

Πιθανότητες & Στατιστική

Γραμμική παλινδρόμηση

Δημήτρης Μητσούδης

Αν η τιμή του συντελεστή συσχέτισης υποδεικνύει ότι οι δύο μεταβλητές συνδέονται με μια γραμμική σχέση, τότε είναι λογικό να προσπαθήσουμε να χαράξουμε μια **ευθεία στο διάγραμμα διασποράς** έτσι ώστε αυτή να διέρχεται όσο το δυνατόν «κοντύτερα» από όλα τα σημεία του γραφήματος.

Δηλαδή, μπορούμε να πούμε ότι η ευθεία αυτή περιγράφει «καλύτερα» τη γραμμική σχέση μεταξύ των δύο αυτών μεταβλητών και, επομένως, μπορούμε να την χρησιμοποιήσουμε για προβλέψεις και εκτιμήσεις.

Η ευθεία αυτή ονομάζεται **ευθεία παλινδρόμησης**.

Εισαγωγή

Στο προηγούμενο μάθημα είδαμε για δύο μεταβλητές X, Y , το συντελεστή γραμμικής συσχέτισης του Pearson

$$r = \frac{S_{XY}}{S_X \cdot S_Y},$$

όπου

$$S_{XY} = \text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y}),$$

και S_X, S_Y οι δειγματικές τυπικές αποκλίσεις των X, Y , αντίστοιχα.

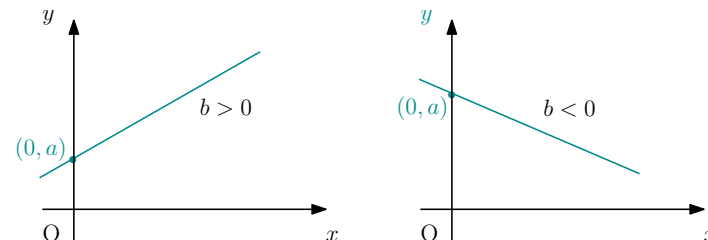
Όπως είδαμε, ο συντελεστής γραμμικής συσχέτισης r είναι ένα μέτρο της **γραμμικής** σχέσης μεταξύ των X και Y , και παίρνει τιμές στο διάστημα $[-1, 1]$.

Υπενθύμιση

Εξίσωση ευθείας: $y = a + bx$.

Η σταθερά b λέγεται **κλίση της ευθείας**, ενώ η σταθερά a είναι η τεταγμένη του σημείου τομής της ευθείας με τον x -άξονα.

Αν $b > 0$ η ευθεία είναι **αύξουσα**, ενώ αν $b < 0$ **φθίνουσα**.



Απλή Γραμμική Παλινδρόμηση

Το απλούστερο μοντέλο που θα μπορούσε να ερμηνεύσει μια γραμμική σχέση μεταξύ των X και Y είναι το απλό γραμμικό μοντέλο.

Συγκεκριμένα, θεωρούμε ότι οι μεταβλητές X (ανεξάρτητη) και Y (εξαρτημένη) συνδέονται με τη σχέση

$$Y = a + bX + \varepsilon.$$

Κάτω από ορισμένες υποθέσεις, όπως ότι τα σφάλματα ε έχουν μέση τιμή μηδέν, και ότι για τις διάφορες τιμές της X , οι αντίστοιχες μέσες τιμές της Y βρίσκονται πάνω σε μια ευθεία ($E(Y|X) = a + bX$), μπορούμε να υπολογίσουμε μια εκτίμηση

$$\hat{Y} = \hat{a} + \hat{b}X,$$

της πληθυσμιακής ευθείας παλινδρόμησης $E(Y|X) = a + bX$.

Ευθεία ελαχίστων τετραγώνων

Η εκτίμηση $\hat{Y} = \hat{a} + \hat{b}X$, της πληθυσμιακής ευθείας παλινδρόμησης ονομάζεται **ευθεία ελαχίστων τετραγώνων** λόγω της μεθόδου που χρησιμοποιούμε για τον υπολογισμό των παραμέτρων της.

