

## Statistica Descrittiva

### Soluzioni 7. Interpolazione: minimi quadrati

#### Introduzione

Siano date le osservazioni del tipo  $(x_i, y_i)$  rilevate sulla  $i$ -esima unità statistica,  $i = 1, \dots, N$ . Supponiamo che sia di interesse valutare la relazione tra le variabili  $Y$  e  $X$  e, in particolare, la dipendenza di  $Y$  da  $X$ . A questo scopo si consideri il modello lineare bivariato in cui  $Y$  è la variabile risposta (dipendente) e  $X$  la variabile esplicativa (indipendente). Come particolare modello di regressione lineare utilizziamo la **retta di regressione**, con intercetta  $a$  e coefficiente angolare  $b$

$$y = a + b \cdot x,$$

dove  $y'$  indica il valore teorico ottenuto tramite il modello. Tra le infinite rette che interpolano i punti  $(x_i, y_i)$ ,  $i = 1, \dots, N$ , cerchiamo la retta che riduce al minimo le distanze tra i valori osservati  $y_i$  ed i valori teorici  $y'_i$ . Tramite il metodo dei minimi quadrati, si ricava che tale retta ha coefficiente angolare

$$\hat{b} = \frac{\sigma_{xy}}{\sigma_x^2}$$

e intercetta

$$\hat{a} = m_y - \hat{b} \cdot m_x.$$

Nell'espressione del calcolo di  $\hat{b}$ ,  $\sigma_{xy}$  indica la **covarianza** tra  $x$  e  $y$ ,

$$\sigma_{xy} = \frac{\sum_{i=1}^N (x_i - m_x) \cdot (y_i - m_y)}{n} = \frac{\sum_{i=1}^N x_i \cdot y_i}{N} - m_x \cdot m_y.$$

La covarianza determina una misura del legame lineare tra  $x$  e  $y$  ed è tale per cui

$$-\sigma_x \cdot \sigma_y \leq \sigma_{xy} \leq \sigma_x \cdot \sigma_y.$$

La sua versione standardizzata porta alla definizione del **coefficiente di correlazione di Bravais-Pearson**

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y},$$

il quale assume valori compresi tra  $-1$  e  $1$ . Più il valore di  $r$  si avvicina a  $1$ , più le osservazioni  $(x_i, y_i)$  sono disposte secondo una retta crescente. Nel caso in cui  $r = 1$ , l'allineamento è perfetto. Viceversa, più il valore di  $r$  si avvicina a  $-1$ , più le osservazioni sono disposte secondo una retta decrescente. Nel caso in cui  $r = -1$ , l'allineamento è perfetto. Nel caso in cui  $r = 0$  si parla di **incorrelazione** (indipendenza lineare) tra  $X$  e  $Y$ . Tale situazione corrisponde ad una retta con coefficiente angolare nullo, quindi parallela all'asse delle ascisse.

Si consideri il caso in cui il ruolo di  $X$  e  $Y$  viene invertito, di modo che sia  $X$  la variabile dipendente e  $Y$  la variabile indipendente. In tal caso, la retta di regressione

$$x' = a^* + b^* \cdot y$$

ha coefficiente angolare

$$\hat{b}^* = \frac{\sigma_{xy}}{\sigma_y^2}$$

ed intercetta

$$\hat{a}^* = m_x - \hat{b}^* \cdot m_y.$$

Ovviamente,  $\hat{b}^*$  ha lo stesso segno (positivo o negativo) di  $\hat{b}$ . Inoltre, la loro media geometrica è pari al coefficiente di correlazione di Bravais-Pearson  $r$ ,

$$r = \sqrt{\hat{b} \cdot \hat{b}^*}.$$

Si consideri ora la misura di variabilità di  $y$  data dalla varianza  $\sigma_y^2$ . La quantità  $N \cdot \sigma_y^2$  è nota come **devianza** (totale)

$$\text{Devianza totale} = \sum_{i=1}^N (y_i - m_y)^2$$

Vale la seguente relazione, nota come **scomposizione della devianza**:

$$\sum_{i=1}^N (y_i - m_y)^2 = \sum_{i=1}^N (y'_i - m_y)^2 + \sum_{i=1}^N (y_i - y'_i)^2$$

$$\text{Devianza totale} = \text{Devianza di regressione} + \text{Devianza residua}.$$

