ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

23-03-2020

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

- 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

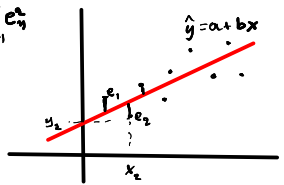
$$y = \overbrace{A + Bx}^{\mu_{y|x}} + \epsilon$$

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

$y = A + Bx + \varepsilon$ $\quad \{(x_1,y_1),(x_2,y_2),\dots,(x_N,y_N)\}$

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

$\min_{a,b} \sum_{n=1}^{N} e_n^2$



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

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

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:

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

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

- 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

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

$$S_{yy} = \sum_{n=1}^{N} (y_n - \bar{y})^2 \qquad SS_{xy} = \sum_{n=1}^{N} (x_n - \bar{x})(y_n - \bar{y})$$

$$\sigma_\epsilon = \sqrt{\frac{SSE}{N}}$$

Συνολικό άθροισμα τετραγώνων

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

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.

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

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

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

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

Συντελεστής Γραμμικής Συσχέτισης - Pearson

- Συμβολίζεται με ρ όταν αφορά τον πληθυσμό.

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

- Συμβολίζεται με r όταν αφορά ένα δείγμα.

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

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

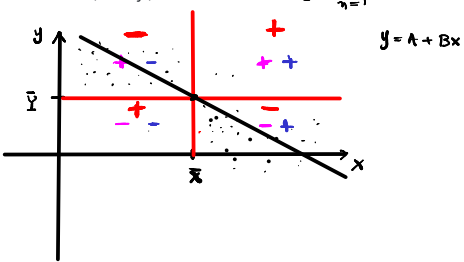
$$R^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}} \quad \text{(Συντελεστής Προσδιορισμού)}$$

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

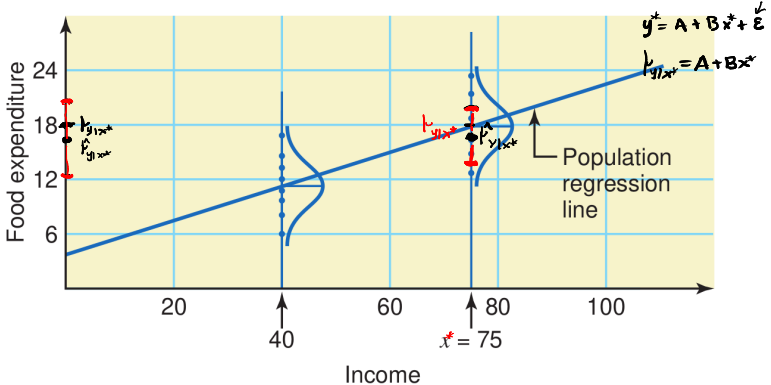
Σχέση μεταξύ συντελεστών γραμμικής συσχέτισης και προσδιορισμού

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

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



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

$y_n = A + Bx_n + \varepsilon_n \;,\; n=1,\dots,N \qquad \frac{1}{N}\sum_{n=1}^{N} y_n = \frac{1}{N}\sum_{n=1}^{N} A + \frac{1}{N}\sum_{n=1}^{N} Bx_n + \frac{1}{N}\sum_{n=1}^{N} \varepsilon_n \Leftrightarrow \boxed{\bar{Y} = A + B\bar{x} + \bar{\varepsilon}}$

$E\{\bar{Y}\} = A + BE\{\bar{x}\} + E\{\bar{\varepsilon}\}^{0}$

$\mu_y = A + B\mu_x$

$\hat{\mu}_{y|x^*} = a + bx^* \;,\; a = \bar{Y} - b\bar{x} \rightarrow \boxed{\hat{\mu}_{y|x^*} = \bar{Y} + b(x^* - \bar{x})}$

$\sigma_x^2 = var\{x\} \;,\; cov\{x,y\} = E\{(x - E(x))(Y - E\{Y\})\} \qquad var\{x+y\} = var\{x\} + var\{y\} + 2cov\{x,y\} \qquad var\{ax\} = a^2 var\{x\}$

$var\{\hat{\mu}_{y|x^*}\} = var\{\bar{Y}\} + (x^* - \bar{x})^2 var\{b\} + 2(x^* - \bar{x})cov\{\bar{Y}, b\}$

$var\{\bar{Y}\} = var\left\{\frac{1}{N}\sum y_n\right\} = \frac{1}{N^2}\sum var\{y_n\} = \frac{1}{N^2}\sigma_\varepsilon^2 N = \frac{\sigma_\varepsilon^2}{N} \;,\; var\{b\} = \frac{\sigma_\varepsilon^2}{SS_{xx}}$

$b = \frac{SS_{xy}}{SS_{xx}} \;,\; SS_{xy} = \sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{Y}) = \sum_{n=1}^{N}(x_n - \bar{x})y_n - \bar{Y}\sum_{n=1}^{N}(x_n - \bar{x})^{0} = \sum_{n=1}^{N}(x_n - \bar{x})y_n$

$\boxed{b = \sum_{n=1}^{N} c_n y_n \;,\; c_n = \frac{x_n - \bar{x}}{SS_{xx}}}$

$\boxed{\sum_{n=1}^{N} c_n = 0 \;,\; \sum_{n=1}^{N} c_n x_n = 1 \quad \text{(άσκηση)}}$

$cov(\bar{Y}, b) = E\{(\bar{Y} - E\{\bar{Y}\})(b - E\{b\})\} = E\{\bar{\varepsilon}\cdot(b - B)\} = E\{\bar{\varepsilon}\cdot(\sum c_n y_n - B)\} = E\{\bar{\varepsilon}\cdot(\sum c_n(A + Bx_n + \varepsilon_n) - B)\} =$

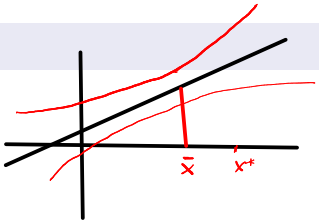
$= E\{\bar{\varepsilon}(A\sum c_n^{0} + B\sum c_n x_n^{1} + \sum c_n\varepsilon_n - B)\} = E\{\bar{\varepsilon}\sum c_n\varepsilon_n\} = E\{\sum c_n\varepsilon_n\bar{\varepsilon}\} = \sum_{n=1}^{N} c_n E\{\varepsilon_n\bar{\varepsilon}\} =$

$= \sum_{n=1}^{N} c_n E\{\varepsilon_n \frac{1}{N}\sum_{m=1}^{N}\varepsilon_m\} = \sum_n c_n E\{\frac{1}{N}\sum\varepsilon_n^2\} = \sigma_\varepsilon^2\cdot\sum_{n=1}^{N} c_n = 0$

$\boxed{var\{\hat{\mu}_{y|x^*}\} = \sigma^2_{\hat{\mu}_{y|x^*}} = \sigma_\varepsilon^2\left(\frac{1}{N} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)}$

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

$\hat{y}_i = a_i + bx^*$

$\hat{y}_{i|x^*}$

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

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

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

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

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

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

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

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

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

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

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

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

Πολλαπλή Γραμμική Παλινδρόμηση

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:

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