CORSO DI STATISTICA (parte 1) - ESERCITAZIONE 6

Dott.ssa Antonella Costanzo

a.costanzo@unicas.it

Esercizio 1. Associazione, correlazione e dipendenza tra caratteri

In un collettivo di 11 famiglie è stata rilevata la distribuzione congiunta dei redditi mensili pro-capite (X) e spesa mensile per tempo libero (Y) espressa in migliaia di euro.

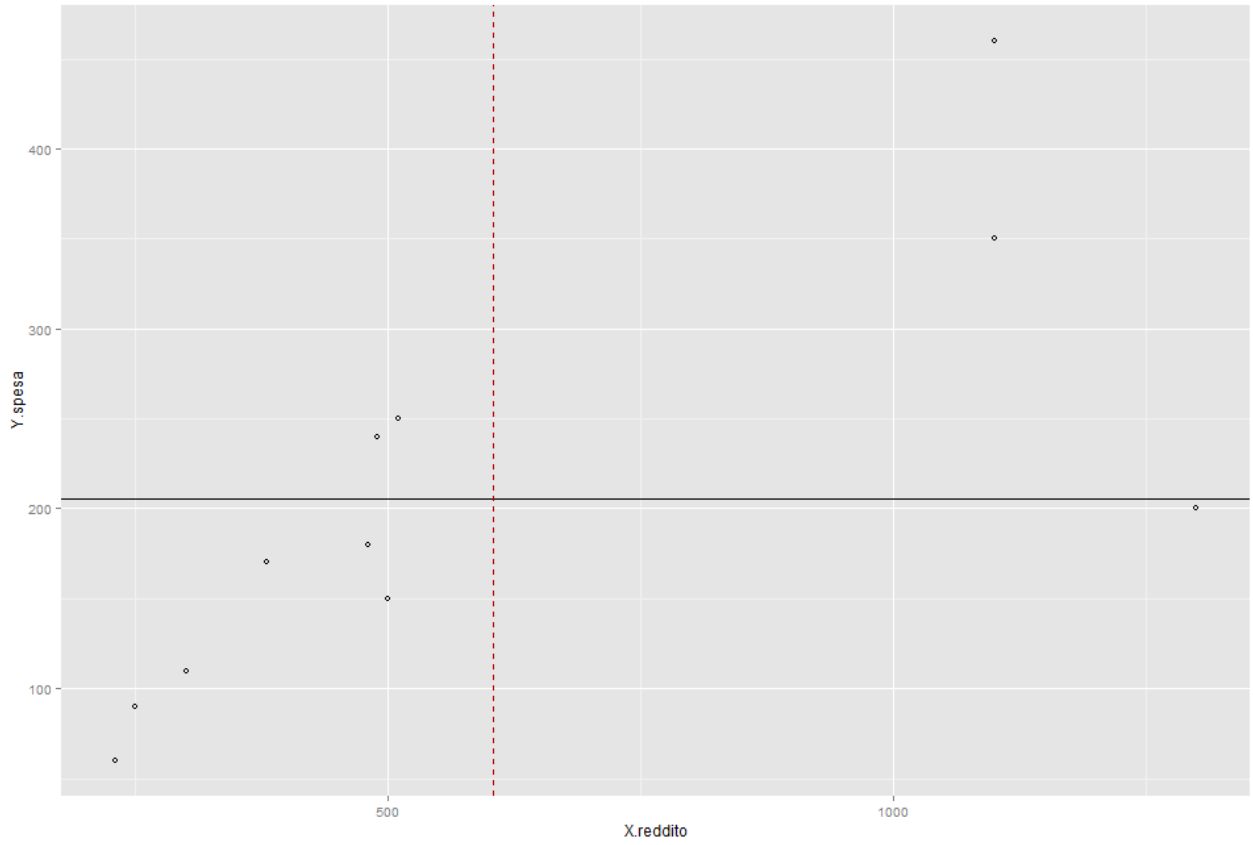
X=reddito	Y=spesa
480	180
500	150
380	170
1100	350
1100	460
230	60
490	240
250	90
300	110
510	250
1300	200

Verificare se esiste una relazione tra X e Y e di che tipo.

