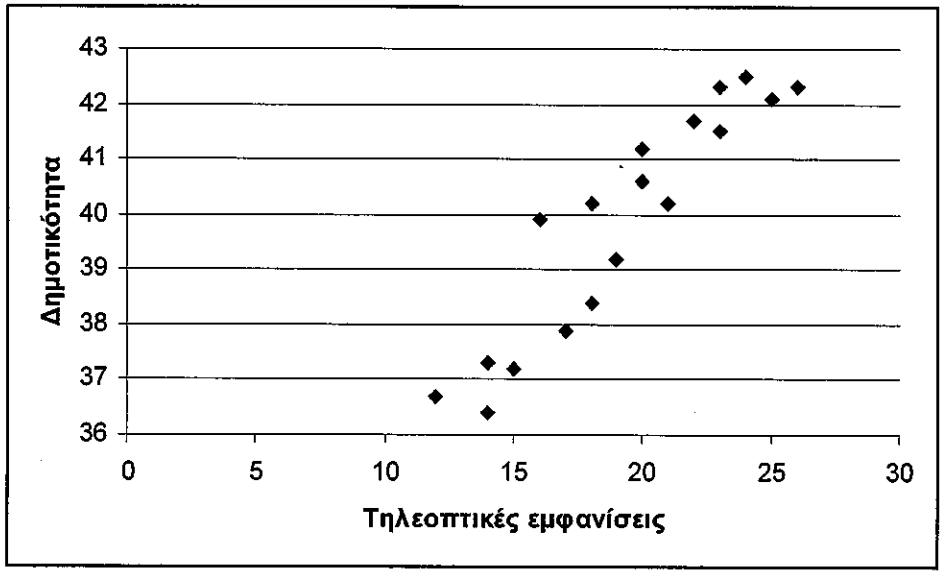
α/α	Y	X	$\left(Y_i-\overline{Y}\right)^2$	$\left(X_i-\overline{X}\right)^2$
1	36,7	12	10,03	52,97
2	37,2	15	7,11	18,30
3	36,4	14	12,02	27,85
4	38,4	18	2,15	1,63
5	37,9	17	3,87	5,19
6	37,3	14	6,59	27,85
7	39,2	19	0,44	0,08
8	40,2	18	0,11	1,63
9	39,9	16	0,00	10,74
10	40,6	20	0,54	0,52
11	40,2	21	0,11	2,97

12	41,5	23	2,67	13,85
13	41,7	22	3,36	7,41
14	41,2	20	1,78	0,52
15	42,3	23	5,92	13,85
16	42,1	25	4,99	32,74
17	42,5	24	6,93	22,30
18	42,3	26	5,92	45,19
Σύνολο	717,60	347,00	74,54	285,61

Από τον παραπάνω πίνακα προκύπτει:

$$\overline{Y} = 39.87, \quad \overline{X} = 19.28, \quad s_Y^2 = 4.38, (s_Y = 2.094), \quad s_X^2 = 16.8,, (s_X = 4.099),$$

Τα παραπάνω στατιστικά μέτρα δεν μπορούν να μας δώσουν καμία πληροφορία για την ύπαρξη ή όχι κάποιας σχέσης ανάμεσα στις δύο μεταβλητές. Μία πρώτη προσέγγιση προς αυτήν την κατεύθυνση θα ήταν μέσω ενός διαγράμματος διασποράς.



Από την παραπάνω γραφική παράσταση φαίνεται ότι σε μεγάλες τιμές της μεταβλητής «τηλεοπτικές εμφανίσεις» αντιστοιχούν μεγάλες τιμές της μεταβλητής «δημοτικότητα» και σε μικρές τιμές της μιας μεταβλητής αντιστοιχούν μικρές τιμές της άλλης μεταβλητής.

