

Παγινόρριση

Στο προηγούμενο κεφάλαιο προσπαθήσαμε να λύσουμε το εξής πρόβλημα:
(i) έχουμε $(n+1)$ σημεία $\{(x_k, y_k)\}_{k=0}^n$ (πειραματικές μετρήσεις/δεδομένα) ή προσπα-
θούμε να τα περιγράψουμε με τη βοήθεια μιας μαθηματικής σχέσης ή (ii) έχουμε
μια συνάρτηση $y=f(x)$ της οποίας ο αναλυτικός τύπος είναι "πολύπλοκος"
(δυσκολία υπολογισμού) ή δεν είναι γνωστός ο αναλυτικός της τύπος, θέλουμε δε
ν' αντιπροσωπεύσουμε την $f(x)$ με μια πιο "απλοϊκή" (στη (ii) θεωρούμε
ότι έχουμε τα σημεία $\{(x_k, f(x_k))\}_{k=0}^n$).

Μια αντιμετώπιση του παραπάνω προβλήματος ήταν με τη βοήθεια της
(θεωρίας) παρεμβολής: προσπαθήσαμε να βρούμε ένα πολυώνυμο n -του
βάθμου (Taylor, Lagrange, Newton) που να διέρχεται από τα σημεία
 $\{(x_k, y_k)\}_{k=0}^n$. Υπάρχουν δύο βασικά μειονεκτήματα στα πολυώνυμα παρεμβολής:
Εάν δίνονται $(n+1)$ -σημεία το πολυώνυμο παρεμβολής είναι n -του βάθμου.
Εάν λοιπόν ο αριθμός των σημείων είναι μεγάλος, τότε n -βάθμος του
πολυωνύμου παρεμβολής είναι μεγάλος, με αποτέλεσμα το πολυώνυμο πέρα
από τα σημεία παρεμβολής να παρουσιάζει ταλαντώσεις ή η παρεμβολή
(σε σημεία διαφορετικά από τους κόμβους) να μην είναι καλή. Ακόμα
τα (x_k, y_k) είναι αποτελέσματα μετρήσεων, επομένως περιέχουν σφάλματα.
Τα σφάλματα επιπλέον σημειώνονται τους συντελεστές των πολυωνύμων
παρεμβολής (ιδιαίτερα του μεγάλου βαθμού) με αποτέλεσμα να επιπηρεάζεται
η ποιότητα της παρεμβολής.
Ακόμα ή η αντιμετώπιση με τις συναρτήσεις σπλίνες, δεν είναι
ικανοποιητική εάν περίπλοκη πολλών σημείων (απαιτούνται πολλά
πολυώνυμα μικρού βαθμού).

Η αντιμετώπιση σε αυτό το κεφάλαιο είναι η εξής: προσπαθούμε
να βρούμε μια μαθηματική σχέση/νόμο - συνήθως πολυώνυμο - που να
προσεγγίζει "καλύτερα" (με κάποια έννοια) τα δεδομένα χωρίς να είναι
απαραίτητο η (μαθηματική) συνάρτηση (που εκφράζει τη σχέση αυτή)
να διέρχεται από τα σημεία (x_k, y_k) , $k=0, 1, 2, \dots, n$
ή n

Θέλουμε ακόμα, αν είναι δυνατόν, να βρούμε το είδος της σχέσης ανάμεσα στα δεδομένα ή το βαθμό συσχέτισής τους.

Μια εικόνα της σχέσης ανάμεσα στα δεδομένα $\{(x_k, y_k)\}_{k=0}^n$ δίνεται με τη βοήθεια του λεγόμενου διαγράμματος διασποράς (scatter plot) (είναι το διάγραμμα των σημείων (x_k, y_k)).

Έστω λοιπόν ότι (έχουμε δει από το διάγραμμα διασποράς) ^{ότι} η σχέση μεταξύ των σημείων $\{(x_k, y_k)\}_{k=0}^n$ περιγράφεται με τη βοήθεια μιας συνάρτησης $y=g(x)$ (προς προσδιορισμό). Η συνάρτηση αυτή είναι η "καλύτερη" δυνατή. Το καλύτερη δυνατή μπορεί να σημαίνει:

(α) $E_\infty = \max_{0 \leq i \leq n} \{ |y_i - g(x_i)| \}$ να είναι το μικρότερο δυνατό ανάμεσα στις "υποψηφίες" $g(x)$.
Θέλουμε ^{δηλ.} να βρούμε το $\min_g \{ \max_{0 \leq i \leq n} |y_i - g(x_i)| \}$ για όλες αυτές τις δυνατές g .

(β) για κάθε $g(x)$ υπολογίζουμε την ποσότητα:

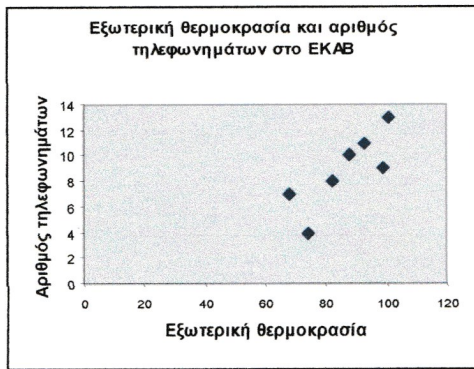
$$E_1 = \sum_{i=0}^n |y_i - g(x_i)| \quad (\text{απόλυτη απόκλιση})$$

Η g επιλέγεται έτσι ώστε να ελαχιστοποιείται η E_1 .

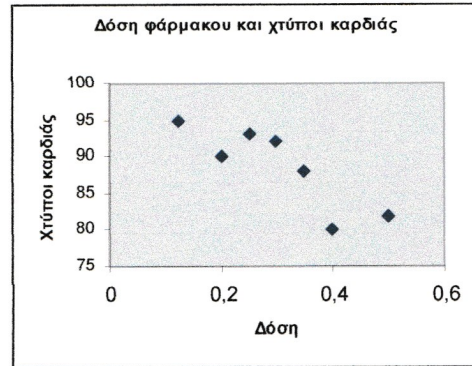
Παρατήρηση.

