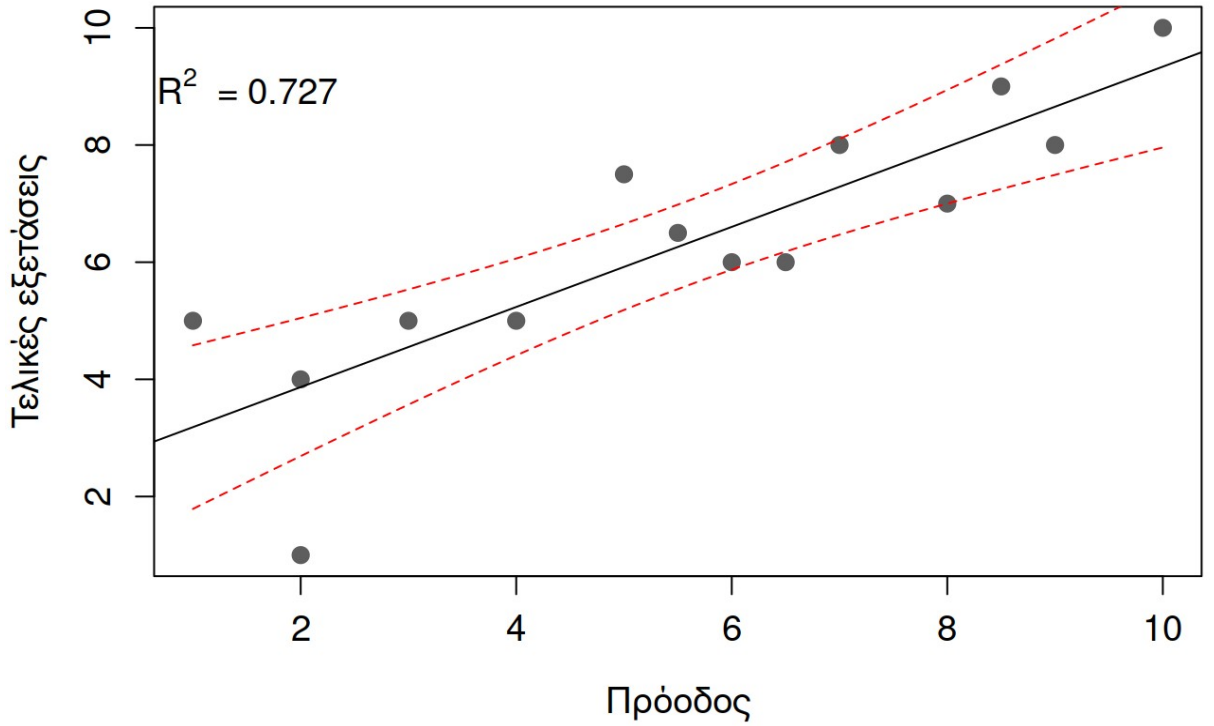
Θεωρία Πιθανοτήτων και Στατιστική



Διδάσκων: Επαμεινώνδας Διαμαντόπουλος
Επικοινωνία: epdiaman@ee.duth.gr

Περιεχόμενα 10ου μαθήματος

- Διάγραμμα Διασποράς (Scatterplot)
- Εμπειρική προσαρμογή ευθείας
- Συντελεστής γραμμικής συσχέτισης Pearson
- Συντελεστής συσχέτισης Spearman
- Απλή γραμμική παλινδρόμηση. Ευθεία ελαχίστων τετραγώνων.
- Εγκυρότητα και αξιοπιστία γραμμικού μοντέλου.

Γνωστικοί στόχοι 10ου μαθήματος

Στο τέλος του μαθήματος ο φοιτητής πρέπει:

- Να μπορεί να ποσοτικοποιεί τη γραμμική συσχέτιση μεταξύ δύο μεταβλητών με το συντελεστή συσχέτισης Pearson.
- Να μπορεί να σχεδιάσει ένα διάγραμμα διασποράς, να υπολογίσει τους συντελεστές της ευθείας ελαχίστων τετραγώνων και να αξιολογήσει τα βασικά της χαρακτηριστικά.

Διάγραμμα διασποράς (Scatterplot)
Εμπειρική προσαρμογή ευθείας

Διάγραμμα Διασποράς (Scatterplot)

Το διάγραμμα διασποράς είναι το κατάλληλο γράφημα που δημιουργούμε ως πρώτο βήμα για να μελετήσουμε τη σχέση που υπάρχει μεταξύ δύο αριθμητικών μεταβλητών.

Με το διάγραμμα διασποράς είναι δυνατόν να ανιχνευθεί η σχέση που ενδεχομένως να υπάρχει μεταξύ των δύο μεταβλητών.

Παράδειγμα

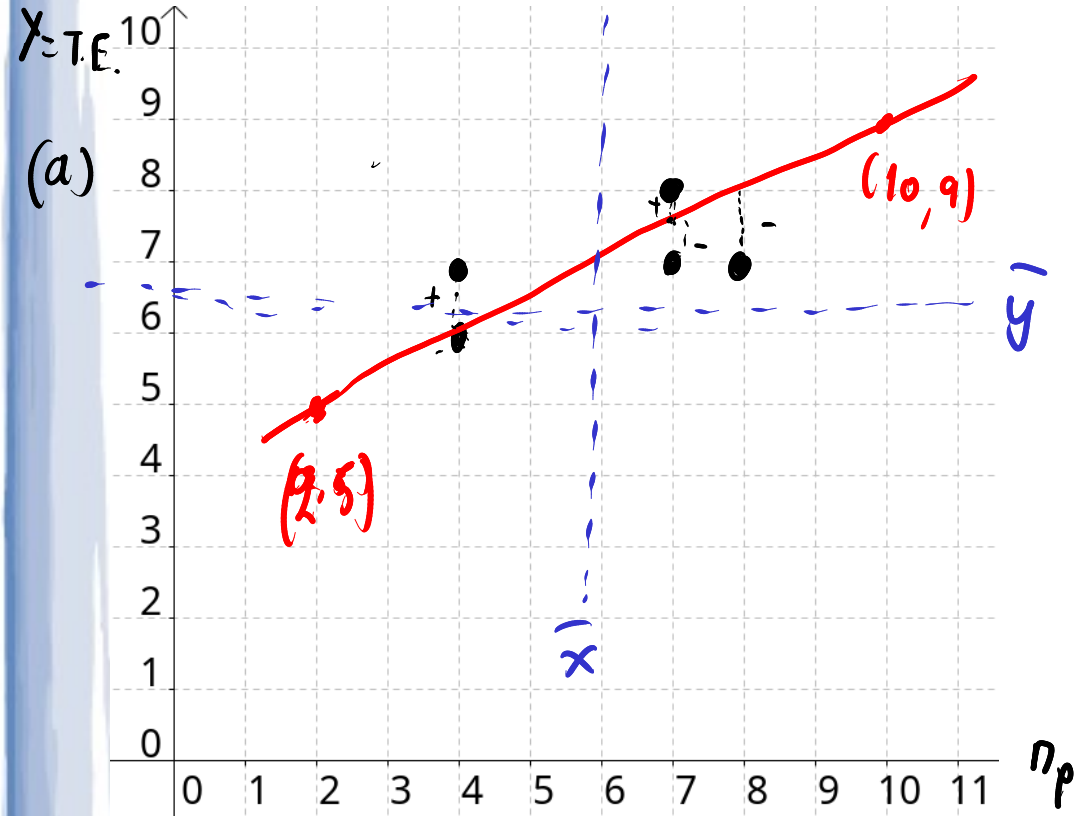
Για 5 φοιτητές καταγράφηκαν οι βαθμοί προόδου και τελικής εξέτασης και βρέθηκαν οι παρακάτω βαθμολογίες.

Πρόοδος	4	4	7	8	7
Τελική εξέταση	7	6	8	7	7

(α) Να γίνει το διάγραμμα διασποράς των βαθμολογιών (X = Πρόοδος, Y = Τελική εξέταση).
(β) Να γίνει προσαρμογή ευθείας στο σύνολο των σημείων με ελεύθερη επιλογή σημείων.
(γ) Ποια η ερμηνεία του σταθερού όρου του μοντέλου και του συντελεστή της Προόδου;
(δ) Να εκτιμηθεί ο βαθμός τελικής εξέτασης που αντιστοιχεί σε βαθμό Προόδου ίσο με 6.
(ε) Να υπολογιστούν τα υπόλοιπα των προβλέψεων και να αναπαρασταθούν διαγραμματικά.

Πρόοδος (X)	4	4	7	8	7
Τελική εξέταση (Y)	7	6	8	7	7
$\hat{y}$	6	6	7,5	8	7,5

(α) Να γίνει το διάγραμμα διασποράς των βαθμολογιών (X = Πρόοδος, Y = Τελική εξέταση).
(β) Να γίνει προσαρμογή ευθείας στο σύνολο των σημείων με ελεύθερη επιλογή σημείων.
(γ) Ποια η ερμηνεία του σταθερού όρου του μοντέλου και του συντελεστή της Προόδου;
(δ) Να εκτιμηθεί ο βαθμός τελικής εξέτασης που αντιστοιχεί σε βαθμό Προόδου ίσο με 6.
(ε) Να υπολογιστούν τα υπόλοιπα των προβλέψεων και να αναπαρασταθούν διαγραμματικά.



(ε) $y_{\text{υπόλοιπα}}$: $y_i - \hat{y}_i$

$i = 1, 2, \ldots, 5$

$\hat{y}_i = 4 + \frac{1}{2} \cdot x_i$.

$y_{\text{υπόλοιπα}}$: 1, 0, 0,5, −1, 0.5 .

(β) $(2,5) \in \varepsilon \Rightarrow 5 = a + b \cdot 2$
$(10,9) \in \varepsilon \Rightarrow 9 = a + b \cdot 10$
$\Big\} \; 8b = 4 \Leftrightarrow b = \frac{1}{2}$, $a = 4$

$Y = 4 + \frac{1}{2} \cdot X$

(δ) Για $X = 6$: $Y = 4 + \frac{1}{2} \cdot 6 = 7$.

πρόοδος = X

Συντελεστής γραμμικής συσχέτισης (Pearson Correlation Coefficient)

Συντελεστής γραμμικής συσχέτισης Pearson

Η προσαρμογή μίας ευθείας σε ένα διάγραμμα διασποράς αναμένεται να είναι αποτελεσματική αν οι δύο μεταβλητές έχουν πράγματι γραμμική σχέση. Δύο στατιστικά που εκφράζουν το μέγεθος της γραμμικότητας είναι η **συνδιακύμανση** και ο **συντελεστής γραμμικής συσχέτισης του Pearson**.
Α. Αν X, Y τυχαίες μεταβλητές:

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = E[(X - EX)(Y - EY)] = E[(X - \mu_X)(Y - \mu_Y)] = E(XY) - EX \cdot EY$$

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Β. Αν (x_i, y_i), ι = 1, 2, …, n ζεύγη παρατηρήσεων: $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}$

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})}}$$

$-1 \le r_{xy} \le +1.$

Σημείωση: Η διαίρεση με το n – 1 εξασφαλίζει πως το s_{XY} είναι αμερόληπτος εκτιμητής του σ_{XY}. Μία απόδειξη είναι διαθέσιμη εδώ: https://math.stackexchange.com/questions/2019122/unbiased-estimate-of-the-covariance

Συντελεστής γραμμικής συσχέτισης Pearson

Ο συντελεστής γραμμικής συσχέτισης είναι η κανονικοποιημένη εκδοχή της συνδιακύμανσης καθώς διατηρεί το ίδιο πρόσημό, έχοντας εξ ορισμού εύρος μεταξύ -1 και +1.

Απόδειξη

Η απόδειξη προκύπτει άμεσα από την παρακάτω σχέση:

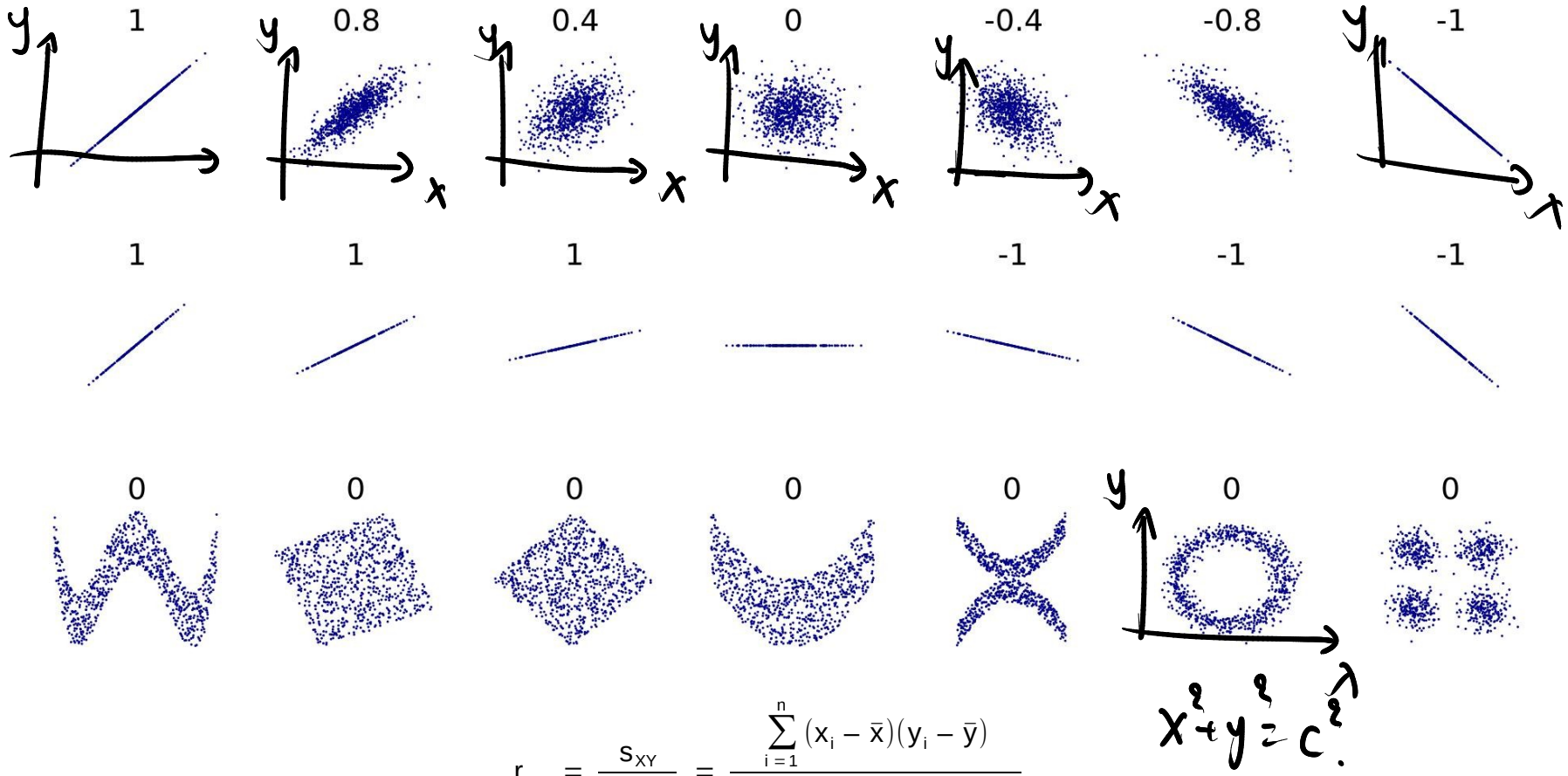
$$0 \leq \mathrm{Var}\left(\frac{X}{\sigma_X} \pm \frac{Y}{\sigma_Y}\right) = \mathrm{Var}\left(\frac{X}{\sigma_X}\right) + \mathrm{Var}\left(\frac{Y}{\sigma_Y}\right) \pm 2\,\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right) = 2 \pm 2\,\mathrm{Cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right)$$

$$Var\left(\frac{1}{\sigma_x} \cdot X\right) = \frac{1}{\sigma_x^2} \cdot Var(X) = \frac{1}{\sigma_x^2} \cdot \sigma_x^2 = 1$$

Πηγές

1. Sheldon Ross's A First Course in Probability 10th edition (p. 346). Διαθέσιμο για λήψη εδώ: https://ebin.pub/a-first-course-in-probability-global-edition-10nbsped-9780134753119-0134753119-9781292269207-1292269200.html
2. https://math.stackexchange.com/q/4110498

Συντελεστής γραμμικής συσχέτισης Pearson



$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})}}$$

$$\sum (x_i-\bar{x})\cdot(y_i-\bar{y}) = \sum x_i\cdot y_i - \bar{x}\cdot y_i - x_i\cdot\bar{y} + \bar{x}\cdot\bar{y} = \sum x_i\cdot y_i - \bar{x}\cdot n\cdot\bar{y} - \bar{y}\cdot n\cdot\bar{x} + n\cdot\bar{x}\cdot\bar{y}$$

Άσκηση 1

Να υπολογιστεί η συνδιακύμανση και ο συντελεστής συσχέτισης των βαθμών προόδου και τελικής εξέτασης.

Πρόοδος x	4	4	7	8	7
Τελική εξέταση y	7	6	8	7	7

$$S_{xy} = Cov(X,Y) = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i\cdot y_i - n\cdot\bar{x}\cdot\bar{y}\right].$$

$$n=5 \;;\; \sum_{i=1}^{5} x_i\cdot y_i = 4\cdot7+4\cdot6+7\cdot8+8\cdot7+7\cdot7 = 28+24+56+56+49 = 213.$$

$$\bar{x} = \frac{30}{5} = 6,\quad \bar{y} = \frac{35}{5} = 7$$

$$S_{xy} = \frac{1}{4}\cdot[213-5\cdot6\cdot7] = \frac{1}{4}(213-210) = 0,75.$$

Άσκηση 1

$$\sum (x_i-\bar{x})^2 = \sum (x_i^2 - 2x_i\cdot\bar{x} + \bar{x}^2) = \sum x_i^2 - 2\bar{x}\cdot n\cdot\bar{x} + n\bar{x}^2 = \sum x_i^2 - n\cdot\bar{x}^2$$

Να υπολογιστεί η συνδιακύμανση και ο συντελεστής συσχέτισης των βαθμών προόδου και τελικής εξέτασης.

	16	16	49	64	49
Πρόοδος x	4	4	7	8	7
Τελική εξέταση y	7	6	8	7	7

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \quad s_x^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n-1}\cdot\left[\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2\right]$$

$$s_x^2 = \frac{1}{4}\cdot\left[194 - 5\cdot 6^2\right] = \frac{1}{4}\cdot(194-180) = \frac{14}{4} \Rightarrow s_x = \frac{\sqrt{14}}{2}.$$

$$s_y^2 = \frac{1}{4}\cdot\left[247 - 5\cdot 7^2\right] = \frac{1}{4}(247-245) = \frac{1}{2} \Rightarrow s_y = \frac{1}{\sqrt{2}}$$

$$r_{xy} = \frac{0,75}{1,87\cdot 0,71} = 0,565$$

Άσκηση 2

Για 5 ζεύγη τιμών (x_i, y_i), γνωρίζουμε ότι

$$\sum_{i=1}^{5} x_i = 25,\ \sum_{i=1}^{5} x_i^2 = 165,\ \sum_{i=1}^{5} y_i = 24,\ \sum_{i=1}^{5} y_i^2 = 134,\ \sum_{i=1}^{5} x_i y_i = 144.$$

Να υπολογιστεί ο συντελεστής γραμμικής συσχέτισης των $\{x_i\}$, $\{y_i\}$.

Σημείωση: $\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$, και $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}, \quad S_{xy} = \frac{1}{4} \cdot \left[144 - 5 \cdot \frac{25}{5} \cdot \frac{24}{5}\right] = 6.$$

$$S_x^2 = \frac{1}{4}\left[165 - 5 \cdot \left(\frac{25}{5}\right)^2\right] = \frac{1}{4} \cdot (165 - 125) = 10 \Rightarrow S_x = \sqrt{10}$$

$$S_y^2 = \frac{1}{4} \cdot \left[134 - 5 \cdot \left(\frac{24}{5}\right)^2\right] = \frac{1}{4} \cdot 18{,}8 = 4{,}7 \Rightarrow S_y = \sqrt{4{,}7}$$

$$r_{xy} = \frac{6}{\sqrt{10} \cdot \sqrt{4{,}7}} = 0{,}875.$$

Άσκηση 3

Μία έρευνα, είχε ως στόχο τον εντοπισμό ενός τύπου που θα εκτιμά την ηλικίας των μεγάλων αιωνόβιων δέντρων (y) από την περιφέρειά τους (x). Τα δεδομένα από 15 δέντρα συνοψίζονται με τις ακόλουθες πληροφορίες:

$$\sum_{i=1}^{15} x_i = 3.368\,,\ \sum_{i=1}^{15} x_i^2 = 917.780\,,\ \sum_{i=1}^{15} y_i = 6.496\,,\ \sum_{i=1}^{15} y_i^2 = 4.260.667\,,\ \sum_{i=1}^{15} x_i y_i = 1.933.219\,.$$

Να υπολογιστεί ο συντελεστής γραμμικής συσχέτισης των $\{x_i\}$, $\{y_i\}$ και να ερμηνευτεί το αποτέλεσμα.

Συντελεστής γραμμικής συσχέτισης Pearson

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

Τα *υποκειμενικά* όρια που έχουν καθιερωθεί στην επιστημονική κοινότητα σχετικά με το χαρακτηρισμό αυτού του συντελεστή είναι τα εξής (|r| είναι η απόλυτη τιμή του συντελεστή συσχέτισης):

0 ≤ |r| < 0,3	Μη γραμμική σχέση		
0,3 ≤ |r| < 0,7	Ασθενής γραμμική σχέση	0,3 ≤ r < 0,7	Ασθενής θετική γραμμική σχέση
		-0,7 ≤ r < -0,3	Ασθενής αρνητική γραμμική σχέση
0,7 ≤ |r| ≤ 1	Ισχυρή γραμμική σχέση	0,7 ≤ r ≤ 1	Ισχυρή θετική γραμμική σχέση
		-1 ≤ r ≤ -0,7	Ισχυρή αρνητική γραμμική σχέση

Ευθεία ελαχίστων τετραγώνων στην απλή γραμμική παλινδρόμηση (Simple Linear Regression)

$f(x,y): \mathbb{R}^2 \to \mathbb{R} \Rightarrow \nabla f(x,y) = \left(\frac{\partial f}{\partial x}(x,y), \frac{\partial f}{\partial y}(x,y)\right)$.

Υπολογισμός συντελεστών 1/3

Έστω (x_i, y_i), i = 1, 2, …, n τα n ζεύγη παρατηρήσεων και έστω

$$\hat{y} = \hat{\alpha} + \hat{b}\,x$$

η εξίσωση της ευθείας που αναζητούμε. Τα σφάλματα της εκτίμησης για κάθε ένα x_i, i = 1, 2, …, n, είναι:

$$\text{resid}_i(\hat{\alpha}, \hat{b}) = \varepsilon_i(\hat{\alpha}, \hat{b}) = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{b}\,x_i$$

Αναζητούμε την εξίσωση ευθείας που ελαχιστοποιεί το άθροισμα τετραγώνων αυτών των σφαλμάτων.

$\hat{\alpha}, \hat{b} = ;$ ώστε $S(\hat{\alpha}, \hat{b}) = \sum_{i=1}^{n} \varepsilon_i^2(\hat{\alpha}, \hat{b}) = \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{b}\,x_i)^2 = \text{Minimum}$

Για τον υπολογισμό των α, b, αρκεί να βρούμε τις κρίσιμες τιμές: $\nabla S(\hat{\alpha}, \hat{b}) = 0 \Leftrightarrow \frac{\partial S}{\partial \hat{\alpha}} = 0$ και $\frac{\partial S}{\partial \hat{b}} = 0$

Είναι:

$$\frac{\partial S}{\partial \hat{\alpha}} = \sum_{i=1}^{n} -2(y_i - \hat{\alpha} - \hat{b}\,x_i) = -2\left[\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\alpha} - \hat{b}\sum_{i=1}^{n} x_i\right] = -2\left[n\bar{y} - n\hat{\alpha} - \hat{b}\,n\bar{x}\right]$$

και:

$$\frac{\partial S}{\partial \hat{\alpha}} = 0 \Leftrightarrow \bar{y} - \hat{\alpha} - \hat{b}\,\bar{x} = 0 \text{ ή } \hat{\alpha} = \bar{y} - \hat{b}\,\bar{x}$$

Σημείωση

Από την τελευταία σχέση προκύπτει ειδικότερα ότι το σημείο $(\bar{x}, \bar{y})$ ανήκει στην ευθεία ελαχίστων τετραγώνων.

Υπολογισμός συντελεστών 2/3

Περαιτέρω, από την σχέση $\frac{\partial S}{\partial \hat{b}} = 0$ παίρνουμε:

$$\sum_{i=1}^{n} x_i(y_i - \hat{\alpha} - \hat{b}x_i) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i(y_i - \bar{y} + \hat{b}\bar{x} - \hat{b}x_i) = 0 \Leftrightarrow \sum_{i=1}^{n} x_i(y_i - \bar{y}) - \hat{b}\sum_{i=1}^{n} x_i(x_i - \bar{x}) = 0$$

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

Υπολογισμός συντελεστών 3/3

Εύκολα, αποδεικνύεται ότι

$$\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})=\sum_{i=1}^{n}x_i(y_i-\bar{y})-\sum_{i=1}^{n}\bar{x}\,y_i+n\bar{x}\,\bar{y}=\sum_{i=1}^{n}x_i(y_i-\bar{y})$$

$$\sum_{i=1}^{n}(x_i-\bar{x})\cdot(x_i-\bar{x})=\sum_{i=1}^{n}x_i(x_i-\bar{x})-\sum_{i=1}^{n}\bar{x}\,x_i+n\bar{x}\,\bar{x}=\sum_{i=1}^{n}x_i(x_i-\bar{x})$$

Από τις τελευταίες δύο ισότητες, συνάγεται ότι μία ισοδύναμη έκφραση των συντελεστών είναι:

$$\hat{\alpha}=\bar{y}-\hat{b}\bar{x}\ ,\ \hat{b}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{s_{XY}}{s_X^2}=r_{XY}\frac{s_Y}{s_X}$$

$$y=\hat{a}+\hat{b}\cdot x\ ,\ \hat{b}=\frac{s_{xy}}{s_x^2}\ ,\ \hat{a}=\bar{y}-\hat{b}\,\bar{x}\ .$$

Ασκήσεις

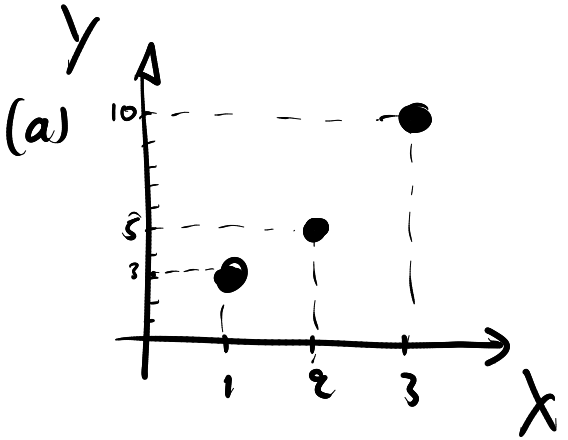
Άσκηση 1

Για τα δεδομένα του πίνακα:

$\bar{x} = 2$

$\bar{y} = 6$

X	1	2	3
Y	3	5	10



α) Να γίνει το διάγραμμα διασποράς.