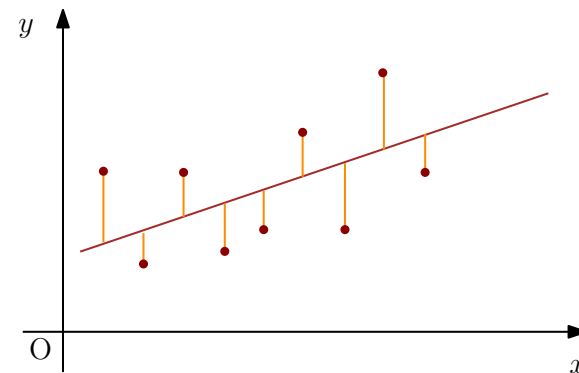
Συγκεκριμένα, αν έχουμε n ζεύγη παρατηρήσεων $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ αναζητούμε προσέγγιση της μορφής

$$y_i = a + bx_i + \varepsilon_i,$$

έτσι ώστε να ελαχιστοποιείται το άθροισμα των τετραγώνων των «υπολοίπων» ε_i . Δηλαδή να γίνεται ελάχιστο το

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

Ευθεία ελαχίστων τετραγώνων



Το άθροισμα των τετραγώνων των **κατακόρυφων** αποστάσεων των σημείων από την ευθεία είναι ελάχιστο.

Αποδεικνύεται ότι η ευθεία ελαχίστων τετραγώνων είναι:

$$\hat{Y} = \hat{a} + \hat{b}X,$$

με

$$\hat{b} = \frac{S_{XY}}{S_X^2}, \quad \hat{a} = \bar{Y} - \hat{b}\bar{X}.$$

Ισοδύναμα,

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n\bar{X} \cdot \bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{X}^2}, \quad \hat{a} = \bar{Y} - \hat{b}\bar{X}.$$

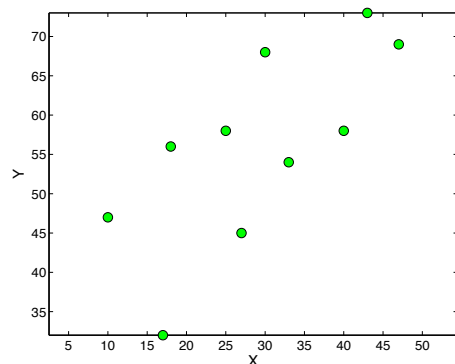
Παράδειγμα

Έστω ότι στον ακόλουθο πίνακα μας δίνονται για ένα δείγμα 10 ατόμων: ο χρόνος μελέτης (X) κάθε φοιτητή, και η επίδοση του καθενός στην τελική εξέταση (Y).

X	40	43	18	10	25	33	27	17	30	47
Y	58	73	56	47	58	54	45	32	68	69

Εδώ στην ουσία έχουμε υποθέσει ότι οι τιμές της X είναι σωστές και οι τιμές της Y δεν ανήκουν όλες σε μια ευθεία λόγω τυχαίων σφαλμάτων.

Το διάγραμμα διασποράς για τα δεδομένα μας



Παρατηρούμε ότι μικρές τιμές της μεταβλητής X σχετίζονται με μικρές τιμές της Y , ενώ μεγάλες τιμές της X σχετίζονται με μεγάλες τιμές της Y .

Κατασκευάζουμε τον πίνακα:

x_i	y_i	$x_i y_i$	x_i^2
40	58	2320	1600
43	73	3139	1849
18	56	1008	324
10	47	470	100
25	58	1450	625
33	54	1782	1089
27	45	1215	729
17	32	544	289
30	68	2040	900
47	69	3243	2209
290	560	17211	9714

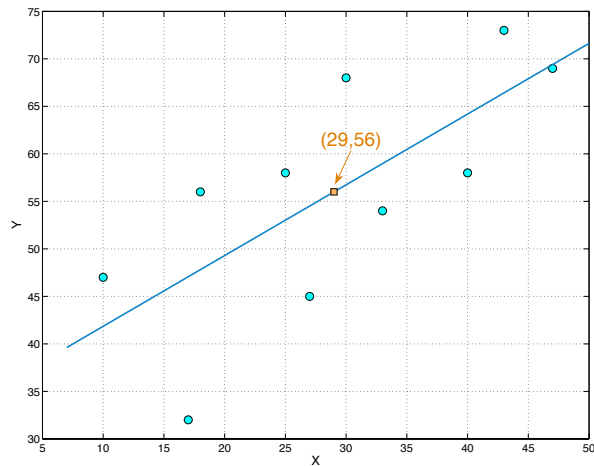
Επομένως:

$$\bar{X} = 29, \quad \bar{Y} = 56, \quad \sum_{i=1}^n x_i y_i = 17211, \quad \sum_{i=1}^n x_i^2 = 9714,$$