L'espressione precedente evidenzia la scomposizione della devianza di  $Y$  in una parte che dipende dalla sua relazione con  $X$  (Devianza di regressione) e in una parte che la relazione con  $X$  non riesce a spiegare (Devianza residua). Nel caso in cui Devianza residua = 0, i punti interpolati coincidono con i punti osservati, vale a dire  $y_i = y'_i$ . In questo caso Devianza totale = Devianza di regressione. Nel caso in cui Devianza di regressione = 0, allora la retta di regressione non è adatta a spiegare la relazione tra  $X$  e  $Y$  e si realizza la massima divergenza tra i valori osservati  $y_i$  ed i valori teorici  $y'_i$ . In questo caso Devianza totale = Devianza residua.

L'introduzione della scomposizione della devianza è utile alla definizione di un indice che misura la bontà di adattamento di un modello di regressione lineare alle osservazioni  $(x_i, y_i)$ . Tale indice è il **coefficiente di determinazione**

$$R^2 = \frac{\text{Dev. regressione}}{\text{Dev. totale}},$$

il quale valuta la quota di variabilità di  $Y$  che viene spiegata dalla sua relazione con  $X$ . Il coefficiente di determinazione  $R^2$  si può anche scrivere come

$$R^2 = 1 - \frac{\text{Dev. residua}}{\text{Dev. totale}}$$

data la scomposizione della devianza totale vista in precedenza, e, solo nel caso della retta di regressione, anche come  $R^2 = r^2$ , vale a dire come il quadrato del coefficiente di correlazione di Bravais-Pearson. Il coefficiente di determinazione assume valori tra 0 e 1,  $0 \leq R^2 \leq 1$ . Più alti sono i valori di  $R^2$  migliore è l'adattamento del modello scelto.

**Esercizio A.**

a) Valutiamo le quantità di interesse per il calcolo delle componenti necessaria all'individuazione dell'intercetta  $a$  e del coefficiente angolare  $b$  della retta di regressione. Si consideri a questo proposito la seguente tabella

Regione	$x_i$	$y_i$	$x_i^2$	$x_i y_i$	$y_i^2$	$y'_i$	$(y'_i - m_y)^2$	$(y'_i - y_i)^2$
Piemonte	2,52	17,18	6,350	43,294	295,152	17,007	3,112	0,030
Lombardia	3,78	19,14	14,288	72,349	366,340	18,688	0,007	0,205
Trentino A. A.	6,96	22,55	48,442	156,948	508,503	22,928	17,280	0,143
Veneto	3,61	17,14	13,032	61,875	293,780	18,461	0,096	1,745
Liguria	2,25	15,45	5,063	34,763	238,703	16,647	4,512	1,433
Emilia-Romagna	4,38	19,72	19,184	86,374	388,878	19,488	0,513	0,054
Toscana	3,40	20,22	11,560	68,748	408,848	18,181	0,349	4,158
Totale	26,90	131,40	117,919	524,350	2500,203	131,400	25,870	7,768

Si ricava che

$$m_x = 3,843 \text{ e } m_y = 18,771,$$

$$\sigma_x^2 = \frac{\sum_{i=1}^N x_i^2}{N} - (m_x)^2 = 117,919/7 - (3,843)^2 = 2,077$$

e

$$\sigma_{xy} = \frac{\sum_{i=1}^N x_i \cdot y_i}{N} - m_x \cdot m_y = 524,35/7 - 3,843 \cdot 18,771 = 2,770.$$

