ΜΕΜ-205 Περιγραφική Στατιστική

Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

09-03-2020

$$\mathcal{Q}(a,b) = (\mathbf{y} - a\mathbf{u} - b\mathbf{x})^T(\mathbf{y} - a\mathbf{u} - b\mathbf{x})$$

$$\bar{X} = \frac{1}{N}\mathbf{u}^T\mathbf{x}, \quad \bar{Y} = \frac{1}{N}\mathbf{u}^T\mathbf{y}$$

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

$$SS_{xy} = (\mathbf{x} - \bar{X}\mathbf{u})^T(\mathbf{y} - \bar{Y}\mathbf{u}), \quad SS_{xx} = (\mathbf{x} - \bar{X}\mathbf{u})^T(\mathbf{x} - \bar{X}\mathbf{u})$$

Παράδειγμα

Βρείτε τη εκτίμηση ελαχίστων τετραγώνων του μοντέλου γραμμικής παλινδρόμησης υποθέτοντας τα παρακάτω δεδομένα κάνοντας χρήση των διανυσματικών εκφράσεων.

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

$N = 3$

$x = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$ $y = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$

$u = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \in \mathbb{R}^3$ $\bar{x} = \frac{1}{3} u^T x = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} = 1$ $\bar{y} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \frac{5}{3}$

$$SS_{xy} = \left(\begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} - 1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)^T \left(\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} - \frac{5}{3} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right) = \frac{1}{3} \begin{bmatrix} -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} = 1$$

$SS_{xx} = 2$

$b = \frac{SS_{xy}}{SS_{xx}} = \frac{1}{2}$ $a = \frac{5}{3} - \frac{1}{2} \cdot 1 = \frac{7}{6}$

$$\hat{y} = \frac{7}{6} + \frac{1}{2} x$$

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

- 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

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

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

Γιατί εμφανίζεται το $N-2$;

$\hat{y}_n = a + bx_n$

$a = a(\bar{x}, \bar{y})$ $\quad \bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$

$\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$

$$\text{SSE} = \sum_{n=1}^{N} (y_n - a - bx_n)^2 \qquad a = \bar{y} - b\bar{x}$$

$$= \sum_{n} (y_n - \bar{y} + b\bar{x} - bx_n)^2$$

$$s_e = \sqrt{\frac{SS_{yy} - b * SS_{xy}}{N-2}}$$

όπου:

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

Υπενθυμίζεται

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

Εάν είχαμε γνώση των δεδομένων του πληθυσμού θα μπορούσαμε να υπολογίσουμε την τυπική απόκλιση των τυχαίων σφαλμάτων από τη σχέση:

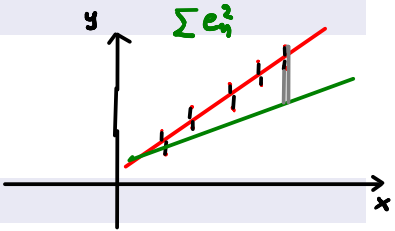
$$\sigma_\epsilon = \sqrt{\frac{SS_{yy} - B * SS_{xy}}{N_p}}$$

όπου σε αυτή την περίπτωση θα είχαμε:

$$SS_{yy} = \sum_{n=1}^{N_p} (y_n - \mu_y)^2, \quad SS_{xy} = \sum_{n=1}^{N_p} (x_n - \mu_x)(y_n - \mu_y)$$

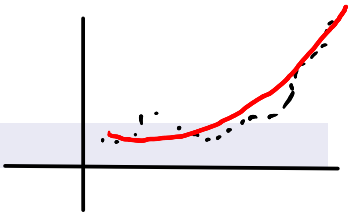
Συνολικό Άθροισμα τετραγώνων

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



Άθροισμα τετραγώνων παλινδρόμησης

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



Συντελεστής Προσδιορισμού

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

- Ποσοτικοποιεί την αποτελεσματικότητα του μοντέλου.

$\hat{y}_n = a + bx_n$

$a = \bar{Y} - b\bar{x}$

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

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

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

$$\sum_n (\hat{y}_n - \bar{Y})(y_n - \hat{y}_n) = \sum_n (a + bx_n - \bar{Y})(y_n - a - bx_n) = \sum_n (\bar{Y} - b\bar{x} + bx_n - \bar{Y}) \cdot$$

$$\cdot (y_n - \bar{Y} + b\bar{x} - bx_n) =$$

$$= b\sum_n (x_n - \bar{x})(y_n - \bar{Y} + b\bar{x} - bx_n) = b\sum_n (x_n - \bar{x})[y_n - \bar{Y} - b(x_n - \bar{x})] =$$

$$= b\left\{\sum_n (x_n - \bar{x})(y_n - \bar{Y}) - b\sum_n (x_n - \bar{x})^2\right\} = b\{SS_{xy} - bSS_{xx}\} = 0$$

