

Το κανονικό γραμμικό μοντέλο

Έστω το μοντέλο πολλαπλής γραμμικής παλινδρόμησης

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i, i = 1, 2, \dots, n$$

$$\epsilon_i \sim N(0, \sigma^2)$$

$$\epsilon_i \perp\!\!\!\perp \epsilon_j \ (i \neq j) \text{ ανεξάρτητες}$$

Το μοντέλο μπορεί να γραφεί στην μορφή πινάκων ως

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n \times 1}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}_{n \times (p+1)}$$

Το κανονικό γραμμικό μοντέλο

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{(p+1) \times 1}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n \times 1},$$

Επομένως,

$$Y = X\beta + \epsilon$$

Παρατηρήσεις

- Στην θεωρία της γραμμικής παλινδρόμησης, ο πίνακας σχεδιασμού (design matrix) θεωρείται σταθερός - εξετάζουμε την κατανομή $Y|X$, άρα $\Rightarrow$
- $E(Y) = X\beta$ και $Var(Y) = Var(\epsilon) = \sigma^2 I_n$ αφού τα σφάλματα είναι ανεξάρτητα μεταξύ τους.
- Έπεται από τα θεωρήματα γραμμικού συνδυασμού κανονικών

$$Y \sim N(X\beta, \sigma^2 I_n).$$

Το κανονικό γραμμικό μοντέλο

Για να βρούμε εκτιμητές των παραμέτρων, αρκεί να ελαχιστοποιήσουμε το άθροισμα των τετραγωνικών αποστάσεων

$$Q(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}^\top \mathbf{Y} + \beta \mathbf{X}^\top \mathbf{X} \beta - 2\beta^\top \mathbf{X}^\top \mathbf{Y}$$

Παραγωγίζουμε ως προς β

$$\begin{aligned}\frac{\partial Q(\beta)}{\partial \beta} &= 2\mathbf{X}^\top \mathbf{X} \beta - 2\mathbf{X}^\top \mathbf{Y} = \mathbf{0} \Rightarrow \\ (\mathbf{X}^\top \mathbf{X}) \hat{\beta} &= \mathbf{X}^\top \mathbf{Y} \Rightarrow \\ \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}\end{aligned}$$

- Πρέπει ο πίνακας $\mathbf{X}^\top \mathbf{X}$ να είναι **θετικά ορισμένος**.
- Αυτό μπορεί να επιβεβαιωθεί εφόσον καμία μεταβλητή (στήλη του $\mathbf{X}$) εκ των $X_1, X_2, \dots, X_p$ δεν είναι γραμμικός συνδυασμός των υπολοίπων (θυμηθείτε το ζήτημα της πολυσυγγραμμικότητας).

Συμπεράσματα

- Η κατανομή του β είναι $(p + 1)$ -διάστατη κανονική ως γραμμικός συνδυασμών κανονικών κατανομών

$$\begin{aligned}E(\hat{\beta}) &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E(\mathbf{Y}) = (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X}) \beta = \beta \\ \text{Var}(\hat{\beta}) &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\sigma^2 I_n) \mathbf{X} \{(\mathbf{X}^\top \mathbf{X})^{-1}\}^\top \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X}) (\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\end{aligned}$$

Επομένως,

$$\hat{\beta} \sim N\{\beta, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}\}$$