και

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n \bar{X} \cdot \bar{Y}}{\sum_{i=1}^n x_i^2 - n \bar{X}^2} = \frac{17211 - 10 \cdot 29 \cdot 56}{9714 - 10 \cdot 29^2} = \frac{971}{1304} = 0.7446$$
$$\hat{a} = \bar{Y} - \hat{b} \bar{X} = 56 - \frac{971}{1304} \cdot 29 = 34.4057$$

Συνεπώς, η ευθεία ελαχίστων τετραγώνων είναι: $\hat{Y} = 34.4057 + 0.7446X$



Μπορούμε να χρησιμοποιήσουμε την ευθεία ελαχίστων τετραγώνων

$$\hat{Y} = 34.4057 + 0.7446X$$

για να βρούμε εκτιμήσεις του $\hat{Y}$ (βαθμός στις εξετάσεις) για διάφορες τιμές του X (χρόνος προετοιμασίας).

Π.χ. κάποιος με χρόνο προετοιμασίας $X = 36$ αναμένεται να έχει μέση επίδοση

$$\hat{Y} = 34.4057 + 0.7446 \cdot 36 = 61.2113.$$

Προσδιορισμός «καλού» μοντέλου παλινδρόμησης

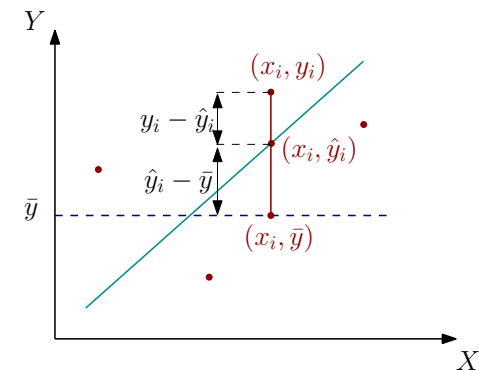
Στον ακόλουθο πίνακα με τα δεδομένα μας προσθέτουμε στην τελευταία γραμμή τις εκτιμήσεις $\hat{Y}$ που προκύπτουν από τη γραμμική παλινδρόμηση.

X	40	43	18	10	25	33	27	17	30	47
Y	58	73	56	47	58	54	45	32	68	69
$\hat{Y}$	64.19	66.42	47.81	41.85	53.02	58.98	54.51	47.06	56.74	64.40

Ερώτημα: Πόσο «καλά» εκτιμά η ευθεία ελαχίστων τετραγώνων $\hat{Y} = \hat{a} + \hat{b}X$ την (πληθυσμιακή) ευθεία παλινδρόμησης;

Προφανώς:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$



Αποδεικνύεται ότι:

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{SSE} + \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{SSR}$$

- $SST = \sum_{i=1}^n (y_i - \bar{y})^2$: ολικό άθροισμα τετραγώνων ή ολική μεταβλητότητα των y_i
- $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$: άθροισμα τετραγώνων των σφαλμάτων ή υπόλοιπο μεταβλητότητας
- $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$: άθροισμα τετραγώνων παλινδρόμησης

Στην πράξη: $0 \leq r^2 \leq 1$.

Όσο πιο κοντά στο 1 βρίσκεται το r^2 , τόσο καλύτερη είναι η ευθεία ελαχίστων τετραγώνων ως εκτίμηση της ευθείας παλινδρόμησης.

Συντελεστής προσδιορισμού

Το SST είναι ένα μέτρο της συνολικής μεταβλητότητας των παρατηρήσεων y_i . Το SSR εκφράζει το μέρος της μεταβλητότητας που μπορεί να οφείλεται στο X , ενώ το $SSE = SST - SSR$ εκφράζει την υπόλοιπη μεταβλητότητα που οφείλεται στην τυχαιότητα και δεν εξηγείται από την παλινδρόμηση.

