Esercitazione 6 del corso di Statistica (parte 1)

Dott.ssa Paola Costantini

20 Novembre 2008

Stipendio percepito	Età	N. di anni di servizio	Qualifica funzionale	Regime di impiego	Genere	Stato Civile	Abitazione di Proprietà
1750	34	12	Impiegato	Tempo pieno	F	Coniugato	Sì
1950	26	6	Impiegato	Tempo pieno	M	Non coniugato	No
3400	34	8	Operaio	Collaboratori esterni	F	Vedovo	Sì
2500	41	10	Operaio	Tempo pieno	F	Non coniugato	Sì
1150	38	9	Impiegato	Collaboratori esterni	M	Coniugato	No
2400	29	11	Operaio	Tempo pieno	F	Vedovo	Sì
2900	36	15	Impiegato	Tempo pieno	M	Non coniugato	Sì
2000	32	10	Impiegato	Tempo pieno	F	Vedovo	Sì
2150	36	7	Impiegato	Part time	M	Non coniugato	Sì
3900	48	8	Impiegato	Tempo pieno	F	Vedovo	Sì
1550	29	13	Dirigente	Collaboratori esterni	M	Coniugato	No
2000	31	7	Operaio	Collaboratori esterni	F	Coniugato	Sì
1800	33	8	Operaio	Tempo pieno	M	Coniugato	No
1850	42	9	Impiegato	Part time	M	Vedovo	Sì
1350	26	12	Impiegato	Collaboratori esterni	F	Coniugato	No
2450	30	15	Operaio	Tempo pieno	M	Non coniugato	Sì
2550	41	13	Impiegato	Collaboratori esterni	F	Non coniugato	Sì
2000	28	7	Impiegato	Collaboratori esterni	F	Vedovo	No
2400	33	8	Operaio	Tempo pieno	M	Non coniugato	Sì
1500	37	12	Impiegato	Collaboratori esterni	F	Coniugato	No

Esercizio 1

Considerando i caratteri "Regime di Impiego" e "Numero anni di servizio", si può affermare che il Regime di Impiego dipende in media dal "Numero anni di servizio"?

Utilizziamo l'indice $\eta_{y|x}$ per misurare il grado di indipendenza in media del carattere qualitativo *y* dal carattere quantitativo *x.*

Ordiniamo i dati del carattere N. anni di servizio e suddividiamolo in 2 classi equiampie

N. anni di servizio	Regime di Impiego
6	Tempo pieno
7	Collaboratori esterni
7	Collaboratori esterni
7	Part time
8	Tempo pieno
8	Tempo pieno
8	Collaboratori esterni
8	Tempo pieno
9	Collaboratori esterni
9	Part time
10	Tempo pieno
10	Tempo pieno
11	Tempo pieno
12	Collaboratori esterni
12	Collaboratori esterni
12	Tempo pieno
13	Collaboratori esterni
13	Collaboratori esterni
15	Tempo pieno
15	Tempo pieno

Range = 15-6 = 9

2 classi equiampie = 9/2= 4,5

n. anni di servizio / Regime di impiego	[6,10.5]	]10.5,15]	**TOTALE**
Tempo pieno	6	4	**10**
Part-time	2	0	**2**
Collaboratore est	4	4	**8**
totale	**12**	**8**	**20**

Quando y è quantitativo:

$$\eta_{Y|X} = \frac{\sigma^2_{EXT_Y}}{\sigma^2_Y} = \frac{\sum_{i=1}^{k} \left(\mu_{Y|X=x_i} - \mu_Y\right)^2 n_{i\bullet}}{\sum_{j=1}^{h} \left(\hat{y}_j - \mu_Y\right)^2 n_{\bullet j}}$$

