ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

26-03-2020

$y = A + Bx + \varepsilon$ $\qquad$ $y = A + B^{(1)}x^{(1)} + B^{(2)}x^{(2)} + \ldots + B^{(K)}x^{(K)} + \varepsilon$

$$y = A + \mathbf{x}^T\mathbf{B} + \epsilon$$

$$\mathbf{x} = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(K)} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} B^{(1)} \\ B^{(2)} \\ \vdots \\ B^{(K)} \end{bmatrix}$$

~~Ευθεία~~ παλινδρόμησης για τον πληθυσμό

για δοσμένο x

$$\mu_{y|\mathbf{x}} = A + \mathbf{x}^T\mathbf{B}$$

$\mathbb{E}\{y|x\} = A + x^T B + \cancel{\mathbb{E}\{\varepsilon|x\}}^{0}$

$\mu_{y|x} = A + x^T B + 0$

Δειγματικό μοντέλο πολλαπλής γραμμικής παλινδρόμησης

$$\hat{y} = a + \mathbf{x}^T \mathbf{b}$$

- a είναι δειγματική προσέγγιση του A
- $\mathbf{b} = [b^{(1)}, b^{(2)}, \ldots, b^{(K)}]^T$ είναι δειγματική προσέγγιση του $\mathbf{B}$
- $\hat{y}$ είναι η εκτιμώμενη τιμή του y για δοσμένο $\mathbf{x} = [x^{(1)}, x^{(2)}, \ldots, x^{(K)}]^T$

Τυχαίο σφάλμα του δειγματικού μοντέλου απλής γραμμικής παλινδρόμησης

$$e = y - \hat{y}$$

Έστω το τυχαίο δείγμα

$$\{(x_1^{(1)}, \ldots, x_1^{(K)}, y_1), (x_2^{(1)}, \ldots, x_2^{(K)}, y_2), \ldots, (x_N^{(1)}, \ldots, x_N^{(K)}, y_N)\}$$

Για το τυχαίο σφάλμα του δειγματικού μοντέλου πολλαπλής γραμμικής παλινδρόμησης έχουμε:

$$e_n = y_n - \hat{y}_n, \quad n = 1, \ldots, N$$

όπου η προσέγγιση του κάθε y_n δίνεται ώς

$$\hat{y}_n = a + \mathbf{x}_n^T \mathbf{b}$$

Άθροισμα τετραγωνικών σφαλμάτων

$$\text{SSE} = \sum_{n=1}^{N} e_n^2$$

$K+1$

$$\mathbf{p} = \begin{bmatrix} a \\ b^{(1)} \\ b^{(2)} \\ \vdots \\ b^{(K)} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(K)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(K)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_N^{(1)} & x_N^{(2)} & \dots & x_N^{(K)} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Προσέγγιση ελαχίστων τετραγώνων

$$\mathcal{Q}(\mathbf{p}) = \mathbf{e}^T\mathbf{e} = \mathbf{y}^T\mathbf{y} - 2\mathbf{p}^T\mathbf{X}^T\mathbf{y} + \mathbf{p}^T\mathbf{X}^T\mathbf{X}\mathbf{p}$$

$$\mathbf{p} = \arg\min_{\mathbf{p}'} \mathcal{Q}(\mathbf{p}')$$

$$\mathbf{p} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$\hat{y}_n = a + \underset{\sim}{x}_n^T \underset{\sim}{b}, \quad n = 1, \dots, N \qquad y_n - \hat{y}_n = e_n = y_n - a - \underset{\sim}{x}_n^T \underset{\sim}{b}, \quad n = 1, \dots, N$

$$\underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}}_{N\times 1} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{N\times 1} - \underbrace{\begin{bmatrix} a \\ a \\ \vdots \\ a \end{bmatrix}}_{N\times 1} - \underbrace{\begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(k)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(k)} \\ \vdots & \vdots & & \vdots \\ x_N^{(1)} & x_N^{(2)} & \cdots & x_N^{(k)} \end{bmatrix}}_{N\times k} \underbrace{\begin{bmatrix} b^{(1)} \\ b^{(2)} \\ \vdots \\ b^{(k)} \end{bmatrix}}_{k\times 1} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{y} - \underbrace{\begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(k)} \\ 1 & x_2^{(1)} & \cdots & x_2^{(k)} \\ \vdots & \vdots & & \vdots \\ 1 & x_N^{(1)} & \cdots & x_N^{(k)} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} a \\ b^{(1)} \\ \vdots \\ b^{(k)} \end{bmatrix}}_{p}$$

e

$$e = y - Xp$$

$$e^T e = (y - Xp)^T (y - Xp) =$$
$$= (y^T - p^T X^T)(y - Xp) =$$
$$= y^T y - \underbrace{y^T X p}_{} - \underbrace{p^T X^T y}_{(Xp)^T y} + p^T X^T X p$$
$$\underbrace{\qquad\qquad}_{-2p^T X^T y}$$

