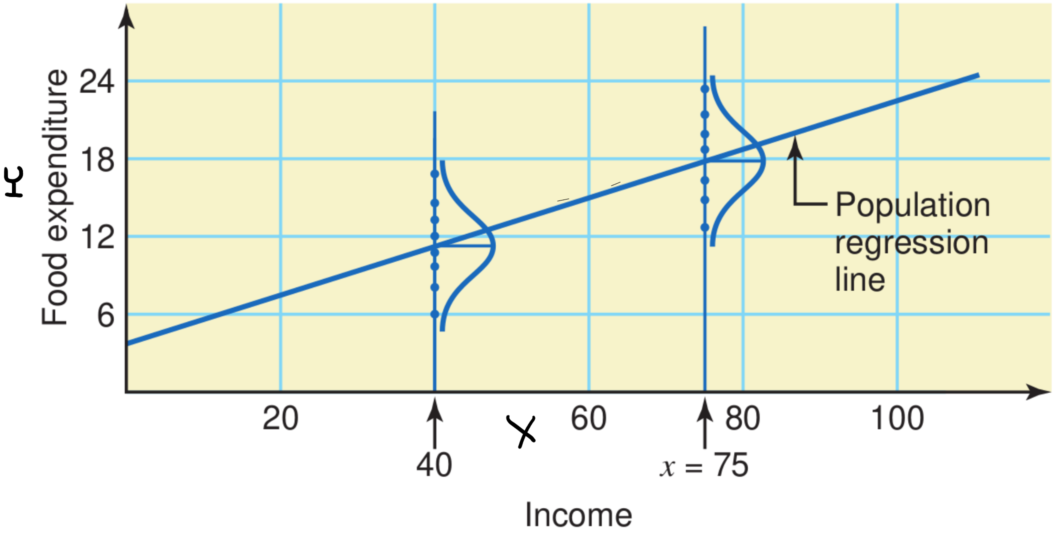
MEM-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@pm.me)

15-03-2023


Y
Food expenditure
24
18
12
6
20
40
X
60
80
x = 75
100
Income
Population
regression
line

Δειγματικό μοντέλο απλής γραμμικής παλινδρόμησης

$Y = A + BX$

$$\hat{y} = a + bx$$

- a είναι δειγματική προσέγγιση του A
- b είναι δειγματική προσέγγιση του B
- $\hat{y}$ είναι η εκτιμώμενη τιμή του y για δοσμένο x

Τυχαίο σφάλμα του δειγματικού μοντέλου απλής γραμμικής παλινδρόμησης

$$e = y - \hat{y}$$

Έστω το τυχαίο δείγμα

ανεξάρτητη $\qquad \alpha, b \qquad \hat{y} = a + bx$

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \longrightarrow (x_i, y_i) \longrightarrow \hat{y}_i$$

εξαρτημένη

Για το τυχαίο σφάλμα του δειγματικού μοντέλου απλής γραμμικής παλινδρόμησης έχουμε:

Πραγματοποίηση της τ.μ. e

$$e_n = y_n - \hat{y}_n, \quad n = 1, \ldots, N$$

όπου η προσέγγιση του κάθε y_n δίνεται ώς

$$\hat{y}_n = a + bx_n \quad , n = 1, \ldots, N$$

Άθροισμα τετραγωνικών σφαλμάτων

$$\text{SSE} = \sum_{n=1}^{N} e_n^2$$

Άθροισμα τετραγωνικών σφαλμάτων συναρτήσει των παραμέτρων του δειγματικού μοντέλου

