ΜΕΜ-205 Περιγραφική Στατιστική

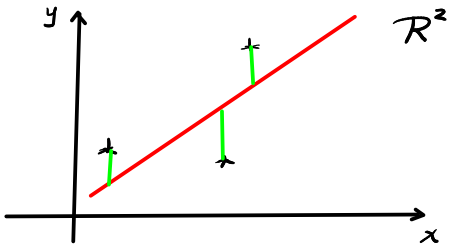
Τμήμα Μαθηματικών και Εφ. Μαθηματικών, Πανεπιστήμιο Κρήτης

Κώστας Σμαραγδάκης (kesmarag@gmail.com)

20-03-2023

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

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



R^2 $(x_i, y_i) \rightarrow [0, 1]$

$$R^2 = \frac{SSR}{SST}$$

κλίση

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

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

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

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

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

$$SS_{xx} = \sum_{n=1}^{N} (x_n - \bar{x})^2$$

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

$$R^2 = \frac{\left[\frac{1}{N-1} SS_{xy}\right]^2}{\frac{1}{N-1} SS_{xx} \cdot \frac{1}{N-1} SS_{yy}}$$

Cov[x,y]

Std(x)

Std(y)

Παράδειγμα

Βρείτε τον συντελεστή προσδιορισμού του συνόλου δεδομένων:

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

$N = 4$

x	y	x^2	y^2	xy
0	1	0	1	0
1	3	1	9	3
2	4	4	16	8
5	4	25	16	20
8	12	30	42	31

$$SS_{xx} = \sum x^2 - \frac{1}{N}\left(\sum x\right)^2 = 30 - \frac{1}{4}8^2 = 30 - 16 = 14$$

$$SS_{yy} = 42 - \frac{1}{4}12^2 = 6$$

$$SS_{xy} = \sum xy - \frac{\sum x \sum y}{N} = 31 - \frac{8 \cdot 12}{4} = 7$$

$$R^2 = \frac{7^2}{6 \cdot 14} = \frac{7}{12}$$

$y = A + Bx + \varepsilon$, $\mu_{y|x} = A + Bx$, A, B άγνωστα

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

Μέση τιμή, τυπική απόκλιση και κατανομή του b

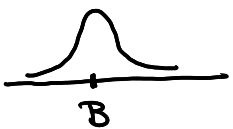
$\hat{\mu}_{y|x} = a + bx$

$y = A + Bx + \varepsilon$

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

$$\mu_{\mathrm{b}} = \mathrm{B}, \quad \sigma_{\mathrm{b}} = \frac{\sigma_\epsilon}{\sqrt{\mathrm{SS_{xx}}}}$$

$$\mathrm{b} \sim \mathcal{N}(\mu_{\mathrm{b}}, \sigma^2_{\mathrm{b}})$$



- Όταν το σ είναι άγνωστο δεν μπορούμε να υπολογίσουμε το σ_{b}

Εκτιμήτρια της τυπικής απόκλιση του b

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

← αυτό μπορούμε να υπολογίσουμε.

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

$$SS_{xy} = \sum_{n=1}^{N}(x_n - \bar{x})(y_n - \bar{y}) = \sum_{n=1}^{N}(x_n - \bar{x})\,y_n - \bar{y}\cancel{\sum_{n=1}^{N}(x_n - \bar{x})}^{\,0}$$

$$\sum_{n=1}^{N}(x_n - \bar{x}) = \sum x_n - N\bar{x} = N\frac{1}{N}\sum x_n - N\bar{x} = 0$$

($\frac{1}{N}\sum x_n = \bar{x}$)

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

Άρα

$$b = \frac{\sum_{n=1}^{N}(x_n - \bar{x})\,y_n}{\sum_{n=1}^{N}(x_n - \bar{x})^2} = \sum_{n=1}^{N}\underbrace{\frac{(x_n - \bar{x})}{\sum_{k=1}^{N}(x_k - \bar{x})^2}}_{C_n}\,y_n = \sum_{n=1}^{N} C_n\,y_n$$