Di qui si ricava che

$$b = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{2,770}{2,077} = 1,334$$

e

$$a = m_y - b \cdot m_x = 18,771 - 3,843 \cdot 1,334 = 13,647.$$

Quindi la retta di regressione ottenuta tramite i minimi quadrati è

$$y' = 13,647 + 1,334 \cdot x.$$

b) Sulla base della tabella precedente si ha che

$$\text{Devianza totale} = 7 \cdot \sigma_y^2 = 7 \cdot \left( \frac{\sum_{i=1}^N y_i^2}{N} - (m_y)^2 \right) = 2500,203 - 7 \cdot (18,771)^2 = 33,638,$$

$$\text{Devianza di regressione} = \sum_{i=1}^N (y'_i - m_y)^2 = 25,870,$$

$$\text{Devianza residua} = \sum_{i=1}^N (y_i - y'_i)^2 = 7,768,$$

e quindi la scomposizione della devianza totale è verificata, essendo  $33,638 = 25,870 + 7,768$ .

Il coefficiente di determinazione  $R^2$  si calcola come rapporto tra devianza di regressione e devianza totale e risulta pari a  $25,870/33,638 = 0,769$ . essendo un valore abbastanza alto, si deduce che la retta di regressione presenta un buon accostamento ai punti osservati.

c) La percentuale teorica di persone soddisfatte del proprio tempo libero per la regione Umbria è  $13,647 + 1,334 \cdot 3,01 = 17,661$ .

d) Nel caso in cui si volesse interpolare  $X$  rispetto a  $Y$ ,  $b$  avrebbe lo stesso segno del coefficiente angolare calcolato per la retta in cui  $Y$  è la variabile dipendente e  $X$  quella indipendente, dato che questo viene

determinato dalla covarianza  $\sigma_{xy}$ . Però il valore risulterebbe diverso. In particolare, per questa nuova interpolazione risulta che  $b = \sigma_{xy}/\sigma_y^2 = 2,770/4,805 = 0,576$ .

**Esercizio B.**

a) Consideriamo la trasformazione data dal logaritmo in base naturale, che conduce a

$$\ln(y) = \ln(\alpha_0) + \alpha_1 \cdot \ln(x).$$

Indichiamo con  $z_i = \ln(x_i)$  e con  $u_i = \ln(y_i)$ . A questo punto facciamo riferimento alla retta di regressione

$$u'_i = \alpha_0^* + \alpha_1 \cdot z_i,$$

dove  $\alpha_0^* = \ln(\alpha_0)$ . Si considerino le quantità riportate nella seguente tabella:

Reddito	$x_i$	$y_i$	$z_i = \ln(x_i)$	$u_i = \ln(y_i)$	$z_i^2$	$z_i u_i$	$u_i^2$	$u'_i$	$(u'_i - u_i)^2$
0-5	2,50	4241	0,916	8,353	0,840	7,653	69,765	9,052	0,490
5-6	5,50	2623	1,705	7,872	2,906	13,420	61,970	7,556	0,100
6-7,5	6,75	1682	1,910	7,428	3,646	14,184	55,171	7,168	0,067
7,5-11	9,25	1013	2,225	6,921	4,949	15,396	47,896	6,570	0,123
11-15	13,00	430	2,565	6,064	6,579	15,553	36,769	5,925	0,019
15-19	17,00	219	2,833	5,389	8,027	15,268	29,042	5,416	0,001
19-25	22,00	125	3,091	4,828	9,555	14,925	23,313	4,927	0,010
25-50	37,50	59	3,624	4,078	13,136	14,778	16,626	3,915	0,026
50-100	75,00	9	4,317	2,197	18,641	9,486	4,828	2,600	0,162
Totale	188,500	10401	23,186	53,129	68,278	120,664	345,380	53,129	0,998