Στο (α), τα προβλήματα $\min$ είναι δύσκολο ν' αντιμετωπιστούν με στοιχειώδεις μαθηματικές τεχνικές. Ανάμεσα, στο (β) απαιτείται (για την εύρεση του $\min$) ο ταυτόχρονος μηδενισμός μέσων παραγώγων. Η απόλυτη τιμή όμως είναι συνάρτηση μη-παραγωγίσιμη στο σημείο 0, οπότε μπορεί να έχουμε πρόβλημα στον προσδιορισμό της $y=g(x)$.

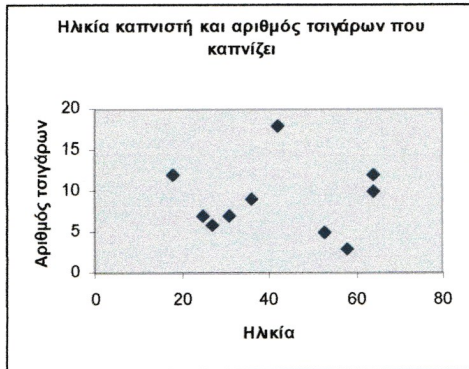
ΔΙΑΓΡΑΜΜΑΤΑ ΔΙΑΣΠΟΡΑΣ



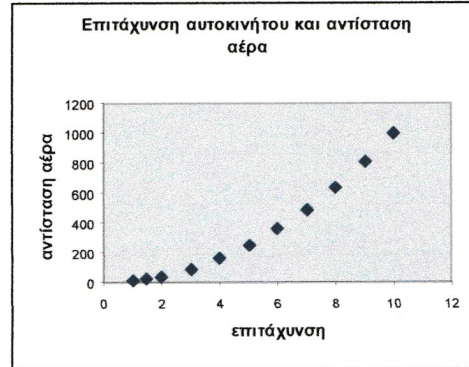
(α) (θετική γραμμική σχέση, ισχυρή)



(β) (αρνητική γραμμική σχέση, ισχυρή)



(γ) (σχεδόν ανύπαρκτη σχέση)



(δ) (μη γραμμική ισχυρή σχέση)

Η συνήθης ή δημοφιλέστερη μέθοδος για τον προσδιορισμό της $y=g(x)$, είναι $(n' g(x))$ να ελαχιστοποιεί την ποσότητα:

$$E_2 = \sum_{i=0}^n (y_i - g(x_i))^2$$

ή καλείται μέθοδος ελαχίστων τετραγώνων (least squares method).

Παρακάτω, αναπτύσσουμε την Μ.Ε.Τ σε μερικές βασικές/χαρακτηριστικές περιπτώσεις.

Παρατήρηση.

(α) Η συνάρτηση $y=g(x)$ καλείται καμπύλη προγνωσης

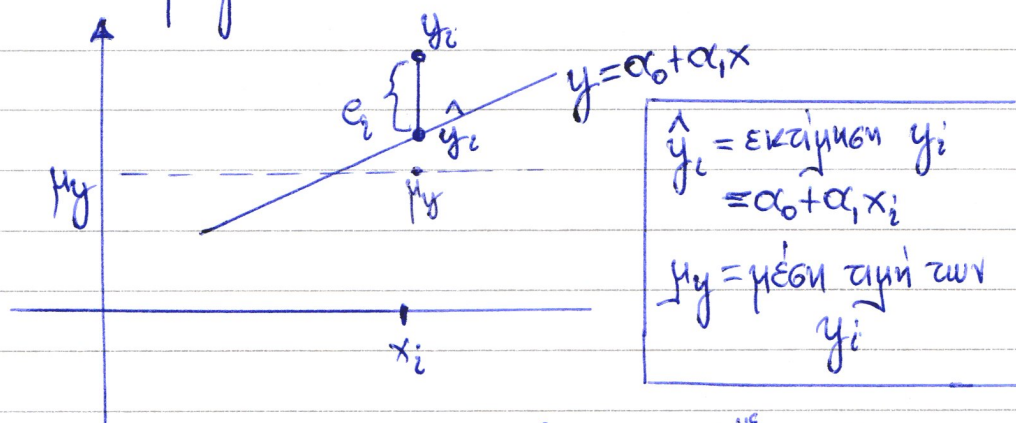
(β) Η προσδιορισμός της $y=g(x)$, που περιγράφει καλύτερα τα δεδομένα $\{(x_k, y_k)\}_{k=0}^n$ χωρίς (το χράφιμα της) να διέρχεται από τα σημεία καλείται προγνωση (regression).

Γραμμική παλινδρόμηση

Έχουμε $\{(x_k, y_k)\}_{k=0}^n$ $(n+1)$ -φευγαρία δεδομένων ή θέλουμε να προσδιορίσουμε μια ευθεία:

$$y = \alpha_0 + \alpha_1 x$$

που προερχή "καλύτερα" τα δεδομένα με τη βοήθεια της μεθόδου των ελαχίστων τετραγώνων.



Θέλουμε, να ελαχιστοποιήσουμε των ποσότητα: (αναφορικά με α_0, α_1)

$$E_2(\alpha_0, \alpha_1) = \sum_{i=0}^n (y_i - (\alpha_0 + \alpha_1 x_i))^2$$

Παρατήρηση

Η ποσότητα $e_i = y_i - \hat{y}_i = y_i - (\alpha_0 + \alpha_1 x_i)$ καλείται εφάλμα/υπόλοιπο (residual),

ενώ η $E_2(\alpha_0, \alpha_1)$ συνολικό τετραγωνικό εφάλμα.

