ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

Θεωρία 8ης εβδομάδας

x
y

Αιτιοκρατικό μοντέλο

$$y = A + Bx$$

Πιθανοθεωρητικό μοντέλο - Μοντέλο απλής γραμμικής παλινδρόμησης

$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$

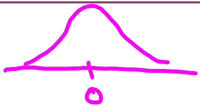
$$y = \underset{\alpha}{A} + \underset{b}{B}x + \underset{e}{\epsilon}, \quad \epsilon : \text{όρος τυχαίου σφάλματος}$$

A : σταθερός όρος (constant term), B : κλίση (slope)

$\hat{y} = \alpha + bx$

$$y = A + Bx + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

$$\hat{y} = a + bx$$

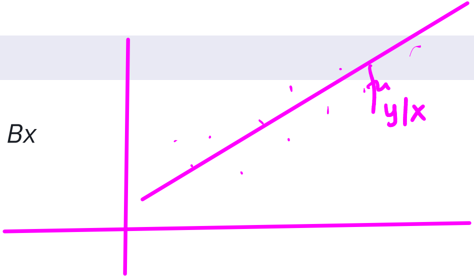


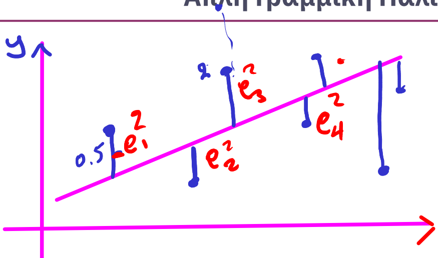
Παραδοχές

- Για δοσμένο x το ϵ ακολουθεί κανονική κατανομή με μηδενική μέση τιμή.
- Τα τυχαία σφάλματα διαφορετικών παρατηρήσεων είναι ανεξάρτητα.
- Για κάθε x οι κατανομές των τυχαίων σφαλμάτων παρουσιάζουν την ίδια τυπική απόκλιση.

Ευθεία παλινδρόμησης για τον πληθυσμό

$$\mu_{y|x} = A + Bx$$





$$\hat{y} = a + bx$$

$$\boxed{b = \frac{SS_{xy}}{SS_{xx}}, \quad a = \bar{Y} - b\bar{X}}$$

$$SS_{yy} = \sum y_n^2 - \frac{(\sum y_n)^2}{N}$$

όπου SS_{xy}, SS_{xx} δίνονται ως:

$$SS_{xy} = \sum_{n=1}^{N} x_n y_n - \frac{(\sum_{n=1}^{N} x_n)(\sum_{n=1}^{N} y_n)}{N}, \quad SS_{xx} = \sum_{n=1}^{N} x_n^2 - \frac{(\sum_{n=1}^{N} x_n)^2}{N}$$

Επιπλέον τα SS_{xy} και SS_{xx} μπορούν ισοδύναμα να υπολογισθούν ως:

$$SS_{xy} = \sum_{n=1}^{N} (x_n - \bar{X})(y_n - \bar{Y}), \quad SS_{xx} = \sum_{n=1}^{N} (x_n - \bar{X})^2$$

$\hat{y}(x)$ $\sum (y_n - \hat{y}(x_n))^2$

$$y = A + Bx + \epsilon, \quad \epsilon : \text{όρος τυχαίου σφάλματος}$$

- Για κάθε x έχουμε υποθέσει ότι το σφάλμα ϵ ακολοθεί την κανονική κατανομή $\mathcal{N}(0, \sigma_\epsilon)$.
- Η τυπική απόκλιση σ_ϵ του τυχαίου σφάλματος αναφέρεται στο πληθυσμό και κατά επέκταση η τιμή της δεν είναι γνωστή στις περισσότερες περιπτώσεις.

Εκτιμήτρια της τυπικής απόκλισης των σφαλμάτων

$$s_e = \sqrt{\frac{\text{SSE}}{N-2}}, \quad \text{SSE} = \sum_{n=1}^{N} (\overbrace{y_n - \hat{y}_n}^{e_n})^2 = SS_{yy} - b\, SS_{xy}$$

$\bar{x}$ μ_x

$\bar{y}$ μ_y

$e_n = y_n - \hat{y}_n$

$\sum (e_n - \overset{0}{\overset{\shortparallel}{\bar{E}}})^2 = SSE$

διαστήματα εμπιστοσύνης.

