

ΑΝΑΛΥΣΗ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Ας υποθέσουμε ότι έχουμε:

- Ένα τυχαίο δείγμα $X_1, X_2, \dots, X_n$ από μία κατανομή X
- Ένα τυχαίο δείγμα $Y_1, Y_2, \dots, Y_n$ από μία κατανομή Y

(Φανταστείτε μία τυχαία επιλογή n στρεμμάτων στα οποία καλλιεργούμε ένα συγκεκριμένο είδος καλλιέργειας και τα εξετάζουμε ως προς:

X : την ποσότητα συγκεκριμένου λιπάσματος (κιλά) που ρίχνουμε στο κάθε στρέμμα

Y : την απόδοση (σε τόνους) που έχει το κάθε στρέμμα

Έχοντας την ένδειξη ότι υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών X, Y υποθέτουμε ότι υπάρχουν σταθερές β_0 και β_1 τέτοιες ώστε:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

όπου

ε_i : το τυχαίο σφάλμα

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

δηλαδή αν υποθέσουμε ότι το i -«άτομο» του δείγματος παίρνει την τιμή x_i , ($X_i = x_i$), το σφάλμα ε_i εκφράζει την απόκλιση της τιμής Y_i από τον γραμμικό όρο $\beta_0 + \beta_1 x_i$,

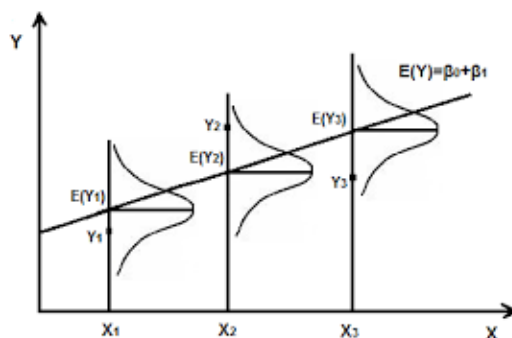
$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i), \quad i = 1, 2, \dots, n$$

Σύμφωνα με το παράδειγμά μας, αυτό σημαίνει ότι αν στο i στρέμμα ρίξουμε λίπασμα x_i , τότε η απόδοση Y_i δεν είναι συγκεκριμένη, αλλά είναι μία τυχαία μεταβλητή που λόγω της γραμμικής συσχέτισης θα έχει τη μορφή

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Μπορούμε μάλιστα να υποθέσουμε ότι $\varepsilon_i \sim N(0, \sigma^2)$ κάτι που με τη σειρά του σημαίνει ότι $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, δηλαδή:

$$E(Y_i | X_i = x_i) = \beta_0 + \beta_1 x_i$$



Οι συντελεστές β_0, β_1 είναι άγνωστοι. Θεωρητικά θα μπορούσαμε να τους προσδιορίσουμε αν είχαμε τις απαντήσεις όλου του πληθυσμού, κάτι που πρακτικά είναι αδύνατον. Αυτό λοιπόν που κάνουμε είναι να εκτιμήσουμε τους παραπάνω συντελεστές. Η μέθοδος με την οποία θα το πετύχουμε ονομάζεται **μέθοδος ελαχίστων τετραγώνων** και βασίζεται στην ελαχιστοποίηση του

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Η ελαχιστοποίηση επιτυγχάνεται με τη μέθοδο των μερικών παραγώγων. Συγκεκριμένα:

$$\frac{\partial}{\partial \beta_0} \left(\sum \varepsilon_i^2 \right) = \frac{\partial}{\partial \beta_0} \left(\sum (y_i - \beta_0 - \beta_1 x_i)^2 \right) = 2 \sum (y_i - \beta_0 - \beta_1 x_i) (-1)$$

και

$$\frac{\partial}{\partial \beta_1} \left(\sum \varepsilon_i^2 \right) = \frac{\partial}{\partial \beta_1} \left(\sum (y_i - \beta_0 - \beta_1 x_i)^2 \right) = 2 \sum (y_i - \beta_0 - \beta_1 x_i) (-x_i)$$