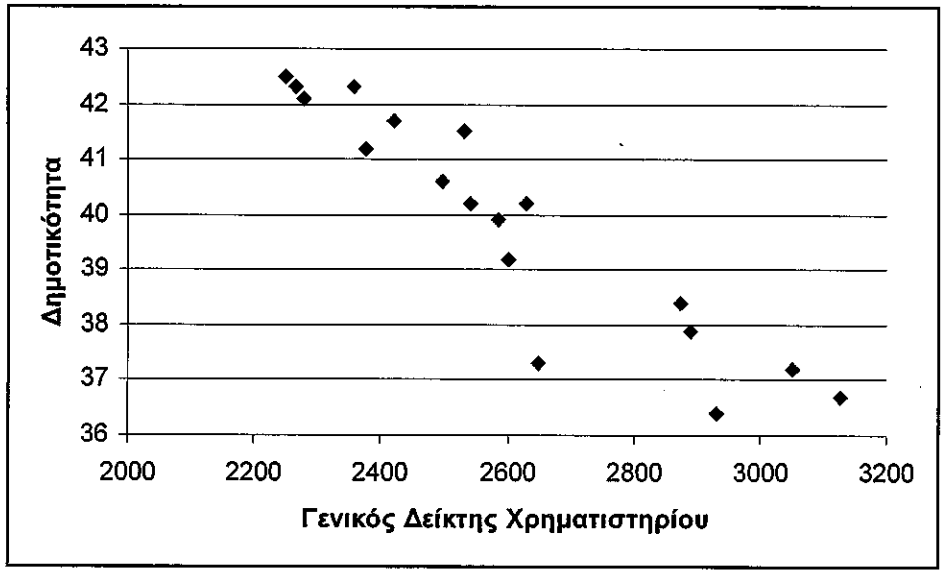
Στο ίδιο διάστημα οι τιμές του γενικού δείκτη του Χρηματιστηρίου διαμορφώθηκαν ως εξής:

Γενικός Δείκτης Χρηματιστηρίου
3128
3052
2932
2875
2890
2649
2602

2630
2585
2497
2540
2532
2420
2376
2357
2280
2250
2267

Η μεταβλητή αυτή έχει μέση τιμή 2603,44 και τυπική απόκλιση 271,445.

Μήπως υπάρχει κάποια σχέση ανάμεσα στη δημοτικότητα του βουλευτή και στις τιμές του γενικού δείκτη χρηματιστηρίου; Μία πρώτη προσέγγιση προς αυτήν την κατεύθυνση θα ήταν μέσω του παρακάτω διαγράμματος διασποράς.



Από την παραπάνω γραφική παράσταση φαίνεται ότι σε μεγάλες τιμές της μιας μεταβλητής αντιστοιχούν μικρές τιμές της άλλης μεταβλητής και το αντίστροφο.

Μάλιστα και από τις δύο γραφικές παραστάσεις φαίνεται να υπάρχει μία γραμμική σχέση ανάμεσα στις δύο μεταβλητές. Στην πρώτη περίπτωση όπου οι μεταβλητές μεταβάλλονται προς την ίδια κατεύθυνση θα λέμε ότι έχουμε θετική γραμμική συσχέτιση και στην δεύτερη περίπτωση θα λέμε ότι έχουμε αρνητική γραμμική συσχέτιση ανάμεσα στις δύο μεταβλητές.

Χρειαζόμαστε λοιπόν ένα μέτρο που να μας βοηθά να διαπιστώσουμε εάν έχουμε θετική ή αρνητική γραμμική συσχέτιση ανάμεσα σε δύο μεταβλητές. Ένα τέτοιο μέτρο θα μπορούσε να είναι η ποσότητα:

$$s_{XY} = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{n-1}$$

η οποία θα είναι θετική αν οι μεταβλητές μεταβάλλονται προς την ίδια κατεύθυνση και αρνητική στην αντίθετη περίπτωση. Το παραπάνω μέτρο το ονομάζουμε συνδιακύμανση. Η συνδιακύμανση προφανώς εξαρτάται από τις τάξεις μεγέθους των δύο μεταβλητών. Προκειμένου να έχουμε ένα μέτρο που θα είναι ανεξάρτητο από τις μονάδες μέτρησης των δύο μεταβλητών διαιρούμε με το γινόμενο των τυπικών αποκλίσεων των δύο μεταβλητών. Έτσι προκύπτει ο συντελεστής γραμμικής συσχέτισης

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Ο συντελεστής αυτός παίρνει τιμές από –1 μέχρι +1. Τιμές κοντά στο –1 σημαίνουν την ύπαρξη σημαντικής αρνητικής γραμμικής συσχέτισης, ενώ τιμές κοντά στο +1 δηλώνουν την ύπαρξη σημαντικής θετικής γραμμικής συσχέτισης. Τέλος αν ο συντελεστής γραμμικής συσχέτισης δύο μεταβλητών έχει τιμή κοντά στο 0, τότε αυτές οι μεταβλητές δεν έχουν σημαντική **γραμμική** συσχέτιση (αλλά θα μπορούσαν να έχουν μία σχέση άλλης μορφής).