Αν r είναι ο συντελεστής γραμμικής συσχέτισης, τότε αποδεικνύεται ότι

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Ο r^2 ονομάζεται **συντελεστής προσδιορισμού** και εκφράζει το ποσοστό της συνολικής μεταβλητότητας των y_i που μπορεί να ερμηνευθεί από την παλινδρόμηση.

Τυπικό σφάλμα της εκτίμησης

Ονομάζουμε **τυπικό σφάλμα της εκτίμησης (s)** τη μέση απόκλιση μεταξύ της πραγματικής και της εκτιμώμενης τιμής της μεταβλητής και την υπολογίζουμε ως:

$$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \sqrt{\frac{SSE}{n-2}}.$$

Εάν το τυπικό σφάλμα της εκτίμησης είναι μικρό ($\Rightarrow$ **οι παρατηρούμενες και οι εκτιμώμενες τιμές δεν διαφέρουν πολύ**), τότε η ευθεία παλινδρόμησης μας δίνει μια καλή περιγραφή της σχέσης μεταξύ των X και Y .

Σε αντίθετη περίπτωση, δεν μπορούμε να ισχυρισθούμε ότι έχουμε καλή περιγραφή της σχέσης.

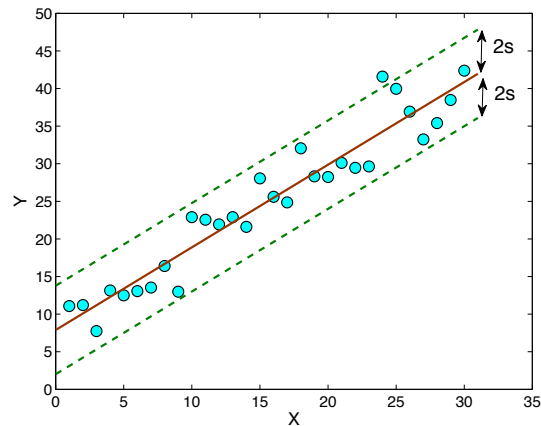
Παράδειγμα*

Το τυπικό σφάλμα της εκτίμησης, είναι ένα μέτρο της διασποράς των (x, y) γύρω από την ευθεία ελαχίστων τετραγώνων, και έχει ιδιότητες ανάλογες με αυτές της τυπικής απόκλισης.

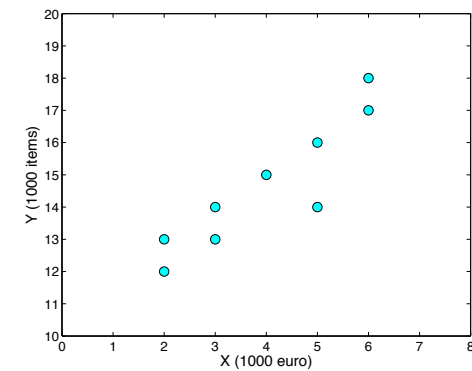
Έτσι, αν φέρουμε δύο ευθείες παράλληλες προς την ευθεία ελαχίστων τετραγώνων και σε κατακόρυφες προς αυτήν αποστάσεις $s, 2s, 3s$, τότε, για μεγάλα n ($n \geq 30$), μεταξύ των δύο αυτών ευθειών θα βρίσκεται περίπου το 68%, το 95% και το 99.7% των σημείων του διαγράμματος διασποράς, αντίστοιχα.

Ο πίνακας που ακολουθεί δίνει τη ζήτηση ενός προϊόντος (Y) σε χιλιάδες τεμάχια, για διάφορα επίπεδα διαφημιστικής δαπάνης (X) σε χιλιάδες €.

X	2	2	3	3	4	4	5	5	6	6
Y	12	13	13	14	15	15	14	16	17	18



Το διάγραμμα διασποράς για τα δεδομένα μας:

