Πολλαπλή Παλινδρόμηση

Δημήτρης Κουγιουμτζής

10 Μαΐου 2011

Μοντέλο γραμμικής πολλαπλής παλινδρόμησης

Μοντέλο πολυωνυμικής γραμμικής παλινδρόμησης βαθμού k

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \epsilon.$$

$$\Downarrow$$

$$x_1 = x, \quad x_2 = x^2, \quad \ldots \quad x_k = x^k,$$

$$\Downarrow$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon,$$

$$\mathrm{E}[\epsilon] = 0, \quad \sigma_e^2 = \mathrm{Var}[\epsilon]$$

Γενικά: $x_1, x_2, \ldots, x_k$ ανεξάρτητες μεταβλητές

μοντέλο γραμμικής πολλαπλής παλινδρόμησης

Μοντέλο προσθετικής πολλαπλής παλινδρόμησης

Για x_1 και x_2:

1. Το μοντέλο πρώτου πολυωνυμικού βαθμού (γραμμικής πολλαπλής παλινδρόμησης)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

2. Το μοντέλο δεύτερου πολυωνυμικού βαθμού χωρίς αλληλεπίδραση

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \epsilon.$$

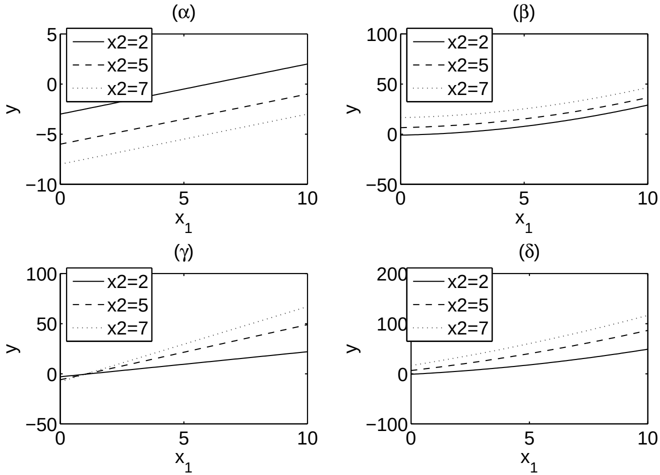
3. Το μοντέλο πρώτου πολυωνυμικού βαθμού με αλληλεπίδραση

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

4. Το πλήρες μοντέλο δευτέρου πολυωνυμικού βαθμού (με αλληλεπίδραση)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon.$$

Γραφική διερεύνηση μοντέλου



(α) $y = -1 + 0.5x_1 - x_2$, (β) $y = -1 + 0.5x_1 + 25x_1^2 - x_2 + 0.5x_2^2$,
(γ) $y = -1 + 0.5x_1 - x_2 + x_1x_2$,
(δ) $y = -1 + 0.5x_1 + 25x_1^2 - x_2 + 0.5x_2^2 + x_1x_2$.

Εκτίμηση μοντέλου πολλαπλής γραμμικής παλινδρόμησης

Πολυ-μεταβλητό δείγμα μεγέθους n: $\{x_{1i}, x_{2i}, \ldots, x_{ki}, y_i\}_{i=1}^{n}$
Εκτίμηση παραμέτρων με τη μέθοδο ελαχίστων τετραγώνων:

$$f(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})\right)^2 .$$

Κανονικές εξισώσεις

$$\begin{aligned}
b_0 n + b_1 \sum x_{1i} + b_2 \sum x_{2i} + \cdots b_k \sum x_{ki} &= \sum y_i \\
b_0 \sum x_{1i} + b_1 \sum x_{1i}^2 + b_2 \sum x_{1i} x_{2i} + \cdots b_k \sum x_{1i} x_{ki} &= \sum x_{1i} y_i \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad & \qquad \vdots \\
b_0 \sum x_{ki} + b_1 \sum x_{1i} x_{ki} + b_2 \sum x_{2i} x_{ki} + \cdots b_k \sum x_{ki}^2 &= \sum x_{ki} y_i
\end{aligned}$$

$\Longrightarrow$ εκτιμήσεις $b_0, b_1, \ldots, b_k$.

Προσαρμογή μοντέλου πολλαπλής γραμμικής παλινδρόμησης

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}$$

$e_i = y_i - \hat{y}_i$.

$$s_e^2 = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$\text{adj}R^2 = 1 - \frac{n-1}{n-(k+1)} \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Συμπερασματολογία για παραμέτρους / προβλέψεις