Svolgimento:

Nello studio della relazione tra due variabili quantitative, risulta di grande utilità la rappresentazione dei valori di X e di Y in una diagramma a dispersione (o scatterplot). Questo grafico mette in evidenza, con una buona approssimazione, il tipo di legame tra Y e X.

Grafico 1. Scatterplot di Y e X



Si può notare un legame positivo (concordanza) tra il reddito mensile e la spesa mensile per il tempo libero; E' presente una maggiore dispersione della spesa per il tempo libero in corrispondenza dei valori medio alti del reddito. Una possibile misura del grado di associazione tra i due caratteri X e Y nel nostro caso è costituita dalla **covarianza**:

$$cov(X,Y) = \sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$\sigma_{xy} > 0$ se a valori crescenti di X corrispondono valori crescenti di Y (concordanza). Inoltre, più forte è la relazione tra le due variabili, più ci aspettiamo che la covarianza diventi grande in valore assoluto

$\sigma_{xy} < 0$ se a valori crescenti di X corrispondono valori decrescenti di Y (discordanza)

$\sigma_{xy} = 0$ se gli scostamenti positivi e negativi di X e Y dalle rispettive medie si compensano. In tal caso non esiste associazione tra i caratteri.

Nota: $\sigma_{xy} = 0$ è condizione sufficiente ma non necessaria per l'indipendenza tra X e Y. Ciò vuol dire che $\sigma_{xy} = 0$ non implica necessariamente indipendenza tra i caratteri. D'altro canto se X e Y sono indipendenti allora certamente $\sigma_{xy} = 0$.

Dati per i calcoli:

X=reddito	Y=spesa	$(x - \mu_x)$	$(x - \mu_x)^2$	$(y - \mu_y)$	$(y - \mu_y)^2$	$(x - \mu_x)(y - \mu_y)$	xy
480	180	-123.64	15285.95	-25.45	647.93	3174.11	86400
500	150	-103.64	10740.50	-55.45	3075.21	5747.11	75000
380	170	-223.64	50013.22	-35.45	1257.02	7928.93	64600
1100	350	496.36	246376.86	144.55	20893.39	71747.11	385000
1100	460	496.36	246376.86	254.55	64793.39	126347.11	506000
230	60	-373.64	139604.13	-145.45	21157.02	54347.11	13800
490	240	-113.64	12913.22	34.33	1193.39	-3925.62	117600
250	90	-353.64	125058.68	-115.45	13329.75	40828.93	22500
300	110	303.64	92195.04	-95.45	9111.57	28983.47	33000
510	250	-93.64	8767.77	44.55	1984.30	-4171.07	127500
1300	200	696.36	484922.31	-5.45	29.75	-3798.35	260000
Totale			**1432254.55**		**137472.73**	**327181.82**	**1691400**

$\mu_x = 603.64$

$\mu_y = 205.45$

$\sigma_{xy} = \frac{1}{11}(327181.82) = 29743.81$ positiva → concordanza

Per la covarianza è disponibile una formula alternativa (equivalente e più operativa):

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n} xy - \mu_x\mu_y = \mu_{xy} - \mu_x\mu_y$$

$\sigma_{xy} = \frac{1}{11}(1691400) - 603.64 * 205.45 = 29743.8$

Nota: questa formulazione alternativa semplifica il calcolo e soprattutto l'interpretazione della covarianza. Se c'è indipendenza la covarianza è zero dato che, nel caso di indipendenza, il valore atteso (o media) del prodotto dei valori di X e Y è pari al prodotto delle rispettive medie, ossia $\mu_{xy} = \mu_x\mu_y$.

