ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

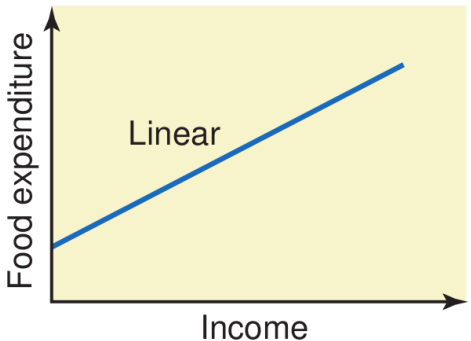
05-03-2020

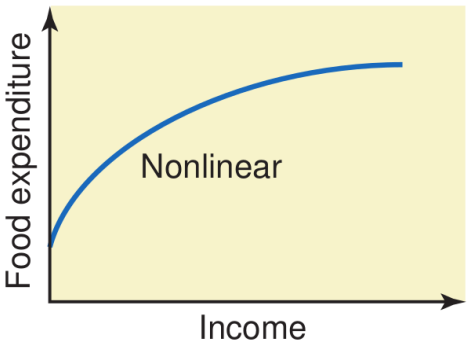
Παλινδρόμηση

Ένα μοντέλο παλινδρόμησης είναι μια μαθηματική εξίσωση που περιγράφει την σχέση μεταξύ δύο ή περισσότερων μεταβλητών. Το μοντέλο παλινδρόμησης με δύο μεταβλητές, μια ανεξάρτητη και μια εξαρτημένη ονομάζεται **μοντέλο απλής παλινδρόμησης**.

Απλή Γραμμική Παλινδρόμηση (Simple Linear Regression)

Ένα μοντέλο παλινδρόμησης το οποίο συνδέει με γραμμικό τρόπο την ανεξάρτητη με την εξαρτημένη μεταβλητή ονομάζεται **μοντέλο απλής γραμμικής παλινδρόμησης**.


Food expenditure
Linear
Income


Food expenditure
Nonlinear
Income

Αιτιοκρατικό μοντέλο

$$y = A + Bx$$

Πιθανοθεωρητικό μοντέλο - Μοντέλο απλής γραμμικής παλινδρόμησης

$$y = A + Bx + \epsilon, \quad \epsilon : \text{όρος τυχαίου σφάλματος}$$

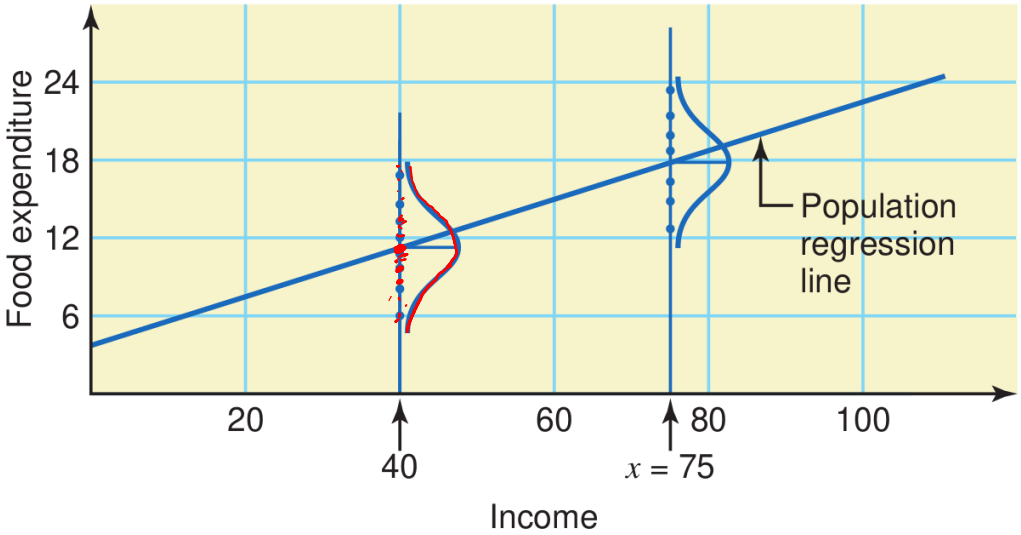
$$A : \text{σταθερός όρος (constant term)}, \quad B : \text{κλίση (slope)}$$

Παραδοχές

- Για δοσμένο x το ϵ ακολουθεί τυπική κανονική κατανομή. $N(0, \sigma_\epsilon)$
- Τα τυχαία σφάλματα διαφορετικών παρατηρήσεων είναι ανεξάρτητα.
- Για κάθε x οι κατανομές των τυχαίων σφαλμάτων παρουσιάζουν την ίδια τυπική απόκλιση.

