ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

Θεωρία 8ης εβδομάδας

Αιτιοκρατικό μοντέλο

$$y = A + Bx$$

Πιθανοθεωρητικό μοντέλο - Μοντέλο απλής γραμμικής παλινδρόμησης

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

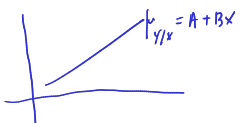
$$A : \text{σταθερός όρος (constant term)}, \quad B : \text{κλίση (slope)}$$

Παραδοχές

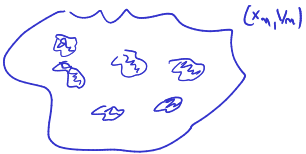
$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$

- 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



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



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

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

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:

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

$$y = \overbrace{A + Bx}^{h_{y|x}} + \epsilon, \quad \epsilon : \text{όρος τυχαίου σφάλματος}$$

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

$$\sigma_\varepsilon = \sqrt{\frac{\sum \varepsilon_n^2}{N}}$$

- Για κάθε x έχουμε υποθέσει ότι το σφάλμα ϵ ακολοθεί την κανονική κατανομή $\mathcal{N}(0, \sigma_\epsilon^2)$.
- 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

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

$$y_n - \hat{y}_n = y_n - (a + b x_n) = y_n - a - b x_n =$$
$$= y_n - \underbrace{y} - b \underbrace{x} - b x_n$$

Συνολικό άθροισμα τετραγώνων

$$SSE = \sum_{n=1}^{N} (y_n - \hat{y})^2$$

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

Άθροισμα τετραγώνων παλινδρόμησης

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

Συντελεστής Προσδιορισμού

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

- Ποσοτικοποιεί την αποτελεσματικότητα του μοντέλου.

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

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

κλιση γραμμης

πως σχηματιζεται γραμμη

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

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

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

Συντελεστής Γραμμικής Συσχέτισης - Pearson

- Συμβολίζεται με ρ όταν αφορά τον πληθυσμό.

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

- Συμβολίζεται με r όταν αφορά ένα δείγμα.

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

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

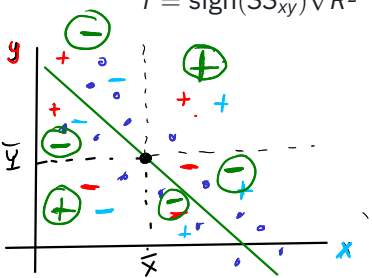
$$R^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} \quad \text{(Συντελεστής Προσδιορισμού)}$$

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

Σχέση μεταξύ συντελεστών γραμμικής συσχέτισης και προσδιορισμού

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

πρόσημο.

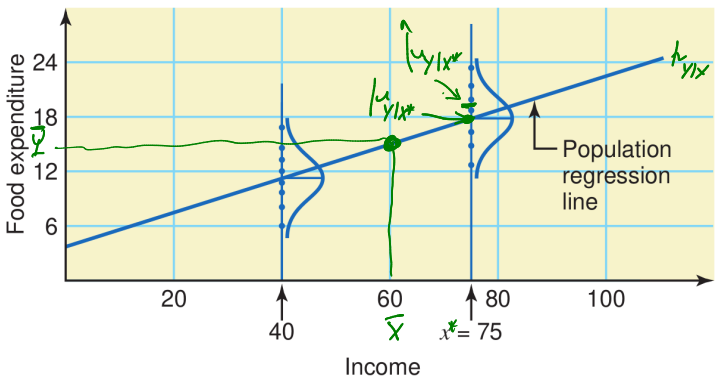


$$SS_{xy} = \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})$$

$$SS_{xx} = \sum (x_n - \bar{x})^2$$

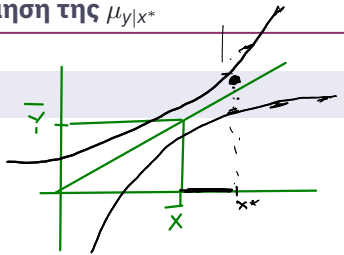
$$SS_{yy} = \sum (y_n - \bar{y})^2$$

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^*;



Διαστήματα εμπιστοσύνης για τις τιμές της εξαρτημένης μεταβλητής

x* δοσμένο



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

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

Διάστημα εμπιστοσύνης

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

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

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

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

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

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

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

$y^* = A + Bx^* + \varepsilon^*$

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

$\sigma^2_{\hat{y}^*} = \sigma^2_\varepsilon \left(\frac{1}{N} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right) + \sigma^2_\varepsilon =$