β) Να βρεθεί η συνδιακύμανση των X, Y.

γ) Να βρεθεί ο συντελεστής συσχέτισης των X, Y.

δ) Να βρεθεί η εξίσωση της ευθείας γραμμικής παλινδρόμησης της Y πάνω στη X.

ε) Να εκτιμηθεί η τιμή του Y όταν το X θα πάρει την τιμή 4.

(ε) Για $X=4$: $\hat{Y} = -1 + 3{,}5 \cdot 4 = 13$.

(γ) $r_{xy} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{3{,}5}{\sqrt{13}} = 0{,}971$.

(β) $S_{xy} = \frac{1}{2} \cdot [(1-2)\cdot(3-6) + (2-2)\cdot(5-6) + (3-2)\cdot(10-6)] = \frac{1}{2}\cdot(3+4) = 3{,}5$.

$S_x^2 = \frac{1}{2}\cdot[(1-2)^2 + (2-2)^2 + (3-2)^2] = 1 \Rightarrow S_x = 1$

$S_y^2 = \frac{1}{2}\cdot[(3-6)^2 + (5-6)^2 + (10-6)^2] = \frac{26}{2} = 13 \Rightarrow S_y = \sqrt{13}$

(δ) $\hat{b} = \frac{S_{xy}}{S_x^2} = \frac{3{,}5}{1} = 3{,}5$

$\hat{a} = \bar{y} - \hat{b}\,\bar{x} = 6 - 3{,}5 \cdot 2 = -1$

$y = -1 + 3{,}5 \cdot x$

Σημαντικά αθροίσματα τετραγώνων

Έστω (x_i, y_i), i = 1, 2, …, n, n ζεύγη παρατηρήσεων και $\hat{y} = \hat{\alpha} + \hat{b}x$ η εξίσωση της ευθείας γραμμικής παλινδρόμησης της Υ πάνω στη Χ. Τότε, ισχύει:

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

Απόδειξη (1/2)

$$(y_i - \bar{y})^2 = (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 + 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

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

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y}) = (y_i - \bar{y}) - \hat{b}(x_i - \bar{x}) \quad \text{(γιατί } y = \hat{\alpha} + \hat{b}\bar{x} \text{ και } \hat{y}_i = \hat{\alpha} + \hat{b}x_i\text{)}$$

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = \hat{b}\sum_{i=1}^{n} (x_i - \bar{x})[(y_i - \bar{y}) - \hat{b}(x_i - \bar{x})]$$

Αρκεί ν. δ. ο. $\sum_{i=1}^{n} (x_i - \bar{x})[(y_i - \bar{y}) - \hat{b}(x_i - \bar{x})] = 0.$

Σημαντικά αθροίσματα τετραγώνων

Έστω (x_i, y_i), i = 1, 2, …, n, n ζεύγη παρατηρήσεων και $\hat{y} = \hat{\alpha} + \hat{b}x$ η εξίσωση της ευθείας γραμμικής παλινδρόμησης της Υ πάνω στη Χ. Τότε, ισχύει:

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

Απόδειξη (2/2)

$$\sum_{i=1}^{n} (x_i - \bar{x})[(y_i - \bar{y}) - \hat{b}(x_i - \bar{x})] = \sum_{i=1}^{n} (x_i - \bar{x}) \left[(y_i - \bar{y}) - \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{n} (x_k - \bar{x})^2} (x_i - \bar{x}) \right]$$

$$\left(\text{είναι } \hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) - \frac{\sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})}{\sum_{k=1}^{n} (x_k - \bar{x})^2} (x_i - \bar{x})^2 = 0$$

Σημείωση: Η απόδειξη είναι διαθέσιμη εδώ: https://en.wikipedia.org/wiki/Explained_sum_of_squares

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

Στα πλαίσια της γραμμικής παλινδρόμησης τα παραπάνω αθροίσματα τετραγώνων έχουν ιδιαίτερη σημασία και ονομασία.

α) $TSS=\sum_{i=1}^{n}(y_i-\bar{y})^2$: συνολικό άθροισμα τετραγώνων (total sum of squares)

Το συνολικό άθροισμα τετραγώνων (TSS) εκφράζει τη συνολική μεταβλητότητα της εξαρτημένης μεταβλητής.

β) $ESS=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$: επεξηγημένο άθροισμα τετραγώνων (explained sum of squares)

Το επεξηγημένο άθροισμα τετραγώνων (ESS) είναι το μέρος της μεταβλητότητας της εξαρτημένης μεταβλητής που εξηγείται από τις επεξηγηματικές μεταβλητές του γραμμικού μοντέλου.

γ) $RSS=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$: άθροισμα τετραγώνων υπολοίπων (residual sum of squares)

Το άθροισμα τετραγώνων υπολοίπων (RSS) εκφράζει το μέρος της μεταβλητότητας της εξαρτημένης μεταβλητής που δεν εξηγείται από τις επεξηγηματικές μεταβλητές του γραμμικού μοντέλου.

Η σχέση $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ γράφεται:

$$TSS = ESS + RSS$$

Διαιρώντας, με το συνολικό άθροισμα τετραγώνων (TSS) παίρνουμε:

$$\frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Συντελεστής Προσδιορισμού (R^2) 1/2

Το πόσο καλά προσαρμόζεται το γραμμικό μοντέλο στα δεδομένα ποσοτικοποιείται και από το συντελεστή προσδιορισμού (coeffficient of determination) ο οποίος ορίζεται ως:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Αν το μοντέλο προβλέπει χωρίς σφάλμα όλες τις παρατηρήσεις, τότε

$$y_i - \hat{y}_i = 0 \Leftrightarrow R^2 = 1 = 100\,\%.$$

Αν τα σφάλματα προβλέψεων του μοντέλου είναι όσο η διαφορά από την μέση τιμή (baseline model), τότε

$$y_i - \hat{y}_i = y_i - \bar{y} \Leftrightarrow R^2 = 0 = 0\,\%.$$

Αν τα σφάλματα προβλέψεων ξεπερνούν την απόσταση των παρατηρήσεων από τη μέση τιμή, τότε $R^2 < 0$.

Συντελεστής Προσδιορισμού (R^2) 2/2

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

Ο συντελεστής R^2 συνήθως παρουσιάζεται ως ποσοστό με μεγαλύτερες τιμές να καταδεικνύουν καλύτερη προσαρμογή στα δεδομένα. Τα όρια αποδοχής της αποτελεσματικότητας του μοντέλου ποικίλουν ανάλογα με την επιστημονική περιοχή.

Στην απλή γραμμική παλινδρόμηση αποδεικνύεται ότι ο συντελεστής προσδιορισμού είναι το τετράγωνο του συντελεστή συσχέτισης Pearson μεταξύ των x_i και των y_i.

$$R^2 = r^2$$

Ειδικότερα: Στην απλή γραμμική παλινδρόμηση είναι αδύνατον για το R^2 να πάρει αρνητικές τιμές.

Σημείωση
Μία απόδειξη μπορεί να βρεθεί εδώ:
https://math.stackexchange.com/questions/129909/correlation-coefficient-and-determination-coefficient

Ασκήσεις

$Y = -1 + 3,5 \cdot X$

Άσκηση 1 (συνέχεια)

Για τα δεδομένα του πίνακα:

X	1	2	3
Y	3	5	10

$\bar{y} = 6$

$\hat{y}$	2,5	6	9,5

ζ) Να βρεθούν τα αθροίσματα τετραγώνων TSS, ESS, RSS.

η) Να βρεθεί ο συντελεστής προσδιορισμού R^2.

$$TSS = \sum_{i=1}^{3} (y_i - \bar{y})^2 = (3-6)^2 + (5-6)^2 + (10-6)^2 = 26.$$

$$ESS = \sum_{i=1}^{3} (\hat{y}_i - \bar{y})^2 = (2,5-6)^2 + (6-6)^2 + (9,5-6)^2 = 24,5$$

$$RSS = \sum_{i=1}^{3} (y_i - \hat{y}_i) = (3-2,5)^2 + (5-6)^2 + (10-9,5)^2 = 0,25 + 1 + 0,25 = 1,5.$$

$$R^2 = \frac{ESS}{TSS} = \frac{24,5}{26} \approx 0,942 \approx 94,2\%.$$

Κατανομή Συντελεστών

$$\hat{\alpha} = \bar{y} - \hat{b}\,\bar{x}\ ,\ \ \hat{b} = \frac{s_{XY}}{s_X^2} = r_{XY}\frac{s_Y}{s_X}$$

Ο υπολογισμός των συντελεστών οδηγεί σε κάποιες αριθμητικές τιμές. Αυτές οι τιμές είναι μία εκτίμηση των συντελεστών που θα υπολογίζαμε από το σύνολο του πληθυσμού αν είχαμε πρόσβαση σε αυτόν.

Συνεπώς, αποτελούν μία εκτίμηση των άγνωστων τιμών και ως εκτίμηση, συνοδεύονται από ένα σφάλμα.

Όπως και στην περίπτωση της μέσης τιμής, αυτό το σφάλμα μπορούμε να το περιγράψουμε από το 95% διάστημα εμπιστοσύνης.

Ο υπολογισμός του 95% δ.ε. προϋποθέτει τη γνώση της κατανομής των τιμών.

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

Κατανομή Συντελεστών

Αποδεικνύεται ότι

$$\frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})} \sim t(n-2), \quad \frac{\hat{b} - b}{SE(\hat{b})} \sim t(n-2)$$

όπου

$$SE(\hat{b}) = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2 / \sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{RSS}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$SE(\hat{\alpha}) = SE(\hat{b})\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Σημειώσεις
1. Οι αποδείξεις των παραπάνω τύπων για την απλή γραμμική παλινδρόμηση είναι διαθέσιμες εδώ:
https://en.wikipedia.org/wiki/Simple_linear_regression

2, Η απόδειξη στη γενική περίπτωση για το γεγονός ότι οι συντελεστές ακολουθούν την κατανομή t μπορεί να βρεθεί εδώ:
https://stats.stackexchange.com/questions/117406/proof-that-the-coefficients-in-an-ols-model-follow-a-t-distribution-with-n-k-d

Κατανομή Συντελεστών

$$\frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})} \sim t(n-2) \Rightarrow 95\%\ \delta.\ \varepsilon.: \left(\hat{\alpha} - t_{n-2;0.025} SE(\hat{\alpha}),\ \hat{\alpha} + t_{n-2;0.025} SE(\hat{\alpha})\right)$$

$$\frac{\hat{b} - b}{SE(\hat{b})} \sim t(n-2) \Rightarrow 95\%\ \delta.\ \varepsilon.: \left(\hat{b} - t_{n-2;0.025} SE(\hat{b}),\ \hat{b} + t_{n-2;0.025} SE(\hat{b})\right)$$

Είναι χρήσιμο να βρούμε το κατά πόσο οι τιμές των συντελεστών διαφοροποιούνται σημαντικά από το 0.

Αυτό, προκύπτει άμεσα από το 95% δ.ε.:

- Αν το 0 ανήκει μέσα στο 95% δ.ε. τότε ο συντελεστής δεν διαφοροποιείται σημαντικά από το 0 και η αντίστοιχη εξάρτηση δεν είναι στατιστικώς σημαντική.
- Αν το 0 δεν ανήκει μέσα στο 95% δ.ε. τότε ο συντελεστής διαφοροποιείται σημαντικά από το 0 και η αντίστοιχη εξάρτηση είναι στατιστικώς σημαντική.

Κατανομή Συντελεστών

Περισσότερο τυπικά, μπορούμε να ελέγξουμε την υπόθεση

H_0: b = 0 έναντι της H_1: b ≠ 0.

Το στατιστικό που υπολογίζουμε είναι το:

$$t_0 = \frac{\hat{b}}{SE(\hat{b})} \sim t(n-2), \quad \text{και} \quad p = P(|t| > |t_0|)$$

Αντίστοιχα, για να ελέγξουμε την υπόθεση

H_0: α = 0, έναντι της H_1: α ≠ 0.

το στατιστικό που υπολογίζουμε είναι το:

$$t_0 = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \sim t(n-2), \quad \text{και} \quad p = P(|t| > |t_0|)$$

Απλή Γραμμική Παλινδρόμηση – Σύνοψη

Συνοψίζοντας, η ευθεία ελαχίστων τετραγώνων έχει εξίσωση $\hat{y} = \hat{\alpha} + \hat{b}\,x$ όπου

$$\hat{b} = \frac{s_{XY}}{s_X^2},\ \ \hat{\alpha} = \bar{y} - \hat{b}\,\bar{x}$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2,\ \ ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,\ \ RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SE(\hat{b}) = \sqrt{\frac{RSS}{(n-2)\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$H_0: b=0:\ \ t = \frac{\hat{b}}{SE(\hat{b})} \sim t(n-2),\ \ p = P(|t| > |t_0|)$$

$$SE(\hat{\alpha}) = SE(\hat{b})\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

$$H_0: \alpha=0:\ \ t = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \sim t(n-2),\ \ p = P(|t| > |t_0|)$$

Ασκήσεις

$\hat{y} = -1 + 3{,}5 \cdot x$. $\hat{a} = -1$, $\hat{b} = 3{,}5$

Άσκηση 1 (συνέχεια)

Για τα δεδομένα του πίνακα:

X	1	2	3
Y	3	5	10

$RSS = 1{,}5$

θ) Να βρεθεί ένα 95% δ.ε. για τους συντελεστές του γραμμικού μοντέλου.