$$SSE = \sum e_n^2 = \sum (y_n - \hat{y}_n)^2 = \sum (y_n - a - b x_n)^2$$

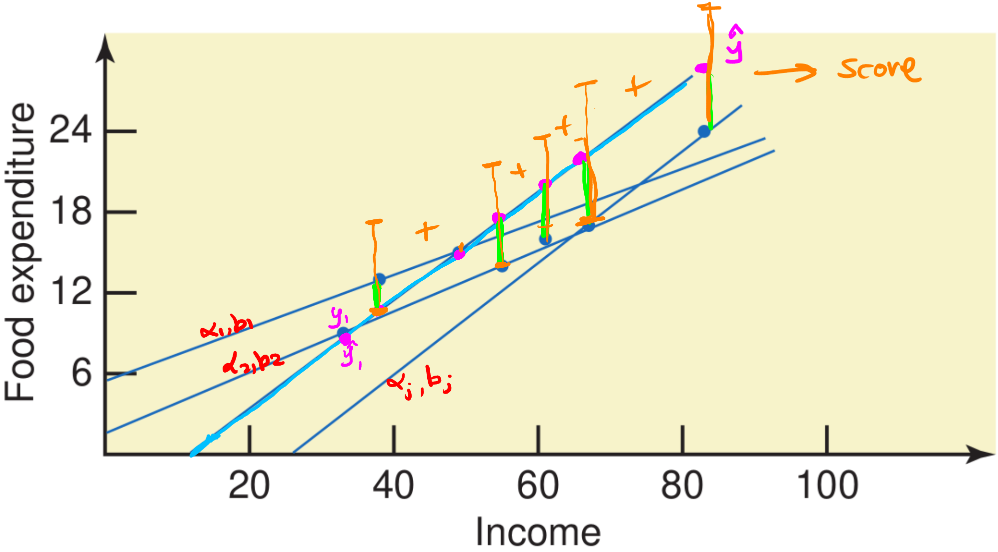
$$\mathcal{Q}(a, b) = \mathsf{SSE} = \sum_{n=1}^{N} (y_n - a - bx_n)^2$$

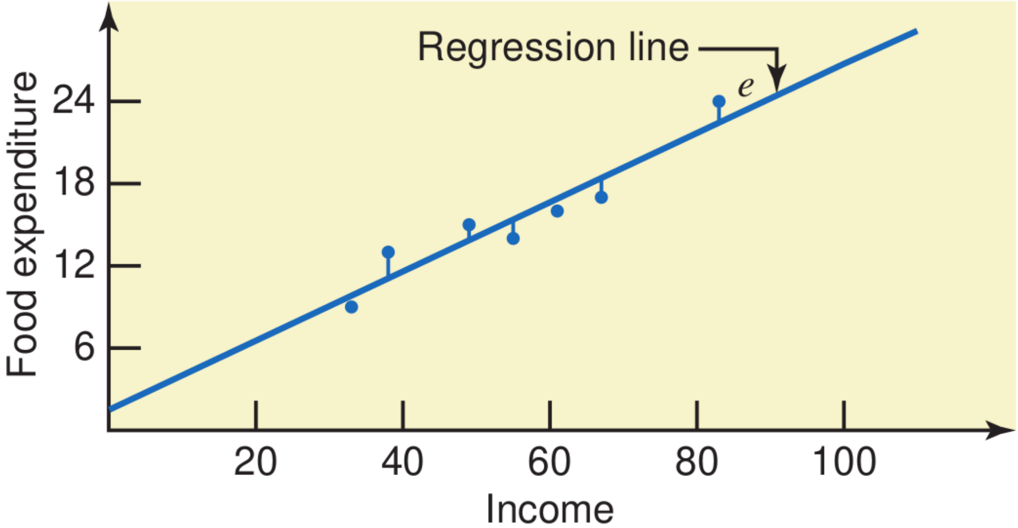
Εκτίμηση ελαχίστων τετραγώνων

Ως εκτίμησεις των a, b λαμβάνουμε τις τιμές a^*, b^* που ελαχιστοποιούν το άθροισμα των τετραγωνικών σφαλμάτων.

$$a, b = \arg\min_{a', b'} \mathcal{Q}(a', b')$$

$a, b \in \mathbb{R}$ τ.ω. $Q(a', b') \geq Q(a, b) \ \forall a', b' \in \mathbb{R}$


Food expenditure
24
18
12
6
20
40
60
80
100
Income
score
α_1, b_1
α_2, b_2
α_j, b_j
y_1
$\hat{y}_1$
$\hat{y}$


