

## A. Πρόβλεψη μέσω ενός απλού γραμμικού μοντέλου

Θεωρητική (ατομική) πρόβλεψη του απλού γραμμικού μοντέλου (πρόβλεψη της τιμής  $y$ ) για κάποια γνωστή τιμή της ανεξάρτητης μεταβλητής, έστω  $x_0$ .

$$y_0 = \beta_1 + \beta_2 x_0 + \varepsilon_0, \quad (1)$$

Μέση πρόβλεψη του απλού γραμμικού μοντέλου (στον πληθυσμό) για την τιμή  $x_0$ :

$$E(y_0) = \beta_1 + \beta_2 x_0, \quad \text{καθώς } E(\varepsilon_0) = 0.$$

Εκτιμητής πρόβλεψης με βάση τις LS εκτιμήσεις  $\hat{\beta}$  και  $\hat{\beta}$  του δείγματος:<sup>1</sup>

$$\hat{y}_0 = \hat{\beta}_1 + \hat{\beta}_2 x_0, \quad (2)$$

Σφάλμα πρόβλεψης:

$$\hat{y}_0 - y_0 = (\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2)x_0 - \varepsilon_0. \quad (3)$$

- Μέση τιμή του σφάλματος  $\hat{y}_0 - y_0$

$$E(\hat{y}_0 - y_0) = E(\hat{\beta}_1 - \beta_1) + x_0 E(\hat{\beta}_2 - \beta_2) - E(\varepsilon_0) = 0,$$

που σημαίνει αμεροληψία προβλέψεων.

- Διακύμανση του σφάλματος  $\hat{y}_0 - y_0$ :

$$\text{Var}(\hat{y}_0 - y_0) = \text{Var}[(\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2)x_0 - \varepsilon_0]$$

---

<sup>1</sup> Η σχέση (2) με βάση την οποία υπολογίζουμε την πρόβλεψη αναφέρεται ως εκτιμητής της πρόβλεψης, καθώς βασίζεται στους εκτιμητές  $\hat{\beta}_1$  και  $\hat{\beta}_2$ .

$$\begin{aligned}
&= \text{Var}(\hat{\beta}_1 - \beta_1) + x_0^2 \text{Var}(\hat{\beta}_2 - \beta_2) + \text{Var}(\epsilon_0) + 2x_0 \text{Cov}(\beta_1 - \beta_1; \hat{\beta}_2 - \beta_2) \\
&= \text{Var}(\hat{\beta}_1) + x_0^2 \text{Var}(\hat{\beta}_2) + \sigma^2 + 2x_0 \text{Cov}(\hat{\beta}_1, \hat{\beta}_2). \tag{4}
\end{aligned}$$

Χρησιμοποιώντας τους τύπους διακυμάνσεων-συνδιακυμάνσεων των  $\hat{\beta}_1$  και  $\hat{\beta}_2$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^N x_i^2}{N \sum_{i=1}^N (x_i - \bar{x})^2}, \quad \text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2}, \quad \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

η σχέση (4) γράφεται ως

$$\begin{aligned}
\text{Var}(\hat{y}_0 - y_0) &= \frac{\sigma^2 \sum_{i=1}^N x_i^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} + \frac{\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2} x_0^2 + \sigma^2 - \frac{2\bar{x}\sigma^2}{\sum_{i=1}^N (x_i - \bar{x})^2} x_0 \\
&= \sigma^2 \left[ \frac{\sum_{i=1}^N x_i^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} + \frac{Nx_0^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} - \frac{2N\bar{x}x_0}{N \sum_{i=1}^N (x_i - \bar{x})^2} + 1 \right] \\
&= \sigma^2 \left[ \frac{\sum_{i=1}^N x_i^2 + Nx_0^2 - 2N\bar{x}x_0 + N\bar{x}^2 - N\bar{x}^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} + 1 \right] \\
&= \sigma^2 \left[ \frac{\sum_{i=1}^N (x_i^2 - \bar{x}^2)}{N \sum_{i=1}^N (x_i - \bar{x})^2} + \frac{N(x_0 - \bar{x})^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} + 1 \right], \quad \text{καθώς} \quad \sum_{i=1}^N (x_i - \bar{x})^2 = \sum_{i=1}^N (x_i^2 - \bar{x}^2) \\
&= \sigma^2 \left[ \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N \sum_{i=1}^N (x_i - \bar{x})^2} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2} + 1 \right] \\
&= \sigma^2 \left[ \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2} + 1 \right] \tag{5}
\end{aligned}$$