ι) Να ελεγχθούν οι υποθέσεις

H_0: α = 0, έναντι της H_1: α ≠ 0.

H_0: b = 0, έναντι της H_1: b ≠ 0.

$\frac{1}{3}\sum_{i=1}^{3} x_i^2 = \frac{1}{3}(1^2+2^2+3^2) = \frac{14}{3}$.

$SE(\hat{a}) = 0{,}866\sqrt{\frac{14}{3}} = 1{,}87$

$\text{Δίνεται}: SE(\hat{b}) = \sqrt{\frac{RSS}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}}$, $SE(\hat{\alpha}) = SE(\hat{b})\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$

95% δ.ε.: $(-1 - 12.71 \cdot 1{,}87, -1 + 12.71 \cdot 1{,}87)$

$= (-24.8, 22.8)$.

(θ) $SE(\hat{b}) = \sqrt{\frac{1{,}5}{1 \cdot 2}} = 0{,}866$: 95% δ.ε. $(\hat{b} - t_{1;0{,}025} \cdot 0{,}866, \hat{b} + t_{1;0.025} \cdot 0{,}866)$ ⊛

(i) $H_0: b=0$ είναι $H_1: b \neq 0$ Όριο απορρίψης H_0: 0,05

$$t_0 = \frac{\hat{b}}{SE(\hat{b})} = \frac{3.5}{0.866} = 4,041 \quad , t \sim t(n-2) = t(1)$$

$$p = P(t > 4.041 \text{ ή } t < -4,041) = 0,154 > 0,05 \Rightarrow$$

$\Rightarrow$ Δεν απορρίπτεται η H_0.

$H_0: a=0$ είναι ως $H_1: a \neq 0$.

$$t_0 = \frac{\hat{a}}{SE(\hat{a})} = \frac{-1}{1.87} = -0,535$$

$p = P(t < -0,535 \text{ ή } t > 0.535) = 0,687. \Rightarrow$ η H_0 δεν απορρίπτεται.

Ασκήσεις

Άσκηση 1 (συνέχεια)

Για τα δεδομένα του πίνακα:

X	1	2	3
Y	3	5	10

θ) Να βρεθεί ένα 95% δ.ε. για τους συντελεστές του γραμμικού μοντέλου.

ι) Να ελεγχθούν οι υποθέσεις

H_0: α = 0, έναντι της H_1: α ≠ 0.

H_0: b = 0, έναντι της H_1: b ≠ 0.

$$\text{Δίνεται: } SE(\hat{b}) = \sqrt{\frac{RSS}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}},\ SE(\hat{\alpha}) = SE(\hat{b})\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

$$(\theta) = (3{,}5 - 12{,}71 \cdot 0{,}866,\ 3{,}5 + 12{,}71 \cdot 0{,}866) = (-7{,}5\ ,\ 14{,}5)$$

F – test για τους συντελεστές του μοντέλου

Το F-test για την γραμμική παλινδρόμηση ελέγχει εάν κάποια από τις ανεξάρτητες μεταβλητές σε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης είναι σημαντικά διαφορετική από το 0.

Ειδικότερα, για το μοντέλο $Y = \alpha + bX$, ελέγχεται η υπόθεση

$$H_0: \alpha = b = 0, \qquad \text{έναντι της} \qquad H_1: \alpha \neq 0 \text{ ή } b \neq 0.$$

Κάτω από την υπόθεση H_0, το αντίστοιχο μοντέλο (restricted ή reduced ή baseline model) είναι το $Y = \alpha$ = ο αριθμητικός μέσος των y_i = αμερόληπτος εκτιμητής της μ_Y.

Η απόκλιση των εκτιμήσεων από το “μηδενικό” μοντέλο εκφράζεται από το άθροισμα

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

Αντίστοιχα, τα σφάλματα του μοντέλου που εξετάζουμε εκφράζονται στο άθροισμα

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

F – test για τους συντελεστές του μοντέλου

Είναι εύκολα αντιληπτό πως το μέγεθος του λόγου ESS / RSS προσδιορίζει και την εγκυρότητα της υπόθεσης H_0 με τις μεγαλύτερες τιμές να καταδεικνύουν απομάκρυνση από αυτήν.
Αποδεικνύεται ότι

$$ESS \sim \chi^2(1), \qquad RSS \sim \chi^2(n-2),$$

συνεπώς, αν

$$F_0 = ESS / \frac{RSS}{n-2}$$

τότε $F \sim F(1, n-2)$, και η πιθανότητα $p = P(F > F_0)$ αποτελεί την πιθανότητα που υπολογίζουμε για να απορρίψουμε ή όχι την μηδενική υπόθεση H_0: $\alpha = b = 0$.

Εύλογο είναι πως αυτή πρέπει να απορριφθεί για να έχει νόημα το γραμμικό μοντέλο που μελετούμε.

Πηγή: http://facweb.cs.depaul.edu/sjost/csc423/documents/f-test-reg.htm

Ασκήσεις

Άσκηση 1 (συνέχεια)

Για τα δεδομένα του πίνακα:

X	1	2	3
Y	3	5	10

κ) Για το γραμμικό μοντέλο που βρέθηκε να ελεγχθεί η υπόθεση

H_0: α = b = 0, έναντι της H_1: όχι η H_0.

Δίνεται: $F_0 = ESS/\frac{RSS}{n-2}$ και $F \sim F(1, n-2)$

$F_0 = \frac{24,5}{\frac{1,5}{1}} = 16,3$. $F \sim F(1,1)$. $p = P(F > 16,3) = 0,154 > 0,05$

άρα η $H_0: a = b = 0$ δεν απορρίπτεται.

Ασκήσεις

Άσκηση 2

Στον παρακάτω πίνακα δίνονται οι βαθμοί προόδου και τελικής εξέτασης για 10 φοιτητές.

Πρόοδος (X)	4	4	7	8	8	2	9	10	5	3
Εξέταση (Y)	7	6	9	10	7	5	8	9	4	5

α) Να γίνει το διάγραμμα διασποράς.

β) Να βρεθεί η συνδιακύμανση των X, Y.

γ) Να βρεθεί ο συντελεστής συσχέτισης των X, Y.

δ) Να βρεθεί η εξίσωση της ευθείας γραμμικής παλινδρόμησης $Y = \alpha + bX$.

ε) Να εκτιμηθεί η τιμή του Y όταν το X θα πάρει την τιμή 4.

ζ) Να υπολογιστεί ένα 95% δ.ε. για τους συντελεστές του γραμμικού μοντέλου.

η) Να ελεγχθούν οι υποθέσεις H_0: $\alpha = 0$ και H_0: $b = 0$ (δίπλευρος έλεγχος).

θ) Να ελεγχθεί η υπόθεση H_0: $\alpha = b = 0$ με τη δοκιμασία F.

Δίνεται: $\bar{x} = 6$, $\bar{y} = 7$, $\sum_{i=1}^{10} x_i y_i = 458$, $\sum_{i=1}^{10} x_i^2 = 428$, $\sum_{i=1}^{10} (x_i - \bar{x})^2 = 68$, $\sum_{i=1}^{10} (y_i - \bar{y})^2 = 36$, $t_{8;0.025} = 2.306$.

Τυπολόγιο

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})}}$$

$$\hat{\alpha} = \bar{y} - \hat{b}\,\bar{x}\ ,\ \hat{b} = \frac{s_{XY}}{s_X^2} = r_{XY}\frac{s_Y}{s_X}$$

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2,\ ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,\ RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,\ R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$\frac{\hat{\alpha} - \alpha}{SE(\hat{\alpha})} \sim t(n-2) \Rightarrow 95\%\ \delta.\varepsilon.: \left(\hat{\alpha} - t_{n-2;0.025}SE(\hat{\alpha}),\ \hat{\alpha} + t_{n-2;0.025}SE(\hat{\alpha})\right)$$

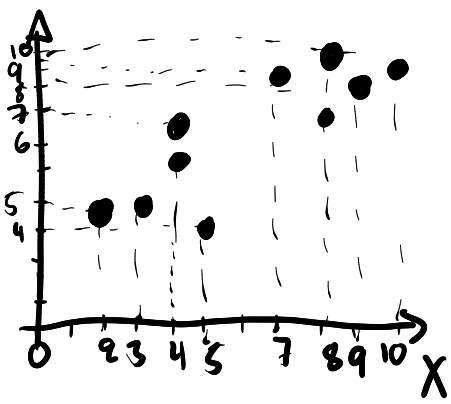
$$\frac{\hat{b} - b}{SE(\hat{b})} \sim t(n-2) \Rightarrow 95\%\ \delta.\varepsilon.: \left(\hat{b} - t_{n-2;0.025}SE(\hat{b}),\ \hat{b} + t_{n-2;0.025}SE(\hat{b})\right)$$

$$SE(\hat{b}) = \sqrt{\frac{RSS}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}},\ SE(\hat{\alpha}) = SE(\hat{b})\sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2}$$

$$F_0 = ESS/\frac{RSS}{n-2},\ F \sim F(1, n-2).$$

$$H_0: b=0:\ t = \frac{\hat{b}}{SE(\hat{b})} \sim t(n-2),\ p = P(|t| > |t_0|)$$

$$H_0: \alpha=0:\ t = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \sim t(n-2),\ p = P(|t| > |t_0|)$$

(a) 

(b) $s_{xy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i \cdot y_i - n\bar{x}\cdot\bar{y}\right] = \frac{1}{9}\cdot(458 - 10\cdot 6\cdot 7) = 4{,}22$

(c) $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{9}\cdot 68 = 7{,}55 \Rightarrow s_x = 2{,}75$

$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{9}\cdot 39 = 4 \Rightarrow s_y = 2$

$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{4{,}22}{2{,}75\cdot 2} = 0{,}767$

(d) $\hat{b} = \frac{s_{xy}}{s_x^2} = \frac{4{,}22}{7{,}55} = 0{,}56$

$\hat{a} = \bar{y} - \hat{b}\cdot\bar{x} = 7 - 0{,}56\cdot 6 = 3{,}65$

$y = 3{,}65 + 0{,}56\cdot x$

(e) Für $x = 4$: $y = 3{,}65 + 0{,}56\cdot 4 = 5{,}9$

(σ) $SE(\hat{b}) = \sqrt{\frac{RSS}{(n-2)\sum(x_i-\bar{x})^2}} = \sqrt{\frac{17{,}12}{8\cdot 68}} = 0{,}18$

$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 17{,}12$ | $SE(\hat{a}) = 0{,}18\cdot\sqrt{\frac{498}{10}} = 1{,}16$

(f) $\hat{b}_0 : \left(0{,}56 - t_{8;0.025}^{2{,}306}\cdot 0{,}18,\ 0{,}56 + t_{8;0.025}^{2{,}306}\cdot 0{,}18\right) = (0{,}14,\ 0{,}97)$.

$\hat{a}$: $(3.65 - 2.306 \cdot 1.16,\ 3.65 + 2.306 \cdot 1.16) = (0.98,\ 6.32)$

(ν) $H_0: b = 0$ έναντι $H_1: b \neq 0$

$t_0 = \frac{\hat{b}}{SE(\hat{b})} = \frac{0.56}{0.18} = 3,1$. $t \sim t(8)$, $p = P(t < -3.1 \text{ ή } t > 3.1) = 0.014$

$p = 0.014 < 0,05 \Rightarrow$ η $H_0: b = 0$ απορρίπτεται

$H_0: a = 0$ έναντι $H_1: a \neq 0$

$t_0 = \frac{\hat{a}}{SE(\hat{a})} = \frac{3,65}{1,16} = 3,14$, $t \sim t(8)$, $p = P(t < -3,14 \text{ ή } t > 3,14) = 0,014$

$p = 0.014 < 0.05 \Rightarrow$ η $H_0: a = 0$ απορρίπτεται.

