Ιδιότητες της ευθείας παλινδρόμησης

Η ευθεία παλινδρόμησης περνάει από το σημείο $\overline{Y}, \overline{X}$ αφού $\hat{a} = \overline{Y} - \hat{b}\overline{X}$

$\overline{\hat{Y}} = \overline{Y}$

$$\hat{Y}_i = \hat{a} + \hat{b}X_i = \left(\overline{Y} - \hat{b}\overline{X}\right) + \hat{b}X_i = \overline{Y} + \hat{b}\left(X_i - \overline{X}\right)$$

$$\sum_{i=1}^{n} \hat{Y}_i = n\overline{Y} + \hat{b}\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = n\overline{Y}$$

$$\frac{\sum_{i=1}^{n} \hat{Y}_i}{n} = \overline{Y} \Rightarrow \quad \overline{\hat{Y}} = \overline{Y}$$

$\sum_{i=1}^{n} \hat{u}_i = 0$

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n}\left(Y_i - \hat{a} - \hat{b}X_i\right) = \sum_{i=1}^{n} Y_i - n\hat{a} - \hat{b}\sum_{i=1}^{n} X_i = 0$$

Αυτό όμως προϋποθέτει την ύπαρξη του $\hat{a}$. Αν δηλαδή υποχρεώσουμε την ευθεία παλινδρόμησης να περάσει από την αρχή των αξόνων τότε η ιδιότητα αυτή δεν ισχύει

$$\hat{y}_i = \hat{b}x_i$$

$$\left.\begin{aligned} Y_i &= \hat{a} + \hat{b}X_i + \hat{u}_i \\ \overline{Y} &= \hat{a} + \hat{b}\overline{X} \end{aligned}\right\} \Rightarrow Y_i - \overline{Y} = \hat{b}\left(X_i - \overline{X}\right) + \hat{u}_i$$

$$y_i = \hat{b}x_i + \hat{u}_i \Rightarrow \hat{y}_i = \hat{b}x_i$$

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$$

$$\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \sum_{i=1}^{n} \left(\hat{b}x_i\right)\hat{u}_i = \hat{b}\sum_{i=1}^{n} x_i \hat{u}_i = \hat{b}\sum_{i=1}^{n} x_i \left(y_i - \hat{b}x_i\right) =$$

$$= \hat{b}\sum_{i=1}^{n} x_i y_i - \hat{b}^2 \sum_{i=1}^{n} x_i^2 = \hat{b}\left(\hat{b}\sum_{i=1}^{n} x_i^2\right) - \hat{b}^2 \sum_{i=1}^{n} x_i^2 = 0$$

Δεν υπάρχει συσχέτιση μεταξύ $\hat{y}_i$ και $\hat{u}_i$

$$\sum_{i=1}^{n} \hat{u}_i X_i = 0$$

$$\sum_{i=1}^{n} \hat{u}_i X_i = \sum_{i=1}^{n} \left(Y_i - \hat{a} - \hat{b}X_i\right)X_i = 0$$