Εκτίμηση διασποράς $s_{b_j}^2$ για $j = 0, 1, \ldots, k$ (είναι πολύπλοκη)
$(1-\alpha)\%$ παραμετρικό διάστημα εμπιστοσύνης για β_j:

$$b_j \pm t_{n-(k+1),1-\alpha/2} s_{b_j}.$$

H_0: $\beta_j = \beta_j^0$ γίνεται με το στατιστικό

$$t = \frac{\beta_j - \beta_j^0}{s_{b_j}} \sim t_{n-(k+1)}.$$

Εκτίμηση διασποράς $s_{\hat{y}}^2$ (είναι πολύπλοκη)
$(1-\alpha)\%$ διάστημα εμπιστοσύνης για τη μέση τιμή της y όταν δίνονται τα $x_1, \ldots, x_k$:

$$\hat{y} \pm t_{n-(k+1),1-\alpha/2} s_{\hat{y}}$$

$(1-\alpha)\%$ διάστημα πρόβλεψης μιας (μελλοντικής) τιμής της y

$$\hat{y} \pm t_{n-(k+1),1-\alpha/2} \sqrt{s_e^2 + s_{\hat{y}}^2}.$$

Παράδειγμα, δείκτης προσρόφησης

A/A	Εξαγώγιμο σίδηρο x_1	Εξαγώγιμο αργίλλιο x_2	Δείκτης προσρόφησης y
1	61	13	4
2	175	21	18
3	111	24	14
4	124	23	18
5	130	64	26
6	173	38	26
7	169	33	21
8	169	61	30
9	160	39	28
10	244	71	36
11	257	112	65
12	333	88	62
13	199	54	40

Παράδειγμα, δείκτης προσρόφησης

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

Παράμετρος	*Εκτιμητής* b_i	*Εκτίμηση SD* s_{b_i}
β_0	-7.351	3.485
β_1	0.11273	0.02969
β_2	0.34900	0.07131

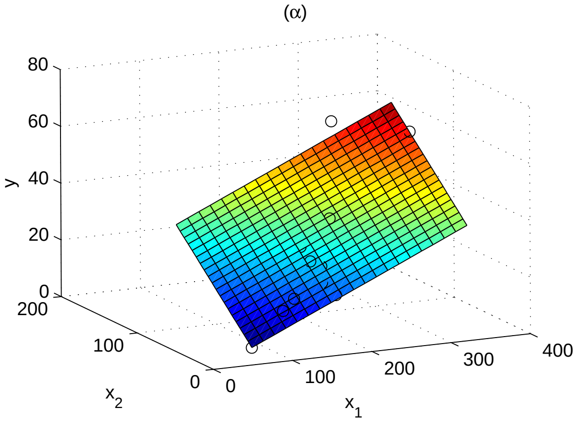
95% διάστημα εμπιστοσύνης για το συντελεστή του εξαγώγιμου σιδήρου β_1 ($t_{10,0.975} = 2.228$)

$$0.11273 \pm 2.228 \cdot 0.02969 = [0.0466, 0.1789]$$

και για β_2

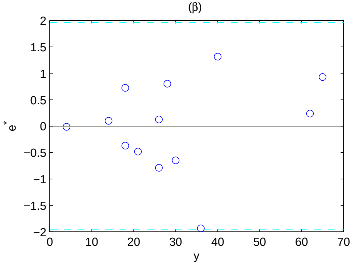
$$0.34900 \pm 2.228 \cdot 0.07131 = [0.1901, 0.5079].$$

Παράδειγμα, δείκτης προσρόφησης



$s_e = 4.616$
$R^2 = 0.948$
$adjR^2 = 0.931$

Διαγνωστικός έλεγχος



Πρόβλεψη για $x_1 = 160$ και $x_2 = 39$

$$\hat{y} = -7.351 + 0.11273 \cdot 160 + 0.34900 \cdot 39 = 24.30.$$

95% διάστημα εμπιστοσύνης για το μέσο δείκτη προσρόφησης y

$$24.30 \pm 2.228 \cdot 1.30 = [21.40, 27.20] \quad s_{\hat{y}} = 1.30$$

95% διάστημα πρόβλεψης για μια μελλοντική τιμή του y

$$24.30 \pm 2.228 \cdot \sqrt{(4.616)^2 + (1.30)^2} = [13.62, 34.98]$$

Επιλογή μεταβλητών

$x_1, x_2, \ldots, x_k$ ανεξάρτητες μεταβλητές, k μεγάλο

Επιλογή του μικρότερου δυνατού υποσύνολου ανεξάρτητων (επεξηγηματικών) μεταβλητών που εξηγεί καλά τη y

Απλή προσέγγιση:
Προσαρμογή όλων των δυνατών μοντέλων
Κριτήριο προσαρμογής, π.χ. adjR^2

Υπολογισμός βέλτιστου μοντέλου πολλαπλής παλινδρόμησης βηματικά:
διαδοχικοί έλεγχοι για H_0: $\beta_j = 0$

- μέθοδος απαλοιφής προς τα πίσω
- επιλογή προς τα μπρος

Πολλαπλή συγγραμικότητα

Πρόβλημα **πολλαπλής συγγραμικότητας**:
Κάποια(ες) από τις $x_1, x_2, \ldots, x_k$ είναι ισχυρά αλληλοεξαρτημένες

Προσαρμογή μοντέλου παλινδρόμησης της x_j ως προς όλες τις υπόλοιπες.

Αν η x_j μπορεί να προβλεφθεί καλά από τις υπόλοιπες $k-1$ $\Rightarrow$ πολλαπλή συγγραμικότητα