$ESS = \sum (\hat{y}_i - \bar{y})^2 = 18.87$

(θ) $H_0: a = b = 0$ έναντι $H_1: a \neq 0$ ή $b \neq 0$. $F_0 = \frac{ESS}{RSS/n-2} = \frac{18.87}{17.12/8} = 8.81$. $F \sim F(1,8)$

$p = P(F > 8.81) = 0.017 < 0.05$

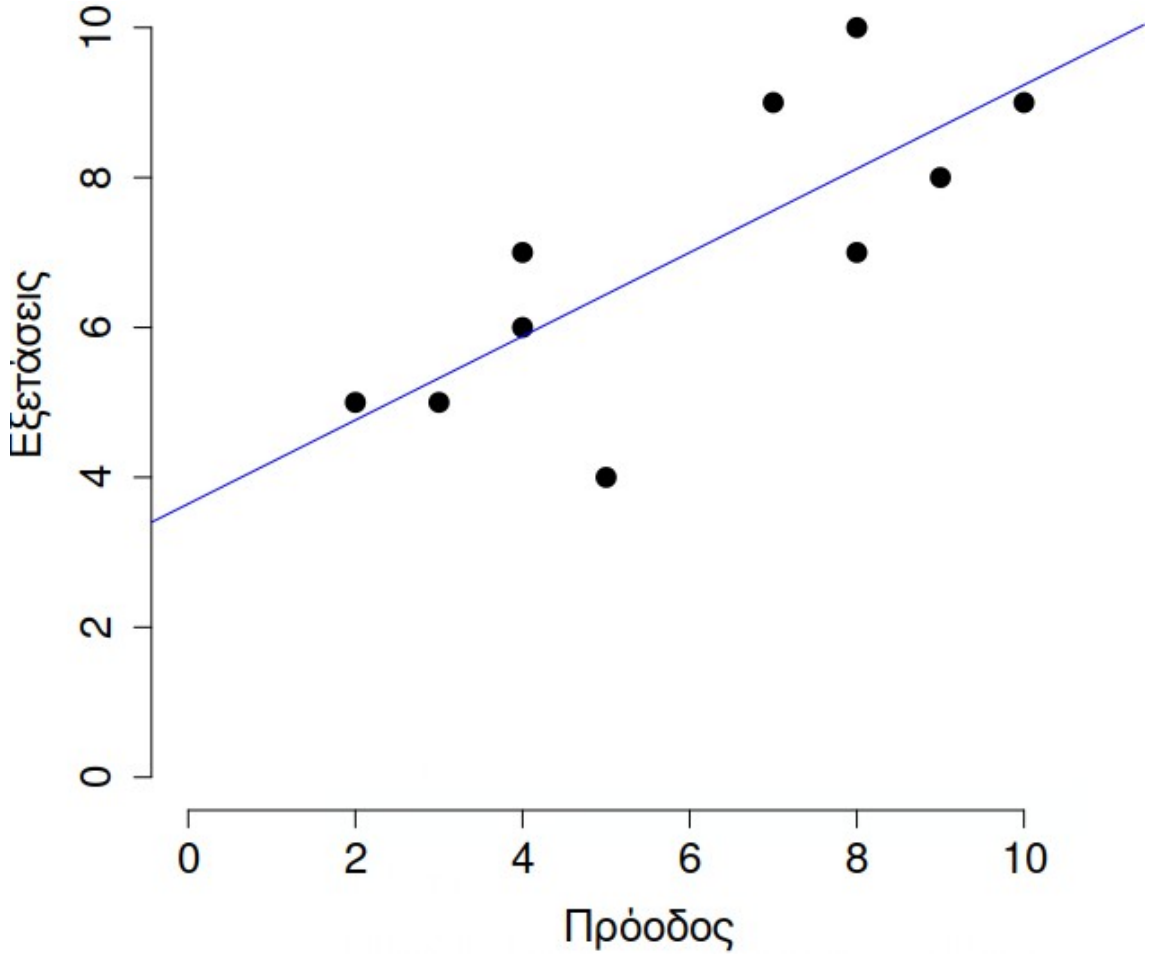
Κώδικας R

```
test = c(4, 4, 7, 8, 8, 2, 9, 10, 5, 3)
exams = c(7, 6, 9, 10, 7, 5, 8, 9, 4, 5)

fit = lm(exams ~ test)

summary(fit)
```

Output

```
Call:
lm(formula = exams ~ test)

Residuals:
    Min      1Q  Median      3Q     Max
-2.44118 -0.58824 -0.05882  0.89706  1.88235

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6471     1.0778   3.384  0.00959 **
test          0.5588     0.1647   3.392  0.00947 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.359 on 8 degrees of freedom
Multiple R-squared:  0.5899,
Adjusted R-squared:  0.5386
F-statistic: 11.51 on 1 and 8 DF,  p-value: 0.009471
```

```
test = c(4, 4, 7, 8, 8, 2, 9, 10, 5, 3); exams = c(7, 6, 9, 10, 7, 5, 8, 9, 4, 5)
plot(test, exams, xlab = "Πρόοδος", ylab = "Εξετάσεις", cex=1.5, cex.axis = 1.5, cex.lab = 1.5,
xlim=c(0,11), ylim=c(0,11), pch = 19, frame = FALSE)
abline(lm(exams ~ test, data = mtcars), col = "blue")
cor(test, exams)
```



Η Απλή Γραμμική Παλινδρόμηση με Άλγεβρα Πινάκων

SLR με Άλγεβρα Πινάκων

Έστω (x_i, y_i), $i = 1, 2, \ldots, n$ τα n ζεύγη παρατηρήσεων. Αναζητούμε συντελεστές b_0, b_1, που ελαχιστοποιούν τα σφάλματα ε_i των εξισώσεων

$$y_1 = b_0 + b_1 x_1 + \varepsilon_1$$

$$y_2 = b_0 + b_1 x_2 + \varepsilon_2$$

$$\ldots$$

$$y_n = b_0 + b_1 x_n + \varepsilon_n$$

Οι εξισώσεις γράφονται $Y = Xb + \varepsilon$, όπου $Y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}$, $X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \ldots & \ldots \\ 1 & x_n \end{bmatrix}$, $b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$, $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}$

Διαστάσεις πινάκων: Y: n x 1, X: n x 2, b: 2 x 1, ε: n x 1.
Y'Y: 1 x 1, b'X'Y: 1 x 1 b'X'Xb: 1 x 1

SLR με Άλγεβρα Πινάκων

Είναι $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$ άρα $\varepsilon'\varepsilon = \begin{bmatrix} \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix} = \sum_{i=0}^{n} \varepsilon_i^2$ και $MSE(b) = \frac{1}{n}\sum_{i=0}^{n} \varepsilon_i^2 = \frac{1}{n}\varepsilon'\varepsilon.$

Επιπλέον, ε'ε = (Y – Xb)'(Y – Xb) = (Y' – b'X')(Y – Xb) = Y'Y – b'X'Y – Y'Xb + b'X'Xb

Είναι: $\frac{\partial(\varepsilon'\varepsilon)}{\partial b} = \frac{\partial}{\partial b}(Y'Y) - \frac{\partial}{\partial b}(b'X'Y) - \frac{\partial}{\partial b}(Y'Xb) + \frac{\partial}{\partial b}(b'X'Xb) = -2X'Y + 2X'Xb$ (2)

και: $\frac{\partial(\varepsilon'\varepsilon)}{\partial b} = 0 \Leftrightarrow \hat{b} = (X'X)^{-1}X'Y$: Εκτιμητής ελαχίστων τετραγώνων.

Σημειώσεις:
(1) Αν α ένας πίνακας 1 x 1, και $b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$, τότε $\frac{\partial \alpha}{\partial b} = \begin{bmatrix} \frac{\partial \alpha}{\partial b_0}, \frac{\partial \alpha}{\partial b_1} \end{bmatrix}$

(2) Ιδιότητες μερικής παραγώγου ως προς πίνακα: https://en.wikipedia.org/wiki/Matrix_calculus#Scalar-by-vector_identities

Υπολογίζουμε $\frac{1}{n}X'Y = \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} y_i \\ \frac{1}{n}\sum_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix}$

$$\frac{1}{n}X'X = \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} = \frac{1}{n}\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix}$$

$$\left(\frac{1}{n}X'X\right)^{-1} = \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix}^{-1} = \frac{1}{\overline{x^2} - \bar{x}^2}\begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} = \frac{1}{s_X^2}\begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

$$b = (X'X)^{-1}X'Y = \left(\frac{1}{n}X'X\right)^{-1}\frac{1}{n}X'Y = \frac{1}{s_X^2}\begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\begin{bmatrix} \bar{y} \\ \overline{xy} \end{bmatrix} = \begin{bmatrix} \bar{y} - \frac{s_{XY}}{s_X^2}\bar{x} \\ \frac{s_{XY}}{s_X^2} \end{bmatrix} \Rightarrow b_0 = \frac{s_{XY}}{s_X^2},\ b_1 = \frac{s_{XY}}{s_X^2}.$$

SLR με Άλγεβρα Πινάκων

$\hat{b} = (X'X)^{-1}X'Y$: Εκτιμητής ελαχίστων τετραγώνων.

Πρόβλεψη (prediction): $\hat{Y} = X\hat{b} = X(X'X)^{-1}X'Y = HY$, όπου $H = X(X'X)^{-1}X'$ (hat matrix)

Ιδιότητες H: $(\alpha)\ H' = H$, $(\beta)\ H^2 = H$.

Σφάλμα πρόβλεψης (residuals): $\text{Res} = e = Y - \hat{Y} = Y - HY = (I - H)Y$

Ιδιότητες $I - H$: $(\alpha)\ (I - H)' = I - H$, $(\beta)\ (I - H)^2 = I - H$.

$$\text{MSE}(\hat{b}) = \frac{1}{n}\sum_{i=0}^{n} e_i^2 = \frac{1}{n}e'e = \frac{1}{n}Y'(I - H)'(I - H)Y = \frac{1}{n}Y'(I - H)Y$$

Άσκηση 1

X	1	2	3
Y	3	5	10

$$3 = b_0 + 1b_1 + \varepsilon_1$$

$$5 = b_0 + 2b_1 + \varepsilon_2$$

$$10 = b_0 + 3b_1 + \varepsilon_n$$

$$Y = Xb + \varepsilon : Y = \begin{bmatrix} 3 \\ 5 \\ 10 \end{bmatrix},\ X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix},\ b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix},\ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix},\ \hat{b} = (X'X)^{-1}X'Y = \begin{bmatrix} -1 \\ 3.5 \end{bmatrix}.$$

Εξίσωση πρόβλεψης: $Y = 3.5\,X - 1$

Κώδικας Octave

```
X = [1, 1;1, 2;1, 3]; Y = [3; 5; 10]
b = inv(transpose(X)*X)*transpose(X)*Y
H = X*inv(transpose(X)*X)*transpose(X)
```

Ασκήσεις

Άσκηση 3

Να βρεθεί η μόνη σωστή από τις παρακάτω προτάσεις που αφορά το γραμμικό μοντέλο Y = α + βX του διαγράμματος (X = Temperature, Y = Thickness).

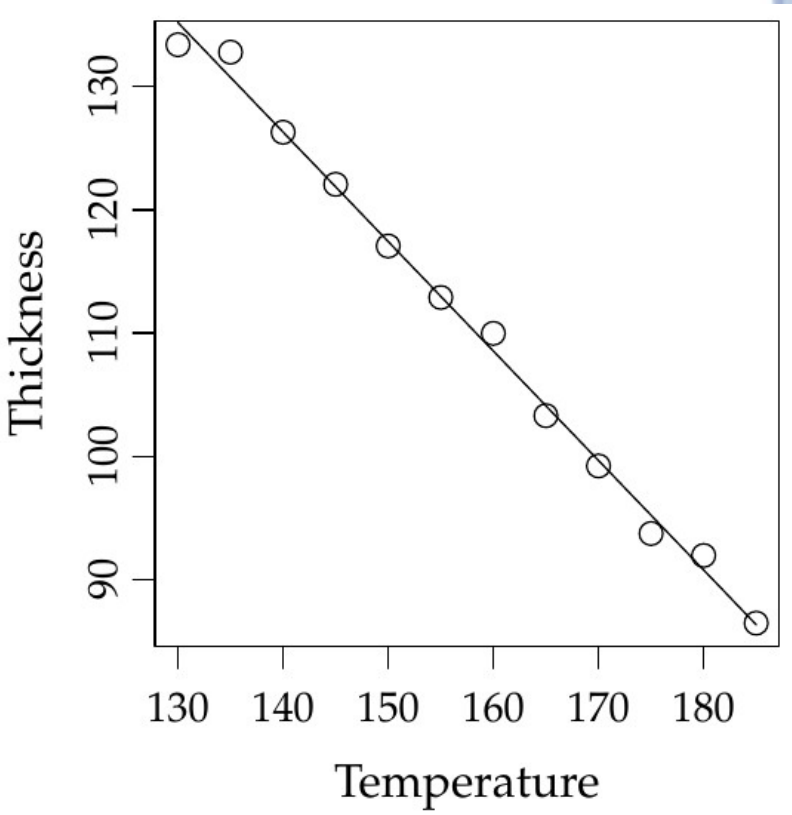
1) α = 0, β = -0,9, RSS = 36

2) α = 0, β = 0,9, RSS = 3,6

3) α = 252, β = -0,9, RSS = 3,6

4) α = -252, β = -0,9, RSS = 36

5) α = 252, β = -0,9, RSS = 36



Ασκήσεις

Άσκηση 4

Να βρεθεί η μόνη σωστή από τις παρακάτω προτάσεις που αφορά το γραμμικό μοντέλο Y = α + βX του διαγράμματος (X = Temperature, Y = Thickness).

