



Εισαγωγή

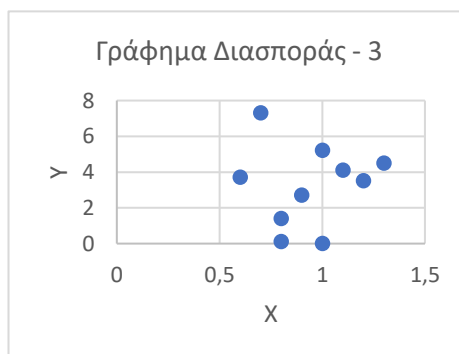
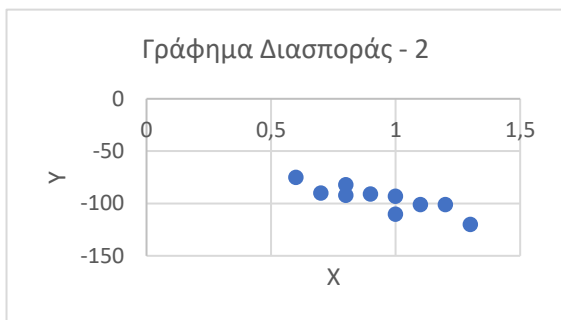
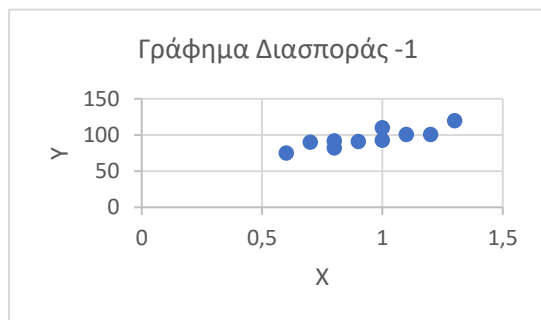
Σε πολλές περιπτώσεις θέλουμε να εξετάσουμε αν μπορούμε να προσαρμόσουμε ανάμεσα στα δεδομένα μια ευθεία γραμμή (γραμμικό μοντέλο). Στη συνέχεια να χρησιμοποιήσουμε αυτό το μοντέλο για την ανάλυση των δεδομένων, τον υπολογισμό ενδιάμεσων τιμών καθώς και την πρόβλεψη τιμών.

Με την συσχέτιση (correlation) μελετάμε εάν υπάρχει σχέση μεταξύ μιας τυχαίας μεταβλητής (**εξαρτημένη μεταβλητή**) και μιας ή περισσότερων άλλων τυχαίων μεταβλητών (**ανεξάρτητες μεταβλητές**). Με την παλινδρόμηση (regression) προσπαθούμε να προβλέψουμε την συμπεριφορά της εξαρτημένης μεταβλητής από τη σχέση της με την ανεξάρτητη ή ανεξάρτητες μεταβλητές.

Όταν η πρόβλεψη για την εξαρτημένη τυχαία μεταβλητή βασίζεται σε μια μόνο τυχαία ανεξάρτητη μεταβλητή τότε μιλάμε για την **απλή παλινδρόμηση**. Στην περίπτωση όπου η πρόβλεψη για την εξαρτημένη τυχαία μεταβλητή βασίζεται σε περισσότερες από μια τυχαίες ανεξάρτητες μεταβλητές τότε έχουμε την **πολλαπλή παλινδρόμηση**.

Γράφημα διασποράς

Στα παρακάτω Γράφηματα Διασποράς 1 και 2 διακρίνουμε μια γραμμική συσχέτιση μεταξύ των τυχαίων μεταβλητών X (ανεξάρτητη) και Y (εξαρτημένη). Αντίθετα, στο Γράφημα Διασποράς 3 φαίνεται ότι δεν υπάρχει καμιά γραμμική συσχέτιση μεταξύ των τυχαίων μεταβλητών X και Y .



Συντελεστής συσχέτισης

Ο βαθμός γραμμικής συσχέτισης δυο τυχαίων μεταβλητών X και Y για τον πληθυσμό συμβολίζεται με το γράμμα ρ . Ως ένας λογικός εκτιμητής για το ρ είναι ο δειγματικός συντελεστής γραμμικής συσχέτισης του Pearson

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

Ισχύει $-1 \leq r \leq 1$. Αν $r = 0$, τότε δεν υπάρχει γραμμική σχέση μεταξύ X και Y . Όσο το $|r|$ πλησιάζει στο 1 τόσο πιο ισχυρή είναι η γραμμική σχέση μεταξύ X και Y . Αν $r = 1$, τότε όλα τα σημεία (x_i, y_i) βρίσκονται πάνω σε μια ευθεία γραμμή με θετική κλίση (Γράφημα Διασποράς 1). Ενώ όταν $r = -1$, τότε όλα τα σημεία (x_i, y_i) βρίσκονται πάνω σε μια ευθεία γραμμή με αρνητική κλίση (Γράφημα Διασποράς 2).

Το Excel διαθέτει τη στατιστική συνάρτηση CORREL για τον υπολογισμό του συντελεστή συσχέτισης.

Γραμμική απλή παλινδρόμηση

Το ζητούμενο είναι να κατασκευάσουμε τη γραμμική ευθεία παλινδρόμησης

$$\hat{y} = \hat{a} + \hat{b}x$$

η οποία προσεγγίζει τα σημεία (x_i, y_i) έτσι ώστε το άθροισμα των τετραγώνων των σφαλμάτων

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

να είναι το ελάχιστο δυνατόν. Αυτή η μέθοδος είναι γνωστή ως η Μέθοδος των Ελαχίστων Τετραγώνων.

Τότε για κάθε x μπορούμε να προβλέψουμε την τιμή y με την αντίστοιχη τιμή $\hat{y}$ με κάποια πιθανότητα σφάλματος.

Στο Excel χρησιμοποιούμε το εργαλείο «Παλινδρόμηση» (Regression) της εργαλειοθήκης «Ανάλυση Δεδομένων» (Data Analysis) για πλήρη ανάλυση γραμμικής παλινδρόμησης.