

6. ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΜΕ ΤΗ ΧΡΗΣΗ ΠΙΝΑΚΩΝ

Η χρησιμοποίηση και ο συμβολισμός πινάκων απλοποιεί σημαντικά τα αποτελέσματα της γραμμικής παλινδρόμησης, ιδίως στην περίπτωση της πολλαπλής παλινδρόμησης. Γενικά, για την χρήση πινάκων υιοθετούμε τα σύμβολα με έντονα γράμματα. Έτσι, με κεφαλαίο έντονο $\mathbf{X}$ συμβολίζουμε τον πίνακα των παρατηρήσεων στις ανεξάρτητες μεταβλητές $X_1, X_2, \dots, X_p$ ενώ με $\boldsymbol{\varepsilon}$ και $\boldsymbol{\beta}$ τα διανύσματα των συντελεστών, των τυχαίων μεταβλητών και των λαθών, αντίστοιχα.

Έστω

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

το μοντέλο της πολλαπλής παλινδρόμησης με p εξαρτημένες μεταβλητές $X_1, X_2, \dots, X_p$.

Θεωρούμε τους πίνακες

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

$$\underset{(n \times 1)}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \underset{(p+1) \times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_p \end{bmatrix}$$

Ο πίνακας $\mathbf{X}$ αναφέρεται στις παρατηρήσεις των ανεξαρτήτων μεταβλητών $X_1, X_2, \dots, X_p$. Η πρώτη στήλη του πίνακα αποτελείται από στοιχεία 1 προκειμένου να περιληφθεί στην εξίσωση πινάκων ο

σταθερός συντελεστής α της γραμμικής παλινδρόμησης. Κάθε στήλη του πίνακα $\mathbf{X}$ αποτελείται από τις n παρατηρήσεις πάνω στη συγκεκριμένη μεταβλητή. Τα στοιχεία της i γραμμής του πίνακα $\mathbf{X}$ αναφέρονται στις παρατηρήσεις πάνω στις ανεξάρτητες μεταβλητές (δοκιμές) $X_1, X_2, \dots, X_p$ στην i δοκιμή ($i = 1, 2, \dots, n$). Το διάνυσμα $\mathbf{y}$ αναφέρεται στις n τιμές της εξαρτημένης μεταβλητής του $Y, Y_1, Y_2, \dots, Y_p$. Το διάνυσμα $\boldsymbol{\beta}$ έχει ως στοιχεία τον σταθερό συντελεστή α και τους συντελεστές $\beta_1, \beta_2, \dots, \beta_p$ των ανεξαρτήτων μεταβλητών $X_1, X_2, \dots, X_p$.

Χρησιμοποιώντας τον συμβολισμό αυτό των πινάκων, μπορούμε να εκφράσουμε την εξίσωση πολλαπλής παλινδρόμησης με μορφή πινάκων ως εξής:

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{(n \times (p+1))}{\mathbf{X}} \underset{((p+1) \times 1)}{\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}$$

Σημείωση: Σε πολλά συγγράμματα για το συμβολισμό του σταθερού συντελεστή της γραμμικής παλινδρόμησης δεν χρησιμοποιείται το γράμμα α αλλά το β_0 . Για το λόγο αυτό και η πρώτη στήλη του πίνακα $\mathbf{X}$ που έχει ως στοιχεία το 1 και που όπως είπαμε αντιστοιχεί στον σταθερό συντελεστή της γραμμικής παλινδρόμησης λέγεται *μηδενική στήλη*.

Σημείωση: Όπως και στην απλή γραμμική παλινδρόμηση, μια ισοδύναμη έκφραση που διευκολύνει τους υπολογισμούς είναι αυτή που κάνει χρήση των διορθωμένων αθροισμάτων τετραγώνων και γινομένων (*corrected sums of squares and cross products*).

Αν $\bar{\mathbf{x}} = (\bar{X}_1, \bar{X}_2, \dots, \bar{X}_p)$ είναι το διάνυσμα διαστάσεων $(p \times 1)$ των μέσων των ανεξαρτήτων μεταβλητών $X_i, i=1, 2, \dots, p$ τότε, εναλλακτικά, μπορούμε να ορίσουμε ως πίνακα των παρατηρήσεων επί των ανεξαρτήτων μεταβλητών, τον πίνακα

$$\mathbf{X} = \begin{bmatrix} 1, & X_{11} - \bar{X}_1, & \dots, & X_{1p} - \bar{X}_p \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1, & X_{n1} - \bar{X}_1, & \dots, & X_{np} - \bar{X}_p \end{bmatrix}.$$