Για την ελαχιστοποίηση των $E_2(\alpha_0, \alpha_1)$ θα πρέπει:

$$(i) \quad \frac{\partial E_2}{\partial \alpha_0} = \frac{\partial}{\partial \alpha_0} \sum_{i=0}^n (y_i - (\alpha_0 + \alpha_1 x_i))^2 = (-1) 2 \sum_{i=0}^n (y_i - \alpha_0 - \alpha_1 x_i) = 0$$

$$(ii) \quad \frac{\partial E_2}{\partial \alpha_1} = \frac{\partial}{\partial \alpha_1} \sum_{i=0}^n (y_i - (\alpha_0 + \alpha_1 x_i))^2 = 2 \sum_{i=0}^n (y_i - \alpha_0 - \alpha_1 x_i)(-x_i) = 0$$

οπότε:

$$\left. \begin{aligned} \sum_{i=0}^n (y_i - \alpha_0 - \alpha_1 x_i) &= 0 \\ \sum_{i=0}^n x_i (y_i - \alpha_0 - \alpha_1 x_i) &= 0 \end{aligned} \right\} \Rightarrow \begin{aligned} \alpha_0 \cdot (n+1) + \alpha_1 \cdot \sum_{i=0}^n x_i &= \sum_{i=0}^n y_i \quad (I) \\ \alpha_0 \cdot \sum_{i=0}^n x_i + \sum_{i=0}^n x_i^2 &= \sum_{i=0}^n x_i y_i \quad (II) \end{aligned}$$

Οι εξισώσεις (I) & (II) καλούνται κανονικές εξισώσεις. Η λύση του συστήματος των δύο αυτών εξισώσεων δίνει:

$$\alpha_1 = \frac{(n+1) \sum_{i=0}^n x_i y_i - \left(\sum_{i=0}^n x_i \right) \left(\sum_{i=0}^n y_i \right)}{(n+1) \sum_{i=0}^n x_i^2 - \left(\sum_{i=0}^n x_i \right)^2}$$

$$\alpha_0 = \frac{\left(\sum_{i=0}^n x_i^2 \right) \left(\sum_{i=0}^n y_i \right) - \left(\sum_{i=0}^n x_i y_i \right) \left(\sum_{i=0}^n x_i \right)}{(n+1) \left(\sum_{i=0}^n x_i^2 \right) - \left(\sum_{i=0}^n x_i \right)^2}$$

Το α_0 υπολογίζεται εύκολως από τη σχέση:

$$\alpha_0 = \frac{\sum_{i=0}^n y_i}{n+1} - \alpha_1 \frac{\sum_{i=0}^n x_i}{n+1}$$

Η ευθεία $y = \alpha_0 + \alpha_1 x$ καλείται ευθεία παλινδρόμησης (regression line).

Παράδειγμα 1

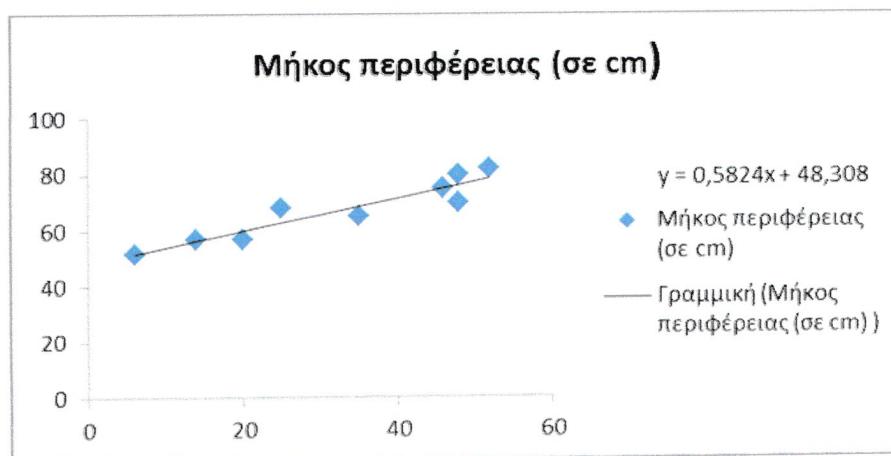
Σ' ένα πρόγραμμα προστασίας του περιβάλλοντος, θα θέλαμε να ξέρουμε εάν υπάρχει σχέση ανάμεσα στην απόσταση ενός δένδρου από τη θάλασσα και το μήκος της περιφέρειας του κορμού του. Τ' αποτελέσματα μετρήσεων, 9 τυχαία επιλεγμένων δένδρων ενός είδους, φαίνονται στον παρακάτω πίνακα.

Απόσταση (σε m)	Μήκος περιφέρειας (σε cm)
6	52
14	57
20	57
25	68
35	65
48	70
46	75
48	80
52	82

- (α) Να κατασκευαστεί *διάγραμμα διασποράς* για τα δεδομένα. Τι μπορείτε να πείτε κοιτάζοντας το διάγραμμα, υπάρχει σχέση ανάμεσα στην απόσταση ενός δένδρου από την θάλασσα και το μήκος της περιφέρειας του κορμού του; Τι είδους; Εξηγείστε.
- (β) Να βρεθεί η *ευθεία παλινδρόμησης* για το μέσο μήκος της περιφέρειας του κορμού ενός δένδρου, ως προς την απόστασή του από την θάλασσα.
- (γ) Ποιος είναι το μήκος της περιφέρειας ενός δένδρου που απέχει 30 m από τη θάλασσα; Ποιο, όταν απέχει 60 m από τη θάλασσα;

Λύση

Το *διάγραμμα διασποράς* είναι



Από το διάγραμμα είναι φανερό ότι υπάρχει μια **(ισχυρή) γραμμική σχέση** ανάμεσα στην απόσταση ενός δένδρου από τη θάλασσα και το μήκος της περιφέρειας του κορμού του **(δεδομένα κοντά σε μια νοητή ευθεία γραμμή)**.

- (β) Με τη βοήθεια των δεδομένων κατασκευάζουμε τον παρακάτω πίνακα

Απόσταση (σε m)	Μήκος περιφέρειας (σε cm)	x ²	xy
6	52	36	312
14	57	196	798
20	57	400	1140
25	68	625	1700
35	65	1225	2275
48	70	2304	3360
46	75	2116	3450
48	80	2304	3840
52	82	2704	4264
294	606	11910	21139
Σx	Σy	Σx ²	Σxy

Η ευθεία παλινδρόμησης $y = a_0 + a_1 x$, του μήκους της περιφέρειας ως προς την απόσταση, υπολογίζεται από τις σχέσεις (n=9):

$$\alpha_1 = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{9 \cdot 21139 - 294 \cdot 606}{9 \cdot 11910 - 294^2} = 0,5824$$

ενώ

$$\alpha_0 = \frac{\sum_{i=1}^n y_i}{n} - \alpha_1 \frac{\sum_{i=1}^n x_i}{n} = \frac{606}{9} + 0,5824 \cdot \frac{294}{9} = 48,3$$

Άρα η ευθεία παλινδρόμησης είναι :

$$y = 48,3 + 0,5824 \cdot x$$

(δείτε την ευθεία στο διάγραμμα διασποράς)

- (γ) Αν η απόσταση ενός δένδρου είναι 30 m από τη θάλασσα, τότε το μήκος της περιφέρειάς του είναι:

$$y = 48,3 + 0,5824 \cdot 30 = 65,77 \text{ cm}$$

Αν η απόσταση ενός δένδρου είναι 60 m από τη θάλασσα, τότε **δεν** μπορούμε να κάνουμε ασφαλή πρόβλεψη για το μήκος της περιφέρειάς του. Ασφαλή πρόβλεψη κάνουμε μόνον για αποστάσεις στο πεδίο ορισμού των δεδομένων (από 6 μέχρι 52 μέτρα απόσταση δηλαδή).