Η τελευταία σχέση δείχνει:

i)  $\text{Var}(\hat{y}_0 - y_0)$  αυξάνεται όταν το μέγεθος του δείγματος  $N$  μειώνεται

ii)  $\text{Var}(\hat{y}_0 - y_0)$  αυξάνεται όταν η απόκλιση  $x_0 - \bar{x}$  μεγαλώνει, δηλ. δεν προβλέπουμε για τη μέση τιμή της ανεξάρτητης μεταβλητής  $\bar{x}$ , και

iii)  $\text{Var}(\hat{y}_0 - y_0)$  αυξάνεται όταν η διακύμανση του τυχαίου όρου, δηλαδή  $\sigma^2$  μεγαλώνει.

### Διάστημα εμπιστοσύνης της πρόβλεψης $\hat{y}_0$

Αν υποθέσουμε ότι  $\varepsilon_i \sim N(0, \sigma^2)$ , τότε το σφάλμα πρόβλεψης, που ορίζεται ως

$$\hat{y}_0 - y_0 = (\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2)x_0 - \varepsilon_0,$$

ακολουθεί και αυτό την κανονική κατανομή, δηλ.

$$\hat{y}_0 - y_0 \sim N[0, \text{Var}(\hat{y}_0 - y_0)],$$

καθώς  $E(\hat{y}_0 - y_0) = 0$  και  $\text{Var}(\hat{y}_0 - y_0)$  δίνεται από τη σχέση (5). Η τυποποιημένη κατανομή του  $\hat{y}_0 - y_0$  είναι:

$$\frac{\hat{y}_0 - y_0}{\sqrt{\text{Var}(\hat{y}_0 - y_0)}} \sim N(0, 1).$$

Με βάση την κατανομή αυτή και θεωρώντας ότι η διακύμανση  $\sigma^2$  εκτιμάται βρίσκουμε το διάστημα εμπιστοσύνης της αληθινής τιμής  $y_0$  ορίζεται ως

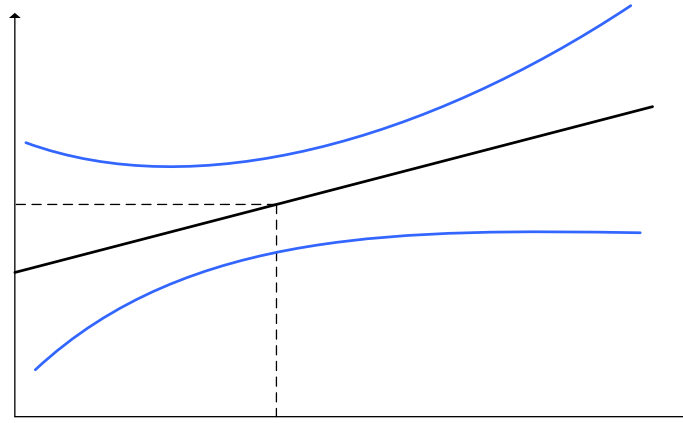
$$\Pr \left[ \hat{y}_0 - t_c \cdot \sqrt{\text{Var}(\hat{y}_0 - y_0)} \leq y_0 \leq \hat{y}_0 + t_c \cdot \sqrt{\text{Var}(\hat{y}_0 - y_0)} \right] = 1 - \alpha$$

και δίνεται ως

$$\hat{y}_0 - t_c \cdot \sqrt{\text{Var}(\hat{y}_0 - y_0)} \leq y_0 \leq \hat{y}_0 + t_c \cdot \sqrt{\text{Var}(\hat{y}_0 - y_0)},$$

όπου  $t_c$  αποτελεί την κριτική τιμή της t-student κατανομής με 2 βαθμούς ελευθερίας για επίπεδο σημαντικότητας  $\alpha$ .