Πολλές φορές διευκολύνει να γράφουμε τον πίνακα $\mathbf{X}$ με τη μορφή διαμερισμένου πίνακα, δηλαδή

$$\mathbf{X} = [\mathbf{j} \ \mathbf{X}_1]$$

όπου $\mathbf{j}$ είναι το διάνυσμα - στήλη που αποτελείται από n μονάδες και $\mathbf{X}_1$ είναι ο πίνακας των κεντροποιημένων παρατηρήσεων

$$\mathbf{X}_1 = \begin{bmatrix} X_{11} - \bar{X}_1, & \dots, & X_{1p} - \bar{X}_p \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ X_{n1} - \bar{X}_1 & \dots, & X_{np} - \bar{X}_p \end{bmatrix}$$

6.1 Πίνακας διακυμάνσεων - συνδιακυμάνσεων του ε

Τα λάθη στην πολλαπλή παλινδρόμηση είναι ουσιαστικά ένα τυχαίο διάνυσμα. Οι υποθέσεις που κάναμε για τα λάθη αυτά μπορούν να συνοψιστούν με μορφή πινάκων ως εξής:

$$E(\varepsilon) = \mathbf{0} \quad \text{και} \quad V(\varepsilon) = \sigma^2 \mathbf{I}_n$$

όπου $V(\varepsilon)$ είναι ο πίνακας διακυμάνσεων - συνδιακυμάνσεων του ε , $\mathbf{I}$ είναι ο $n \times n$ ταυτοτικός πίνακας και $\mathbf{0}$ είναι ένα $n \times 1$ διάνυσμα από 0. Αν προσθέσουμε την υπόθεση ότι κάθε ένα από τα ε_i ακολουθεί την κανονική κατανομή, μπορούμε να γράψουμε

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

6.2 Εκτιμήτριες ελαχίστων τετραγώνων (*Least squares estimators*)

Η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\beta}$ του β επιλέγεται και πάλι έτσι ώστε να ελαχιστοποιεί το άθροισμα των τετραγώνων των λαθών. Αν επομένως θεωρήσουμε τη συνάρτηση που εκφράζει το άθροισμα των τετραγώνων των λαθών, δηλαδή τη συνάρτηση

$$SSE(a, b_1, b_2, \dots, b_p) = \sum_{i=1}^n (y_i - a - b_1 x_{i1} - \dots - b_p x_{ip})^2$$

ή, αντίστοιχα, την συνάρτηση των διορθωμένων λαθών

$$SSE(a, b_1, b_2, \dots, b_p) = \sum_{i=1}^n (y_i - a - b_1 (x_{i1} - \bar{x}_1) - \dots - b_p (x_{ip} - \bar{x}_p))^2$$

θα πρέπει, προκειμένου να βρούμε τις εκτιμήσεις των συντελεστών, να ελαχιστοποιήσουμε τη συνάρτηση (με μορφή πινάκων)

$$\begin{aligned} SSE(\mathbf{b}) &= \sum_{i=1}^n (y_i - \mathbf{x}_i^t \mathbf{b})^2 \\ &= (\mathbf{y} - \mathbf{X} \mathbf{b})' (\mathbf{y} - \mathbf{X} \mathbf{b}) \end{aligned}$$

όπου $\mathbf{b}^t = (a, b_1, b_2, \dots, b_p)$ και $\mathbf{X}$ είναι ο πίνακας των παρατηρήσεων στις ανεξάρτητες μεταβλητές όπως αυτός δόθηκε νωρίτερα ή αντίστοιχα, ο πίνακας των παρατηρήσεων στις κεντροποιημένες μεταβλητές.

Το διάνυσμα των μερικών παραγώγων του $SSE(\mathbf{b})$ ως προς το διάνυσμα $\mathbf{b}$ είναι

$$\frac{\partial SSE(\mathbf{b})}{\partial \mathbf{b}} = -2 \mathbf{X}^t \mathbf{y} + 2 \mathbf{X}^t \mathbf{X} \mathbf{b}$$

Μια επαρκής συνθήκη για να επιτευχθεί το ελάχιστο της συνάρτησης αυτής είναι η συνθήκη

$$\frac{\partial SSE(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0}$$

ή, ισοδύναμα, η συνθήκη

$$\mathbf{X}^t \mathbf{X} \mathbf{b} = \mathbf{X}^t \mathbf{y}$$

Με την προϋπόθεση ότι υπάρχει ο αντίστροφος του πίνακα $\mathbf{X}^t \mathbf{X}$ η εκτίμηση ελαχίστων τετραγώνων $\mathbf{b}$ του $\boldsymbol{\beta}$ δίνεται από τη σχέση

$$\mathbf{b} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}$$

Το $\mathbf{b}$ ως εκτίμηση του $\boldsymbol{\beta}$ είναι προφανώς μια από τις δυνατές τιμές της εκτιμήτριας $\hat{\boldsymbol{\beta}}$ του $\boldsymbol{\beta}$.