Δεν υπάρχει συσχέτιση μεταξύ X_i και $\hat{u}_i$

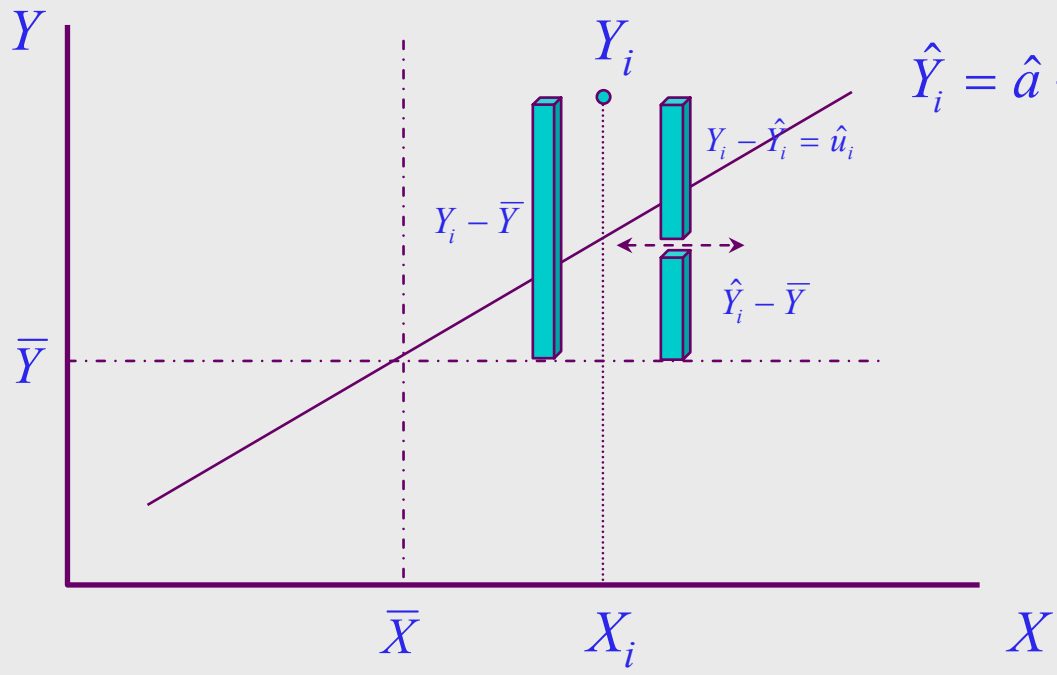
Συντελεστής προσδιορισμού

Ο συντελεστής προσδιορισμού *(coefficient of determination)* είναι ένα μέτρο του βαθμού προσαρμογής *(goodness of fit)* της ευθείας παλινδρόμησης στις παρατηρήσεις του δείγματος.

Συμβολίζεται με r^2 (στην περίπτωση της πολλαπλής παλινδρόμησης με R^2).

Ο αριθμητικός μέσος μιας μεταβλητής $(\overline{Y})$ είναι η καλλίτερη πρόβλεψη όταν η μόνη διαθέσιμη πληροφορία είναι οι τιμές της ίδιας της μεταβλητής.

Η X μπορεί να θεωρηθεί ότι ερμηνεύει την Y στον βαθμό που συμβάλλει στην πρόβλεψή της πέρα από τον μέσο $\overline{Y}$



$\sum \left(Y_i - \overline{Y}\right)^2$

Συνολικό Άθροισμα Τετραγώνων *(Total Sum of Squares)* - ***TSS*** .
Αντιπροσωπεύει την συνολική διακύμανση του ***Y***.

$$\sum \left(\hat{Y}_i - \overline{Y}\right)^2$$

Ερμηνευόμενο Άθροισμα Τετραγώνων *(Regression Sum of Squares)* - ***RSS*** . Αντιπροσωπεύει την διακύμανση του ***Y*** που ερμηνεύεται από την ευθεία παλινδρόμησης.

$$\sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \hat{u}_i^2$$

Το Άθροισμα Τετραγώνων των σφαλμάτων *(Error Sum of Squares)* - ***ESS*** . Αντιπροσωπεύει την διακύμανση του ***Y*** που δεν ερμηνεύεται από την ευθεία παλινδρόμησης.

Αποδεικνύεται ότι

TSS=RSS+ESS

Συντελεστής Προσδιορισμού

$$r^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

$$r^2 = \frac{\sum \left(\hat{Y}_i - \bar{Y}\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{\sum \left(Y_i - \bar{Y}\right)^2}$$

$$0 \le r^2 \le 1$$

$$r^2 = \frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n} \left(\hat{b} x_i\right)^2}{\sum_{i=1}^{n} y_i^2} = \hat{b}^2 \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \hat{b}^2 \frac{\sum_{i=1}^{n} x_i^2 \Big/ (n-1)}{\sum_{i=1}^{n} y_i^2 \Big/ (n-1)} = \hat{b}^2 \frac{S_X^2}{S_Y^2}$$

Παράδειγμα:

X_i	Y_i	$\left(Y_i - \hat{Y}_i\right)^2$	$\left(Y_i - \overline{Y}\right)^2$
150	100	1.907	367.361
160	112	12.283	51.361
170	114	2.590	26.694
180	129	39.391	96.694
190	115	220.170	17.361
200	145	64.764	667.361
		341.105	1226.8

$$r^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{341.105}{1226.8} = 0.722$$

$$r^2 = \hat{b}^2 \frac{S_X^2}{S_Y^2} = 0.711^2 \frac{350}{245.37} = 0.722$$

Συντελεστής συσχέτισης

Ο συντελεστής συσχέτισης *(correlation coefficient)* μας δίνει τον βαθμό γραμμικής συσχέτισης ανάμεσα σε δύο μεταβλητές χωρίς να ενδιαφέρεται για την αιτιώδη σχέση που μπορεί να υπάρχει μεταξύ τους.