Συντελεστής Προσδιορισμού (Coefficient of Determination)

Συνολικό άθροισμα τετραγώνων

$$\mathrm{SST} = \sum_{n=1}^{N} (y_n - \bar{Y})^2 = SS_{yy}$$

Άθροισμα τετραγώνων παλινδρόμησης

$$\mathrm{SSR} = \sum_{n=1}^{N} (\hat{y}_n - \bar{Y})^2$$

Συντελεστής Προσδιορισμού

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}}, \quad 0 \leq R^2 \leq 1$$

- Ποσοτικοποιεί την αποτελεσματικότητα του μοντέλου.

$\sum (y_n - \bar{Y})^2 =$

$= \sum (\underbrace{y_n - \hat{y}_n} + \underbrace{\hat{y}_n - \bar{Y}})^2$

SS_{yy} $\downarrow$ ≥ 0

$$\text{SST} = \text{SSR} + \text{SSE}$$

$b = \frac{SS_{xy}}{SS_{xx}}$

$$R^2 = \frac{\text{SST} - \text{SSE}}{\text{SST}} = \frac{b\, SS_{xy}}{SS_{yy}}, \quad 0 \leq R^2 \leq 1$$

$b \frac{SS_{xy}}{SS_{yy}}$

$= \sum (y_n - \hat{y}_n)^2 + \sum (\hat{y}_n - \bar{Y})^2 + 2\sum (y_n - \hat{y}_n)(\hat{y}_n - \bar{Y})$

Αντικαθιστώντας τη τιμή του b έχουμε το R^2 στη μορφή:

$$\boxed{R^2 = \frac{SS_{xy}^2}{SS_{xx} SS_{yy}}}$$

$\sum (x_n - \bar{X})^2$ $\sum (y_n - \bar{Y})^2$

Συντελεστής Γραμμικής Συσχέτισης - Pearson

- Συμβολίζεται με ρ όταν αφορά τον πληθυσμό.

$$\rho \in [-1, 1]$$

$$\sum (x_n - \bar{x})(y_n - \bar{y})$$

- Συμβολίζεται με r όταν αφορά ένα δείγμα.

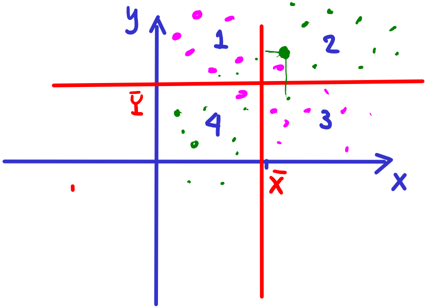
$$r \in [-1, 1]$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

$$R^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} \quad \text{(Συντελεστής Προσδιορισμού)}$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \quad \text{(Συντελεστής Γραμμικής Συσχέτισης)}$$

Σχέση μεταξύ συντελεστών γραμμικής συσχέτισης και προσδιορισμού



$$r = \underline{\mathrm{sign}(SS_{xy})}\sqrt{R^2}$$

$$\mathrm{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

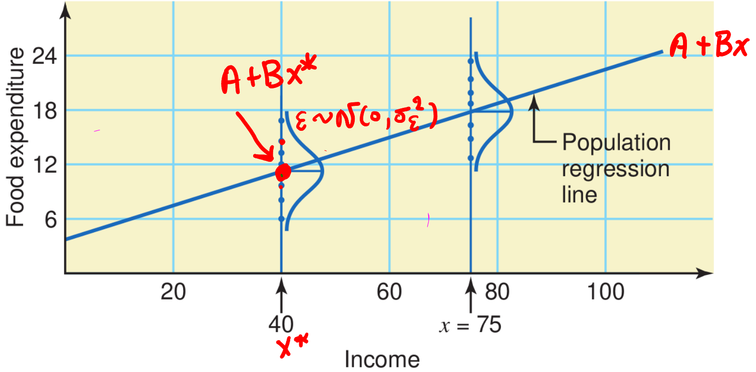
$$x = \mathrm{sign}(x) \cdot |x|$$

$$\sum (x_n - \bar{X})(y_n - \bar{Y})$$

$\alpha = 0.05$

1. Για δοσμένο x^* ποιο είναι το διάστημα εμπιστοσύνης (1-α)*100% για τη μέση τιμή $\mu_{y|x^*}$;
2. Για δοσμένο x^* ποιο είναι το διάστημα εμπιστοσύνης (1-α)*100% για την τιμή μιας συγκεκριμένης παρατήρησης y^*;