Επομένως, η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\boldsymbol{\beta}}$ του διανύσματος $\boldsymbol{\beta}$ δίνεται από τη σχέση

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}$$

Σημείωση: Στην περίπτωση που χρησιμοποιούμε για τον πίνακα $\mathbf{X}$ την έκφραση εκείνη, σύμφωνα με την οποία τα στοιχεία του πίνακα αυτού αναφέρονται στις αποκλίσεις των παρατηρήσεων από τους αντίστοιχους μέσους τους, η εκτιμήτρια του $\boldsymbol{\beta}$ μπορεί να απλοποιηθεί. Αυτό γιατί

$$\begin{aligned} \underset{(p+1) \times n}{\mathbf{X}^t} \underset{n \times (p+1)}{\mathbf{X}} &= [\mathbf{j} \quad \mathbf{X}_1]^t [\mathbf{j} \quad \mathbf{X}_1] \\ &= \begin{bmatrix} \mathbf{j}^t \\ \mathbf{X}_1^t \end{bmatrix} [\mathbf{j} \quad \mathbf{X}_1] \\ &= \begin{bmatrix} \mathbf{j}^t \mathbf{j} & \mathbf{j}^t \mathbf{X}_1 \\ \mathbf{X}_1^t \mathbf{j} & \mathbf{X}_1^t \mathbf{X}_1 \end{bmatrix} \\ &= \begin{bmatrix} n & \mathbf{0}^t \\ \mathbf{0} & \underset{(p+1) \times (p+1)}{\mathbf{X}_1^t \mathbf{X}_1} \end{bmatrix} \end{aligned}$$

δοθέντος ότι κάθε στήλη του πίνακα $\mathbf{X}$ θα πρέπει να αθροίζει στο μηδέν.

Το j διαγώνιο στοιχείο του πίνακα $\mathbf{X}_1^t \mathbf{X}_1$ είναι το άθροισμα τετραγώνων

$$\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2$$

της j ανεξάρτητης μεταβλητής γύρω από το μέσο της, ενώ το j μη-διαγώνιο στοιχείο είναι το άθροισμα των γινομένων

$$\sum_{i=1}^n (X_{ij} - \bar{X}_j)(X_{ih} - \bar{X}_h)$$

Επομένως ο πίνακας $\mathbf{X}_1^t \mathbf{X}_1$ θα έχει τη μορφή

$$\mathbf{X}_1^t \mathbf{X}_1 = \begin{bmatrix} \sum (X_{i1} - \bar{X}_1)^2 & \dots & \sum (X_{i1} - \bar{X}_1)^2 (X_{ip} - \bar{X}_p) \\ \vdots & \vdots & \vdots \\ \sum (X_{i1} - \bar{X}_1)^2 (X_{ip} - \bar{X}_p) & \dots & \sum (X_{ip} - \bar{X}_p)^2 \end{bmatrix}$$

Επίσης θα έχουμε

$$\underset{(p+1) \times n}{\mathbf{X}^t} \underset{n \times 1}{\mathbf{Y}} = \begin{bmatrix} \mathbf{j}^t \\ \mathbf{X}_1^t \end{bmatrix} \mathbf{Y} = \begin{bmatrix} n\bar{Y} \\ \mathbf{X}_1^t \mathbf{Y} \end{bmatrix}$$

ή, ισοδύναμα,

$$\mathbf{X}^t \mathbf{Y} = \begin{bmatrix} \sum (X_{i1} - \bar{X}_1)(Y_i - \bar{Y}) \\ \sum (X_{ip} - \bar{X}_p)(Y_i - \bar{Y}) \end{bmatrix} = \begin{bmatrix} \sum (X_{i1} - \bar{X}_1)Y_i \\ \sum (X_{ip} - \bar{X}_p)Y_i \end{bmatrix}$$