Συμβολίζεται με το r και συνδέεται άμεσα με τον συντελεστή προσδιορισμού αφού $r = \pm\sqrt{r^2}$

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2}}$$

$$r = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}}$$

Ιδιότητες

1. $-1 \le r \le 1$

2. $r_{X,Y} = r_{Y,X}$

3. Είναι ένα μέτρο ***γραμμικής συσχέτισης***. Έτσι όταν r=0 δεν συνεπάγεται αναγκαστικά ότι οι δύο μεταβλητές είναι *ανεξάρτητες*.

$$\hat{Y}_i = \hat{a} + \hat{b} X_i \qquad r = \sqrt{\hat{b}^2 \frac{S_X^2}{S_Y^2}}$$

$$\hat{X}_i = \hat{a}' + \hat{b}' Y_i \qquad r = \sqrt{\hat{b}'^2 \frac{S_Y^2}{S_X^2}}$$

$$\Rightarrow \quad r = \sqrt{\hat{b}\hat{b}'}$$

Παράδειγμα:

X_i	Y_i	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
150	100	-25	-19.17	479.167	625	367.361
160	112	-15	-7.17	107.500	225	51.361
170	114	-5	-5.17	25.833	25	26.694
180	129	5	9.83	49.167	25	96.694
190	115	15	-4.17	-62.500	225	17.361
200	145	25	25.83	645.833	625	667.361
				1245	1750	1226.833

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} = \frac{1245}{\sqrt{1750 \cdot 1226.833}} = 0.8496$$

$$r = \sqrt{\hat{b}\hat{b}'} = \sqrt{1.015 \cdot 0.7114 = 0.8496} = 0.8496$$

Έλεγχος υποθέσεων και διαστήματα εμπιστοσύνης για τα *a* και *b*

Υπόθεση: $u_t \sim N(0, \sigma^2)$

Επειδή κάθε ***γραμμικός μετασχηματισμός*** μιας τυχαίας μεταβλητής που ακολουθεί την κανονική κατανομή είναι επίσης τυχαία μεταβλητή με κανονική κατανομή, το Y_i ακολουθεί την ***κανονική κατανομή*** αφού $Y_i = a + bX_i + u_i$

Για τον ίδιο λόγο τόσο το $\hat{a}$ όσο και το $\hat{b}$ είναι ***τυχαίες μεταβλητές*** που ακολουθούν την ***κανονική κατανομή***

$$\hat{a} \sim N\left(a, \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}\right)\sigma^2\right)$$

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}\right)$$

Στατιστικός έλεγχος του *b*

$$Z = \frac{\hat{b} - b_0}{\sigma_{\hat{b}}} = \frac{\hat{b} - b_0}{\sqrt{\sigma^2 \Big/ \sum_{i=1}^{n} x_i^2}} \qquad Z \sim N(0,1)$$

$$t_{n-2} = \frac{\hat{b} - b_0}{S_{\hat{b}}} = \frac{\hat{b} - b_0}{\sqrt{\hat{\sigma}^2 \Big/ \sum_{i=1}^{n} x_i^2}} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i}{n-2}$$

Παράδειγμα: $\hat{Y}_i = -5.333 + 0.7114 X_i$

$S_{\hat{a}} = 38.81$

$S_{\hat{b}} = 0.2207$

H_0 : b=0 (α=0.05)

$$t = \frac{\hat{b} - b}{S_{\hat{b}}} = \frac{0.7114 - 0}{0.2207} = 3.222$$

Από τους πίνακες $t_{0.025/4} = 2.776$

Η υπόθεση H_0 απορρίπτεται

Στατιστικός έλεγχος του *a*

$$Z = \frac{\hat{a} - a}{\sigma_{\hat{a}}} = \frac{\hat{a} - a}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}}} \qquad Z \sim N(0,1)$$

$$t = \frac{\hat{a} - a}{S_{\hat{b}}} = \frac{\hat{a} - a}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} x_i^2}}}$$