Regression line
e
Food expenditure
24
18
12
6
20
40
60
80
100
Income

$$\mathcal{Q}(\mathrm{a},\mathrm{b}) = \sum_{\mathrm{n}=1}^{\mathrm{N}} (\mathrm{y_n} - \mathrm{a} - \mathrm{bx_n})^2$$

$$\frac{\partial}{\partial a}Q = -\sum_{n=1}^{N} 2(y_n - a - bx_n) \quad , \text{θέλουμε} \quad \frac{\partial Q}{\partial a} = 0 \Rightarrow$$

$$\Rightarrow \sum_{n=1}^{N}(y_n - a - bx_n) = 0 \Rightarrow \sum_{n=1}^{N} y_n - Na - b\sum_{n=1}^{N} x_n = 0 \xRightarrow{/N}$$

$$\Rightarrow \frac{1}{N}\sum_{n=1}^{N} y_n - b\frac{1}{N}\sum_{n=1}^{N} x_n = \alpha \Rightarrow \boxed{\alpha = \bar{Y} - b\bar{X}}$$

$$\frac{\partial Q}{\partial b} = -\sum_{n=1}^{N} 2(y_n - a - bx_n)x_n = 0$$

$$\sum_{n=1}^{N}(x_n y_n - \alpha x_n - b x_n^2) = 0 \Rightarrow \sum_{n=1}^{N}(x_n y_n - \bar{Y}x_n + b\bar{X}x_n - b x_n^2) = 0$$

$$\Rightarrow \quad \sum_{n=1}^{N} x_n y_n - \bar{y}\sum_{n=1}^{N} x_n + b\bar{x}\sum_{n=1}^{N} x_n - b\sum_{n=1}^{N} x_n^2 = 0$$

$$\sum_{n=1}^{N} x_n y_n - \frac{1}{N}\sum_{n=1}^{N} x_n \sum_{n=1}^{N} y_n + b\left[\frac{1}{N}\left(\sum_{n=1}^{N} x_n\right)^2 - \sum_{n=1}^{N} x_n^2\right] = 0$$

$$\Rightarrow \quad \boxed{b = \frac{\sum_{n=1}^{N} x_n y_n - \frac{1}{N}\sum_{n=1}^{N} x_n \sum_{n=1}^{N} y_n}{\sum_{n=1}^{N} x_n^2 - \frac{1}{N}\left(\sum_{n=1}^{N} x_n\right)^2}}$$

$$\hat{y} = a + bx$$

$$b = \frac{SS_{xy}}{SS_{xx}}, \quad a = \bar{Y} - b\bar{X}$$

όπου SS_{xy}, SS_{xx} δίνονται ως:

$$SS_{xy} = \sum_{n=1}^{N} x_n y_n - \frac{(\sum_{n=1}^{N} x_n)(\sum_{n=1}^{N} y_n)}{N}, \quad SS_{xx} = \sum_{n=1}^{N} x_n^2 - \frac{(\sum_{n=1}^{N} x_n)^2}{N}$$

Επιπλέον τα SS_{xy} και SS_{xx} μπορούν ισοδύναμα να υπολογισθούν ως:

$$SS_{xy} = \sum_{n=1}^{N}(x_n - \bar{X})(y_n - \bar{Y}), \quad SS_{xx} = \sum_{n=1}^{N}(x_n - \bar{X})^2$$

$$\sum_{n=1}^{N}(x_n - \bar{X})(y_n - \bar{Y}) \rightarrow \sum_{n=1}^{N} x_n y_n - \frac{(\sum_{n=1}^{N} x_n)(\sum_{n=1}^{N} y_n)}{N}$$

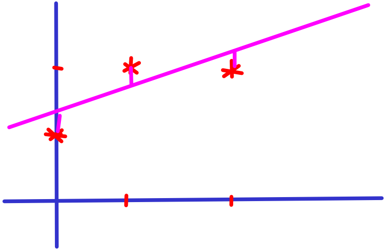
$$\parallel$$

$$\sum x_n y_n - \bar{Y} \sum x_n - \bar{X} \sum y_n + N \bar{X} \bar{Y} =$$

$$\sum x_n y_n - \frac{1}{N} \sum y_n \sum x_n - \cancel{\frac{1}{N} \sum x_n \sum y_n} + \cancel{N \frac{1}{N} \sum x_n \frac{1}{N} \sum y_n}$$