Θέτουμε τις παραπάνω μερικές παραγώγους ίσες με 0 και έχουμε:

$$\frac{\partial}{\partial \beta_0} \left(\sum \varepsilon_i^2 \right) = 0 \quad \Leftrightarrow \quad -2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Leftrightarrow \sum (y_i - \beta_0 - \beta_1 x_i) = 0 \Leftrightarrow \sum y_i = n\beta_0 + \beta_1 \sum x_i \quad (1)$$

Ομοίως:

$$\frac{\partial}{\partial \beta_1} \left(\sum \varepsilon_i^2 \right) = 0 \Leftrightarrow 2 \sum (y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

$$\Leftrightarrow \sum (-x_i y_i + \beta_0 x_i + \beta_1 x_i^2) = 0 \Leftrightarrow \sum x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2 \quad (2)$$

Επομένως προκύπτει το γραμμικό σύστημα:

$$\begin{cases} n \cdot \beta_0 + \sum x_i \cdot \beta_1 = \sum y_i \\ \sum x_i \cdot \beta_0 + \sum x_i^2 \cdot \beta_1 = \sum x_i y_i \end{cases}$$

Για τις ορίζουσες του παραπάνω συστήματος έχουμε ότι:

$$D = \begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix} = n \sum x_i^2 - \left(\sum x_i \right)^2$$

$$D_{\beta_1} = \begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix} = n \sum x_i y_i - \left(\sum x_i \right) \left(\sum y_i \right)$$

$$D_{\beta_0} = \begin{vmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i^2 \end{vmatrix} = \left(\sum y_i \right) \left(\sum x_i^2 \right) - \left(\sum x_i \right) \left(\sum x_i y_i \right)$$

Η μοναδική λύση του παραπάνω συστήματος είναι η επόμενη:

$$\hat{\beta}_1 = \frac{v \sum x_i y_i - (\sum x_i)(\sum y_i)}{v \sum x_i^2 - (\sum x_i)^2} = \frac{v \sum x_i y_i - v\bar{x} \cdot v\bar{y}}{v \sum x_i^2 - (v\bar{x})^2} = \frac{v \sum x_i y_i - v^2 \bar{x}\bar{y}}{v \sum x_i^2 - v^2 \bar{x}^2}$$

ή

$$\hat{\beta}_1 = \frac{v(\sum x_i y_i - v\bar{x}\bar{y})}{v(\sum x_i^2 - v\bar{x}^2)} = \frac{\sum x_i y_i - v\bar{x}\bar{y}}{\sum x_i^2 - v\bar{x}^2} = \frac{\text{αριθμητής του } r}{\text{1ο υπόριζο του } r}$$