$= \sigma^2_\varepsilon \left[1 + \frac{1}{N} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$

$$y = A + \sum_{k=1}^{K} B^{(k)} x^{(k)}$$

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

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

Ευθεία παλινδρόμησης για τον πληθυσμό

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

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

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

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

$$\hat{y}_n = a + \mathbf{x}_n^T \mathbf{b} = \alpha + \sum_{k=1}^{K} b^{(k)} x_n^{(k)}$$

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

$$\mathbf{e}^T\mathbf{e} = \sum_{n=1}^{N} e_n^2$$

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

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

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

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

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

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

$$\underset{\sim}{e} \qquad \underset{\sim}{y}$$

$$\begin{bmatrix} 1 & x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(K)} \\ 1 & x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(K)} \\ \vdots & & & & \\ 1 & x_N^{(1)} & x_N^{(2)} & \cdots & x_N^{(K)} \end{bmatrix} \begin{bmatrix} a \\ b^{(1)} \\ \vdots \\ b^{(K)} \end{bmatrix}$$

$$\underset{\sim}{X} \in \mathbb{R}^{N \times (K+1)} \qquad \underset{\sim}{P} \in \mathbb{R}^{(K+1)}$$

$$\underset{\sim}{e} = \underset{\sim}{y} - \underset{\approx}{X}\,\underset{\sim}{P}$$

$$\underset{\sim}{e}^T\underset{\sim}{e} = (\underset{\sim}{y} - \underset{\sim}{\underset{\sim}{X}}\,\underset{\sim}{p})^T(\underset{\sim}{y} - \underset{\sim}{\underset{\sim}{X}}\,\underset{\sim}{p}) = \underset{\sim}{y}^T\underset{\sim}{y} - \underset{\sim}{y}^T\underset{\sim}{\underset{\sim}{X}}\,\underset{\sim}{p} - \underset{\sim}{p}^T\underset{\sim}{\underset{\sim}{X}}^T\underset{\sim}{y} + \underset{\sim}{p}^T\underset{\sim}{\underset{\sim}{X}}^T\underset{\sim}{\underset{\sim}{X}}\,\underset{\sim}{p} =$$

$$y\cdot(Xp) \qquad (Xp)^T\cdot y$$

$$= \underset{\sim}{y}^T\underset{\sim}{y} - 2\,\underset{\sim}{p}^T\underset{\sim}{\underset{\sim}{X}}^T\underset{\sim}{y} + \underset{\sim}{p}^T\underset{\sim}{\underset{\sim}{X}}^T\underset{\sim}{\underset{\sim}{X}}\,\underset{\sim}{p} = Q(\underset{\sim}{p})$$

$$Q(p) = e^Te = y^Ty - 2p^TX^Ty + (Xp)^T(Xp) =$$

$$= y^Ty - 2p^TX^Ty + \sum_{m=1}^{N}([Xp]_m)^2$$

$$\frac{\partial Q}{\partial p_j} = -2\frac{\partial}{\partial p_j}(p^TX^Ty) + \frac{\partial}{\partial p_j}\sum_{m=1}^{N}([Xp]_m)^2 = 0$$

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

$$\frac{\partial}{\partial p_j}\sum_{m=1}^{N}\left(\sum_{k=1}^{k+1}X_{mk}p_k\right)^2 =$$

$$= 2\sum_{m=1}^{N}\sum_{k=1}^{k+1}X_{mk}p_k\cdot\frac{\partial}{\partial p_j}\left(\sum_{k=1}^{k+1}X_{mk}p_k\right) =$$

$$= 2\sum_{m=1}^{N}\sum_{k=1}^{k+1}X_{mk}p_kX_{mj} = 2\sum_{m=1}^{N}X_{mj}\sum_{k=1}^{k+1}X_{mk}p_k =$$

$$= 2\sum_{m=1}^{N}X_{mj}[Xp]_m = 2[\underset{\sim}{\underset{\sim}{X}}^T\underset{\sim}{\underset{\sim}{X}}\,\underset{\sim}{p}]_j$$

$$2[X^T X p]_j = 2[X^T y]_j \quad , j = 1, \dots k+1$$

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

Παράδειγμα

Να βρεθεί το δειγματικό μοντέλο γραμμικής παλινδρόμησης για το σύνολο δεδομένων

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

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