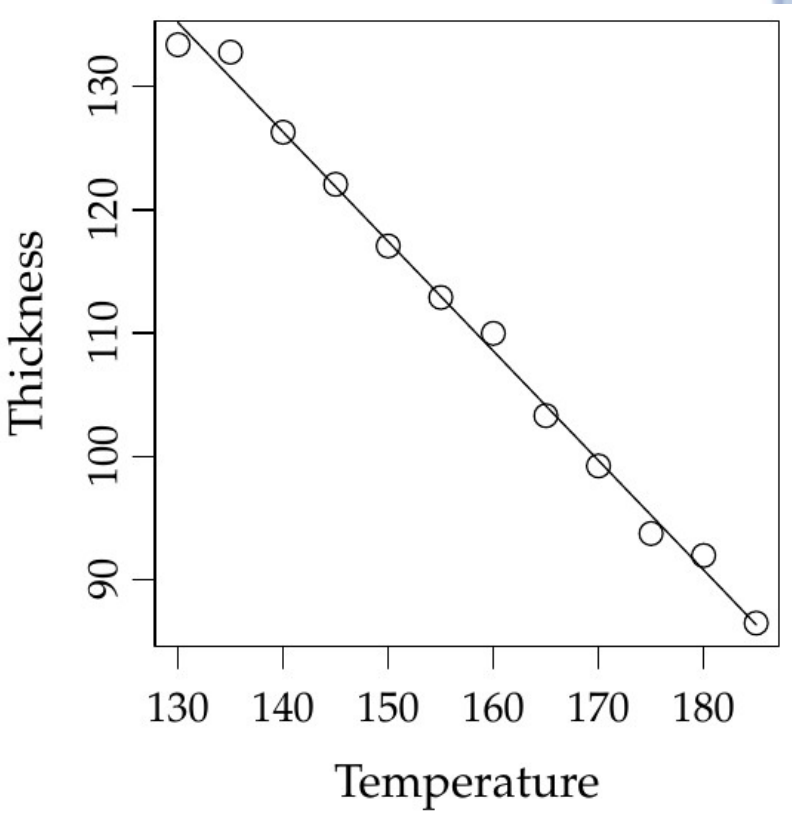
1) Το ποσοστό της επεξηγημένης μεταβλητότητας της Y είναι 50% και η συσχέτιση είναι 0,98

2) Το ποσοστό της επεξηγημένης μεταβλητότητας της Y είναι 0% και η συσχέτιση είναι −0,98

3) Το ποσοστό της επεξηγημένης μεταβλητότητας της Y είναι 96% και η συσχέτιση είναι −1

4) Το ποσοστό της επεξηγημένης μεταβλητότητας της Y είναι 96% και η συσχέτιση είναι 0,98

5) Το ποσοστό της επεξηγημένης μεταβλητότητας της Y είναι 96% και η συσχέτιση είναι −0,98



$R^2 = \frac{ESS}{TSS}$

$R^2 = r_{xy}^2$

Άσκηση 5: Μια εταιρεία κατασκευάζει μια ηλεκτρονική συσκευή για χρήση σε πολύ μεγάλο εύρος θερμοκρασιών. Η εταιρεία γνωρίζει ότι η αυξημένη θερμοκρασία μειώνει τη διάρκεια ζωής της συσκευής και, ως εκ τούτου, πραγματοποίησε μια μελέτη στην οποία μετρήθηκε ο χρόνος ζωής σε συνάρτηση με τη θερμοκρασία. Βρέθηκαν τα ακόλουθα δεδομένα:

Θερμοκρασία (T)	10	20	30	40	50	60	70	80	90
Διάρκεια y	420	365	285	220	176	117	69	34	5

Η προσαρμογή του γραμμικού μοντέλου στην R έγινε με τις εξής εντολές

```
D <- data.frame(t=c(10,20,30,40,50,60,70,80,90), y=c(420,365,285,220,176,117,69,34,5))
fit <- lm(y ~ t, data=D)
summary(fit)
```

Ζητούμενο

Αν L = α + bT, να βρεθεί ένα 95% δ.ε. για το συντελεστή b.

Δίνεται $t_{7;0.025} = 2.365$

Υπόδειξη: $t = \frac{\hat{b}}{SE(\hat{b})} \sim t(n-2)$

Output

```
Call:
lm(formula = y ~ t, data = D)
Residuals:
   Min     1Q Median     3Q    Max
-21.022 -12.622  -9.156  17.711  29.644
Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept) 453.5556    14.3936    31.51   8.38e-09 ***
T            -5.3133     0.2558   -20.77   1.51e-07 ***
---
Multiple R-squared:  0.984,    Adjusted R-squared:  0.9818
F-statistic: 431.5 on 1 and 7 DF,  p-value: 1.505e-07
```

Άσκηση 6. Η απόδοση y μιας χημικής διεργασίας είναι μια τυχαία μεταβλητή της οποίας η τιμή θεωρείται ότι είναι γραμμική συνάρτηση της θερμοκρασίας x. Βρίσκονται τα ακόλουθα δεδομένα αντίστοιχων τιμών των x και y

Θερμοκρασία (T)	0	25	50	75	100
Απόδοση (y)	14	38	54	76	95

Η εκτέλεση του κώδικα

```
exercise5.data <- data.frame(
x=c(0,25,50,75,100),
y=c(14,38,54,76,95))
fit <- lm(y ~ x, data=exercise5.data)
summary(fit)
```

Έδωσε το εξής output:

(α) Τεκμηριώνεται γραμμική σχέση μεταξύ της θερμοκρασίας και της απόδοσης;

(β) Να βρείτε την αναμενόμενη απόδοση στη θερμοκρασία των 80 C^o.

Output

```
Call:
lm(formula = y ~ x, data = exercise5.data)
Residuals:
  1     2     3     4     5
-1.4   2.6  -1.4   0.6  -0.4
Coefficients:
            Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 15.4      1.49666    10.29     0.00196 **
x            0.80     0.02444    32.73     6.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.932 on 3 degrees of freedom
Multiple R-squared:  0.9972,   Adjusted R-squared:  0.9963
F-statistic:  1071 on 1 and 3 DF,  p-value: 6.267e-05
```

Προϋποθέσεις εγκυρότητας και αξιοπιστίας

Προϋποθέσεις εγκυρότητας και αξιοπιστίας

Πριν την αξιοποίηση ενός γραμμικού μοντέλου, υπάρχουν δύο ερωτήματα που πρέπει να απαντηθούν:

1. Είναι ένα έγκυρο εργαλείο για την πρόβλεψη των τιμών της εξαρτημένης μεταβλητής;

2. Είναι η πρόβλεψη αξιόπιστη;

3. Υπάρχουν σημεία με σημαντική ικανότητα “μόχλευσης” της ευθείας;

Και οι τρεις αυτές ερωτήσεις βρίσκουν την απάντησή τους από την παρατήρηση της κατανομής των υπολοίπων.

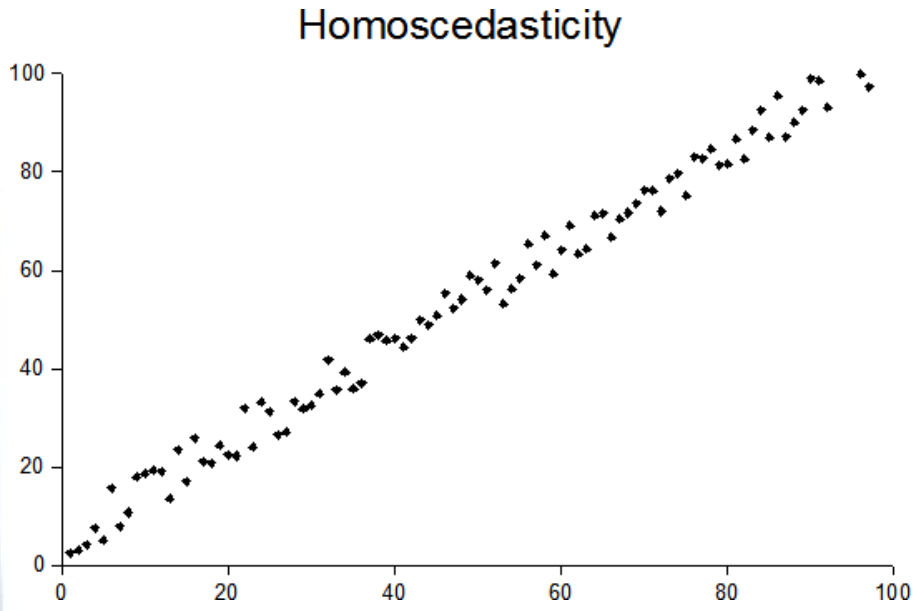
Η εγκυρότητα εξαρτάται από την ομοσκεδαστικότητα των τιμών της Y σε όλο το εύρος της X.

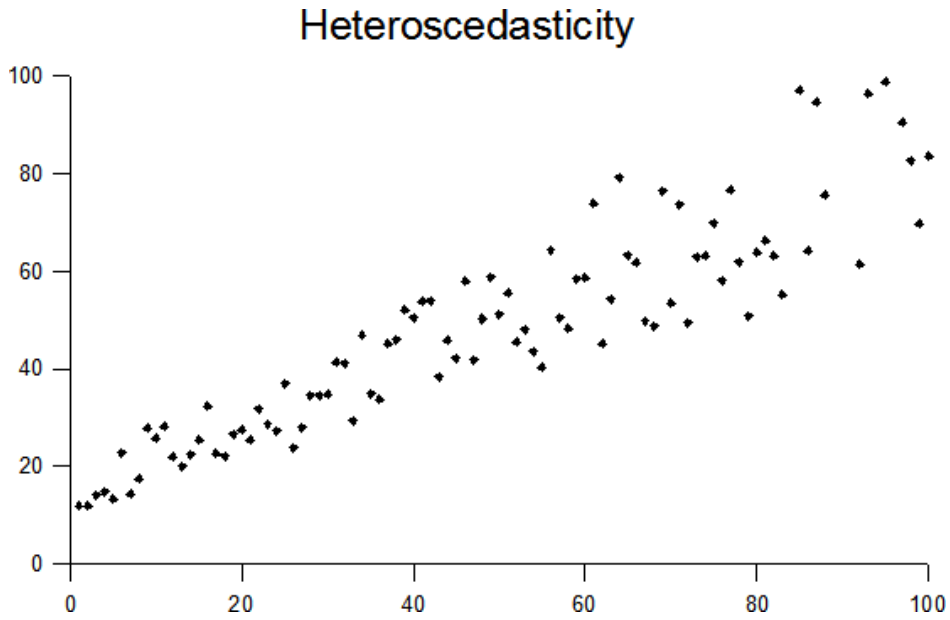
Η αξιοπιστία εξαρτάται από την κανονικότητα της κατανομής των υπολοίπων.

Η ικανότητα μόχλευσης εξαρτάται από τη θέση του σημείου σε σχέση με το “κέντρο” της ευθείας, δηλαδή το σημείο $(\bar{x}, \bar{y})$

Αναπαράσταση: http://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html

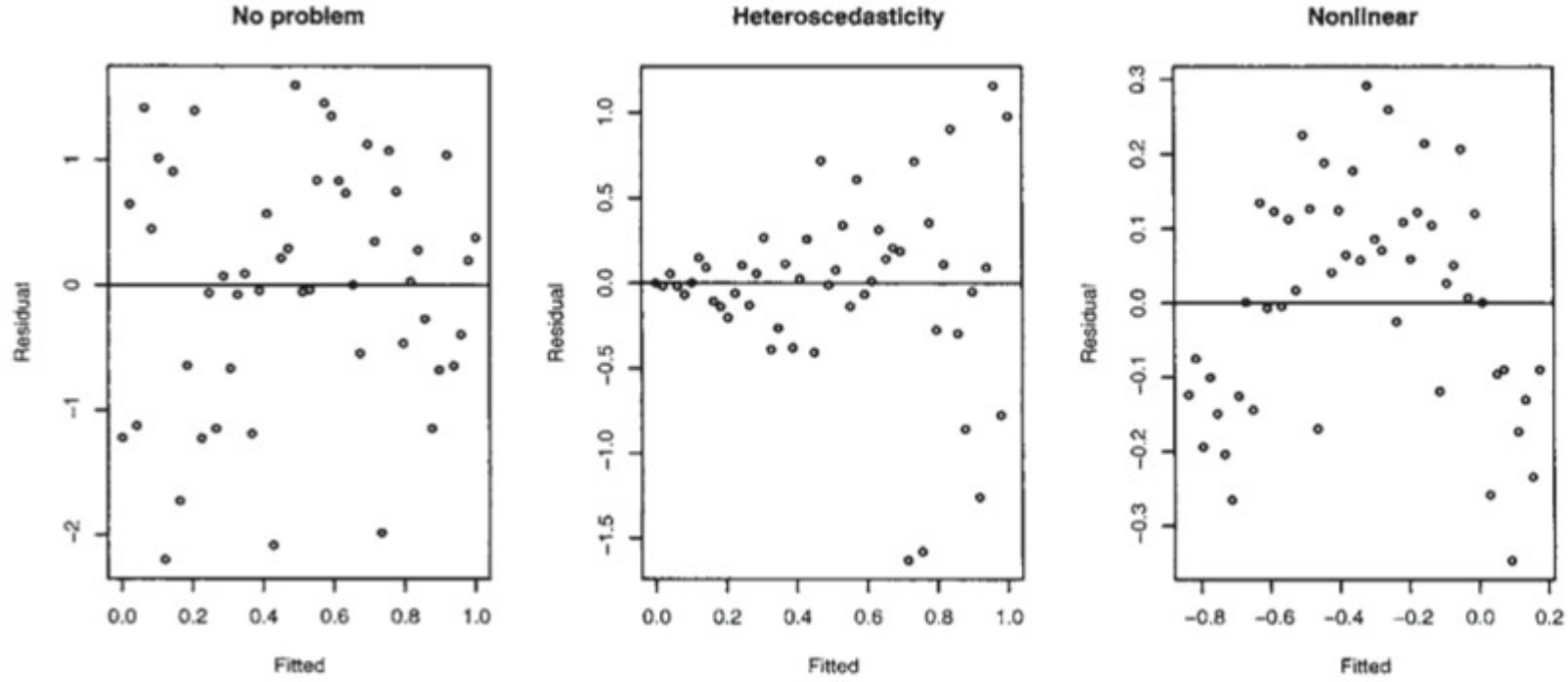
Ομοσκεδαστικότητα (Εγκυρότητα)





Στην περίπτωση της ετεροσκεδαστικότητας υποδεικνύεται πως υπάρχουν και άλλοι παράγοντες που επηρεάζουν την τιμή της Υ και ως εκ τούτου δεν έχει νόημα η χρήση του γραμμικού μοντέλου για την πρόβλεψη των τιμών.