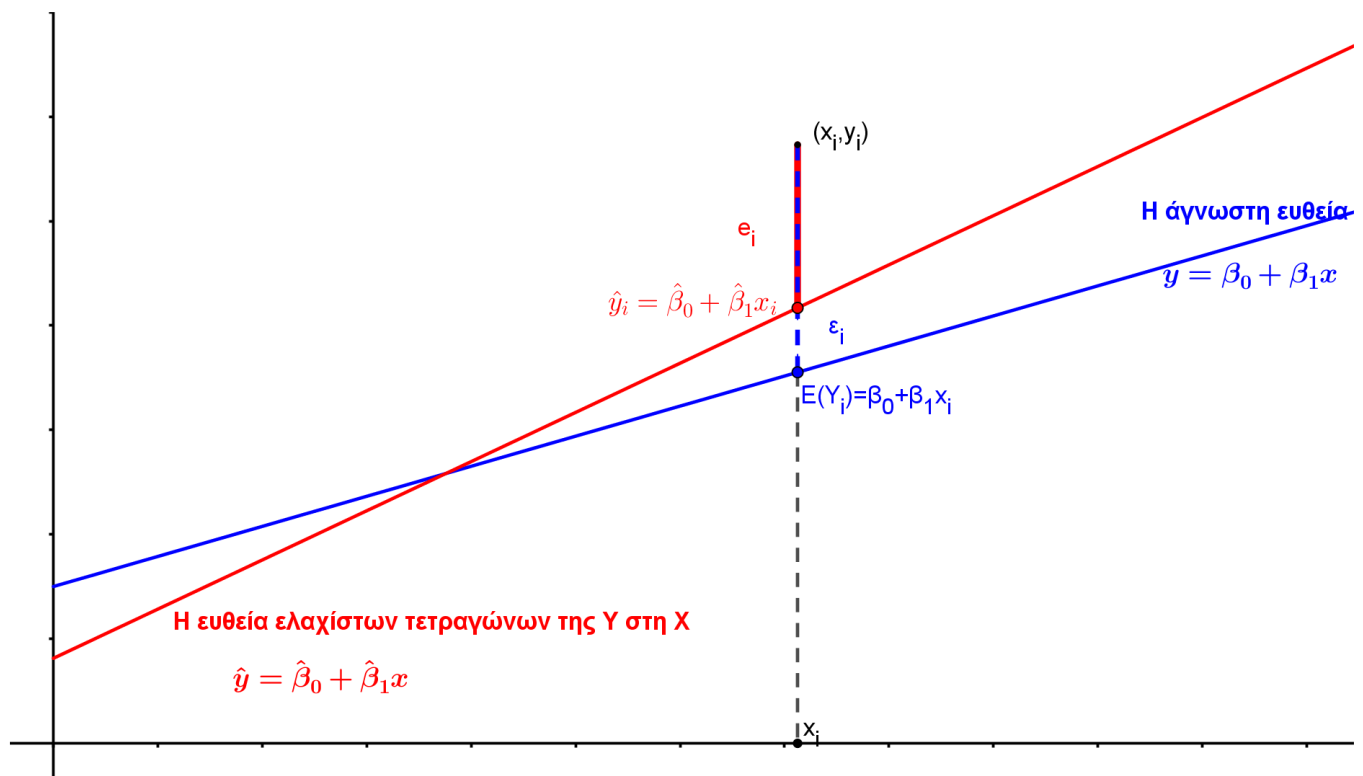
και

$$\hat{\beta}_0 = \frac{(\sum y_i)(\sum x_i^2) - (\sum x_i)(\sum x_i y_i)}{v \sum x_i^2 - (\sum x_i)^2} = \dots \dots \dots = \bar{y} - \hat{\beta}_1 \cdot \bar{x}$$

Η εκτιμώμενη ευθεία

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

ονομάζεται ευθεία ελαχίστων τετραγώνων της Y πάνω στη X και αποτελεί εκτίμηση της άγνωστης ευθείας $y = \beta_0 + \beta_1 x$.



Η εξίσωση $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ αποτελεί εκτίμηση της $E(Y_i) = \beta_0 + \beta_1 x_i$, δηλαδή ο αριθμός $\hat{y}_i$ που υπολογίζεται αν αντικαταστήσουμε τη μεταβλητή x στην ευθεία ελαχίστων τετραγώνων με x_i , αποτελεί εκτίμηση της άγνωστης $E(Y_i|X = x_i)$.

Υπόλοιπα (Residuals) $e_i = y_i - \hat{y}_i$

Τα υπόλοιπα e_i αποτελούν εκτιμήσεις των τυχαίων σφαλμάτων ε_i , και υπολογίζονται από την παραπάνω διαφορά, όπου:

y_i : η πραγματική τιμή του ζεύγους (x_i, y_i)

$\hat{y}_i$: η εκτίμηση της $E(Y_i|X = x_i)$

Μέσο τετραγωνικό σφάλμα s_ε^2

Υπολογίζεται από τον τύπο

$$s_\varepsilon^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

και αποτελεί εκτίμηση της διασποράς σ^2 των τυχαίων σφαλμάτων ε_i .