A partire dalle quantità in tabella, si ricava che  $m_z = 2,576$ ,  $m_u = 5,903$ ,  $\sigma_z^2 = 68,278/9 - (23,186/9)^2 = 0,9495$ ,  $\sigma_u^2 = 345,380/9 - (53,129/9)^2 = 3,527$  e  $\sigma_{zu} = 120,664/9 - (23,186/9) * (53,129/9) = -1,801$ . Di conseguenza  $\alpha_1 = -1,897$  e  $\alpha_0^* = \ln(\alpha_0) = 10,790$ . Si ha quindi che la retta di regressione è

$$u' = \ln(y') = 10,790 - 1,897 \cdot \ln(x),$$

o, in modo equivalente,

$$y' = 48555,5 \cdot x^{-1,897}.$$

Il coefficiente di determinazione  $R^2$  per l'interpolazione di  $u$  rispetto a  $z$  si può ottenere come

$$R^2 = 1 - \frac{\text{Devianza residua}}{\text{Devianza totale}} = 1 - \frac{0,998}{9 \cdot 3,527} = 0,969.$$

Il valore molto alto indica un accostamento molto buono della retta alle osservazioni  $(z_i, u_i)$ .

b) Interpolando i punti originari  $(x_i, y_i)$ ,  $i = 1, \dots, 9$ , con una retta si ottiene un valore più elevato della devianza residua (11.064.157) a cui corrisponde un valore del coefficiente di determinazione notevolmente più basso (0,347).

**Esercizio C.**

a) Consideriamo la trasformazione  $z = x^2$ , in modo da lavorare con la retta di regressione

$$y' = \beta_0 + \beta_1 \cdot z.$$

Si ottiene

Mese	$x_i$	$y_i$	$z_i = x_i^2$	$z_i^2$	$z_i y_i$	$y_i^2$	$y_i'$	$(y_i' - y_i)^2$
gennaio	-2	107,2	4	16	428,8	11491,840	109,723	6,365
febbraio	-1	112,5	1	1	112,5	12656,250	109,659	8,074
marzo	0	115,2	0	0	0,0	13271,040	109,637	30,945
aprile	1	99,4	1	1	99,4	9880,360	109,659	105,238
maggio	2	114,1	4	16	456,4	13018,810	109,723	19,159
	0	548,4	10	34	1097	60318,300	548,400	169,782

da cui  $m_z = 2$ ,  $m_y = 109,68$ ,  $\sigma_z^2 = 2,8$ ,  $\sigma_y^2 = 33,9576$  e  $\sigma_{zy} = 0,04$  e quindi  $\beta_2 = 0,014$ ,  $\beta_0 = 109,652$ . Il valore del coefficiente di determinazione  $r^2 = 0,000353$  indica un adattamento dell'interpolazione praticamente nullo.

b) I valori teorici per giugno e luglio sono rispettivamente  $y_6' = 109,652 + 0,014 \cdot (3)^2 = 109,763$  e  $y_7' = 109,652 + 0,014 \cdot (4)^2 = 109,861$  e risultano entrambi lontani dai valori osservati, coerentemente con l'osservazione dell'indice di determinazione prossimo a 0.

### Esercizio D.

a) Si dispone in totale di  $N = 25$  osservazioni  $(x_i, y_i)$ . Si ricava quanto segue:

$$m_x = 6085,4 \text{ e } m_y = 5176,84;$$

$$\sigma_x^2 = 69815947, \sigma_y^2 = 53325804 \text{ e } \sigma_{xy} = 60002024.$$

Si ottiene allora l'intercetta della retta di regressione pari a  $a = -53,14444$  ed il coefficiente angolare pari a  $b = 0,8594$ . Quindi, la retta di regressione è

$$y' = -53,14444 + 0,8594 \cdot x.$$

b) Il coefficiente di correlazione lineare di Bravais-Pearson è pari a  $r = 0,9834$ , evidenziando una notevole relazione lineare (positiva) tra  $X$  e  $Y$ .

c) Calcoliamo  $R^2$  come quadrato di  $r$ ,  $R^2 = 0,9670$ . Il valore è molto alto, evidenziando un adattamento molto buono della retta alle osservazioni. Dal modello, quindi, emerge che al crescere della produzione di grano cresce anche il valore aggiunto dei prezzi della produzione agricola.