Για το πρώτο ζευγάρι μεταβλητών ο συντελεστής συσχέτισης υπολογίζεται ως εξής:

$$r_{XY} = \frac{8.004}{2.094 * 4.099} = 0.93$$

Για το πρώτο ζευγάρι μεταβλητών ο συντελεστής συσχέτισης υπολογίζεται ως εξής:

$$r_{XY} = \frac{-528.984}{2.094 * 271.445} = -0.93$$

Σημειώνουμε ότι η συσχέτιση δεν σημαίνει κατά ανάγκη και αιτιότητα. Το γεγονός ότι υπάρχει μία σημαντική γραμμική συσχέτιση ανάμεσα στη δημοτικότητα του βουλευτή και στο γενικό δείκτη του χρηματιστηρίου δεν σημαίνει κατά ανάγκη ότι η δημοτικότητα του βουλευτή οφείλεται στις τιμές του χρηματιστηρίου, ούτε φυσικά και το αντίστροφο. Η στατιστική μπορεί να μας απαντήσει μόνο στο ερώτημα της ύπαρξης συσχέτισης ή όχι. Από εκεί και πέρα μόνο η κοινή λογική μπορεί να μας βοηθήσει να αποφασίσουμε αν αυτή η συσχέτιση αναδεικνύει και αιτιότητα. Θα μπορούσε μία συσχέτιση να οφείλεται στην τύχη ή σε μία τρίτη μεταβλητή της οποίας η μεταβολές επηρεάζουν και τις δύο μεταβλητές που μελετούμε. (το επιστημονικό συμπέρασμα του βάτραχου, ηλικία – ρυτίδες – φυσική κατάσταση)

Στις περιπτώσεις που έχει νόημα να θεωρήσουμε τη μία μεταβλητή ως αιτία και την άλλη ως αποτέλεσμα μπορούμε να προχωρήσουμε στην κατασκευή ενός μοντέλου που θα μας δώσει την δυνατότητα να προβλέπουμε τις τιμές της μίας μεταβλητής για συγκεκριμένες τιμές την άλλης μεταβλητής. Η μεταβλητή που θέλουμε να προβλέψουμε θα λέγεται εξαρτημένη και θα συμβολίζεται με Y και η μεταβλητή που αποτελεί την αιτία θα λέγεται ανεξάρτητη και θα συμβολίζεται με X.

Αν προσπαθήσουμε να προσαρμόσουμε μία ευθεία στα δεδομένα του πρώτου ζεύγους μεταβλητών αυτή θα ήταν της μορφής $Y=\beta_0+\beta_1 X$. Έστω ότι προσαρμόσαμε μία τέτοια ευθεία και ότι για κάθε τιμή X_i της εξαρτημένης μεταβλητής μπορούμε να υπολογίσουμε μία εκτίμηση $\hat{Y}_i$ για την εξαρτημένη μεταβλητή. Προφανώς μας ενδιαφέρει οι διαφορές $\hat{Y}_i$-Y να είναι οι μικρότερες δυνατές. Θα πρέπει λοιπόν να υπολογίσουμε τα β_0 και β_1 έτσι ώστε οι διαφορές $\hat{Y}_i$-Y να είναι οι ελάχιστες. Με τη

βοήθεια της μεθόδου των ελαχίστων τετραγώνων αποδεικνύεται ότι η ποσότητα $\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2$ είναι ελάχιστη για τις παρακάτω τιμές των β_0 και β_1:

$$\beta_1 = \frac{S_{XY}}{S_X^2}, \quad \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Μετά την κατασκευή του μοντέλου σημαντικό είναι να υπολογίσουμε την ποσότητα r_{XY}^2 η οποία ονομάζεται συντελεστής προσδιορισμού και αντιστοιχεί στο ποσοστό της μεταβλητότητας που οφείλεται στο μοντέλο.

Για παράδειγμα για το δεύτερο ζευγάρι των μεταβλητών έχουμε β_1= -0,00718 και β_0=58,56. Άρα το μοντέλο $Y=\beta_0+\beta_1 X$ γίνεται Y=58,56-0,00718 X και για τιμή του X=2000 το Y που θα έχουμε από το μοντέλο είναι 44,20. Αν πάρουμε για τιμή του X=10000 το Y που θα έχουμε από το μοντέλο είναι -13,24. Η μη αποδεκτή τιμή που μας έδωσε το μοντέλο οφείλεται στο γεγονός ότι επιλέξαμε για τιμή του X μία τιμή η οποία απέχει πολύ από τις τιμές του X βάσει των οποίων κατασκευάσαμε το μοντέλο.