Considerando che:

valori centrali: $y_1 = 8,25$; $y_2 = 12,75$

$$\mu = \frac{\sum_{j=1}^{3} y_j n_{.j}}{N} = \frac{8,25*12+12,75*8}{20} = 10,05$$ n. anni di servizio medio

$$\mu_1 = FullTime = \frac{\sum_{j=1}^{2} y_j n_{1j}}{n_{1.}} = \frac{8,25*6+12,75*4}{10} = 10,05$$ n. anni si servizio Full Time

$$\mu_2 = PartTime = \frac{\sum_{j=1}^{2} y_j n_{1j}}{n_2} = \frac{8,25*2+12,75*0}{2} = 8,25$$ n. anni di servizio Part Time

$$\mu_3 = CollEsterno = \frac{\sum_{j=1}^{2} y_j n_{1j}}{n_3} = \frac{8,25*4+12,75*4}{8} = 10,5$$ n. anni di servizio Coll. Est.

Commento: si può vedere che le medie delle distribuzioni condizionate differiscono dalla media generale di Y, quindi i due caratteri **non sono indipendenti in media**.

Ma quanto è forte il legame di dipendenza in media?

Calcolo del numeratore dell'indice

$$\sum (\mu_{y|x} - \mu_y)^2 * n_{i.} = (10,05 - 10,05)^2 * 10 + (8,25 - 10,5)^2 * 2 + (10,5 - 10,05)^2 * 8 = \mathbf{8,1}$$

Calcolo del denominatore dell'indice

$$\sum (\acute{y}_j - \mu_y)^2 * n_{.j} = (8,25 - 10,05)^2 * 12 + (12,75 - 10,05)^2 * 8 = \mathbf{97,2}$$

Calcolo dell'indice

$$\eta_{Y|X} = \frac{\sigma^2_{EXT_Y}}{\sigma^2_Y} = \frac{\sum_{i=1}^{k} \left(\mu_{Y|X=x_i} - \mu_Y\right)^2 n_{i.}}{\sum_{j=1}^{h} \left(\hat{y}_j - \mu_Y\right)^2 n_{.j}} = \frac{8,1}{97,2} = 0,0833$$

Considerando che $\mathbf{0} \leq \eta_{y|x} \leq \mathbf{1}$ abbiamo un bassissimo grado di dipendenza in media, piuttosto siamo vicini ad un'ipotesi di *indipendenza in media*.

Esercizio 2

Vogliamo scoprire se esiste una qualche relazione (e se esiste di che natura è) tra il prezzo di vendita (espresso in migliaia di dollari) e la superficie abitabile di 10 abitazioni, espressa in piedi quadrati (0,0929 metri quadrati).
Poichè sospettiamo che i prezzi di vendita possano dipendere dalla superficie abitabile, il prezzo di vendita diventa la variabile dipendente e la superficie abitabile la variabile esplicativa.

SqFt (x)	Prezzo (y)	x_i^2	y_i^2	$x \cdot y$
521	26	271441	676	13546
661	31	436921	961	20491
694	37,4	481636	1398,76	25955,6
743	34,8	552049	1211,04	25856,4
787	39,2	619369	1536,64	30850,4
825	38	680625	1444	31350
883	39,6	779689	1568,16	34966,8
920	31,2	846400	973,44	28704
965	37,2	931225	1383,84	35898
1011	38,4	1022121	1474,56	38822,4
8010	**352,8**	**6621476**	**12627,44**	**286440,6**

$$\mu_x = \frac{352,8}{10} = 35,52 \qquad \mu_y = \frac{8010}{10} = 801 \qquad \sum_{1=1}^{n} x_i^{\,2} = 6621476 \qquad \sum_{1=1}^{n} y_i^{\,2} = 12627,44$$

$$\sigma^2_x = VAR(X) = E[X^2] - \mu^2 = \frac{6621476}{10} - 801^2 = 20546,6 \rightarrow \sigma = \sqrt{20546,6} = 143,34$$

$$\sigma^2_y = VAR(YX) = E[Y^2] - \mu^2 = \frac{12627,44}{10} - 35,28^2 = 18,06 \rightarrow \sigma = \sqrt{18,06} = 4,25$$