Παράδειγμα

Βρείτε τη εκτίμηση ελαχίστων τετραγώνων του μοντέλου γραμμικής παλινδρόμησης υποθέτοντας τα παρακάτω δεδομένα.



$$\{(0,1),(1,2),(2,2)\}$$

	x	y	xy	x^2
	0	1	0	0
	1	2	2	1
	2	2	4	4
Sum	3	5	6	5

$$S_{xx} = \sum x_n^2 - \frac{1}{N}\left(\sum x_n\right)^2 = 5 - \frac{1}{3}3^2 = 2$$

$$S_{xy} = \sum x_n y_n - \frac{1}{N}\left(\sum x_n\right)\left(\sum y_n\right) = 6 - \frac{1}{3}3 \cdot 5 = 1$$

$$\left.\begin{matrix} S_{xx} = 2 \\ S_{xy} = 1 \end{matrix}\right\} b = \frac{1}{2}$$

$$a = \bar{Y} - b\bar{X} = \frac{5}{3} - \frac{1}{2}\frac{3}{3} = \frac{7}{6}$$

$$\boxed{\hat{y} = 7/6 + 1/2\, x}$$

$\hat{y}(0) = 7/6 \qquad y(0) = 1$

$\hat{y}_1 = 7/6 \qquad y_1 = 1$

$e_1 = 1 - 7/6 = -1/6$

$\hat{y}_2 = \hat{y}(1) = 7/6 + 1/2 = \frac{10}{6}$

$y_2 = 2 \qquad e_2 = 2 - \frac{10}{6} = \frac{1}{3}$

$\hat{y}_3 = \hat{y}(2) = 7/6 + 1 = 13/6$

$y_3 = 2 \qquad e_3 = 2 - \frac{13}{6} = -\frac{1}{6}$

Άσκηση

Βρείτε τη εκτίμηση ελαχίστων τετραγώνων του μοντέλου γραμμικής παλινδρόμησης υποθέτοντας τα παρακάτω δεδομένα.

$$\{(0,2),(1,1),(1,2),(2,4)\}$$

X, Y τυχαίες μεταβλητές

$$Y = A + BX + \epsilon, \quad \epsilon : \text{όρος τυχαίου σφάλματος}$$

$\hat{Y} = \alpha + bX + e$ $\quad S_e^2$ – διασπορά του e.

- Για κάθε x έχουμε υποθέσει ότι το σφάλμα ϵ ακολουθεί την κανονική κατανομή $\varepsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
- Η τυπική απόκλιση σ_ϵ του τυχαίου σφάλματος αναφέρεται στο πληθυσμό και κατά επέκταση η τιμή της δεν είναι γνωστή στις περισσότερες περιπτώσεις.

Εκτιμήτρια της τυπικής απόκλισης των σφαλμάτων

$\mu_x \to \bar{X}$

$\mu_y \to \bar{Y}$

e_n εξαρτάται από $\bar{X}, \bar{Y}$

χάνω 2 βαθμούς ελευθερ..

$$s_e = \sqrt{\frac{\text{SSE}}{N-2}}, \quad \text{SSE} = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$$

$\alpha = \bar{Y} - b\bar{X}$

$$s_e = \sqrt{\frac{\mathsf{SSE}}{N-2}}, \quad \mathsf{SSE} = \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$$

Γιατί εμφανίζεται το $N-2$;

$$SSE = \sum (e_n - 0)^2$$

Τυπική Απόκλιση των Τυχαίων Σφαλμάτων

$$s_e = \sqrt{\frac{SSE}{N-2}}$$

$$\mathrm{s_e} = \sqrt{\frac{\mathrm{SS_{yy}} - \mathrm{b} * \mathrm{SS_{xy}}}{\mathrm{N}-2}} = \sqrt{\frac{SSE}{N-2}}$$

όπου:

$$\mathrm{SS_{yy}} = \sum_{\mathrm{n}=1}^{\mathrm{N}} (\mathrm{y_n} - \bar{\mathrm{Y}})^2 = \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{y_n^2} - \frac{(\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{y_n})^2}{\mathrm{N}}$$