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

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

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

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

Παράδειγμα

Βρείτε τον συντελεστή προσδιορισμού του συνόλου δεδομένων:

$$\{(0,1),(1,3),(2,4),(5,4)\}$$

	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
	0	1	-2	-2	4	4	4
	1	3	-1	0	0	1	0
	2	4	0	1	0	0	1
	5	4	3	1	3	9	1
Σ	8	12	0	0	7	14	6

$\bar{x} = \frac{8}{4} = 2 \quad \bar{y} = 3$

$$r^2 = \frac{7^2}{14 \cdot 6} = 0.5833\ldots$$

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

Μέση τιμή, τυπική απόκλιση και κατανομή του b

$$\mu_b = B, \quad \sigma_b = \frac{\sigma_\epsilon}{\sqrt{SS_{xx}}}$$

$$b \sim \mathcal{N}(\mu_b, \sigma_b)$$

- Όταν το σ_ϵ είναι άγνωστο δεν μπορούμε να υπολογίσουμε το σ_b

Εκτιμήτρια της τυπικής απόκλιση του b

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

Το $(1 - a) * 100\%$ διάστημα εμπιστοσύνης για το B είναι:

$$[b - ts_b, b + ts_b]$$

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

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

- Περιθώριο σφάλματος: $E = ts_b$

Παράδειγμα

Για επτά νοικοκυριά μιας πόλης έχουμε τα ακόλουθα ζεύγη ετήσιου εισοδήματος και εξόδων σίτισης

$$\{(55, 14), (83, 24), (38, 13), (61, 16), (33, 9), (49, 15), (67, 17)\}$$

1. Βρείτε την προσεγγιστική ευθεία γραμμικής παλινδρόμησης ($\hat{y} = a + bx$) χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων.
2. Υπολογίστε το 95% διάστημα εμπιστοσύνης για την παραμετρο B του πληθυσμού ($y = A + Bx$).

x	y	xy	x^2	y^2
55	14	770	3025	196
83	24	1992	6889	576
38	13	494	1444	169
61	16	976	3721	256
33	9	297	1089	81
49	15	735	2401	225
67	17	1139	4489	289
386	108	6403	23858	1792

$\bar{x} = \frac{386}{7} = 55.14$

$\bar{y} = \frac{108}{7} = 15.42$

$SS_{xy} = \sum_n x_n y_n - \frac{\sum x_n \sum y_n}{N} = 447.57$

$SS_{xx} = \sum_n x_n^2 - \frac{(\sum x_n)^2}{N} = 1772.85$

$b = \frac{SS_{xy}}{SS_{xx}} = 0.2525$

$\hat{y} = 1.505 + 0.2525\,x$

$a = \bar{y} - b\bar{x} = 1.505$

$S_e = \sqrt{\frac{SS_{yy} - b SS_{xy}}{N-2}} = 1.5939$

$S_b = \frac{S_e}{\sqrt{SS_{xx}}} = \frac{1.5939}{\sqrt{1772.85}} = 0.0378$

$df = 7 - 2 = 5$ $\quad t = 2.571$

$B \in [b - tS_b, b + tS_b] = [0.155, 0.35]$

Διάστημα Εμπιστοσύνης του **B**

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
						Confidence Level					

Διάστημα Εμπιστοσύνης του **B**

- Όταν τα x και y δεν συνδέονται με γραμμικό τρόπο.
- Αλλά υπάρχει μετασχηματισμός $g : y \to y'$ τέτοιος ώστε x και y' να μπορούν να περιγραφούν με ένα γραμμικό μοντέλο.

Παραδείγματα

- $y = e^x$ $\xrightarrow{g(y) = \ln y}$ $\ln y = x$
- $y = x^2$ $\xrightarrow{g(y) = \sqrt{y}}$ $\sqrt{y} = x$
- $y = \frac{1}{x}$ $\xrightarrow{g(y) = y^{-1}}$ $y^{-1} = x$
- $y = \log(x)$ $\xrightarrow{g(y) = e^y}$ $e^y = x$

Παράδειγμα

Βρείτε κατάλληλο μετασχηματισμό για το παρακάτω σύνολο δεδομένων ώστε να είναι εφαρμόσιμο το μοντέλο γραμμικής παλινδρόμησης.

$$\{(0,1),(1,2),(4,14),(5,25),(6,35)\}$$

$g(y) = \sqrt{y}$

$\{(0,1),(1,1.4145),(4,3.74),(5,5),(6,5.91)\}$