$$\mu(x \cdot y) = 286440,6/10 = 28644,06$$

$$Cov_{x,y} = \mu(x \cdot y) - (\mu_x \cdot \mu_y) = 28644,06 - (801 \cdot 35,28) = 384,78$$

$$Corr_{x,y} = \rho_{x,y} = \frac{Cov_{x,y}}{\sigma_x \cdot \sigma_y} = \frac{384,78}{143,34 \cdot 4,25} = 0,63 \quad \underline{\text{Correlazione positiva}}$$

$$\rho^{2}_{x,y} = R^2 = 0{,}63^2 = 0{,}3989$$

Per rispondere alla domanda: "quanto è affidabile la relazione?" di solito si analizza il valore di R^2 che misura la proporzione della variazione della variabile dipendente y, che viene spiegata utilizzando la variabile esplicativa x. In questo caso l'R^2 vale 0,3989 e indica che approssimativamente il 40% della variazione dei prezzi di vendita può essere spiegato da un modello lineare utilizzando la superficie abitabile.

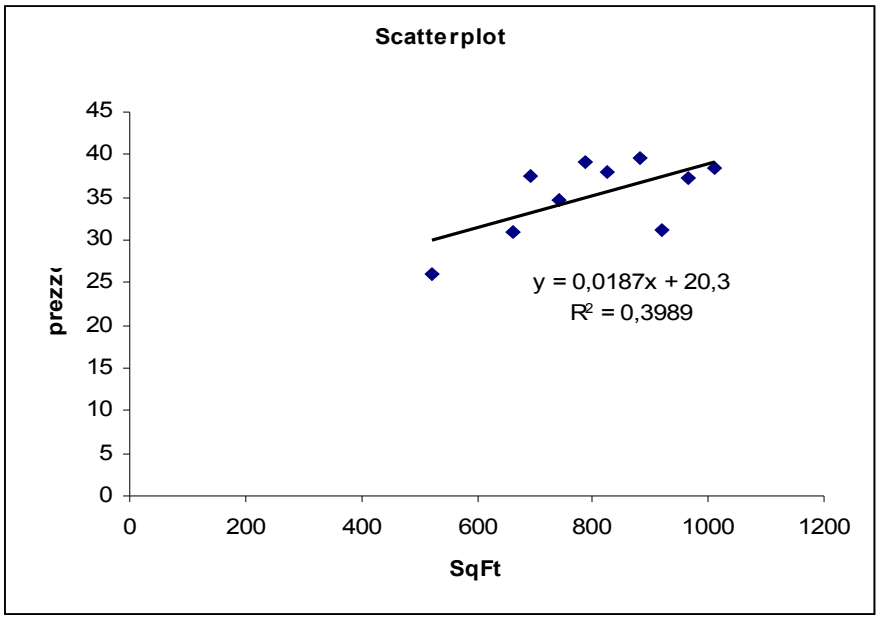
STIMA DELLA RETTA DI REGRESSIONE

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

$$\hat{\beta} = \frac{Cov_{x,y}}{\sigma^{2}_{x}} = \frac{384{,}78}{20546{,}6} = 0{,}0187$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 35{,}28 - (0{,}0187 \cdot 801) = 20{,}3$$

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_i = 20 + 0{,}0187 \cdot x_i$$



L'intercetta di y o termine costante è 20,3, espresso nella stessa unità della variabile y. L'inclinazione (o coefficiente di regressione) è 0,0187 e indica la variazione media della variabile y corrispondente alla variazione di una unità della variabile x. L'unità di misura per questo esempio è 0,0187 migliaia di dollari per piede quadrato quadrato,(1 piede q. = 0,0929 metri quadrati) o

circa 19 euro per metro quadrato. Se due proprietà differiscono in superficie di 100 piedi quadrati ci aspettiamo che i prezzi previsti di vendita siano di 0,019 *100 = 1,9 migliaia di dollari, ovvero 1.900 dollari.