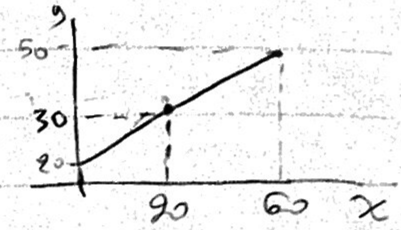
Γραμμική Παλινδρόμηση - Συσχέτιση

Σχέσεις μεταβλητών:
- Συναρτησιακές Σχέσεις: μη στοχαστικές (προσδιοριστικές)
- Στατιστικές Σχέσεις: στοχαστικές (τυχαίες)

Συναρτ. Σχέσεις

$y = f(x)$

$y = 20 + 0{,}5 \cdot x$, y σε €, x σε km σχήμα →

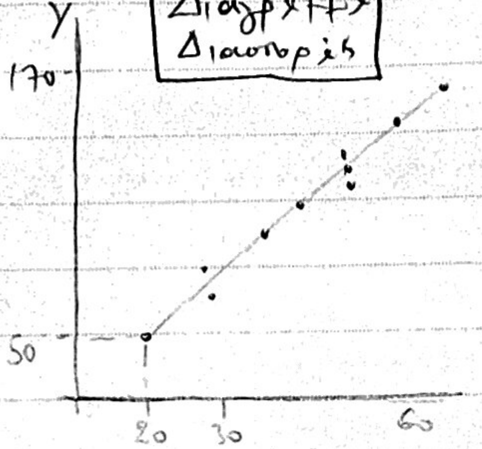


Είναι τέλεια σχέση.

Πχ. 5.2 (ΕΒΟ)

Μήνες (i) παραγωγής	Μέγεθος (x) παραγγελίας	Ώρες Εργασίας
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

Διάγραμμα Διασποράς



Απλή Γραμμική Παλινδρόμηση

Έστω (X_i, Y_i), $i=1,\dots,n$ ώστε οι X και Y σχετίζονται γραμμικά

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

ε_i : τυχαία σφάλματα, ασυσχέτιστες ανά δύο τ.μ. με

$E(\varepsilon_i) = 0$ και $Var(\varepsilon_i) = \sigma^2$

τα β_0 και β_1 : παράμετροι του μοντέλου

$E(Y_i) = \beta_0 + \beta_1 X_i$ ← άγνωστη συνάρτηση παλινδρόμησης

$Var(Y_i) = \sigma^2 \quad \forall x$

Εκτίμηση της Συνάρτησης Παλινδρόμησης με τη μέθοδο ελαχ. τετραγώνων

$\varepsilon_i = y_i - E(y_i) = y_i - \beta_0 - \beta_1 x_i$

βρες τα $\hat{\beta}_0$ κ $\hat{\beta}_1$ που ελαχ. $\sum_{i=1}^{n} \varepsilon_i^2$

$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \; (=0 \to) \; \sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \quad (1)$$

$\boxed{\text{Κανον. Εξισώσεις}}$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) \; (=0 \to) \; \sum_{i=1}^{n} x_i y_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \quad (2)$$

$(1) \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$(2) \Rightarrow \sum_{i=1}^{n} x_i y_i = (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \Rightarrow$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n} + \hat{\beta}_1 \left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \right]$$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

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

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$: Εκτιμ. συνάρτηση παλινδρόμησης

<u>Για το πχ με την ΕΒΟ:</u>

$\hat{\beta}_1 = 2.0$ κ $\hat{\beta}_0 = 10.0$ Άρα $\hat{Y} = 10 + 2X$

για παράδ. για $X = X_0 = 55 \Rightarrow \hat{Y}_0 = 10 + 2 \cdot 55 = 120$ ώρες εργ.

Από τα δεδομένα Y_i μπορούμε να υπολογίσουμε την απόκλιση

$e_i = Y_i - \hat{Y}_i$: υπόλοιπα $\Rightarrow \sum e_i = 0$

$$A = \begin{pmatrix} \frac{\partial^2 Q}{\partial \beta_0^2} = 2n & 2\sum x_i \\ \frac{\partial^2 Q}{\partial \beta_0 \partial \beta_1} = 2\sum x_i & \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2 \end{pmatrix} \to \det A = 4n\sum x_i^2 - 4\left(\sum x_i\right)^2$$

$= 4n\sum(x_i - \bar{x})^2 \geq 0$

άρα ελάχ.

Ιδιότητες της Εκτιμώμενης Συνάρτησης Παλινδρόμησης ($\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X_i$)

(i) Το κέντρο βάρους των παρατηρήσεων $(\bar{X},\bar{Y})$ είναι σημείο της ευθείας, γιατί αν $X=\bar{X} \Rightarrow \hat{Y}=\hat{\beta}_0+\hat{\beta}_1\bar{X}=\bar{Y}-\hat{\beta}_1\bar{X}+\hat{\beta}_1\bar{X}=\bar{Y}$

(ii) $\sum e_i=0$ γιατί από (1) (κανον. εξ.) έχουμε

$$\sum e_i=\sum(Y_i-\hat{\beta}_0-\hat{\beta}_1 X_i)=\sum Y_i-n\hat{\beta}_0-\hat{\beta}_1\sum X_i=0$$

(iii) Τα υπόλοιπα ικανοποιούν επίσης τη σχέση $\sum X_i e_i=0$ όπως εύκολα προκύπτει από τη κανον. εξ. (2) της προηγ. σελίδας

(iv) Το άθροισμα των παρατηρήσεων Y_i ισούται με το άθροισμα των εκτιμώμενων τιμών $\hat{Y}_i$, $\sum Y_i=\sum \hat{Y}_i$ όπως προκύπτει από την κανον. εξ. (1) της προηγ. σελίδας