Γραμμική παλινδρόμηση

Απλή γραμμική παλινδρόμηση

Γενικά

Μία γραμμική σχέση μεταξύ δύο μεταβλητών Y και X μπορεί να περιγραφεί από την εξίσωση:

$$Y = \beta_0 + \beta_1 X$$

Για παράδειγμα, το κόστος εγγραφής σε ένα πρόγραμμα σταθερής τηλεφωνίας μπορεί να περιλαμβάνει ένα αρχικό κόστος σύνδεσης 50 ευρώ και 30 ευρώ πάγιο κάθε μήνα. Τότε το συνολικό κόστος Y για X μήνες εγγραφής θα δίνεται από τη σχέση:

$$Y = 50 + 30X$$

Καθώς, πολύ συχνά μία ευθεία γραμμή μπορεί να είναι βολική για να περιγράψει τη σχέση δύο μεταβλητών, η **απλή γραμμική παλινδρόμηση** έχει αναπτυχθεί ως η στατιστική μέθοδος που καθορίζει την καλύτερη δυνατή ευθεία γραμμή που περιγράφει τη σύγχρονη εξέλιξη των δύο μεταβλητών. Τι εννοούμε, όμως, όταν μιλάμε για την «καλύτερη δυνατή γραμμή»;

Υπολογισμός της ευθείας

Ας υποθέσουμε ότι θέλουμε να μελετήσουμε πώς η μεταβλητή Y (π.χ. το βάρος των μαθητών του νηπιαγωγείου) εξελίσσεται ως συνάρτηση της μεταβλητής X (π.χ. το ύψος των μαθητών του νηπιαγωγείου). Για το σκοπό αυτό έχουμε συλλέξει ένα δείγμα από $n = 100$ μαθητές και έχουμε μετρήσει τα ύψη (X) και τα βάρη τους (Y). Θα συμβολίζουμε με κεφαλαία X και Y τις μεταβλητές ως έννοιες (π.χ. X: είναι η έννοια του ύψους, Y: είναι η έννοια του βάρους) και με πεζά x και y τις μετρήσεις τους στο δείγμα. Άρα για το δείγμα μεγέθους $n = 100$ συμβολίζουμε τις 100 μετρήσεις του ύψους X με $x_1, x_2, ..., x_{100}$ και τις 100 μετρήσεις του βάρους Y με $y_1, y_2, ..., y_{100}$.

Η λογική λέει ότι θέλουμε να διαλέξουμε τη γραμμή εκείνη (δηλαδή τα κατάλληλα β_0 και β_1), ώστε το σύνολο των αποστάσεων των πραγματικών μετρήσεων των X και Y από αυτή να είναι η ελάχιστη (Figure 1). Αν, λοιπόν, συμβολίσουμε με $\hat{y}$ τις προβλέψεις που θα μας έδινε η εξίσωση μίας γραμμής $\hat{\beta}_0 + \hat{\beta}_1 x$, γραμμική παλινδρόμηση θα ελαχιστοποιήσει τις αποστάσεις $(y - \hat{y})$. Τις αποστάσεις αυτές συνήθως τις ονομάζουμε *σφάλματα*, αφού μας δείχνουν πόσο μακριά βρίσκεται η αλήθεια (y) από την πρόβλεψή μας με βάση τη γραμμή ($\hat{y}$). Στο συγκεκριμένο παράδειγμα, θα χρειαστεί να βρούμε τα $\hat{\beta}_0$ και $\hat{\beta}_1$ τα οποία θα ελαχιστοποιήσουν το άθροισμα των 100 δυνατών σφαλμάτων $(y_1 - \hat{y}), (y_2 - \hat{y}), ..., y_{100} - \hat{y}$. Για συντομία, όταν θέλουμε να αναφερθούμε σε όλα τα σφάλματα, όπως κάναμε πριν, χρησιμοποιούμε $(y_i - \hat{y}), i = 1, ..., 100$.