που περιγράφει καλύτερα τα δεδομένα, υπό την έννοια ότι το συνολικό τετράγωνο σφάλμα:

$$E_2(\alpha_0, \dots, \alpha_m) = \sum_{i=1}^n (y_i - (\alpha_0 + \dots + \alpha_m x_i^m))^2$$

είναι το μικρότερο δυνατό.

Парашютизм.

Параметры.

То $E_2(\alpha_0, \dots, \alpha_m) = \sum_{i=1}^n (y_i - (\alpha_0 + \dots + \alpha_m x_i^m))^2 = \sum_{i=1}^n (y_i - (\sum_{j=0}^m \alpha_j x_i^j))^2$

$$= \sum_{i=1}^n y_i^2 - 2 \sum_{i=1}^n P_m(x_i) y_i + \sum_{i=1}^n (P_m(x_i))^2$$

Για να ελαχιστοποιήσουμε το $E_2(a_0, \dots, a_m)$ θα πρέπει:

$$\frac{\partial E_2}{\partial \alpha_0} = 0, \quad \frac{\partial E_2}{\partial \alpha_1} = 0, \quad \dots, \quad \frac{\partial E_2}{\partial \alpha_m} = 0. \quad \text{Παραγωγίζοντας έχουμε τις εξισώσεις}$$

$$\left. \begin{aligned} &\alpha_0 \sum_{i=1}^n x_i^0 + \alpha_1 \sum_{i=1}^n x_i^1 + \dots + \alpha_m \sum_{i=1}^n x_i^m = \sum_{i=1}^n x_i^0 y_i \\ &\alpha_0 \sum_{i=1}^n x_i^1 + \alpha_1 \sum_{i=1}^n x_i^2 + \dots + \alpha_m \sum_{i=1}^n x_i^{m+1} = \sum_{i=1}^n x_i^1 y_i \\ &\vdots \\ &\alpha_0 \sum_{i=1}^n x_i^m + \alpha_1 \sum_{i=1}^n x_i^{m+1} + \dots + \alpha_m \sum_{i=1}^n x_i^{m+m} = \sum_{i=1}^n x_i^m y_i \end{aligned} \right\} \begin{array}{l} \text{NAGYONKÉS} \\ \text{EGENBÉLIS} \end{array}$$

Από τις εξισώσεις αυτές προκύπτουν τα $\alpha_0, \dots, \alpha_m$ (το σύστημα έχει μοναδική λύση, εάν τα σημεία είναι διαφορετικά).

Παράδειγμα 2

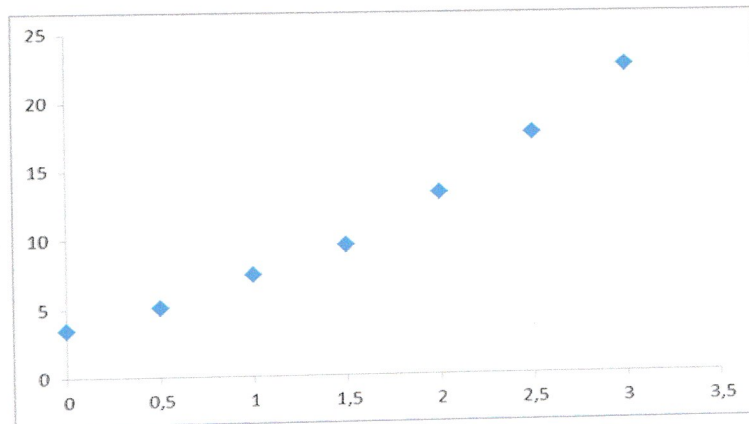
Δίνονται τα παρακάτω δεδομένα:

x	y
0	3,5
0,5	5,1
1	7,4
1,5	9,5
2	13,25
2,5	17,6
3	22,5

- (α) Να κατασκευαστεί *διάγραμμα διασποράς* για τα δεδομένα. Τι μπορείτε να πείτε κοιτάζοντας το διάγραμμα, υπάρχει σχέση ανάμεσα στα x και y; Τι είδους; Εξηγείστε.
- (β) Να προσαρμοστεί στα δεδομένα ένα πολυώνυμο (παλινδρόμησης) 2^{ου} βαθμού.

Λύση

Το *διάγραμμα διασποράς* είναι



Από το διάγραμμα είναι φανερό ότι υπάρχει μια (ισχυρή) μη - γραμμική σχέση ανάμεσα στα x και y. Τα δεδομένα δεν φαίνεται να κοντά σε μια νοητή ευθεία γραμμή, φαίνεται να *προσεγγίζονται καλύτερα από μια παραβολή (πολυώνυμο 2^{ου} βαθμού)*.