C_n: αρ. σταθ., y_n: Τυχαίες μεταβλητές

$$y_n = \overline{\alpha + b x_n} + \varepsilon$$

X_1, X_2 ανεξάρτητες τότε $Var(X_1+X_2) = Var(X_1) + Var(X_2)$

$X = \alpha Y \Rightarrow Var(X) = \alpha^2 Var(Y)$

$$Var(b) = \sum_{n=1}^{N} c_n^2 \, Var(y_n) = \sigma_\varepsilon^2 \sum_{n=1}^{N} c_n^2$$

$$\sum_{n=1}^{N} c_n^2 = \sum_{n=1}^{N} \frac{(x_n - \bar{x})^2}{\left(\sum_{k=1}^{N} (x_k - \bar{x})^2\right)^2} = \frac{\sum_{n=1}^{N} (x_n - \bar{x})^2}{\left[\sum_{n=1}^{N} (x_n - \bar{x})^2\right]^2} = \frac{1}{SS_{xx}}$$

$$Var(b) = \frac{\sigma_\varepsilon^2}{SS_{xx}} \quad \text{ή} \quad \sigma_b^2 = \frac{\sigma_\varepsilon^2}{SS_{xx}} \quad \text{ή} \quad \boxed{\sigma_b = \frac{\sigma_\varepsilon}{\sqrt{SS_{xx}}}}$$

$b \sim t(B, s_b^2)$ ή $b \sim N(B, \sigma_b^2)$

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

$$B \in [b - ts_b, b + ts_b]$$

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 ($y = A + Bx$).

① $b = \frac{SS_{xy}}{SS_{xx}}$, $a = \bar{y} - b\bar{x}$

$n = 7$

$df = n - 2 = 5$

② $B \in [b - t S_b , b + t S_b]$ (2.571 under t, 2.571 under $+t$)

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**

$R^2 << 1$

$y = e^x \Rightarrow \ln y = x \ln e \Rightarrow \ln y = x$

$y_n = A + Bx_n + \varepsilon_n$

$\ln y_n = A + Bx_n + \varepsilon_n$

$y_n^* = A + Bx_n + \varepsilon_n$

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

- 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

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

$\exp(\ln \hat{y}) = e^{\alpha + bx}$

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

Παράδειγμα

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

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

$(5, 5^2)$, $(6, 6^2-1)$

$(0, 0^2+1)$

$(1, 1^2+1)$

$(4, 4^2-2)$

$y = x^2$ ή $\sqrt{y} = x$

$\{(0,1), (1,\sqrt{2}), (4,\sqrt{14}), (5,5), (6,\sqrt{35})\}$ $\rightarrow R^2 \approx 1$

↳ α, b τ.ω SSE να γίνει ελάχιστο.

$\sqrt{y} = \alpha x + b \Rightarrow y = (\alpha x + b)^2$

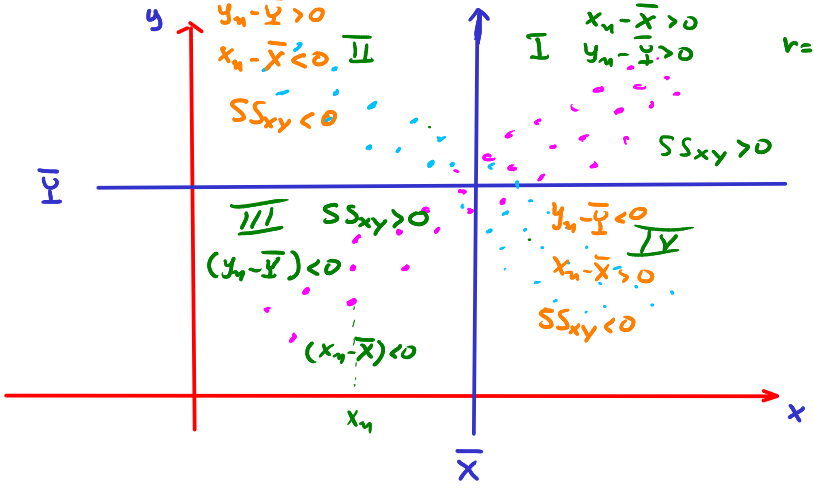
$$R^2 = \frac{SS_{xy}^2}{SS_{xx}SS_{yy}}$$ (Συντελεστής Προσδιορισμού)