$y = A + Bx + \varepsilon$

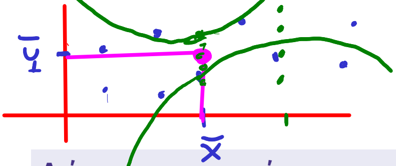


Διαστήματα εμπιστοσύνης για τις τιμές της εξαρτημένης μεταβλητής

$\mu_{y|x^*} = A + Bx^*$, $\sigma_{\mu_{y|x^*}}$

$\hat{\mu}_{y|x^*} = \alpha + bx^*$

Εκτιμήτρια της τυπικής απόκλιση του $\hat{\mu}_{y|x^*}$



$$s_{\hat{\mu}_{y|x^*}} = s_e \sqrt{\frac{1}{N} + \frac{(x^* - \bar{X})^2}{SS_{xx}}}$$

$s_{\hat{\mu}_{y|x^*}}$

Διάστημα εμπιστοσύνης

Το $(1-a) * 100\%$ διάστημα εμπιστοσύνης για την $\mu_{y|x^*}$ είναι: 95%

$$[\hat{\mu}_{y|x^*} - ts_{\hat{\mu}_{y|x^*}}, \hat{\mu}_{y|x^*} + ts_{\hat{\mu}_{y|x^*}}]$$

όπου το t λαμβάνεται από την t_{df}, $df = N - 2$ έτσι ώστε

$$P(T < t) = 1 - a/2$$

- Περιθώριο σφάλματος: $E = ts_{\hat{\mu}_{y|x^*}}$

Διάστημα Εμπιστοσύνης για την εκτίμηση συγκεκριμένης τιμής της y

$y = A + Bx^* + \varepsilon = \mu_{y|x^*} + \varepsilon$

Εκτιμήτρια της τυπικής απόκλιση του $\hat{y}^*$

$$s_{\hat{y}^*} = s_e \sqrt{1 + \frac{1}{N} + \frac{(x^* - \bar{X})^2}{SS_{xx}}}$$

X, Y ανεξάρτητες τ.μ.

$Z = X + Y$

$E\{Z\} = E\{X\} + E\{Y\}$

$\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$

$\sigma_y^2 = \sigma_\varepsilon^2 \left(1/N + \frac{(x^* - \bar{X})^2}{SS_{xx}} \right) + \sigma_\varepsilon^2$

Διάστημα εμπιστοσύνης

Το $(1 - a) * 100\%$ διάστημα εμπιστοσύνης για την y^* είναι:

$$[\hat{y}^* - ts_{\hat{y}^*}, \hat{y}^* + ts_{\hat{y}^*}]$$

όπου το t λαμβάνεται από την t_{df}, $df = N - 2$ έτσι ώστε

$$P(T < t) = 1 - a/2$$

- Περιθώριο σφάλματος: $E = ts_{\hat{y}^*}$

$(x_1, y_1), (x_2, y_2) \ldots, (x_N, y_N)$ απλή (1D)

y

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

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

$$x^T B = \sum_{j=1}^{K} x^{(j)} B^{(j)}$$

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

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

Ευθεία παλινδρόμησης για τον πληθυσμό

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

Δειγματικό μοντέλο απλής 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

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

Πολλαπλή Γραμμική Παλινδρόμηση

Έστω το τυχαίο δείγμα

$\hat{y}_1 \qquad \hat{y}_2 \qquad \alpha, \underset{\sim}{b} \to \hat{y} \qquad \hat{y}_N$

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

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

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

όπου η προσέγγιση του κάθε y_n δίνεται ώς

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

Άθροισμα τετραγωνικών σφαλμάτων

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

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

Προσέγγιση ελαχίστων τετραγώνων

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

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

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

$$e_n = y_n - \hat{y}_n = y_n - \alpha - \underset{\sim}{x}_n^T \underset{\sim}{b}, \quad n = 1, \ldots, N$$

$$\underbrace{\begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}}_{\underset{\sim}{e}} = \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\underset{\sim}{y}} - \begin{bmatrix} \alpha \\ \alpha \\ \vdots \\ \alpha \end{bmatrix} - \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(k)} \\ x_2^{(1)} & \cdots & x_2^{(k)} \\ \vdots & & \vdots \\ x_N^{(1)} & \cdots & x_N^{(k)} \end{bmatrix} \begin{bmatrix} b^{(1)} \\ \vdots \\ b^{(k)} \end{bmatrix}$$

$$\underset{\sim}{X}^* = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(k)} \\ 1 & x_2^{(1)} & \cdots & x_2^{(k)} \\ \vdots & \vdots & & \vdots \\ 1 & x_N^{(1)} & \cdots & x_N^{(k)} \end{bmatrix} \in \mathbb{R}^{N, k+1}$$