Ευθεία παλινδρόμησης για τον πληθυσμό

$$\mu_{y|x} = A + Bx$$


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

Δειγματικό μοντέλο απλής γραμμικής παλινδρόμησης

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

- 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

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$$

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

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

όπου η προσέγγιση του κάθε $\hat{y}_n$ δίνεται ώς

$$\hat{y}_n = a + bx_n$$

Άθροισμα τετραγωνικών σφαλμάτων

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

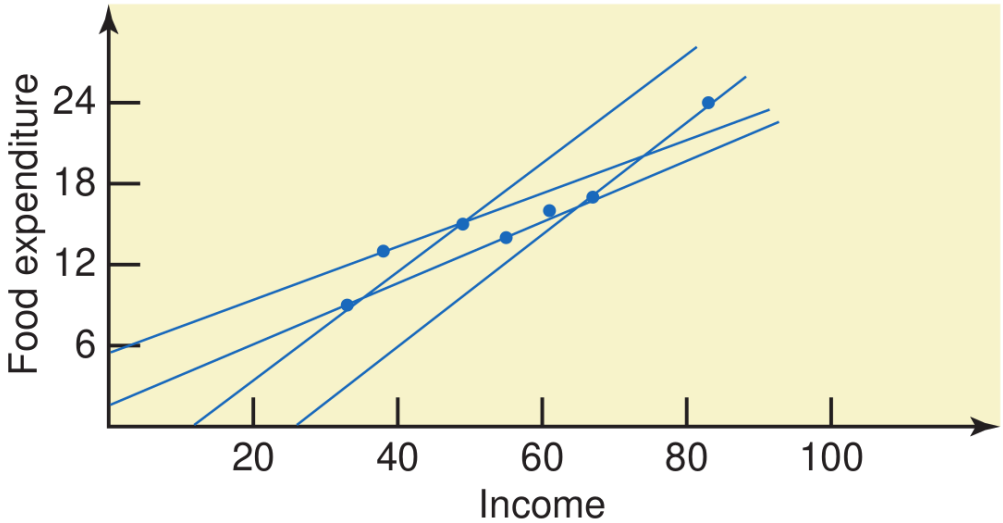
Άθροισμα τετραγωνικών σφαλμάτων συναρτήσει των παραμέτρων του δειγματικού μοντέλου

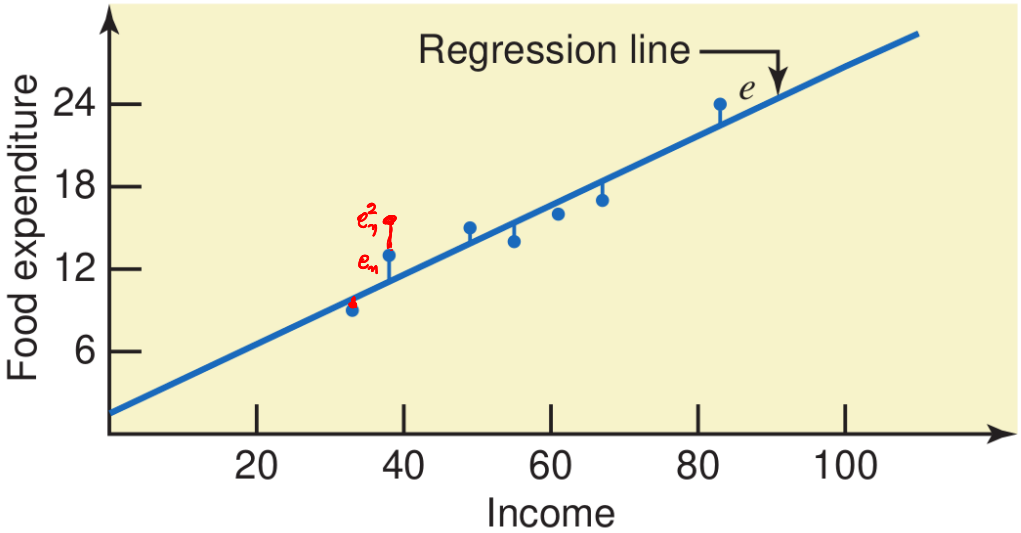
$$\mathcal{Q}(a,b) = \text{SSE} = \sum_{n=1}^{N} (y_n - a - bx_n)^2$$

Εκτίμηση ελαχίστων τετραγώνων

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

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


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


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

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

$\sum_{n=1}^{N} e_n = 0$

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

$$a = -\frac{b}{N}\sum_{n=1}^{N} x_n + \frac{1}{N}\sum_{n=1}^{N} y_n \Leftrightarrow \boxed{a = \bar{Y} - b\bar{X}}$$

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

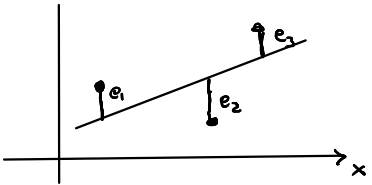
$$\Leftrightarrow \sum_{n=1}^{N} x_n y_n - \bar{Y}\sum_{n=1}^{N} x_n + b\bar{X}\sum_{n=1}^{N} x_n - b\sum_{n=1}^{N} x_n^2 = 0$$

$$b\left(\bar{x}\sum_{n=1}^{N} x_n - \sum_{n=1}^{N} x_n^2\right) = \bar{Y}\sum_{n=1}^{N} x_n - \sum_{n=1}^{N} x_n y_n$$