Παράδειγμα: $\hat{Y}_i = -5.333 + 0.7114 X_i$ $\quad S_{\hat{a}} = 38.81$ $\quad S_{\hat{b}} = 0.2207$

H_0: a=0 (α=0.05)

$$t = \frac{\hat{a} - a}{S_{\hat{a}}} = \frac{-5.33 - 0}{38.81} = -0.1374$$ Από τους πίνακες $t_{0.025/4} = 2.776$

Η υπόθεση H_0 δεν απορρίπτεται

Διαστήματα εμπιστοσύνης για το *a* και *b*

$$P\left(-t_{\alpha/2,n-2} \le t \le t_{\alpha/2,n-2}\right) = 1-\alpha$$

$$P\left(\hat{b} - S_{\hat{b}} \cdot t_{\alpha/2,n-2} \le b \le \hat{b} + S_{\hat{b}} \cdot t_{\alpha/2,n-2}\right) = 1-\alpha$$

$$\hat{b} \pm S_{\hat{b}} \cdot t_{\alpha/2,n-2} \qquad \hat{a} \pm S_{\hat{a}} \cdot t_{\alpha/2,n-2}$$

Παράδειγμα: $\hat{Y}_i = -5.333 + 0.7114 X_i$ $\quad S_{\hat{a}} = 38.81$ $\quad S_{\hat{b}} = 0.2207$

$$\hat{b} - S_{\hat{b}} \cdot t_{\alpha/2,n-2} \le b \le \hat{b} + S_{\hat{b}} \cdot t_{\alpha/2,n-2}$$

$$0.7114 - 0.2207 \cdot 2.776 \le b \le 0.7114 + 0.2207 \cdot 2.776$$

$$0.0987 \le b \le 1.324$$

$$\hat{a} - S_{\hat{a}} \cdot t_{\alpha/2,n-2} \le a \le \hat{a} + S_{\hat{a}} \cdot t_{\alpha/2,n-2}$$

$$-5.33 - 38.81 \cdot 2.776 \le a \le -5.33 + 38.81 \cdot 2.776$$

$$-113.07 \le a \le 102.41$$

Προβλέψεις

Ένα από τα πεδία εφαρμογής της ανάλυσης παλινδρόμησης είναι και η πρόβλεψη *(Forecasting)* της εξαρτημένης μεταβλητής για συγκεκριμένη τιμή της ανεξάρτητης. Διακρίνουμε δύο είδη προβλέψεων:

(α) Την πρόβλεψη της ***μέσης τιμής*** του Y όταν $X=X_0$. Αυτό ισοδυναμεί με κάποιο σημείο πάνω στην γραμμή παλινδρόμησης του πληθυσμού.

(β) Την πρόβλεψη της ***συγκεκριμένης τιμής*** του Y όταν $X=X_0$.

Και στις δύο περιπτώσεις υπολογίζεται ένα ***διάστημα εμπιστοσύνης*** μέσα στο οποίο βρίσκεται η πραγματική τιμή του πληθυσμού.

Πρόβλεψη της *μέσης τιμής* του *Y*

Όταν $X = X_0$ $\quad \hat{Y}_0 = \hat{a} + \hat{b}X_0$

$E\left(\hat{Y} \middle| X_0\right) = E\left(Y \middle| X_0\right)$

Αρα το $\hat{Y}_0$ αποτελεί μια *αμερόληπτη* εκτίμηση του $E\left(Y \middle| X_0\right)$

$$Var\left(\hat{Y}_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2} \right)$$

$$x_i = X_i - \overline{X}$$

$$x_0 = X_0 - \overline{X}$$

όσο περισσότερο απέχει η $\boldsymbol{X_0}$ *από την μέση τιμή τόσο μεγαλύτερη είναι η διακύμανση και λιγότερο ακριβής η πρόβλεψη*