- (β) Η *καμπύλη παλινδρόμησης* $y = a_0 + a_1x + a_2x^2$, υπολογίζεται από τις σχέσεις (n=7):

$$\begin{aligned} \alpha_0 \sum_{i=1}^n x_i^0 + \alpha_1 \sum_{i=1}^n x_i^1 + \alpha_2 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n x_i^0 y_i & \alpha_0 \cdot n + \alpha_1 \sum_{i=1}^n x_i + \alpha_2 \sum_{i=1}^n x_i^2 &= \sum_{i=1}^n y_i \\ \alpha_0 \sum_{i=1}^n x_i^1 + \alpha_1 \sum_{i=1}^n x_i^2 + \alpha_2 \sum_{i=1}^n x_i^3 &= \sum_{i=1}^n x_i^1 y_i & \text{ή} & \alpha_0 \sum_{i=1}^n x_i + \alpha_1 \sum_{i=1}^n x_i^2 + \alpha_2 \sum_{i=1}^n x_i^3 &= \sum_{i=1}^n x_i y_i \\ \alpha_0 \sum_{i=1}^n x_i^2 + \alpha_1 \sum_{i=1}^n x_i^3 + \alpha_2 \sum_{i=1}^n x_i^4 &= \sum_{i=1}^n x_i^2 y_i & \alpha_0 \sum_{i=1}^n x_i^2 + \alpha_1 \sum_{i=1}^n x_i^3 + \alpha_2 \sum_{i=1}^n x_i^4 &= \sum_{i=1}^n x_i^2 y_i \end{aligned}$$

Με τη βοήθεια των δεδομένων κατασκευάζουμε τον παρακάτω πίνακα

x	y	x ²	x ³	x ⁴	xy	x ² y
0	3,5	0	0	0	0	0
0,5	5,1	0,25	0,125	0,0625	2,55	1,275
1	7,4	1	1	1	7,4	7,4
1,5	9,5	2,25	3,375	5,0625	14,25	21,375
2	13,25	4	8	16	26,5	53
2,5	17,6	6,25	15,625	39,0625	44	110
3	22,5	9	27	81	67,5	202,5

10,5	78,85	22,75	55,125	142,1875	162,2	395,55
Σx	Σy	Σx ²	Σx ³	Σx ⁴	Σxy	Σx ² y

Το σύστημα των κανονικών εξισώσεων γίνεται:

$$\begin{aligned} 7 \cdot \alpha_0 + 10,5 \cdot \alpha_1 + 22,75 \cdot \alpha_2 &= 78,85 \\ 10,5 \cdot \alpha_0 + 22,75 \cdot \alpha_1 + 55,125 \cdot \alpha_2 &= 162,2 \\ 22,75 \cdot \alpha_0 + 55,125 \cdot \alpha_1 + 142,1875 \cdot \alpha_2 &= 395,55 \end{aligned}$$

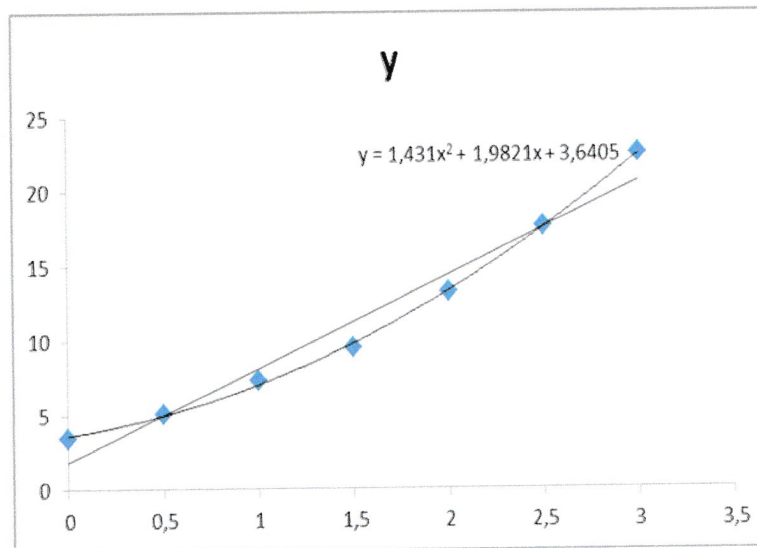
Λύνοντας το σύστημα έχουμε:

$$\alpha_0 = 3,6405 \quad \alpha_1 = 1,9821, \quad \alpha_2 = 1,4310$$

Οπότε η καμπύλη παλινδρόμησης (πολυώνυμο 2^{ου} βαθμού) είναι η:

$$y = 3,6405 + 1,9821 \cdot x + 1,4310 \cdot x^2$$

(δείτε τη στο διάγραμμα διασποράς // εκεί μπορεί να δει κανείς και την ευθεία παλινδρόμησης)



Εκθετική παλινδρόμηση

Μερικές φορές, η σχέση μεταξύ των σημείων είναι της μορφής:

$$g(x) = a e^{bx}$$

(α, β) προς προσδιορισμό.

Η μέθοδος που συνήθως χρησιμοποιείται σ' αυτή την περίπτωση είναι η εξής: λογαριθμίζοντας ή τα 2 μέλη της $y = a e^{bx}$ έχουμε:

$$\ln g = \ln(a e^{bx}) = \ln a + \ln e^{bx} = \ln a + bx.$$

Εάν θέσουμε $y_i = \ln y_i$ ή $a = \ln a$, αναζητάμε μια γραμμική σχέση της μορφής

$$y = a + bx \quad \text{που συνδέει τα σημεία } (x_i, \ln y_i)$$

εφαρμόζονοι ανάλογα όπως πριν, ή ελαχιστοποιώντας την

$$E_{\text{α}} = \sum_{i=1}^n (\ln y_i - \ln g(x_i))^2 = \sum_{i=1}^n (\ln y_i - a - bx_i)^2$$

έχουμε τις κανονικές εξισώσεις:

$$\begin{cases} a \cdot n + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ a \cdot \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases}$$

απ' όπου υπολογίζονται τα α, β. Τέλος $\alpha = e^a$.

Παρατήρηση

Η καμπύλη παλινδρόμησης που παίρνουμε μ' αυτό τον τρόπο δεν είναι η καμπύλη παλινδρόμησης της μεθόδου ελαχίστων τετραγώνων για το αρχικό πρόβλημα. Εάν προσπαθήσουμε να χρησιμοποιήσουμε τη ΜΕΤ στο αρχικό πρόβλημα παίρνουμε βύσματα εξισώσεων από το οποίο (συνήθως) δεν έχουμε ακριβή λύση για τα α, β.