Ο αριθμός

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

ονομάζεται τυπικό σφάλμα όσο το $s_\varepsilon \rightarrow 0$, τόσο η ευθεία $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ δίνει μία καλή περιγραφή της στοχαστικής σχέσης μεταξύ των μεταβλητών X, Y . Αυτό φαίνεται και στα επόμενα:

$$(n-2)s_\varepsilon^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad \{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x}\}$$

$$= \sum (y_i - \bar{y} + \hat{\beta}_1 \cdot \bar{x} - \hat{\beta}_1 x_i)^2 = \sum [(y_i - \bar{y}) - \hat{\beta}_1 \cdot (x_i - \bar{x})]^2$$

$$= \sum (y_i - \bar{y})^2 + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 - 2\hat{\beta}_1 \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\{S_x^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}, \text{ οπότε } \sum (x_i - \bar{x})^2 = (n-1)S_x^2\}$$

$$\{S_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}, \text{ οπότε } \sum (x_i - \bar{x})(y_i - \bar{y}) = (n-1)S_{xy}\}$$

$$= \sum (y_i - \bar{y})^2 + \hat{\beta}_1^2 (n-1)S_x^2 - 2\hat{\beta}_1 (n-1)S_{xy}$$

$$\{\sum (y_i - \bar{y})^2 = (n-1)S_y^2\}$$

$$\{\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{(n-1)S_{xy}}{(n-1)S_x^2} = \frac{S_{xy}}{S_x^2}\}$$

$$= (n-1)S_y^2 + \frac{S_{xy}^2}{S_x^4} (n-1)S_x^2 - 2\frac{S_{xy}}{S_x^2} (n-1)S_{xy}$$

$$= (n-1) \left(S_y^2 + \frac{S_{xy}^2}{S_x^2} - 2\frac{S_{xy}}{S_x^2} S_{xy} \right) = (n-1) \left(S_y^2 - \frac{S_{xy}^2}{S_x^2} \right)$$

Άρα:

$$s_\varepsilon^2 = \frac{n-1}{n-2} \left(S_y^2 - \frac{S_{xy}^2}{S_x^2} \right),$$

οπότε

$$\begin{aligned} s_\varepsilon \rightarrow 0 &\Leftrightarrow s_\varepsilon^2 \rightarrow 0 \Leftrightarrow S_y^2 - \frac{S_{xy}^2}{S_x^2} \rightarrow 0 \Leftrightarrow S_{xy}^2 \rightarrow S_x^2 S_y^2 \Leftrightarrow \left(\frac{S_{xy}}{S_x S_y} \right)^2 \rightarrow 1 \\ &\Leftrightarrow r^2 \rightarrow 1 \Leftrightarrow r \rightarrow 1. \end{aligned}$$

Παρατηρήσεις για την ευθεία ελαχίστων τετραγώνων

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

1. Διέρχεται πάντα από το σημείο $(\bar{x}, \bar{y})$.
2. Για κάθε τιμή x της X προβλέπουμε την

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = E(Y|X = x)$$
 Αποδεικνύεται ότι η πρόβλεψη είναι ασφαλής όταν η τιμή της X είναι μέσα στις τιμές του δείγματος.
3. Η παράμετρος $\hat{\beta}_0$ εκφράζει την $E(Y|X = 0)$.
4. Η παράμετρος $\hat{\beta}_1$ εκφράζει την $E(Y|X = x + 1) - E(Y|X = x)$.
 Δηλαδή τη μεταβολή της μέσης τιμής της εξαρτημένης μεταβλητής Y όταν η τιμή της ανεξάρτητης μεταβλητής X αυξηθεί κατά 1.

Διαστήματα Εμπιστοσύνης σε επίπεδο σημαντικότητας $(1-\alpha)\%$

1. Ένα διάστημα εμπιστοσύνης για την άγνωστη παράμετρο β_0 με συντελεστή εμπιστοσύνης $(1-\alpha)\%$, είναι το:

$$(\hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(v-2) \cdot s_{\varepsilon} \cdot \sqrt{\frac{1}{v} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}})$$

2. Ένα διάστημα εμπιστοσύνης για την άγνωστη παράμετρο β_1 με συντελεστή εμπιστοσύνης $(1-\alpha)\%$, είναι το:

$$(\hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(v-2) \cdot \frac{s_{\varepsilon}}{\sqrt{\sum (x_i - \bar{x})^2}})$$