Με βάση τις παρατηρήσεις αυτές, η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\beta}$ του β για την περίπτωση των κεντροποιημένων παρατηρήσεων, γίνεται

$$\begin{aligned}\hat{\beta} &= \begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} n & 0^t \\ 0 & X_1^t X_1 \end{bmatrix}^{-1} \begin{bmatrix} n & \bar{Y} \\ X_1^t & Y \end{bmatrix} = \\ &= \begin{bmatrix} \hat{Y} \\ (X_1^t X_1)^{-1} X_1^t Y \end{bmatrix}.\end{aligned}$$

Παράδειγμα: Στο τέλος κάθε εξαμήνου οι φοιτητές ενός Πανεπιστημιακού Τμήματος συμπληρώνουν ερωτηματολόγια με τα οποία αξιολογούν τις διάφορες πλευρές του μαθήματος που διδάχτηκαν όπως επίσης και το διδάσκοντα. Ο Πρόεδρος του Τμήματος ενδιαφέρεται να μελετήσει την συνολική αξιολόγηση του διδάσκοντα Y σε σχέση με τις ανεξάρτητες μεταβλητές

X_1 = ικανότητα του διδάσκοντα να παρουσιάσει τη διδακτέα ύλη και

X_2 = ικανότητα του διδάσκοντα να προκαλέσει το ενδιαφέρον των φοιτητών.

Οι τιμές των τριών μεταβλητών για ένα δείγμα $n=20$ διδασκόντων δίδονται στον πίνακα που ακολουθεί. (Η επιλογή των 20 διδασκόντων έγινε με τρόπο ώστε τα αντίστοιχα δεδομένα να ανταποκρίνονται σε μεγάλη ποικιλία τιμών των μεταβλητών X_1 και X_2 . Οι αξιολογήσεις γίνονται σε κλίμακα 1-4).

Μέσες Αξιολογήσεις				
X_{i1}	X_{i2}	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1.1	1.2	1.1	1.04	0.06
1.3	1.7	1.4	1.41	-0.01
1.4	1.6	1.5	1.38	0.12
2.0	1.5	1.6	1.49	0.11
2.0	1.4	1.7	1.43	0.27
1.6	1.4	1.2	1.31	-0.11
2.3	2.1	2.0	1.96	0.04
1.6	2.2	1.6	1.82	-0.22
1.5	2.3	1.6	1.85	-0.25
2.5	2.2	2.1	2.09	0.01
2.6	2.7	2.4	2.43	-0.03
2.1	2.4	2.0	2.09	-0.09
3.2	3.1	3.2	2.86	0.34
2.6	3.0	2.7	2.62	0.08
2.8	1.8	2.0	1.92	0.08
3.1	2.4	2.0	2.39	-0.39
2.2	3.1	2.6	2.56	0.04
3.1	3.4	3.0	3.02	-0.02
3.1	3.8	3.5	3.27	0.23
3.9	3.2	2.9	3.13	-0.23
Μέσοι 2.30	2.325	2.105	2.104	0.0015

Για τα στοιχεία αυτά έχουμε ότι ο πίνακας $\mathbf{X}_1$ έχει τα στοιχεία:

$$\mathbf{X}_1 = \begin{bmatrix} 1.1 & 1.2 \\ 1.3 & 1.7 \\ 1.4 & 1.6 \\ 2.0 & 1.5 \\ 2.0 & 1.4 \\ 1.6 & 1.4 \\ 2.3 & 2.1 \\ 1.6 & 2.2 \\ 1.5 & 2.3 \\ 2.5 & 2.2 \\ 2.6 & 2.7 \\ 2.1 & 2.4 \\ 3.2 & 3.1 \\ 2.6 & 3.0 \\ 2.8 & 1.8 \\ 3.1 & 2.4 \\ 2.2 & 3.1 \\ 3.1 & 3.4 \\ 3.1 & 3.8 \\ 3.9 & 3.2 \end{bmatrix}$$

Η πολλαπλή παλινδρόμηση που θα πρέπει να χρησιμοποιηθεί για την ανάλυση του προβλήματος αυτού έχει τη μορφή

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

ή, ισοδύναμα,

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2}, i = 1, 2, \dots, 20.$$

Με πολλαπλασιασμό πινάκων για τα συγκεκριμένα δεδομένα του προβλήματος βρίσκουμε ότι

$$\mathbf{X}_1^t \mathbf{X}_1 = \begin{bmatrix} 10.86 & 8.04 \\ 8.04 & 10.8375 \end{bmatrix}$$

Ο αντίστροφος του πίνακα αυτού δίνεται από τη σχέση

$$(\mathbf{X}_1^t \mathbf{X}_1)^{-1} = \begin{bmatrix} 0.204274 & -0.151545 \\ -0.151545 & 0.204698 \end{bmatrix}$$

Εξάλλου

$$\mathbf{X}_1^t \mathbf{Y} = \begin{bmatrix} 8.31 \\ 9.2375 \end{bmatrix}$$

Θεωρώντας το γινόμενο

$$(\mathbf{X}_1^t \mathbf{X}_1)^{-1} \mathbf{X}_1^t \mathbf{y}$$

καταλήγουμε στις εκτιμήτριες ελαχίστων τετραγώνων για τους συντελεστές $\hat{\beta}_1$ και $\hat{\beta}_2$ της πολλαπλής παλινδρόμησης.