Παράδειγμα 3 (συνέχεια)

Δίνονται τα παρακάτω δεδομένα

x	y
0	3,5
0,5	5,1
1	7,4
1,5	9,5
2	13,25
2,5	17,6
3	22,5

Να προσαρμοστεί στα δεδομένα μια συνάρτηση της μορφής : $y = a \cdot e^{\beta x}$

Λύση

Με τη βοήθεια των δεδομένων κατασκευάζουμε τον παρακάτω πίνακα

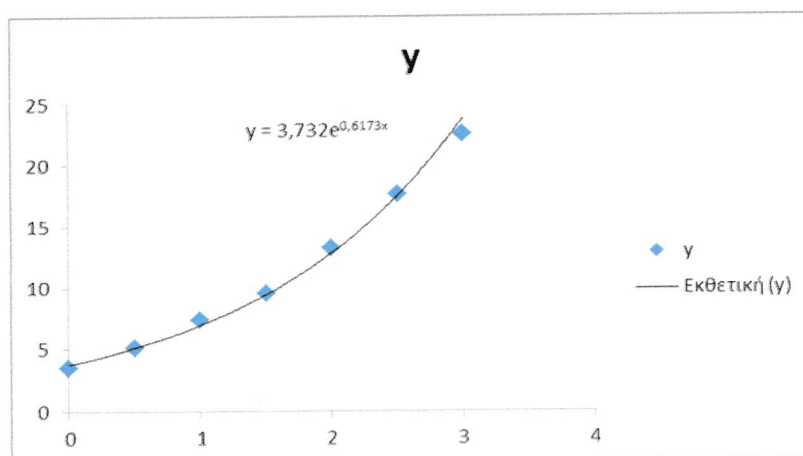
x	y	lny	x ²	xlny
0	3,5	1,253	0	0,000
0,5	5,1	1,629	0,25	0,815
1	7,4	2,001	1	2,001
1,5	9,5	2,251	2,25	3,377
2	13,25	2,584	4	5,168
2,5	17,6	2,868	6,25	7,170
3	22,5	3,114	9	9,341

10,5	78,85	15,700	22,75	27,871
Σx	Σy	Σlny	Σx ²	Σxlny

Οπότε έχουμε το σύστημα των εξισώσεων:

$$\begin{aligned} 7 \cdot a + 10,5 \cdot \beta &= 15,7 \\ 10,5 \cdot a + 22,75 \cdot \beta &= 27,871 \end{aligned} \quad \text{απ' όπου} \quad a = 1,31693, \quad \beta = 0,617286$$

Αλλά τότε $a = e^a = e^{1,3169} = 3,7318$ οπότε $y = 3,7318 \cdot e^{0,6173 \cdot x}$



Γεωμετρική Παλινδρόμηση

Μερικές φορές, η σχέση μεταξύ των σημείων $\{(x_k, y_k)\}_{k=1}^n$ είναι για εκθετική σχέση της μορφής:

$$y(x) = \alpha x^b \quad \text{όπου } \alpha, b \text{ προς προσδιορισμό}$$

Η μέθοδος που κ' εδώ χρησιμοποιείται, συνήδως, σ' αυτή την περίπτωση είναι η εξής: λογαριθμίζοντας κ' τα 2 μέλη της $y = \alpha \cdot x^b$ έχουμε:

$$\ln y(x) = \ln(\alpha \cdot x^b) = \ln \alpha + b \ln x.$$

Εάν θέσουμε, εάν $y_i = \ln y(x_i)$, $a = \ln \alpha$ κ' $x_i = \ln x_i$, αναμτάμε μια γραμμική σχέση της μορφής

$$y = a + b x.$$

Εφαρμόζουμε ανάλογα όπως πριν, κ' ελαχιστοποιώντας την:

$$E_{\text{co}} = \sum_{i=1}^n (\ln y_i - \ln y(x_i))^2 = \sum_{i=1}^n (\ln y_i - a - b \ln x_i)^2$$

έχουμε τις κανονικές εξισώσεις:

$$\begin{cases} a \cdot n + b \sum_{i=1}^n \ln x_i = \sum_{i=1}^n y_i = \sum_{i=1}^n \ln y_i \\ a \sum_{i=1}^n \ln x_i + b \sum_{i=1}^n (\ln x_i)^2 = \sum_{i=1}^n x_i y_i = \sum_{i=1}^n \ln x_i \cdot \ln y_i \end{cases}$$

απ' όπου υπολογίζονται τα a, b . Τέλος $\alpha = e^a$.

Παράδειγμα 4

Δίνονται τα παρακάτω δεδομένα

x	y
0,1	3,5
0,5	5,1
1	7,4
1,5	9,5
2	13,25
2,5	17,6
3	22,5

Να προσαρμοστεί στα δεδομένα μια συνάρτηση της μορφής : $y = a \cdot x^\beta$

Λύση

Με τη βοήθεια των δεδομένων κατασκευάζουμε τον παρακάτω πίνακα

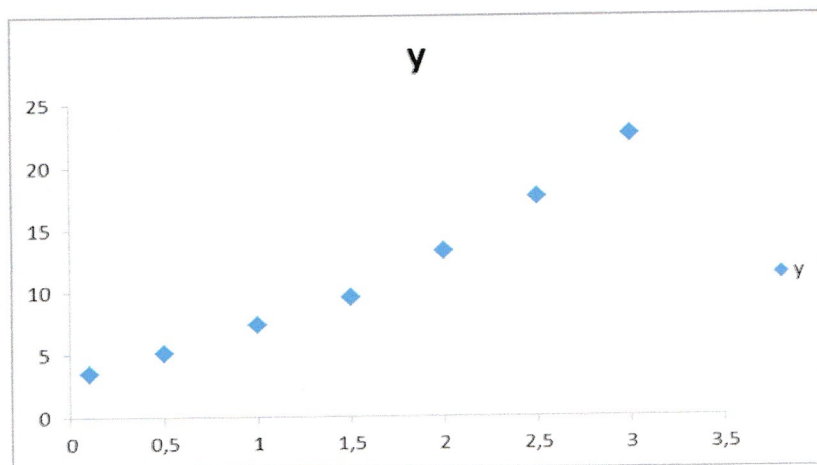
x	y	lnx	(lnx)^2	lny	lnxlny
0,1	3,5	-2,30259	5,30190	1,25276	-2,88459
0,5	5,1	-0,69315	0,48045	1,62924	-1,12930
1	7,4	0,00000	0,00000	2,00148	0,00000
1,5	9,5	0,40547	0,16440	2,25129	0,91282
2	13,25	0,69315	0,48045	2,58400	1,79109
2,5	17,6	0,91629	0,83959	2,86790	2,62783
3	22,5	1,09861	1,20695	3,11352	3,42055