$\bar{x} = \frac{1}{N}\sum x_n \cdot \sum x_n$

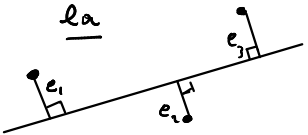
$\frac{1}{N}\left(\sum x_n\right)^2$

$\frac{1}{N}\sum x_n \sum y_n$



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

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

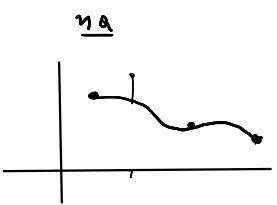


όπου SS_{xy}, SS_{xx} δίνονται ως:

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

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

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



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

$$\sum_{n=1}^{N} x_n y_n - \bar{y}\sum_{n=1}^{N} x_n - \bar{x}\sum_{n=1}^{N} y_n + \bar{x}\bar{y}\sum_{n=1}^{N} 1$$

$$= \sum_{n} x_n y_n - \frac{1}{N}\sum_{n} x_n \sum_{n} y_n - \frac{1}{N}\sum_{n} x_n \sum_{n} y_n + \frac{1}{N}\sum_{n} x_n \sum_{n} y_n$$

$$= \sum_{n=1}^{N} x_n y_n - \frac{\sum x_n \sum y_n}{N}$$

Παράδειγμα

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

$$\{(\overset{x}{0},\overset{y}{1}), (1,2), (2,2)\}$$

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

$\sum_{n=1}^{3} x_n = 3$ $\quad \sum_{n=1}^{3} y_n = 5$ $\quad \sum_{n=1}^{3} x_n y_n = 6$ $\quad \sum_{n=1}^{3} x_n^2 = 5$

$\bar{X} = \frac{3}{3} = 1$ $\quad \bar{Y} = \frac{5}{3}$

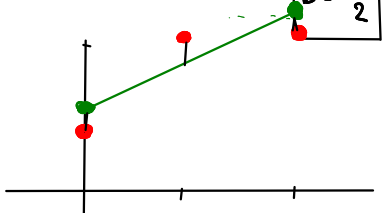
$SS_{xy} = 6 - \frac{3 \cdot 5}{3} = 6 - 5 = 1$ $\quad S_{xx} = 5 - \frac{9}{3} = 5 - 3 = 2$

$b = \frac{1}{2}$

$a = \frac{5}{3} - \frac{1}{2} = \frac{7}{6}$

$\hat{y} = \frac{7}{6} + \frac{1}{2}x$

$x = 0 \quad \hat{y} = \frac{7}{6}$ $\quad x = 2 \quad \hat{y} = \frac{7}{6} + 1$



Άσκηση

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

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

Διανυσματική μορφή

Έστω διανύσματα $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$ στήλες με στοιχεία της παρατηρήσεις της ανεξάρτητης και της εξαρτημένης μεταβλητής αντίστοιχα. Το μοντέλο απλής γραμμικής παλινδρόμησης δίνει εκτιμήσεις για τις τιμές της εξαρτημένης μεταβλητής που αντιστοιχούν στις παρατηρήσεις της ανεξάρτητης μεταβλητής που περιέχονται στο $\mathbf{x}$:

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

όπου $\mathbf{u} \in \mathbb{R}^N$ διάνυσμα στήλη με στοιχεία άσους.
Κατά επέκταση έχουμε:

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

καθώς και

$$\text{SSE} = \mathbf{e}^T\mathbf{e} = \begin{bmatrix} e_1 & e_2 & \cdots & e_N \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix} = \sum_{n=1}^{N} e_n^2$$

$$\mathcal{Q}(a,b) = (\mathbf{y} - a\mathbf{u} - b\mathbf{x})^T(\mathbf{y} - a\mathbf{u} - b\mathbf{x})$$

$$\bar{X} = \frac{1}{N}\mathbf{u}^T\mathbf{x}, \quad \bar{Y} = \frac{1}{N}\mathbf{u}^T\mathbf{y}$$

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

$$SS_{xy} = (\mathbf{x} - \bar{X}\mathbf{u})^T(\mathbf{y} - \bar{Y}\mathbf{u}), \quad SS_{xx} = (\mathbf{x} - \bar{X}\mathbf{u})^T(\mathbf{x} - \bar{X}\mathbf{u})$$

Παράδειγμα

Βρείτε τη εκτίμηση ελαχίστων τετραγώνων του μοντέλου γραμμικής παλινδρόμησης υποθέτοντας τα παρακάτω δεδομένα κάνοντας χρήση των διανυσματικών εκφράσεων.

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