$$(Xp)^T Xp = \sum_{n=1}^{N} ([Xp]_n)^2$$

$$Q(p) = e^T e = y^T y - 2p^T X^T y + \overbrace{p^T X^T X p} = y^T y - 2p^T X^T y + \sum_{n=1}^{N} ([Xp]_n)^2$$

$$\frac{\partial Q}{\partial p_j} = -2 \boxed{\frac{\partial}{\partial p_j}(p^T X^T y)} + \boxed{\frac{\partial}{\partial p_j} \sum_{n=1}^{N} ([Xp]_n)^2} = 0$$

$$\frac{\partial}{\partial p_j}(p^T X^T y) = \frac{\partial}{\partial p_j} \sum_{k=1}^{k+1} p_k [X^T y]_k = [X^T y]_j$$

$$\frac{\partial}{\partial p_j} \sum_{n=1}^{N} \left(\sum_{k=1}^{k+1} X_{nk} p_k \right)^2 = 2 \sum_{n=1}^{N} \sum_{k=1}^{k+1} x_{nk} p_k \frac{\partial}{\partial p_j} \left(\sum_{k=1}^{k+1} x_{nk} p_k \right) = 2 \sum_{n=1}^{N} \sum_{k=1}^{k+1} x_{nk} p_k x_{nj} = 2 \sum_{n=1}^{N} x_{nj} \sum_{k=1}^{k+1} x_{nk} p_k = 2 \sum_{n=1}^{N} x_{nj} [Xp]_n = 2 [X^T X p]_j$$

$$\cancel{2}[X^T X p]_j = \cancel{2}[X^T y]_j, \quad j = 1, \dots, k+1 \qquad X^T X p = X^T y \iff \boxed{p = (X^T X)^{-1} X^T y}$$

Πολλαπλή Γραμμική Παλινδρόμηση

$\hat{y} = a + bx$ $\quad p = \begin{bmatrix} a \\ b \end{bmatrix}$ $\quad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix}$ $\quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$ $\quad p = \underbrace{(X^T X)^{-1}}_{2\times 2} X^T y$

$N \times 2$

απλή παλινδρόμηση

$$S_e^2 = \frac{\sum (y_n - \hat{y}_n)^2}{N-2}$$

$y - 1$

$x_1, \dots, x_k - k$

Πολλαπλή παλινδρόμηση

$$\boxed{S_e^2 = \frac{\sum (y_n - \hat{y}_n)^2}{N-k-1}}$$

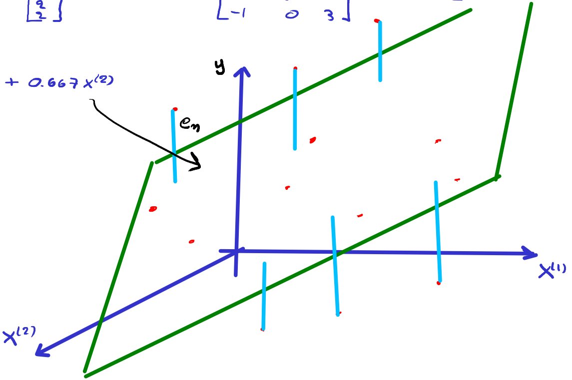
Παράδειγμα

Να βρεθεί το δειγματικό μοντέλο γραμμικής παλινδρόμησης για το σύνολο δεδομένων

$$\{(\overset{x_1^{(1)}}{1}, \overset{x_1^{(2)}}{-1}, \overset{y_1}{1}), (0, -1, -1), (2, 0, 2), (\overset{x_4^{(1)}}{1}, \overset{x_4^{(2)}}{1}, \overset{y_4}{2})\}$$

$$X = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 0 & -1 \\ 1 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad y = \begin{bmatrix} 1 \\ -1 \\ 2 \\ 2 \end{bmatrix} \quad X^T X = \begin{bmatrix} 4 & 4 & -1 \\ 4 & 6 & 0 \\ -1 & 0 & 3 \end{bmatrix} \quad \beta = [1.5, 0.833, 0.667]$$

$$\hat{y} = 1.5 + 8.833x^{(1)} + 0.667x^{(2)}$$



- Στη μέχρι τώρα παρουσίαση της γραμμικής παλινδρόμησης, χρησιμοποιήθηκαν μόνο ονομαστικές μεταβλητές.
- Πολλές φορές τα στοιχεία του δείγματος μπορούν να ανοίκουν σε διακριτές κατηγορίες (κλάσεις).
- Θα λάβουμε υπόψη την επιπλέον πληροφορία εισάγοντας στο μοντέλο ψευδομεταβλητές.