10,6	78,85	0,117783	8,473744	15,70019	4,73839
Σx	Σy	Σlnx	Σ(lnx)^2	Σlny	Σlnxlny

Οπότε έχουμε το σύστημα των εξισώσεων:

$$\begin{aligned} 7 \cdot a + 0,117783 \cdot \beta &= 15,70019 \\ 0,117783 \cdot a + 8,473744 \cdot \beta &= 4,73839 \end{aligned} \quad \text{απ' όπου} \quad a=2,234, \quad \beta=0,5281$$

Αλλά τότε $a = e^a = e^{2,234} = 9,33714$ οπότε $y = 9,33714 \cdot x^{0,5281}$

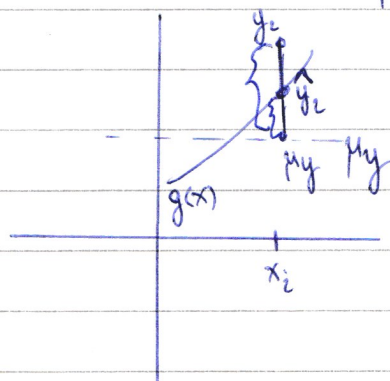


Ποιότητα προσαρμογής - συντελεστής προσδιορισμού

Στις προηγούμενες παραγράφους είδαμε πως μια καμπύλη $y=g(x)$ προσεγγίζει τα δεδομένα με τον καλύτερο δυνατό τρόπο αναφορικά με το κριτήριο των ελαχίστων τετραγώνων. Μετά τον υπολογισμό της $y=g(x)$ θα θέλαμε να έχουμε έναν δείκτη, ένδειξη του πόσο καλά προσαρμόζεται η $y=g(x)$ στα δεδομένα μας (βαθμός προσαρμογής). Ένας τέτοιος δείκτης είναι η ποσότητα:

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

$$\hat{y}_i = g(x_i), \mu_y$$



Η ποσότητα $SST = \sum_{i=1}^n (y_i - \mu_y)^2$ είναι η συνολική διασπορά των $\{y_i\}_{i=1}^n$.

Η ποσότητα $SSR = \sum_{i=1}^n (\hat{y}_i - \mu_y)^2$ είναι η διασπορά των $\{\hat{y}_i\}$ που ερμηνεύεται από την $g(x)$.

Ο δείκτης R^2 καλείται συντελεστής προσδιορισμού (coefficient of determination) κ' φανερώνει το ποσοστό της διασποράς των $\{y_i\}$ που ερμηνεύεται από την $g(x)$.

Ο R^2 παίρνει τιμές στο διάστημα $[0,1]$. Όσο πιο κοντά είναι ο R^2 στο 1 (τόσο περισσότερη διασπορά των τιμών $\{y_i\}$ ερμηνεύεται από την $g(x)$), τόσο καλύτερη είναι η προσαρμογή της $y=g(x)$ στα δεδομένα (τόσο πιο κοντά βρίσκονται τα σημεία (x_i, y_i) στην $g(x)$).

Απόсюда:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \mu_y)^2} \rightarrow SSE = \frac{\text{συνολικό τετραγωνικό σφάλμα}}{\text{συνολικό τετραγωνικό σφάλμα}}$$

Αν $R^2 = 1 \Rightarrow \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 0$ οπότε όλα τα σημεία βρίσκονται πάνω στην $y=g(x)$

Παράδειγμα 5 (συνέχεια)

Για τα δεδομένα του παρακάτω πίνακα

Απόσταση (σε m)	Μήκος περιφέρειας (σε cm)
6	52
14	57
20	57
25	68
35	65
48	70
46	75
48	80
52	82

Να υπολογιστεί ο **συντελεστής προσδιορισμού** (γραμμική παλινδρόμηση).

Λύση

Με τη βοήθεια των δεδομένων κατασκευάζουμε τον παρακάτω πίνακα

Απόσταση (σε m)	Μήκος περιφέρειας (σε cm)	$(y - \mu_y)^2$	y_{est}	$(y - est - \mu_y)^2$
6	52	215,1209	51,8024	220,9563332
14	57	93,45089	56,4616	104,1501892
20	57	93,45089	59,956	45,037521
25	68	1,776889	62,868	14,432401
35	65	2,778889	68,692	4,100625
48	70	11,10889	76,2632	92,08705444
46	75	69,43889	75,0984	71,08850596
48	80	177,7689	76,2632	92,08705444
52	82	235,1009	78,5928	142,2247056
294	606	899,996	605,9976	786,1643898

$$\mu_y = 606/9 = 66,34$$

$$R^2 = 786,16/899,99 = 0,873$$

Ο συντελεστής προσδιορισμού $R^2 = 0,873$ (που σημαίνει ότι 87,3% της διασποράς των τιμών του μήκους της περιφέρειας ενός δένδρου εξηγείται από την παλινδρόμηση).

Το ότι $R^2 = 0,873$ (κοντά στο 1) δείχνει ότι η προσαρμογή της ευθείας παλινδρόμησης στα δεδομένα είναι πολύ καλή.

Επιλογή καμπύλης ταυνοποίησης

Είδαμε πως προσαρμόζεται μια καμπύλη $y=g(x)$ σε δεδομένα με τη βοήθεια της μεθόδου των ελαχίστων τετραγώνων. Ακόμα υπολογίσαμε τον συντελεστή προσδιορισμού R^2 που μας δείχνει πόσο καλά προσαρμόζεται η καμπύλη $y=g(x)$ στα δεδομένα. Είχαμε όμως υποδείξει γι' αυτούς τους υπολογισμούς ότι η μορφή της $y=g(x)$ (ευθεία, πολυώνυμο κλπ) είναι γνωστή. Θα δείξουμε όμως να γέρουμε: απ' όλες τις δυνατές συναρτήσεις $y=g(x)$ ποιά είναι αυτή που προσεγγίζει καλύτερα τα δεδομένα.