Ομοσκεδαστικότητα (Εγκυρότητα)

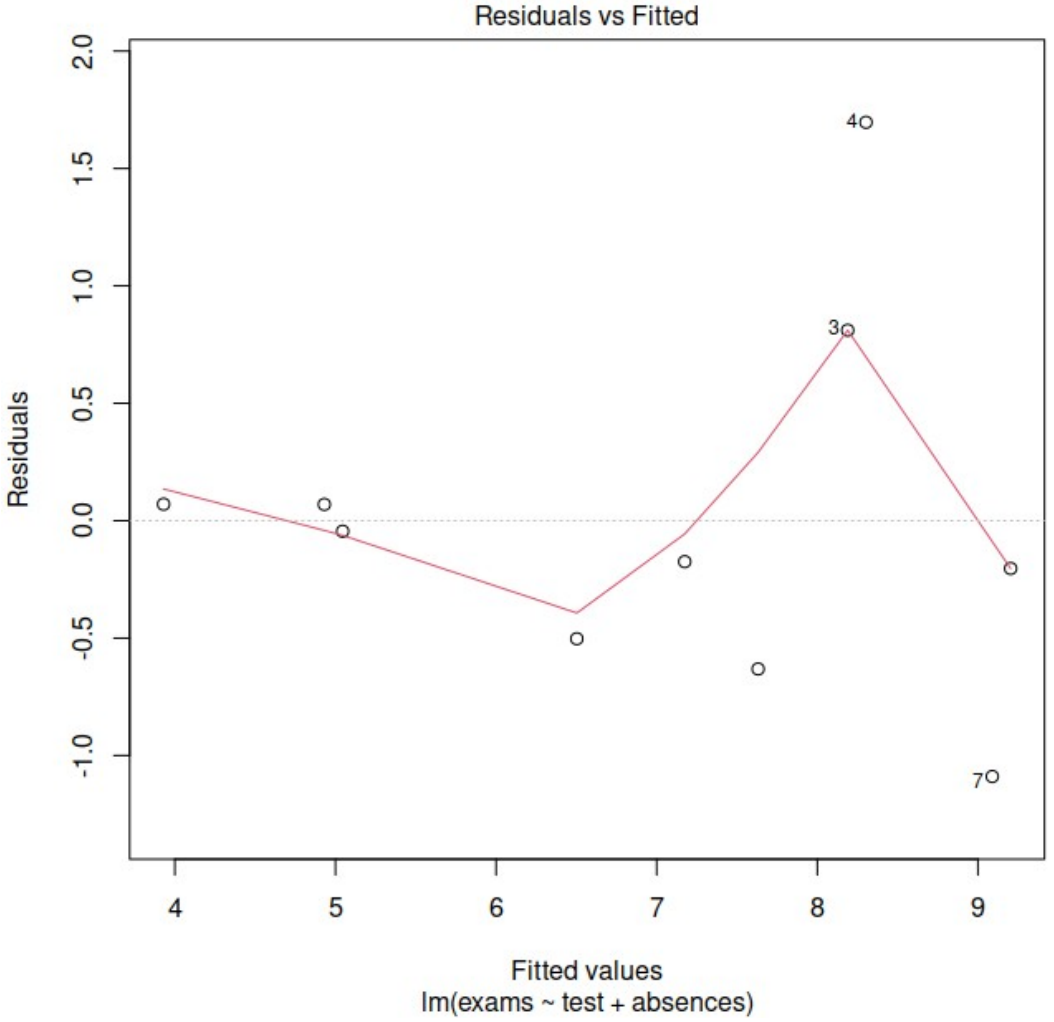


Στο πρώτο διάγραμμα παρατηρούμε πως τα υπόλοιπα έχουν την ίδια κατανομή σε όλο το εύρος των τιμών πρόβλεψης του μοντέλου. Στο δεύτερο διάγραμμα, το εύρος των υπολοίπων αυξάνεται καθώς οι προσαρμοσμένες τιμές (ή ισοδύναμα το x) αυξάνει. Στο τρίτο διάγραμμα τα υπόλοιπα είναι αρνητικά όταν η προσαρμοσμένη τιμή είναι μικρή, θετικά στη μέση και αρνητικά όταν η προσαρμοσμένη τιμή είναι μεγάλη. Αυτό σημαίνει πως οι πραγματικές τιμές της εξαρτημένης είναι κατά σειρά κάτω - πάνω - κάτω από την ευθεία πρόβλεψης υποδεικνύοντας μία μη γραμμική (πιθανώς δεύτεροβάθμια) σχέση με την ανεξάρτητη μεταβλητή. Η ερμηνεία των συντελεστών του μοντέλου και στην περίπτωση αυτή είναι επισφαλής.

Παράδειγμα

Για 10 φοιτητές καταγράφηκε ο βαθμός προόδου, οι απουσίες από το μάθημα και ο βαθμός της τελικής εξέτασης.

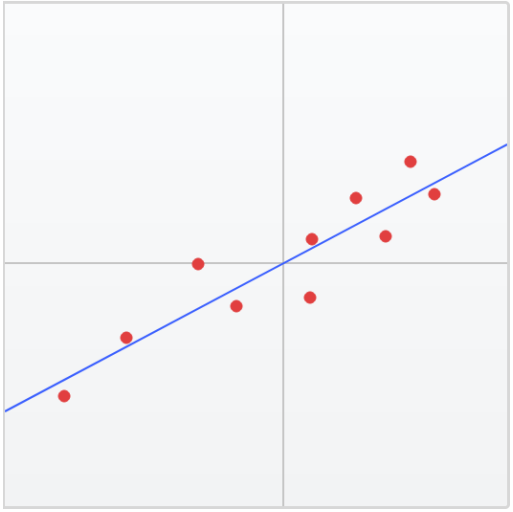
```
test = c(4, 4, 7, 8, 8, 2, 9, 10, 5, 3)
exams = c(7, 6, 9, 10, 7, 5, 8, 9, 4, 5)
absences = c(2, 3, 1, 1, 2, 5, 0, 0, 7, 5)

fit = lm(exams ~ test + absences)
summary(fit)
plot(fit)
```



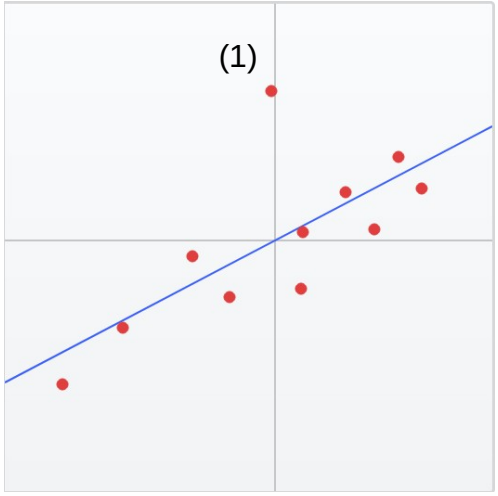
Ανίχνευση σημείων με σημαντική ικανότητα “μόχλευσης”

Ως ικανότητα μόχλευσης περιγράφεται η ιδιότητα κάποιων σημείων (x_i, y_i) να μεταβάλλουν σημαντικά την κλίση της ευθείας και ως εκ τούτου να μειώνουν την αξιοπιστία του υπολογισμού του συντελεστή του x.

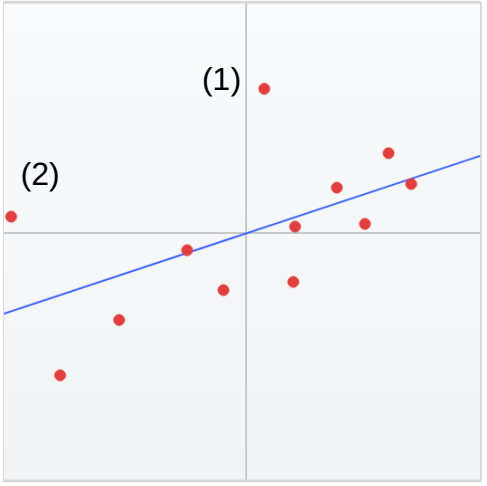
Σημαντική ικανότητα μόχλευσης έχουν τα ιδιάζοντα σημεία που βρίσκονται μακριά από το “κέντρο” της ευθείας ελαχίστων τετραγώνων.



Μικρή ικανότητα μόχλευσης



(1) Ιδιάζον σημείο με μικρή μόχλευση.



(1) Ιδιάζον με μικρή μόχλευση.

(2) Ιδιάζον με μεγάλη μόχλευση.

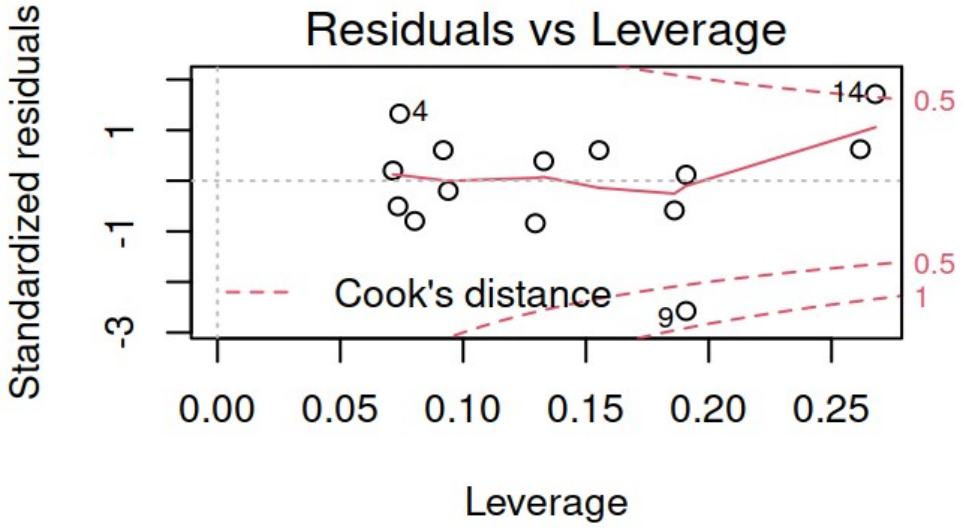
Αναπαράσταση: http://digitalfirst.bfwpub.com/stats_applet/stats_applet_5_correg.html

Ανίχνευση σημείων με σημαντική ικανότητα “μόχλευσης”

Η ικανότητα μόχλευσης μετριέται με διάφορα στατιστικά όπως

α) Το αντίστοιχο στατιστικό (Leverage)[1]

α) Η απόσταση Cook (Cook, 1977)



(1) https://en.wikipedia.org/wiki/Leverage_(statistics)

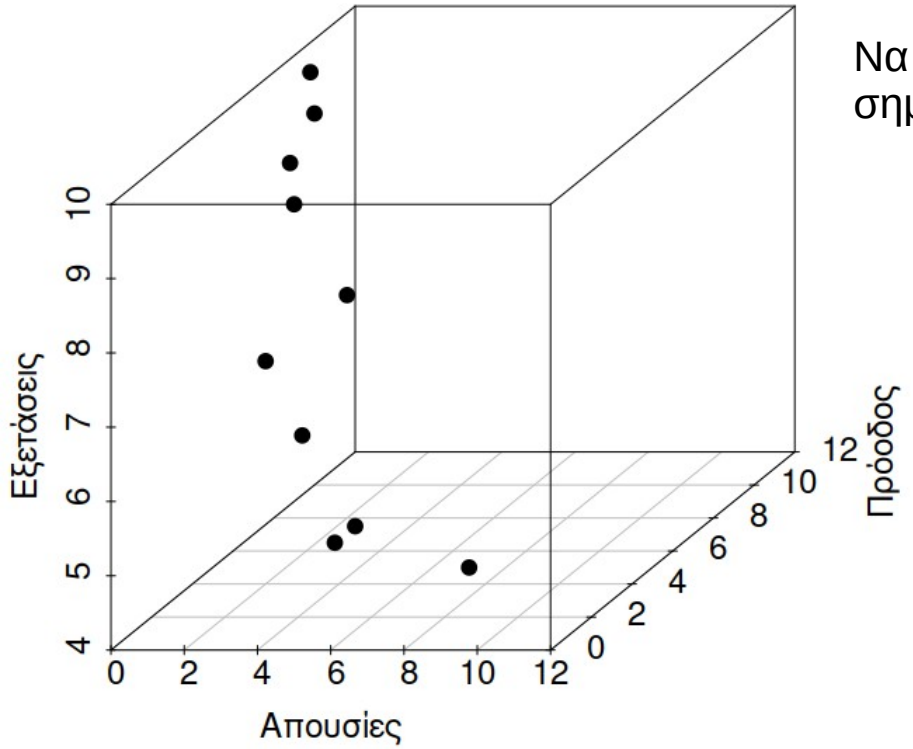
(2) Cook, R. Dennis (February 1977). "Detection of Influential Observations in Linear Regression". Technometrics. American Statistical Association. 19 (1): 15–18. doi:10.2307/1268249. JSTOR 1268249. MR 0436478.

Ευθεία ελαχίστων τετραγώνων στην πολλαπλή γραμμική παλινδρόμηση
(Multivariate Linear Regression)

Παράδειγμα

Για 10 φοιτητές καταγράφηκε ο βαθμός προόδου, οι απουσίες από το μάθημα και ο βαθμός της τελικής εξέτασης.

Πρόοδος	4	4	7	8	8	2	9	10	5	3
Απουσίες	2	3	1	1	2	5	0	0	7	5
Τελική εξέταση	7	6	9	10	7	5	8	9	4	5



Να βρεθεί η εξίσωση του επιπέδου που διέρχεται από τα σημεία (Π, Α, Τ) = (2, 8, 2), (8, 1, 7) και (5, 4, 6).