Ψευδομεταβλητές (Dummy variables)

m ονομαστική μεταβλητή

$m \in \{A, B, C\}$

Z^k $\quad d^{(1)} \; d^{(2)}$ $\quad Z^{k-1}$

$A \to [1, 0, 0]$

$B \to [0, 1, 0]$

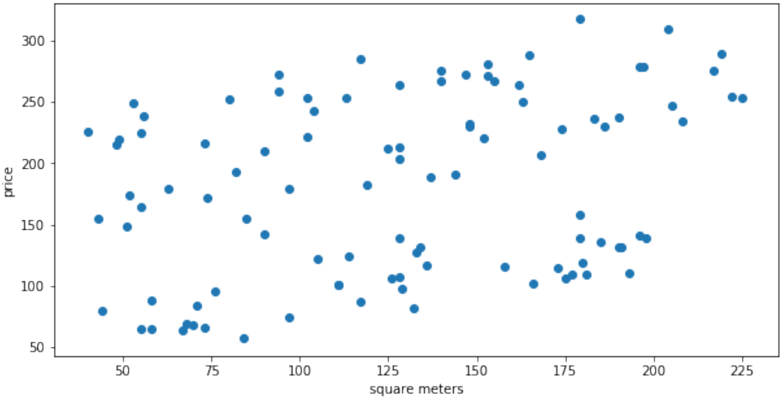
$C \to [0, 0, 1]$

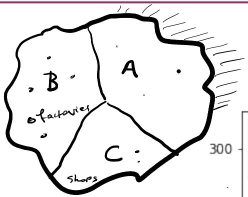
$A \to d^{(1)} = 1, \; d^{(2)} = 0$

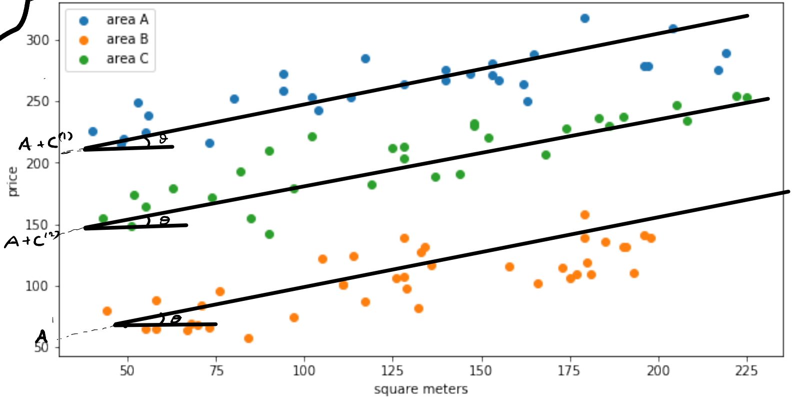
$B \to d^{(1)} = 0, \; d^{(2)} = 1$

$C \to d^{(1)} = d^{(2)} = 0$

Γραμμική Παλινδρόμηση και χρήση Ονομαστικών Μεταβλητών




A
B
C
factories
shops


area A
area B
area C
price
square meters
300
250
200
150
100
50
50
75
100
125
150
175
200
225
A + C(1)
A + C(2)
A
θ

$B^{(1)}$ $x^{(1)}$ $B^{(2)}$ $x^{(2)}$ $B^{(3)}$ $x^{(3)}$

$$\boxed{y = A + Bx + c^{(1)} d^{(1)} + c^{(2)} d^{(2)} + \varepsilon}$$

$A \to [1 \quad 0]$ $\quad d^{(1)} = 1, \; d^{(2)} = 0$

$B \to [0, \; 1]$ $\quad d^{(1)} = 0 \quad d^{(2)} = 1$

$C \to [0, 0]$ $\quad d^{(1)} = d^{(2)} = 0$

$$\hat{y} = a + bx + c^{(1)} d^{(1)} + c^{(2)} d^{(2)}$$

$\underline{\underline{A}}$ $\quad y = A + Bx + c^{(1)} + \varepsilon = \left(A + c^{(1)}\right) + Bx + \varepsilon$

$\underline{\underline{B}}$ $\quad y = \qquad = \left(A + c^{(2)}\right) + Bx + \varepsilon$

$\underline{\underline{C}}$ $\quad y = \qquad = A + Bx + \varepsilon$

$$P = \begin{bmatrix} a \\ b \\ c^{(1)} \\ c^{(2)} \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 & d_1^{(1)} & d_1^{(2)} \\ 1 & x_2 & d_2^{(1)} & d_2^{(2)} \\ 1 & x_3 & d_3^{(1)} & d_3^{(2)} \\ 1 & x_4 & d_4^{(1)} & d_4^{(2)} \end{bmatrix} \qquad P = (X^T X)^{-1} X^T y$$

$$\hat{y} = 145.4 + 0.45x + 59.9d^{(1)} - 98.6d^{(2)}$$