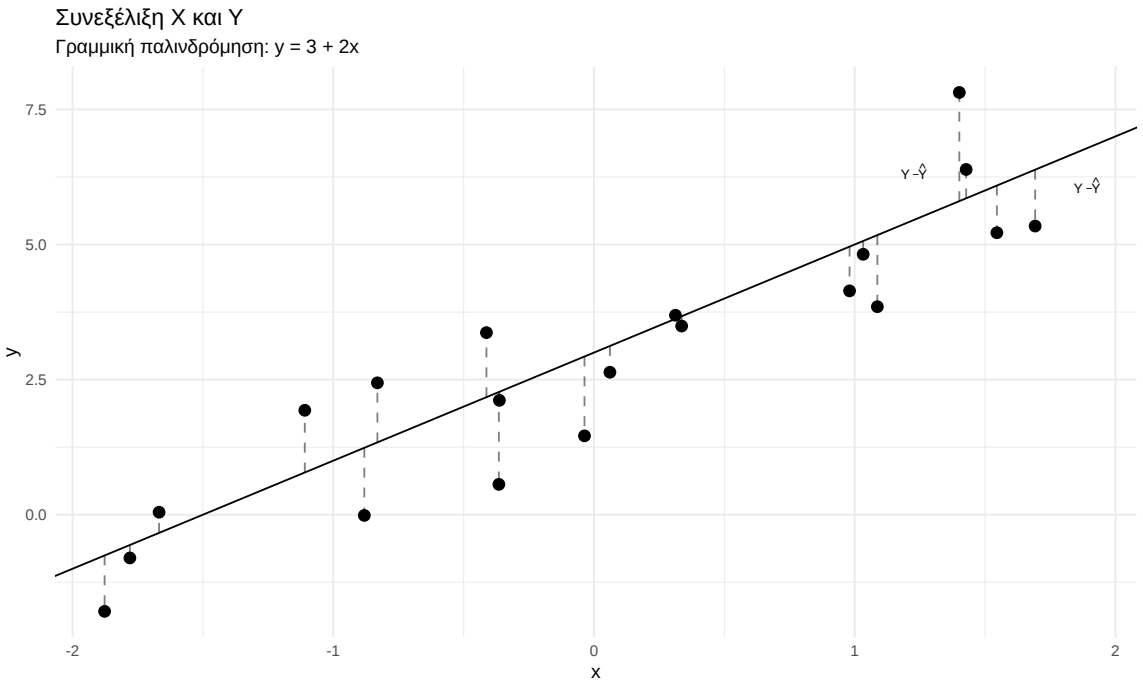

Figure 1

Καθώς τα σφάλματα $y_i - \hat{y}$ μπορούν να είναι είτε θετικά είτε αρνητικά, αναλόγως αν η γραμμή περνά πάνω ή κάτω από το πραγματικό σημείο (x_i, y_i), το άθροισμα αυτό δε θα μας δώσει το σύνολο των αποστάσεων. Οι αποστάσεις ως μήκος χρειάζεται να είναι πάντα θετικές. Για το λόγο αυτό, είναι σύνηθες να υψώνουμε στο τετράγωνο τα σφάλματα ώστε να διασφαλίσουμε ότι θα δουλεύουμε με θετικούς αριθμούς. Επομένως, αναζητούμε εκείνα τα $\hat{\beta}_0$ και $\hat{\beta}_1$ τα οποία θα ελαχιστοποιήσουν το άθροισμα των τετραγώνων των σφαλμάτων $(y - \hat{y})^2$. Επειδή το άθροισμα αυτό θα το συναντάμε συχνά, θα το συμβολίζουμε για συντομία με SS_{RES}, από την αγγλική ορολογία *sum of squares residuals* (residual: υπόλοιπο – σφάλμα)

$$SS_{RES} = (y_1 - \hat{y})^2 + (y_2 - \hat{y})^2 + ... + (y_n - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Με βάση τα παραπάνω μπορεί να αποδειχθεί ότι τα $\hat{\beta}_0$ και $\hat{\beta}_1$ που ζητάμε μπορούν να υπολογιστούν από τους τύπους:

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

και

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

Στις παραπάνω εξισώσεις, τα $\bar{x}$ και $\bar{y}$ συμβολίζουν τη μέση τιμή (μέσος όρος) των μετρήσεων της X (ύψος) και της Y (βάρος) στο δείγμα των 100 μαθητών, αντίστοιχα.

Για παράδειγμα, έστω ότι είχαμε συλλέξει ένα δείγμα 5 μαθητών με τις μετρήσεις του ύψους και του βάρους να δίνονται παρακάτω:

Table 1: Μετρήσεις ύψους και βάρους πέντε μαθητών ενός νηπιαγωγείου.

Ύψος (εκατοστά)	Βάρος (κιλά)
92	16
98	18
100	17
98	20
94	19

Για να υπολογίσουμε την εξίσωση της γραμμής παλινδρόμησης στο συγκεκριμένο παράδειγμα, πρέπει πρώτα από όλα να υπολογίσουμε τα $\bar{x}$ και $\bar{y}$. Αν γυρίσετε πίσω στις προηγούμενες εξισώσεις θα δείτε ότι οι ποσότητες αυτές χρησιμοποιούνται αρκετές φορές. Το $\bar{x}$ είναι ο μέσος όρος του ύψους των 5 μαθητών και υπολογίζεται ως

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{92 + 98 + 100 + 98 + 94}{5} = 96.4$$

Αντίστοιχα, ο μέσος όρος του βάρους των μαθητών είναι:

$$\bar{y} = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5} = \frac{16 + 18 + 17 + 20 + 19}{5} = 18$$

Τώρα μπορούμε να κατασκευάσουμε προοδευτικά τα πιο πολύπλοκα αθροίσματα:

$\mathbf{x}$	$\mathbf{y}$	$(x - \bar{x})$	$(x - \bar{x})^2$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
92	16	$92 - 96.4 = -4.4$	$(-4.4)^2 = 19.36$	$16 - 18 = -2$	$(-4.4) \times (-2) = 8.8$
98	18	$98 - 96.4 = 1.6$	$1.6^2 = 2.56$	$18 - 18 = 0$	$1.6 \times 0 = 0$
100	17	$100 - 96.4 = 3.6$	$3.6^2 = 12.96$	$17 - 18 = -1$	$3.6 \times (-1) = -3.6$
98	20	$98 - 96.4 = 1.6$	$1.6^2 = 2.56$	$20 - 18 = 2$	$1.6 \times 2 = 3.2$
94	19	$94 - 96.4 = -2.4$	$(-2.4)^2 = 5.76$	$19 - 18 = 1$	$(-2.4) \times 1 = -2.4$
Άθροισμα		$\sum = 0$	$\sum = 43.2$	$\sum = 0$	$\sum = 5.8$

Πίνακας 1: Υπολογισμός βοηθητικών ποσοτήτων

Τώρα μπορούμε να υπολογίσουμε τους συντελεστές της γραμμής παλινδρόμησης που μπορούμε με βάση αυτό το μικρό δείγμα των πέντγε παιδιών:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{5.8}{43.2} \approx 0.13$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 18 - 0.13 \times 96.4 \approx 5.47$$

Επομένως, η εξίσωση της γραμμής παλινδρόμησης είναι:

$$\text{βάρος} = 5.47 + 0.13 \times \text{ύψος}$$

Ανάλυση διασποράς