Α Β Γ

$\vec{AB}$, $\vec{B\Gamma}$ $\vec{n} = \vec{AB} \times \vec{B\Gamma}$

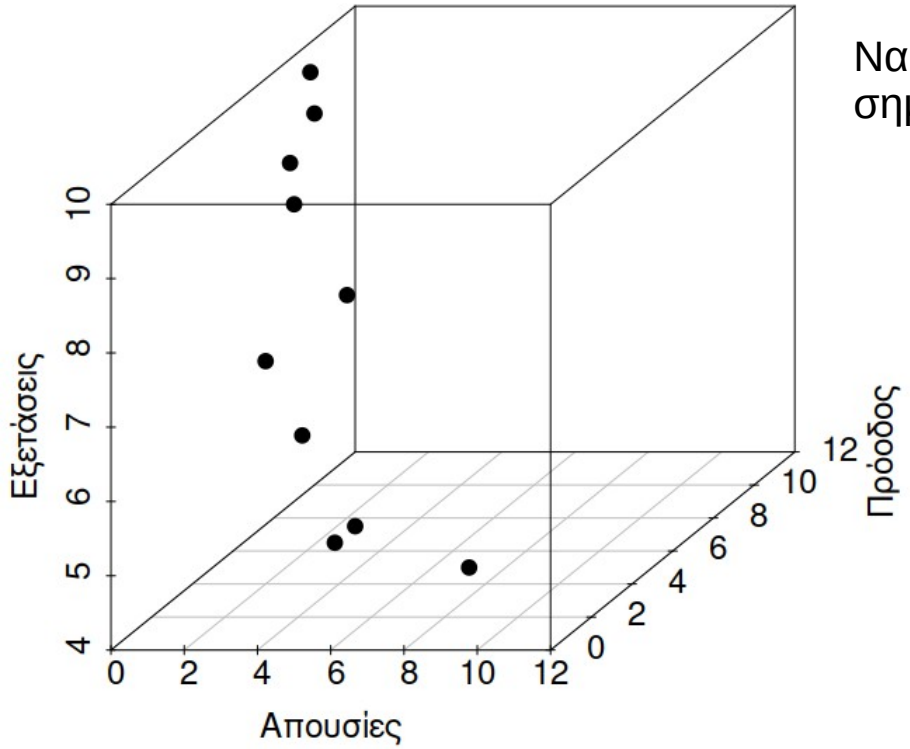
$\vec{AX} \cdot \vec{n} = 0$

$X = (x, y, z)$

Παράδειγμα

Για 10 φοιτητές καταγράφηκε ο βαθμός προόδου, οι απουσίες από το μάθημα και ο βαθμός της τελικής εξέτασης.

Πρόοδος	4	4	7	8	8	2	9	10	5	3
Απουσίες	2	3	1	1	2	5	0	0	7	5
Τελική εξέταση	7	6	9	10	7	5	8	9	4	5



Να βρεθεί η εξίσωση του επιπέδου που διέρχεται από τα σημεία (Π, Α, Τ) = (2, 8, 2), (8, 1, 7) και (5, 4, 6).

Maxima

```
A : matrix([2,8,2,1],[8,6,7,1], [5,4,6,1],[x,y,z,1])
determinant(%);
ratsimp(%);
```

Παράδειγμα

Για 10 φοιτητές καταγράφηκε ο βαθμός προόδου, οι απουσίες από το μάθημα, το φύλο και ο βαθμός της τελικής εξέτασης.

Πρόοδος	4	4	7	8	8	2	9	10	5	3
Απουσίες	2	3	1	1	2	5	0	0	7	5
Φύλο (0 = Γ, 1 = Α)	0	1	0	0	0	1	0	0	1	1
Τελική εξέταση	7	6	9	10	7	5	8	9	4	5

Είναι φανερό πως σε αυτήν την περίπτωση δεν είναι δυνατή η αναπαράσταση του διαγράμματος διασποράς των τιμών, ωστόσο εξακολουθεί να είναι χρήσιμη η πρόβλεψη του βαθμού της τελικής εξέτασης από ένα γραμμικό μοντέλο της μορφής:

$$\text{Τελική εξέταση} = \alpha \cdot \text{Πρόοδος} + \beta \cdot \text{Απουσίες} + \gamma \cdot \text{Φύλο} + \delta$$

Πολλαπλή Γραμμική Παλινδρόμηση

Έστω $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, 2, \ldots, n$ οι n παρατηρήσεις. Αναζητούμε συντελεστές b_0, $b_1, \ldots, b_k$, που ελαχιστοποιούν τα σφάλματα ε_i των εξισώσεων

$$y_1 = b_0 + b_1x_{11} + b_2x_{12} + \ldots + b_kx_{1k} + \varepsilon_1$$

$$y_2 = b_0 + b_1x_{21} + b_2x_{22} + \ldots + b_kx_{2k} + \varepsilon_2$$

$$\ldots$$

$$y_n = b_0 + b_1x_{n1} + b_2x_{n2} + \ldots + b_kx_{nk} + \varepsilon_n$$

Οι εξισώσεις γράφονται $Y = Xb + \varepsilon$, όπου

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ b_1 \\ \ldots \\ b_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}$$

Διαστάσεις πινάκων: Y: $n \times 1$, X: $n \times (k + 1)$, b: $(k+1) \times 1$, ε: $n \times 1$.
$Y'Y$: 1×1, $b'X'Y$: 1×1 $b'X'Xb$: 1×1

Πολλαπλή Γραμμική Παλινδρόμηση

Είναι $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$ άρα $\varepsilon'\varepsilon = \begin{bmatrix} \varepsilon_1, \varepsilon_2, \dots, \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix} = \sum_{i=0}^{n} \varepsilon_i^2$ και $MSE(b) = \frac{1}{n} \sum_{i=0}^{n} \varepsilon_i^2 = \frac{1}{n} \varepsilon'\varepsilon .$

Επιπλέον, ε'ε = (Y – Xb)'(Y – Xb) = (Y' – b'X')(Y – Xb) = Y'Y – b'X'Y – Y'Xb + b'X'Xb

Είναι: $\frac{\partial(\varepsilon'\varepsilon)}{\partial b} = \frac{\partial}{\partial b}(Y'Y) - \frac{\partial}{\partial b}(b'X'Y) - \frac{\partial}{\partial b}(Y'Xb) + \frac{\partial}{\partial b}(b'X'Xb) = -2X'Y + 2X'Xb$

και: $\frac{\partial(\varepsilon'\varepsilon)}{\partial b} = 0 \Leftrightarrow \hat{b} = (X'X)^{-1}X'Y$: Εκτιμητής ελαχίστων τετραγώνων.

Πολλαπλή Γραμμική Παλινδρόμηση

$\hat{b} = (X'X)^{-1}X'Y$: Εκτιμητής ελαχίστων τετραγώνων.

Πρόβλεψη (prediction): $\hat{Y} = X\hat{b} = X(X'X)^{-1}X'Y = HY$, όπου $H = X(X'X)^{-1}X'$ (hat matrix)

Ιδιότητες H: $(\alpha)\ H' = H$, $(\beta)\ H^2 = H$.

Σφάλμα πρόβλεψης (residuals): $\text{Res} = e = Y - \hat{Y} = Y - HY = (I - H)Y$

Ιδιότητες $I - H$: $(\alpha)\ (I - H)' = I - H$, $(\beta)\ (I - H)^2 = I - H$.

$$\text{MSE}(\hat{b}) = \frac{1}{n}\sum_{i=0}^{n} e_i^2 = \frac{1}{n}e'e = \frac{1}{n}Y'(I - H)'(I - H)Y = \frac{1}{n}Y'(I - H)Y$$

Παράδειγμα

Πρόοδος	Απουσίες	Τελική εξέταση
4	2	7
4	3	6
7	1	9
8	1	10
8	2	7
2	5	5
9	0	8
10	0	9
5	7	4
3	5	5

$$Y = \begin{bmatrix} 7 \\ 6 \\ 9 \\ 10 \\ 7 \\ 5 \\ 8 \\ 9 \\ 4 \\ 5 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 4 & 2 \\ 1 & 4 & 3 \\ 1 & 7 & 1 \\ 1 & 8 & 1 \\ 1 & 8 & 2 \\ 1 & 2 & 5 \\ 1 & 9 & 0 \\ 1 & 10 & 0 \\ 1 & 5 & 7 \\ 1 & 3 & 5 \end{bmatrix}, \quad X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 4 & 4 & 7 & 8 & 8 & 2 & 9 & 10 & 5 & 3 \\ 2 & 3 & 1 & 1 & 2 & 5 & 0 & 0 & 7 & 5 \end{bmatrix}$$

$$b = (X'X)^{-1}X'Y = \begin{bmatrix} 8,0612 \\ 0,1142 \\ -0,6718 \end{bmatrix}$$

Κώδικας Octave

```
X = [1, 4, 2;1, 4, 3;1, 7, 1;1, 8, 1;1, 8, 2;1, 2, 5;1, 9, 0;1, 10, 0;1, 5, 7;1, 3, 5]
Y = [7; 6; 9; 10; 7; 5; 8; 9; 4; 5]
b = inv(transpose(X)*X)*transpose(X)*Y
```

Παράδειγμα

Πρόοδος	4	4	7	8	8	2	9	10	5	3
Απουσίες	2	3	1	1	2	5	0	0	7	5
Τελική εξέταση	7	6	9	10	7	5	8	9	4	5

Χρησιμοποιήθηκε η μέθοδος της γραμμικής παλινδρόμησης για την εκτίμηση του βαθμού της τελικής εξέτασης από το βαθμό προόδου και το πλήθος απουσιών.

Κώδικας R

```
test = c(4, 4, 7, 8, 8, 2, 9, 10, 5, 3)
exams = c(7, 6, 9, 10, 7, 5, 8, 9, 4, 5)
absences = c(2, 3, 1, 1, 2, 5, 0, 0, 7, 5)

fit = lm(exams ~ test + absences)
summary(fit)

cor(test, absences)
```

Ζητούμενο

Να σχολιάσετε τα ευρήματα της μεθόδου

Παράδειγμα

Output

```
Call:
lm(formula = exams ~ test + absences)
Residuals:
     Min       1Q   Median       3Q      Max
-1.08943 -0.42801 -0.10976  0.06996  1.69662
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.0612     1.4572   5.532 0.000876 ***
test          0.1142     0.1674   0.682 0.516901
absences     -0.6718     0.1945  -3.455 0.010623 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.883 on 7 degrees of freedom
Multiple R-squared:  0.8484,   Adjusted R-squared:  0.8051
F-statistic: 19.58 on 2 and 7 DF,  p-value: 0.001357

> cor(test, absences)
[1] -0.7686751
```

Προσαρμοσμένος συντελεστής προσδιορισμού (R^2_{adj})

Όσο περισσότερες μεταβλητές χρησιμοποιούνται για την πρόβλεψη της εξαρτημένης μεταβλητής, τόσο καλύτερη θα είναι η προσαρμογή του μοντέλου και τόσο πιο μεγάλη θα είναι η τιμή του R^2. (1) Για την αντιμετώπιση αυτής της αυξητικής τάσης, έχει επικρατήσει η χρήση και αναφορά και του προσαρμοσμένου συντελεστή προσδιορισμού (adjusted R^2) ο οποίος ορίζεται ως (k: το πλήθος των επεξηγηματικών μεταβλητών)(2)

$$R^2_{adj} = 1 - \frac{RSS/df_{RSS}}{TSS/df_{TSS}} = 1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Περίπτωση SLR: $R^2_{adj} = 1 - \frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$

(1) https://en.wikipedia.org/wiki/Coefficient_of_determination#Inflation_of_R2
(2) Μία απόδειξη είναι διαθέσιμη εδώ: https://statproofbook.github.io/P/rsq-der.html

Συγγραμικότητα

Η συγγραμικότητα των ανεξάρτητων μεταβλητών δεν μειώνει τη δυνατότητα πρόβλεψης της εξαρτημένης μεταβλητής ωστόσο αφαιρεί τη δυνατότητα της ερμηνείας του μεγέθους των συντελεστών του γραμμικού μοντέλου. Το παρακάτω απλό παράδειγμα εξηγεί το λόγο. Ας υποθέσουμε πως:

$$Y = 4 + 2*X_1 + 5*X_2$$

είναι ένα γραμμικό μοντέλο πρόβλεψης της Y από τις ανεξάρτητες X_1, X_2 . Από αυτό παίρνουμε τις πληροφορίες:

- Αν $X_1 = 0$, $X_2 = 0$ τότε $Y = 4$.
- Κάθε μία μονάδα αύξηση στη X_1 αντιστοιχεί σε 2 μονάδες αύξηση στο Y
- Κάθε μία μονάδα αύξηση στη X_2 αντιστοιχεί σε 5 μονάδες αύξηση στο Y

Συγγραμικότητα

Ας υποθέσουμε επιπλέον πως οι X_1, X_2 είναι συγγραμικές, δηλαδή υπάρχει κάποια γραμμική σχέση που τις συνδέει όπως η:

$$X_2 = 2 + X_1,$$

από όπου εύκολα γράφουμε $2 = X_2 - X_1$, ή ακόμα $2*\kappa = \kappa*X_2 - \kappa*X_1$, για κάθε ακέραιο αριθμό κ. Με την εξίσωση αυτή και προσθαφαιρώντας πολλαπλάσια του 2 στην αρχική εξίσωση πρόβλεψης της Υ μπορούμε να τη μεταβάλλουμε ισοδύναμα με απεριόριστους τρόπους οι οποίοι θα έχουν τη γενική μορφή:

$Y = 4 - 2*\kappa + (2 - \kappa) *X_1 + (5 + \kappa)*X_2$, όπως για παράδειγμα:

$Y = 2 + X_1 + 6*X_2$, $(\kappa = 1)$

$Y = 6 + 3* X_1 + 4*X_2$, $(\kappa = -1)$.

Φανερά, στο παραπάνω πλαίσιο *δεν έχει νόημα* η ερμηνεία των συντελεστών των ανεξάρτητων μεταβλητών. Στην πράξη, όταν ανιχνεύεται συγγραμικότητα μεταξύ των ανεξάρτητων μεταβλητών σε ένα γραμμικό μοντέλο πρόβλεψης προτείνεται η εξαγωγή μίας από τις συγγραμικές μεταβλητές και η επανάληψη της διαδικασίας με τις υπόλοιπες.