Γράφημα: Διάστημα πρόβλεψης



## B. Πρόβλεψη βάσει πολλαπλού γραμμικού μοντέλου

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_K x_{iK} + \varepsilon_i$$

Πρόβλεψη με βάση τις LS εκτιμήσεις των συντελεστών του μοντέλου:

$$\underset{(1 \times 1)}{\hat{y}_0} = \underset{(1 \times K)}{[1, x_{02}, \dots, x_{0K}]} \underset{\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_K \end{bmatrix}}{} = \underset{(K \times 1)}{\mathbf{x}_0'} \underset{}{\hat{\boldsymbol{\beta}}},$$

για το σημείο “0” του διανύσματος των ανεξάρτητων μεταβλητών:  $\underset{(K \times 1)}{\mathbf{x}_0'} = [1, x_{02}, \dots, x_{0K}]$ .

**Σφάλμα της πρόβλεψης  $\hat{y}_0$  :**

$$\hat{y}_0 - y_0 = \mathbf{x}'_0 \hat{\boldsymbol{\beta}} - (\mathbf{x}'_0 \boldsymbol{\beta} + \varepsilon_0) = \mathbf{x}'_0 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon_0.$$

- Μέση τιμή σφάλματος πρόβλεψης:  $E(\hat{y}_0 - y_0) = E[\mathbf{x}'_0 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon_0] = 0$
- Διακύμανσή του σφάλματος πρόβλεψης:

$$\begin{aligned} \text{Var}(\hat{y}_0 - y_0) &= E(\hat{y}_0 - y_0)^2 \\ &= E[\mathbf{x}'_0 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \varepsilon_0]^2 = E[\mathbf{x}'_0 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2 + E(\varepsilon_0)^2 - 2E[\mathbf{x}'_0 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\varepsilon_0] \\ &= \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + \sigma^2 = \sigma^2 [\mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 + 1], \end{aligned} \quad (6)$$

καθώς ισχύουν τα ακόλουθα αποτελέσματα:

$$\begin{aligned} E[\mathbf{x}'_0 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})]^2 &= E[\mathbf{x}'_0 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{x}_0] \\ &= \mathbf{x}'_0 \text{Var}(\hat{\boldsymbol{\beta}}) \mathbf{x}_0 = \sigma^2 \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0, \quad \text{όπου } \text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}, \end{aligned}$$

$$E[\varepsilon_0]^2 = \sigma^2, \quad \text{καθώς } E[\varepsilon_i]^2 = \sigma^2 \text{ για όλα τα } i, \text{ και}$$

$$E[\mathbf{x}'_0 (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\varepsilon_0] = E[\mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\varepsilon} \varepsilon_0] = \mathbf{x}'_0 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' E(\boldsymbol{\varepsilon} \varepsilon_0) = \mathbf{0},$$

όπου  $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\varepsilon}$ . Σημείωσε ότι  $E(\boldsymbol{\varepsilon} \varepsilon_0) = \mathbf{0}$ , καθώς οι παρατηρήσεις του διανύσματος των τιμών του διαταρακτικού όρου  $\boldsymbol{\varepsilon}$  θεωρούνται ως ανεξάρτητες μεταξύ τους και θεωρούμε ότι κάνουμε προβλέψεις σε κάποιο σημείο “0”, που δεν συμπεριλαμβάνεται στις παρατηρήσεις του δείγματος. Το σημείο αυτό αναφέρεται και ως εκτός του δείγματος, και μας δίνει προβλέψεις για τιμές της μεταβλητής  $y_i$  εκτός του δείγματος. Αυτές αναφέρονται ρητά ως **εκτός του δείγματος προβλέψεις**.