3. Ένα διάστημα εμπιστοσύνης για την άγνωστη μέση τιμή $E(Y|X = x_0)$

- της οποίας η πρόβλεψη για $X = x_0$ προσδιορίζεται από την ευθεία ελαχίστων τετραγώνων και ισούται με $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ -

με συντελεστή εμπιστοσύνης $(1-\alpha)\%$, είναι το:

$$(\hat{y}_0 \pm t_{\frac{\alpha}{2}}(n-2) \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}})$$

4. Η εκτιμώμενη, από την ευθεία ελαχίστων τετραγώνων, τιμή $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, δεν χρησιμοποιείται μόνο για την σημειακή πρόβλεψη της $E(Y|X = x_0)$ αλλά και για την πρόβλεψη της μεμονωμένης τιμής του Y , όταν $X = x_0$. Στην περίπτωση όμως αυτή το διάστημα εμπιστοσύνης τροποποιείται ελαφρώς. Συγκεκριμένα, **ένα διάστημα εμπιστοσύνης για την μεταβλητή Y όταν $X = x_0$, με συντελεστή εμπιστοσύνης $(1-\alpha)\%$** , είναι το:

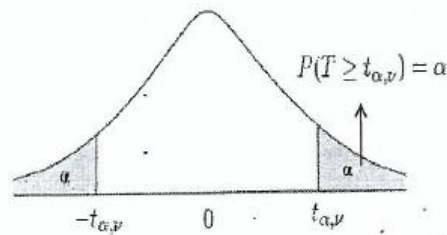
$$(\hat{y}_0 \pm t_{\frac{\alpha}{2}}(n-2) \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}})$$

Χρήσιμες παρατηρήσεις για τον αλγεβρικό υπολογισμό των παραπάνω διαστημάτων εμπιστοσύνης.

Για την εύρεση των παραπάνω διαστημάτων εμπιστοσύνης, εκτός των παραμέτρων $\hat{\beta}_0, \hat{\beta}_1$ που έχουν υπολογιστεί στην εύρεση της ευθείας ελαχίστων τετραγώνων, χρειάζονται επίσης:

- A. Οι τιμές $t_{\frac{\alpha}{2}}(n-2)$ που εξαρτώνται από το επίπεδο σημαντικότητας $1-\alpha$ και από το μέγεθος n του δείγματος. Δίνονται από τον επόμενο πίνακα:

Γ'.2 Πίνακας των τιμών $t_{\alpha,\nu}$ της t -κατανομής με ν βαθμούς ελευθερίας για τις οποίες $P(T \geq t_{\alpha,\nu}) = \alpha$



ν	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.025$	$\alpha = 0.01$	$\alpha = 0.005$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

B. Το άθροισμα $\sum (x_i - \bar{x})^2$ για το οποίο γνωρίζουμε ότι:

$$\sum_{i=1}^v (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - v\bar{x}^2$$

C. Το τυπικό σφάλμα όσο το s_ε για το οποίο γνωρίζουμε ότι:

$$\begin{aligned} s_\varepsilon^2 &= \frac{v-1}{v-2} \left(S_y^2 - \frac{S_{xy}^2}{S_x^2} \right) = \frac{v-1}{v-2} S_y^2 \left(1 - \frac{S_{xy}^2}{S_y^2 S_x^2} \right) = \frac{v-1}{v-2} S_y^2 (1 - r^2) \\ &= \frac{(v-1) S_y^2}{v-2} (1 - r^2) = \frac{\sum_{i=1}^v (y_i - \bar{y})^2}{v-2} (1 - r^2) \\ &= \frac{\sum_{i=1}^n y_i^2 - v\bar{y}^2}{v-2} (1 - r^2) \end{aligned}$$

όπου,

r : ο συντελεστής Pearson.

Οπότε:

$$s_\varepsilon = \sqrt{\frac{\sum y_i^2 - v\bar{y}^2}{v-2} (1 - r^2)}.$$