Υπενθυμίζεται

$$\mathrm{SS_{xy}} = \sum_{\mathrm{n}=1}^{\mathrm{N}} (\mathrm{x_n} - \bar{\mathrm{X}})(\mathrm{y_n} - \bar{\mathrm{Y}}) = \sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{x_n y_n} - \frac{(\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{x_n})(\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{y_n})}{\mathrm{N}}$$

Εάν είχαμε γνώση των δεδομένων του πληθυσμού θα μπορούσαμε να υπολογίσουμε την τυπική απόκλιση των τυχαίων σφαλμάτων από τη σχέση:

$$\sigma_{\epsilon} = \sqrt{\frac{\mathrm{SS_{yy}} - \mathrm{B} * \mathrm{SS_{xy}}}{\mathrm{N_p}}}$$

όπου σε αυτή την περίπτωση θα είχαμε:

$$\mathrm{SS_{yy}} = \sum_{\mathrm{n}=1}^{\mathrm{N_p}} (\mathrm{y_n} - \mu_{\mathrm{y}})^2, \quad \mathrm{SS_{xy}} = \sum_{\mathrm{n}=1}^{\mathrm{N_p}} (\mathrm{x_n} - \mu_{\mathrm{x}})(\mathrm{y_n} - \mu_{\mathrm{y}})$$

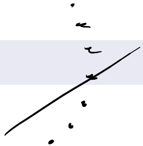
Συντελεστής Προσδιορισμού (Coefficient of Determination)

Συνολικό Άθροισμα τετραγώνων

$\{(x_j, y_j)\}_{j=1}^{N} \to \alpha, b \to \hat{y}_j$

y_j

$$\mathsf{SST} = \sum_{n=1}^{N} (y_n - \bar{Y})^2$$



Άθροισμα τετραγώνων παλινδρόμησης

$$\mathsf{SSR} = \sum_{n=1}^{N} (\hat{y}_n - \bar{Y})^2$$

Συντελεστής Προσδιορισμού

$$R^2 = \frac{\mathsf{SSR}}{\mathsf{SST}}, \quad 0 \leq R^2 \leq 1 \text{ (γιατί;)}$$

- Ποσοτικοποιεί την αποτελεσματικότητα του μοντέλου.

Συντελεστής Προσδιορισμού (Coefficient of Determination)

$$\mathsf{SST} = \mathsf{SSR} + \mathsf{SSE}$$

$$R^2 = \frac{SSR}{SSR + SSE}$$

$$\mathsf{SST} = \sum_{n=1}^{N}(y_n - \bar{Y})^2, \quad \mathsf{SSR} = \sum_{n=1}^{N}(\hat{y}_n - \bar{Y})^2, \quad \mathsf{SSE} = \sum_{n=1}^{N}(y_n - \hat{y}_n)^2$$

Συντελεστής Προσδιορισμού (Coefficient of Determination)

$$R^2 = \frac{\mathsf{SST} - \mathsf{SSE}}{\mathsf{SST}} = \frac{b * SS_{xy}}{SS_{yy}}, \quad 0 \leq R^2 \leq 1$$

Αντικαθιστώντας τη τιμή του b έχουμε το R^2 στη μορφή:

$$R^2 = \frac{SS_{xy}^2}{SS_{xx} * SS_{yy}}$$

Παράδειγμα

Βρείτε τον συντελεστή προσδιορισμού του συνόλου δεδομένων:

$$\{(0,1),(1,3),(2,4),(5,4)\}$$

Μέση τιμή, τυπική απόκλιση και κατανομή του b

$$\mu_{\mathrm{b}} = \mathrm{B}, \quad \sigma_{\mathrm{b}} = \frac{\sigma_\epsilon}{\sqrt{\mathrm{SS_{xx}}}}$$

$$\mathrm{b} \sim \mathcal{N}(\mu_{\mathrm{b}}, \sigma_{\mathrm{b}})$$

- Όταν το σ είναι άγνωστο δεν μπορούμε να υπολογίσουμε το σ_{b}

Εκτιμήτρια της τυπικής απόκλιση του b

$$\mathrm{s_b} = \frac{\mathrm{s_e}}{\sqrt{\mathrm{SS_{xx}}}}$$