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

Σχέση μεταξύ συντελεστών γραμμικής συσχέτισης και προσδιορισμού

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

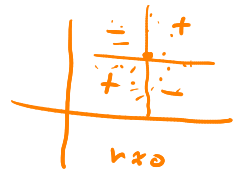
$$\text{Sign}(SS_{xy}) = \begin{cases} 1, & SS_{xy} \geq 0 \\ -1, & SS_{xy} < 0 \end{cases}$$



$r \in [-1,1]$ $\quad \epsilon [0,1]$

$r = sign(SS_{xy}) \sqrt{R^2} \in [-1,1]$

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



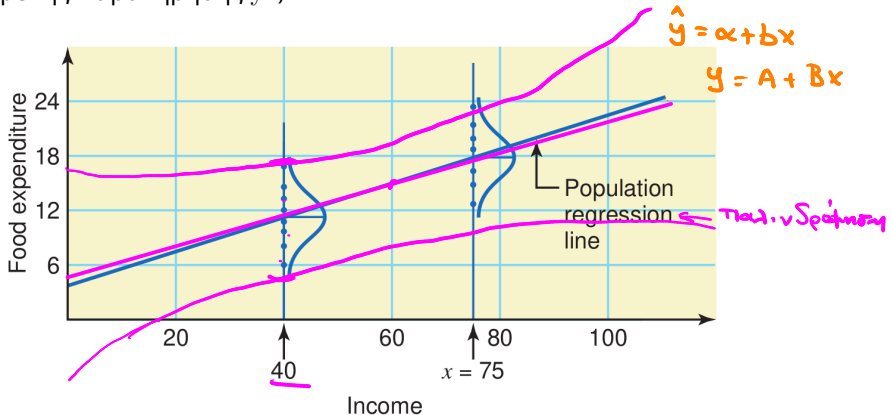
1. Για δοσμένο x^* ποιο είναι το διάστημα εμπιστοσύνης (1-α)*100% για τη μέση τιμή $\mu_{y|x^*}$;
2. Για δοσμένο x^* ποιο είναι το διάστημα εμπιστοσύνης (1-α)*100% για την τιμή μιας συγκεκριμένης παρατήρησης y^*;

$x_m \to \hat{y}_m$

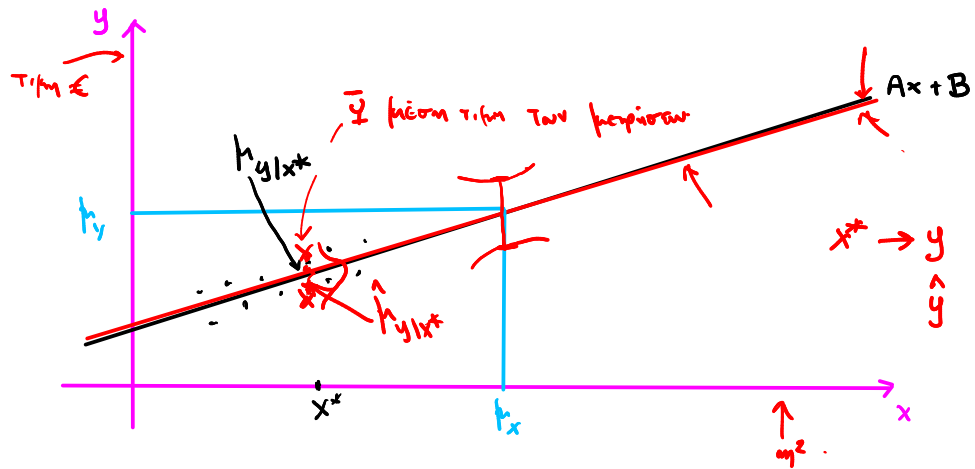
(x_m, y_m)