Η αντίμετώπιση στο ερώτημα είναι η εξής:

(I) Μερικές φορές υπάρχει μια θεωρία που περιγράφει το νόμο/σχέση των μεταβλητών (π.χ. εκθετική σχέση $y=ab^x$). Στην περίπτωση αυτή χρειαζόμαστε καμπύλες αυτής της μορφής.

(II) Εάν δεν υπάρχει προηγούμενη θεωρία, τότε κατασκευάζουμε το διάγραμμα διασποράς. Το διάγραμμα, πολλές φορές, μας υποδεικνύει το είδος της σχέσης μεταξύ των μεταβλητών (π.χ. γραμμική). Δοκιμάζουμε καμπύλες μορφής ανάλογης αυτής που υποδεικνύεται από το διάγραμμα.

Στην πράξη η επιλογή γίνεται ως εξής: (εκ των υστέρων) επιλέχουμε μερικές καμπύλες (ανάλογης μορφής από αυτές που υποδεικνύονται από το διάγραμμα διασποράς) ή για κάθε μια από αυτές υπολογίζεται ο συντελεστής προσδιορισμού R^2 . Για καμπύλες με $R^2=0.90$ η προσέγγιση θεωρείται ικανοποιητική. Επιλέγουμε εκείνη των καμπύλων η οποία έχει τον μεγαλύτερο R^2 .

Παρατήρηση

Μεταξύ μιας πομπητικής $y=g(x)$ που παρέχει μικρή αύξηση στο R^2 ή μιας απλούστερης, προσιμότερη η απλούστερη. Πομπητικές $g(x)$ προτιμούνται όταν αυξάνουν ουδιστικά το R^2 .

Διαφορές παρεμβολής - παρεμβολές

Έστω ότι έχουμε n σημεία $(x_1, y_1), \dots, (x_n, y_n)$ (ίσως $y_i = f(x_i)$), κι έστω $P_{n-1}(x)$ το πολυώνυμο παρεμβολής (π.χ. Lagrange, Newton) ή $y = g(x)$ καμπύλη παρεμβολής που προερχήκει τα δεδομένα. Οι διαφορές μεταξύ των $P_{n-1}(x)$ ή $g(x)$ (δηλ. παρεμβολής ή παρεμβολές) είναι οι εξής:

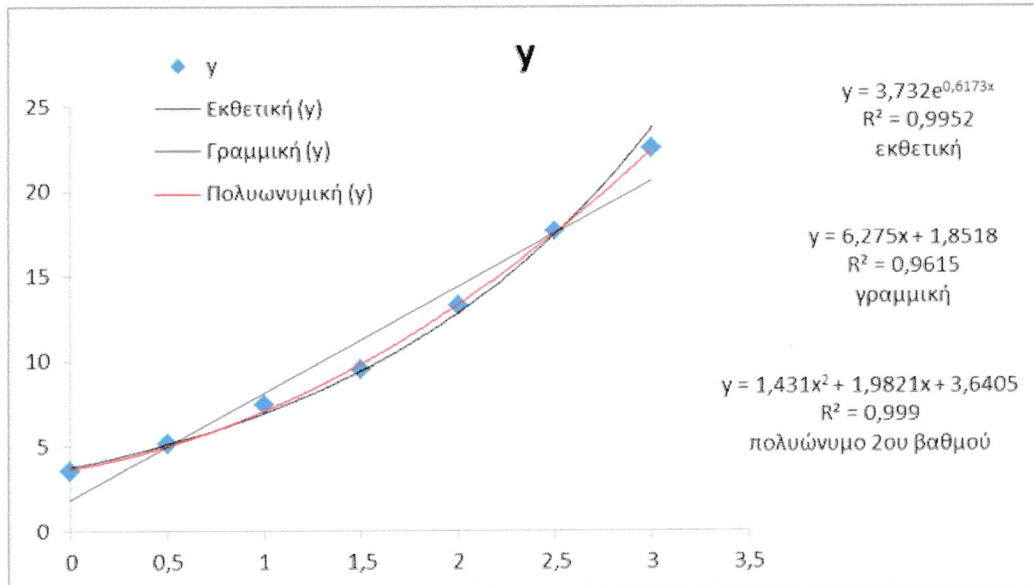
- (I) Το πολυώνυμο παρεμβολής διέρχεται απ' όλα τα σημεία (x_i, y_i) δηλ. $P_{n-1}(x_i) = y_i$, η συνάρτηση παρεμβολής όχι απαραιτείται.
- (II) Εάν έχουμε n σημεία παρεμβολής το πολυώνυμο παρεμβολής $P_{n-1}(x)$ είναι βαθμού $n-1$. Η καμπύλη παρεμβολής προσαρμόζεται και αν πρόκειται για πολυώνυμο είναι μικρού σχεδόν βαθμού (ανεξαρτήτως από το πλήθος των σημείων).

Επί

- (α) Η καμπύλη παρεμβολής προσαρμόζεται όταν έχουμε μεγάλο αριθμό σημείων (πολυώνυμα μεγάλου βαθμού παρουσιάζουν ταλαντώσεις).
- (β) Η καμπύλη παρεμβολής, ίσως, μειονεκτεί στο ότι μπορεί να μην περνά από κανένα από τα σημεία.

Παράδειγμα 6

Για τα δεδομένα των προβλημάτων 2,3 (ίδια δεδομένα), δίνεται παρακάτω ο συντελεστής προσδιορισμού στην περίπτωση (α) γραμμικής (β) πολυωνυμικής με πολυώνυμο 2^{ου} βαθμού και (γ) εκθετικής παλινδρόμησης



Συντελεστής προσδιορισμού:

(α) γραμμικής παλινδρόμησης (ευθεία): **0.9615**

(β) πολυωνυμικής παλινδρόμησης με πολυώνυμο 2^{ου} βαθμού: **0.999**

(γ) εκθετικής παλινδρόμησης: **0.9952**

Η καλύτερη καμπύλη που προσαρμόζεται στα δεδομένα είναι το πολυώνυμο 2^{ου} βαθμού (μεγαλύτερος συντελεστής προσδιορισμού). Όλα σχεδόν τα σημεία είναι πάνω στην καμπύλη