$$\underset{\sim}{p} = \begin{bmatrix} \alpha \\ b^{(1)} \\ \vdots \\ b^{(k)} \end{bmatrix} \in \mathbb{R}^{k+1, 1}$$

$$\boxed{e = y - Xp}$$

$$Q = \sum e_n^2 = e^T e = (y - Xp)^T (y - Xp) =$$

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

$$p_j \to \begin{cases} j=1, & a \\ j \neq 1, & b^{(i)} \end{cases}$$

$$= y^T y - 2 p^T X^T y + p^T X^T X p$$

$$\frac{\partial Q}{\partial p_j} = -2 \frac{\partial}{\partial p_j} \left(p^T \boxed{X^T y} \right) + \frac{\partial}{\partial p_j} \left(p^T X^T X p \right)$$

$$p^T X^T y = p \cdot (X^T y) = \sum_{i=1}^{N} p_i (X^T y)_i$$

$$\frac{\partial}{\partial p_j} \left(p^T X^T y \right) = \frac{\partial}{\partial p_j} \left[\sum p_i (X^T y)_i \right] = (X^T y)_j \cdot \frac{\partial p_j}{\partial p_j} = \underline{(X^T y)_j}$$

$$\frac{\partial}{\partial p_j}\left(p^T X^T X p\right) = \begin{bmatrix} \text{—} \end{bmatrix}\begin{bmatrix} | \end{bmatrix}$$

$$p^T X^T X p = (Xp)^T X p = (Xp)\cdot(Xp) = \sum_{i=1}^{N} (Xp)_i^2$$

$$= \frac{\partial}{\partial p_j}\left(\sum_{i=1}^{N}\left(\sum_{q=1}^{K} X_{iq} p_q\right)^2\right) =$$

$$2\sum_{i=1}^{N}\left(\sum_{q=1}^{K} X_{iq} p_q\right) X_{ij} = 2\left[\left(X^T X\right) p\right]_j$$

$$-2\left(X^T y\right) + 2\left(X^T X\right) p = 0$$

$$\left(X^T X\right) p = X^T y \Rightarrow \boxed{p = \left(X^T X\right)^{-1} X^T y}$$

Παράδειγμα

Να βρεθεί το δειγματικό μοντέλο γραμμικής παλινδρόμησης για το σύνολο δεδομένων

$$\{(1,-1,1),(0,-1,-1),(2,0,2),(1,1,2)\}$$

$$y = \begin{bmatrix} 1 \\ -1 \\ 2 \\ 2 \end{bmatrix} \in \mathbb{R}^4 \qquad X = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 0 & -1 \\ 1 & 2 & 0 \\ 1 & 1 & 1 \end{bmatrix} \in \mathbb{R}^{4\times 3} \qquad p = \begin{bmatrix} \alpha \\ b^{(1)} \\ b^{(2)} \end{bmatrix} \in \mathbb{R}^3$$

$$p = \overbrace{(X^T X)}^{\in \mathbb{R}^{3\times 3}}{}^{-1} \underbrace{X^T}_{\in \mathbb{R}^{3\times 4}} \cdot \underbrace{y}_{\in \mathbb{R}^{4\times 1}} \in \mathbb{R}^3$$

$(0, 2, ?)$

Άσκηση

Δείξτε ότι η εκτίμηση ελαχίστων τετραγώνων

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

στη περίπτωση της απλής γραμμικής παλινδρόμησης οδηγεί, όπως περιμένουμε, στις εκτιμήσεις των παραμέτρων:

$$b = \frac{SS_{xy}}{SS_{xx}}, \quad a = \bar{Y} - b\bar{X}$$

Εκτιμήτρια της τυπικής απόκλιση του $\hat{y}^*$

$$s_{\hat{y}^*} = s_e \sqrt{1 + \frac{1}{N} + \frac{(x^* - \bar{X})^2}{SS_{xx}}}$$

Διάστημα εμπιστοσύνης

Το $(1 - a) * 100\%$ διάστημα εμπιστοσύνης για την y^* είναι:

$$[\hat{y}^* - ts_{\hat{y}^*}, \hat{y}^* + ts_{\hat{y}^*}]$$

όπου το t λαμβάνεται από την t_{df}, $df = N - 2$ έτσι ώστε

$$P(T < t) = 1 - a/2$$

- Περιθώριο σφάλματος: $E = ts_{\hat{y}^*}$