$\hat{y} = \alpha + \beta x$

$y = A + Bx$



Διαστήματα εμπιστοσύνης για τις τιμές της εξαρτημένης μεταβλητής



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

τυπική απόκλιση των σφαλμάτων

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

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

$a + bx$

$B: \sigma_B = \frac{\sigma_e}{\sqrt{SS_{xx}}}$ $\quad s_B = \frac{s_e}{\sqrt{SS_{xx}}}$

Διάστημα εμπιστοσύνης

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

$\bar{y}$ $\quad \bar{y}$

$\mu_{y|x^*} \in [\hat{\mu}_{y|x^*} - t s_{\hat{\mu}_{y|x^*}}, \hat{\mu}_{y|x^*} + t s_{\hat{\mu}_{y|x^*}}]$

t - κατανομή

σ_e άγνωστο.

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

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

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

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

$$\mathrm{s}_{\hat{\mathrm{y}}^*} = \mathrm{s}_\mathrm{e}\sqrt{1 + \frac{1}{\mathrm{N}} + \frac{(\mathrm{x}^* - \bar{\mathrm{X}})^2}{\mathrm{SS}_{\mathrm{xx}}}}$$

Διάστημα εμπιστοσύνης

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

$$[\hat{\mathrm{y}}^* - \mathrm{ts}_{\hat{\mathrm{y}}^*}, \hat{\mathrm{y}}^* + \mathrm{ts}_{\hat{\mathrm{y}}^*}]$$

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

$$\mathrm{P}(\mathrm{T} < \mathrm{t}) = 1 - \mathrm{a}/2$$

- Περιθώριο σφάλματος: $\mathrm{E} = \mathrm{ts}_{\hat{\mathrm{y}}^*}$

$$s^2_{\hat{y}_*} = s_e^2\left(\frac{1}{N} + \frac{(x_* - \bar{X})^2}{SS_{xx}}\right) + s_e^2 = s_e^2\left[1 + \frac{1}{N} + \frac{(x_* - \bar{X})^2}{SS_{xx}}\right]$$

---

y^* $\quad$ $\hat{y}^*$ $\quad$ $y^* = \underbrace{A + Bx^*}_{\mu_{Y|x^*}} + \varepsilon_{x^*}$

$\mu_{Y|x^*}$ $\quad$ $\hat{\mu}_{Y|x^*}$ $\quad$ $\mu_{Y|x^*} = \mathbb{E}[y^*]$

y^* $\quad$ $s_e\sqrt{1 + \frac{1}{N} + \frac{(x^* - \bar{X})^2}{SS_{xx}}}$ $\quad$ $\hat{y}^* = \underbrace{a + bx^*}_{\hat{\mu}_{Y|x^*}} + e_{x^*}$

$\mu_{Y|x^*}$ $\quad$ $s_e\sqrt{\frac{1}{N} + \frac{(x^* - \bar{X})^2}{SS_{xx}}}$

Παράδειγμα

$\{(1,1), (1,2), (2,3), (2,4)\}$ $N = 4$ $x^* = 1.5$

$$S_{\hat{y}^*} = S_e \sqrt{1 + \frac{1}{4} + \frac{(x^* - \bar{x})^2}{SS_{xx}}}$$

$df = N - 2 = 2$ t

$$y^* \in \left[\hat{y}^* - t\, S_{\hat{y}^*},\ \hat{y}^* + t\, S_{\hat{y}^*}\right]$$