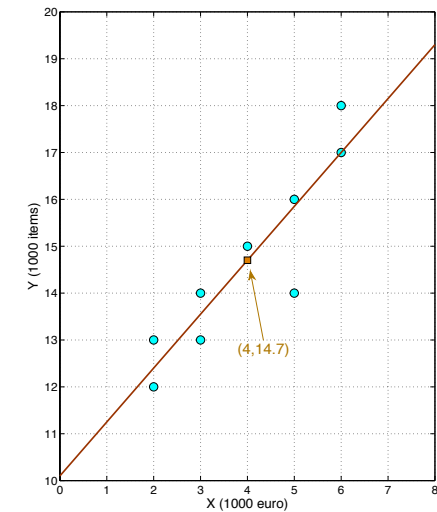


Κατασκευάζουμε τον πίνακα:

x_i	y_i	$x_i y_i$	x_i^2
2	12	24	4
2	13	26	4
3	13	39	9
3	14	42	9
4	15	60	16
4	15	60	16
5	14	70	25
5	16	80	25
6	17	102	36
6	18	108	36
40	147	611	180

Συνεπώς, η ευθεία ελαχίστων τετραγώνων είναι:

$$\hat{Y} = 10.1 + 1.15X$$



Άρα

$$\bar{X} = \frac{40}{10} = 4, \quad \bar{Y} = \frac{147}{10} = 14.7$$

και

$$\hat{b} = \frac{\sum_{i=1}^n x_i y_i - n \bar{X} \cdot \bar{Y}}{\sum_{i=1}^n x_i^2 - n \bar{X}^2} = \frac{611 - 10 \cdot 4 \cdot 14.7}{180 - 10 \cdot 4^2} = \frac{23}{20} = 1.15$$

$$\hat{a} = \bar{Y} - \hat{b} \bar{X} = 14.7 - 1.15 \cdot 4 = 10.1$$

- **Ερμηνεία του $\hat{b}$:** Επειδή $\hat{b} = 1.15 > 0$, συμπεραίνουμε ότι η αύξηση της διαφημιστικής δαπάνης $\Rightarrow$ αύξηση της ζήτησης του προϊόντος.
Συγκεκριμένα, αν η διαφημιστική δαπάνη αυξηθεί κατά 1000€, η μέση ζήτηση του προϊόντος **εκτιμάται** ότι θα αυξηθεί κατά 1150 τεμάχια.
- **Ερμηνεία του $\hat{a}$:** Για μηδενική διαφημιστική δαπάνη, η μέση ζήτηση του προϊόντος εκτιμάται ότι θα είναι 10100 τεμάχια. Επειδή η τιμή 0 είναι μακριά από το διάστημα μελέτης, η ερμηνεία του $\hat{a}$ δεν έχει ιδιαίτερη πρακτική αξία.

Για να υπολογίσουμε το συντελεστή προσδιορισμού r^2 κατασκευάζουμε τον πίνακα:

x_i	y_i	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
2	12	12.40	-2.30	5.2900	-2.7	7.29
2	13	12.40	-2.30	5.2900	-1.7	2.89
3	13	13.55	-1.15	1.3225	-1.7	2.89
3	14	13.55	-1.15	1.3225	-0.7	0.49
4	15	14.70	0.00	0.0000	0.3	0.09
4	15	14.70	0.00	0.0000	0.3	0.09
5	14	15.85	1.15	1.3225	-0.7	0.49
5	16	15.85	1.15	1.3225	1.3	1.69
6	17	17.00	2.30	5.2900	2.3	5.29
6	18	17.00	2.30	5.2900	3.3	10.89
40	147	147.00		26.4500		32.10

Το τυπικό σφάλμα της εκτίμησης (s) είναι ίσο με:

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

$$= \sqrt{\frac{SST - SSR}{n-2}} = \sqrt{\frac{32.10 - 26.45}{8}} = 0.8404.$$

Άρα:

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{26.45}{32.10} = 0.824.$$

Η τιμή αυτή ερμηνεύεται ως εξής: Οι μεταβολές του ύψους της διαφημιστικής δαπάνης ερμηνεύουν το 82% της μεταβλητότητας της ζήτησης του προϊόντος.

Το υπόλοιπο 18% οφείλεται σε τυχαία σφάλματα.

Τέλος, ας κάνουμε μια πρόβλεψη χρησιμοποιώντας το μοντέλο

$$\hat{Y} = 10.1 + 1.15X.$$

Π.χ., αν το ύψος της διαφημιστικής δαπάνης είναι 3500 €, τότε θα εκτιμήσουμε τη μέση ζήτηση του προϊόντος ως

$$10.1 + 1.15 \cdot 3.5 = 14.125 \text{ χιλιάδες τεμάχια,}$$

δηλ. 14125 τεμάχια.