Έτσι, βρίσκουμε ότι

$$\hat{\beta}_1 = 0.298 \text{ και } \hat{\beta}_2 = 0.632.$$

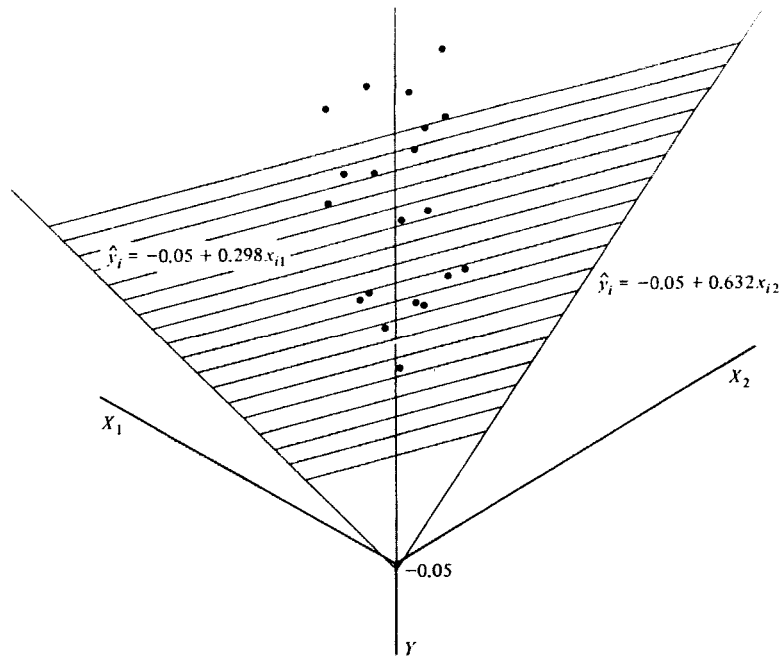
Επομένως η εκτιμώμενη συνάρτηση παλινδρόμησης είναι η συνάρτηση

$$\begin{aligned} \hat{y}_i &= 2.105 + 0.298 (x_{i1} - 2.30) + 0.632 (x_{i2} - 2.325) \\ &= -0.050 + 0.298 x_{i1} + 0.632 x_{i2}. \end{aligned}$$

Τιμές για τον γενικό μέσο αξιολόγησης των διδασκόντων όπως αυτές προκύπτουν από τη χρησιμοποίηση της εκτιμηθείσας εξίσωσης παλινδρόμησης δίνονται στην 4η στήλη του πίνακα.

Στο γράφημα που ακολουθεί παρουσιάζονται οι παρατηρήσεις του προβλήματος. Η αρχή των αξόνων για τα X_1 και X_2 έχει επιλεγεί ώστε να δείχνει πως το εφαρμοσμένο επίπεδο παλινδρόμησης κλίνει προς τα πάνω (slopes upwards) από το σημείο τομής -0.05. Οι άκρες της γραμμοσκιασμένης επιφάνειας είναι οι γραμμές που δίνονται όταν θεωρήσουμε το X_1 ή το X_2 ίσο με το μηδέν στη δειγματική

συνάρτηση παλινδρόμησης. Οι προβλεπόμενες τιμές $\hat{y}_i$ είναι οι κάθετες προβολές των y_i πάνω στο επίπεδο. Τα κατάλοιπα $e_i = y_i - \hat{y}_i$ που δίνονται στην τελευταία στήλη του πίνακα είναι οι αποστάσεις των y_i πάνω ή κάτω από το επίπεδο προσαρμογής. Το άθροισμα των τετραγώνων των καταλοίπων είναι το μικρότερο δυνατό αφού το επίπεδο προσαρμογής έχει επιλεγεί με κριτήριο την ελαχιστοποίηση των τετραγώνων των αποκλίσεων.



Γράφημα της παλινδρόμησης μέσω αξιολογήσεων