d) Sulla base della retta di regressione ottenuta, il valore teorico previsto per l'Italia (per la quale  $x = 25026$ ) è

$$y'_{Italia} = -53,14444 + 0,8594 \cdot 25016 = 21454,2.$$

Tale valore è prossimo al valore vero osservato pari a 25019, come ci si poteva attendere dato il valore molto alto di  $R^2$ .

e) Data la relazione

$$R^2 = \frac{\text{Devianza di regressione}}{\text{Devianza totale}},$$

si ricava che Devianza di regressione =  $R^2 \cdot \text{Devianza totale} = 0,9670 \cdot (25 \cdot 53325804) = 1289151301$ . Infine, dalla scomposizione della devianza totale, si ricava che Devianza residua =  $(25 \cdot 53325804) - 1289151312 = 43993788$ .

### Esercizio E.

a) Per interpolare gli Investimenti Fissi Lordi ( $Y$ ) rispetto al Prodotto Interno Lordo Totale ( $X$ ) è utile ottenere la seguente tabella:

Paesi	PIL	Investimenti			
	$x_i$	$y_i$	$x_i^2$	$x_i y_i$	$y_i^2$
Belgio	190,2	33,6	36.176,04	6.390,72	1.128,86
Danimarca	99,6	14,9	9.920,16	1.484,04	222,01
Germania	1.492,1	337,9	2.226.362,41	504.180,59	114.176,41
...	...	...	...	...	...
Austria	76,8	12,1	5.898,24	929,28	146,41
Svezia	142,5	20,6	20.306,25	2.935,50	424,36
	6.188,9	1.174,8	5.649.200	1.092.316	216.835

da cui, essendo  $N = 15$ , si ricava  $m_x = 412,59$ ,  $m_y = 78,32$ ,  $\sigma_x^2 = 206382,83$  e  $\sigma_{xy} = 40507,02$  con cui si ottengono il coefficiente angolare  $b = 0,1963$  e l'intercetta  $a = -2,6714$ ; si ottiene inoltre  $r^2 = 0,962$ , indicante un adattamento molto buono della retta ai valori osservati.

b) La classificazione richiesta fornisce la tabella

X	Y				
	0-50	50-150	150-250	250-350	
0-300	9	1	0	0	10
300-600	0	1	0	0	1
600-1200	0	1	2	0	3
1200-2000	0	0	0	1	1
	9	3	2	1	15

a cui corrispondono i valori centrali di classe  $x_1 = 150$ ,  $x_2 = 450$ ,  $x_3 = 900$ ,  $x_4 = 1.600$  e  $y_1 = 25$ ,  $y_2 = 100$ ,  $y_3 = 200$ ,  $y_4 = 300$ . Da questa si ricavano le tabelle

$x_i$	$f_i$	$x_i f_i$	$x_i^2 f_i$	$y_j$	$f_j$	$y_j f_j$	$y_j^2 f_j$
150	10	1.500	225.000	25	9	225	5.625
450	1	450	202.500	100	3	300	30.000
900	3	2.700	2.430.000	200	2	400	80.000
1.600	1	1.600	2.560.000	300	1	300	90.000
	15	6.250	5.417.500	15	1.225	205.625	

da cui si possono ricavare le quantità  $m_x = 416,67$ ,  $m_y = 81,67$ ,  $\sigma_x^2 = 2.813.291,67/15$ ,  $\sigma_y^2 = 105.575,17/15$ , e la tabella

$x_i y_j f_{ij}$	25	100	200	300	
150	33.750	15.000	0	0	
450	0	45.000	0	0	
900	0	90.000	360.000	0	
1.600	0	0	0	480.000	
					1.023.750

da cui si ricava  $\sigma_{xy} = 1.023.750/15 - 416,67 \cdot 81,67 = 513.308,42/15$ ,  $b = 0,1825$ ,  $a = 5,6277$  e  $r^2 = 0,887$ , indicante un buon adattamento della retta ai valori osservati.