Τα όρια μέσα στα οποία βρίσκεται το $E\left(Y \middle| X_0\right)$

$$\hat{Y}_0 \pm Z_{\alpha/2}\sigma \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2}} \quad \text{ή} \quad \hat{Y}_0 \pm t_{\alpha/2,n-2}\hat{\sigma} \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2}}$$

Παράδειγμα: $\hat{Y}_i = -5.333 + 0.7114 X_i$

$\overline{X} = 175 \quad \hat{\sigma} = 9.234 \quad \sum_{i=1}^{n} x_i^2 = 1750$

Ζητείται η προβλεπόμενη μέση τιμή του Y στον πληθυσμό όταν $X=165$ και $X=210$. Η πρώτη τιμή βρίσκεται κοντά στην μέση τιμή ενώ η δεύτερη βρίσκεται σε μεγαλύτερη απόσταση.

$X_0 = 165$

$$\hat{Y}_i = -5.333 + 0.7114 \cdot 165 = 112.048$$

$$E\left(Y \middle| X_0 = 165\right) = \pm 112.048 - 2.776 \cdot 9.234 \sqrt{\frac{1}{6} + \frac{(175-165)^2}{1750}}$$

$$99.921 \le E\left(Y \middle| X_0 = 165\right) \le 124.175$$

$X_0 = 210$

$$\hat{Y}_i = -5.333 + 0.7114 \cdot 210 = 144.061$$

$$E\left(Y \middle| X_0 = 210\right) = \pm 144.061 - 2.776 \cdot 9.234 \sqrt{\frac{1}{6} + \frac{(175-210)^2}{1750}}$$

$$120.197 \le E\left(Y \middle| X_0 = 210\right) \le 167.924$$

Πρόβλεψη *συγκεκριμένης τιμής* του Y

Όταν $X = X_0$ $\hat{Y}_0 = \hat{a} + \hat{b}X_0$

και $Y_0 - \hat{Y}_0$ το λάθος της πρόβλεψης

$$E\left(Y_0 - \hat{Y}_0\right) = E(a - \hat{a}) + E\left(b - \hat{b}\right)X_0 + E\left(u_0\right) = 0$$

Αρα το $\hat{Y}_0$ αποτελεί μια *αμερόληπτη* εκτίμηση του Y_0

$$\sigma^2_{Y_0 - \hat{Y}_0} = \sigma^2\left(1 + \frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2}\right) \quad \text{ή} \quad S^2_{Y_0 - \hat{Y}_0} = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^{n} x_i^2}\right)$$

όσο περισσότερο απέχει η X_0 από την μέση τιμή τόσο μεγαλύτερη είναι η διακύμανση και λιγότερο ακριβής η πρόβλέψη.

$$-t_{\alpha/2,n-2}\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{x_0^2}{\sum_{i=1}^{n}x_i^2}}\le Y_0-\hat{Y}\le t_{\alpha/2,n-2}\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{x_0^2}{\sum_{i=1}^{n}x_i^2}}$$

$$\hat{Y}-t_{\alpha/2,n-2}\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{x_0^2}{\sum_{i=1}^{n}x_i^2}}\le Y_0\le\hat{Y}+t_{\alpha/2,n-2}\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{x_0^2}{\sum_{i=1}^{n}x_i^2}}$$

Η πρόβλεψη συγκεκριμένης τιμής γίνεται με μικρότερη ακρίβεια απ΄ότι η πρόβλεψη της μέσης τιμής

Παράδειγμα:

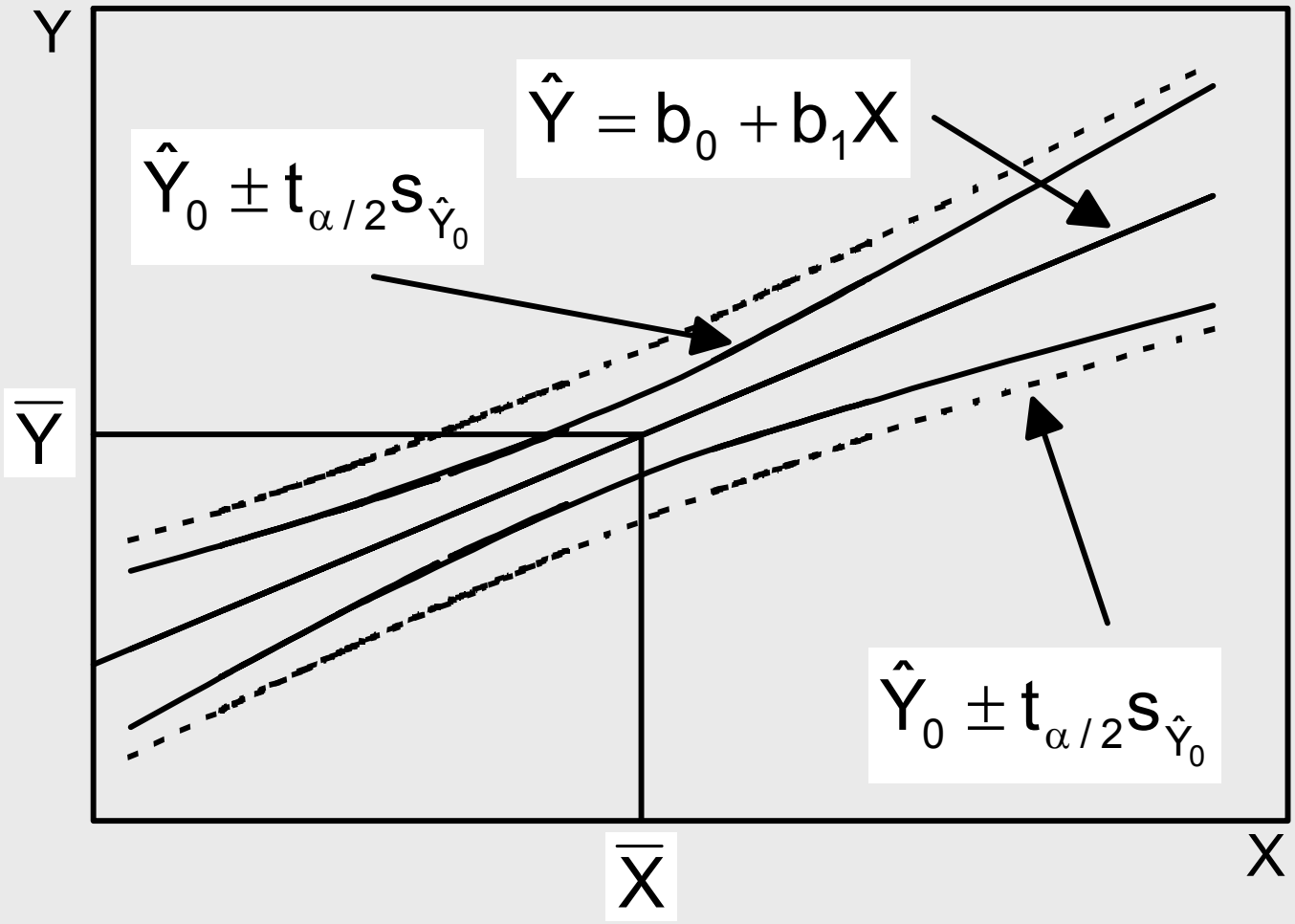
$X_0 = 165$

$$\hat{Y}_i = -5.333 + 0.7114\cdot 165 = 112.048$$

$$Y_0 = \pm 112.048 - 2.776\cdot 9.234\sqrt{1+\frac{1}{6}+\frac{(175-165)^2}{1750}}$$

$$83.69 \le Y_0 \le 140.40$$

Διάγραμμα 8.9:
Διαστήματα Εμπιστοσύνης των προβλέψεων της Y



Ελαστικότητα

Η ελαστικότητα (της Υ ως προς την Χ) ισούται με τον λόγο μεταξύ της ποσοστιαίας μεταβολής της Υ ως προς τη ποσοστιαία μεταβολή της Χ δηλαδή:

$$n_{Y/X} = \frac{\Delta Y / Y}{\Delta X / X} = \frac{\Delta Y}{\Delta X}\frac{X}{Y}$$

$$n_{Y/X} = b_1 \frac{X}{Y} = b_1 \frac{X}{\hat{Y}}$$

Μέση ελαστικότητα

$$n_{Y/X} = b_1 \frac{\overline{X}}{\overline{Y}}$$