La covarianza è un indice assoluto, dipendente dall'unità di misura, non ha limiti predefiniti. Tuttavia è legata alla variabilità dei due caratteri nel senso che non può superare, in valore assoluto, il prodotto degli sqm di X e Y.

Ricordando la **disuguaglianza di Cauchy-Schwarz**:

$$[Cov(X,Y)]^2 \leq \sigma^2(Y)\sigma^2(X)$$

la covarianza al quadrato è inferiore o uguale al prodotto delle varianze delle distribuzioni marginali di X e Y. Avremo quindi che: $|\sigma_{xy}| \leq \sigma_x \sigma_y$

per cui la covarianza è compresa tra i due limiti:

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

Nota: σ_{xy} può "raggiungere" uno dei due limiti se e solo se gli scarti di X e Y dai corrispondenti valori medi sono tra loro proporzionali (per cui se tra X e Y esiste un legame lineare). Sfruttando questa informazione possiamo costruire, a partire dalla covarianza, un indice standardizzato e normalizzato: il coefficiente di correlazione lineare.

Il **coefficiente di correlazione di Bravais-Pearson** misura l'intensità del legame lineare che sussiste tra due variabili X e Y:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

ρ_{xy} è un indice che varia tra –1 e 1 e non dipende dall'unità di misura dei due caratteri. Il segno del coefficiente di correlazione lineare corrisponde al segno della covarianza.

• Se tra due caratteri non vi è correlazione lineare, si ha che $\rho_{xy} = 0$

• Se tra i due caratteri sussiste un legame lineare perfetto (Y= a$\pm$ bX) allora $\rho_{xy} = \pm 1$

Per calcolare ρ_{xy} abbiamo bisogno dei valori σ_x e σ_y che risultano rispettivamente pari a:

$\sigma_x = 360.84$

$\sigma_y = 111.79$

$\rho_{xy} = \frac{29743.81}{360.84(111.79)} = 0.737$ esiste una correlazione positiva tra reddito mensile e spesa per il tempo libero.

L'esistenza di una correlazione, per quanto intensa, non implica una relazione di causa-effetto. In effetti ρ_{xy} misura solo la co-variazione tra i valori di X e Y standardizzati.

Esercizio 2. Stima del modello di regressione per serie grezze

Siamo interessati a valutare se la spesa per il tempo libero (Y) dipende dal reddito mensile pro-capite (X). Vogliamo studiare il legame di **dipendenza lineare (causa-effetto)** tra Y e X dove X è la variabile indipendente e Y la variabile dipendente.

Stima di un **modello di regressione lineare** del tipo:

$$\boldsymbol{Y = \beta_0 + \beta_1 X}$$

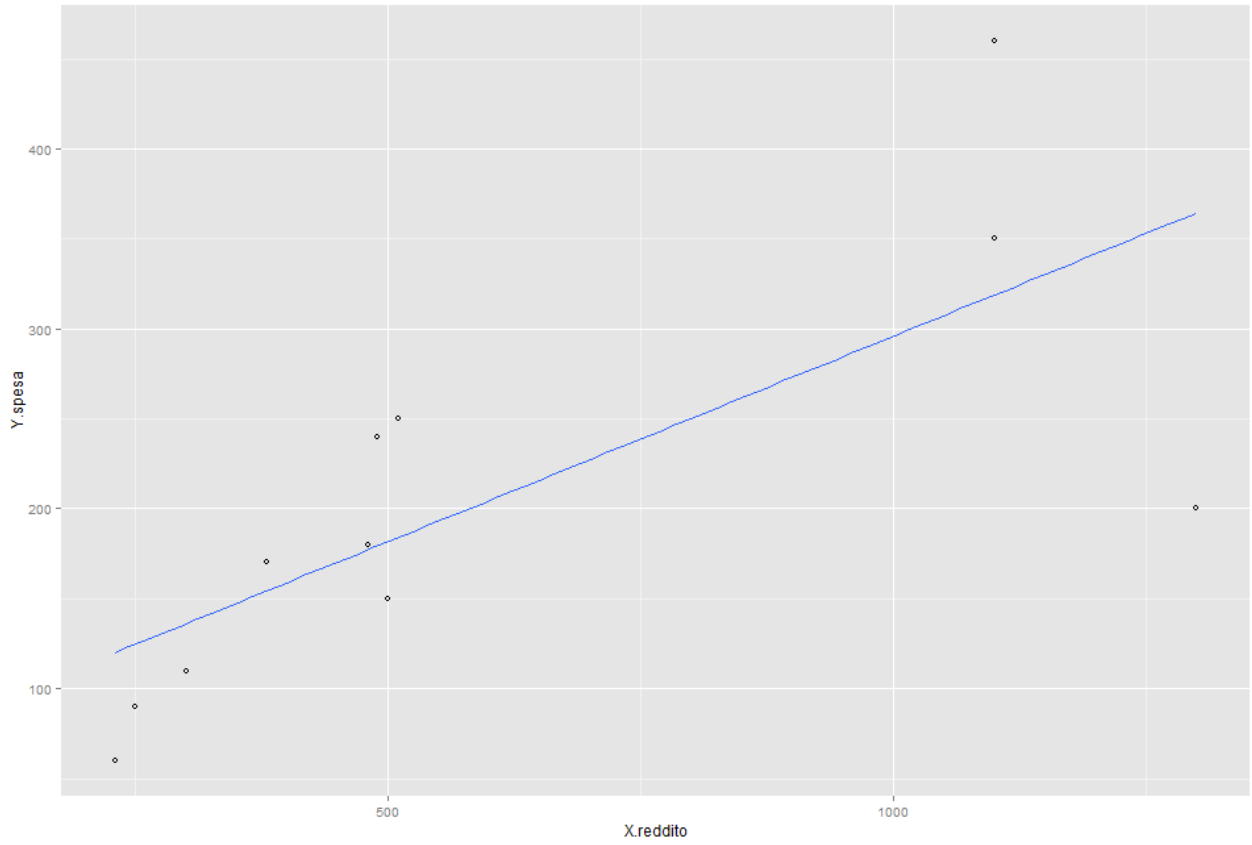
Stima della retta di regressione (criterio dei minimi quadrati):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{29743.81}{130205.50} = 0.23$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 205.45 - 0.23(603.64) = 66.61$$

Grafico 2. La retta di regressione stimata



β_1 è il coefficiente angolare che quindi esprime la pendenza della retta di regressione; la sua stima $\widehat{\beta_1}$ esprime l'impatto in media su Y di un incremento unitario della X. β_0 è l'intercetta e rappresenta il valore atteso di Y (e quindi cosa accade in media a Y) quando X=0. Non sempre risulta "interessante" da interpretare.

Bontà di adattamento del modello: il coefficiente di determinazione R^2

La devianza totale di un modello di regressione può essere scomposta in due termini:

1. **devianza della regressione** attribuibile cioè alla relazione che sussiste fra X ed Y, calcolata come differenza dalla retta di regressione dal valore medio.
2. **devianza dell'errore** (devianza residua) che non è imputabile alla relazione fra X ed Y ma ad altri fattori. Calcolata come differenza tra il valore osservato di Y e quello stimato.

In formula:

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

Il coefficiente di determinazione R^2 :

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

$0 \leq R^2 \leq 1$

R^2 misura quanta parte della devianza complessiva del modello viene "spiegata" dal modello di regressione stimato. Tanto maggiore è il valore di r-quadro tanto più il modello "spiega" bene (si adatta bene) ai dati. Al massimo, se $R^2 = 1$ il modello di regressione stimato si adatta perfettamente ai dati.

Nella regressione lineare il coefficiente di determinazione può essere ottenuto a partire dal coefficiente di correlazione. In particolare, vale la seguente:

$$R^2 = \rho_{xy}^2$$

Nel nostro caso avremo:

$$R^2 = (0.737)^2 = 0.543$$

Alternativamente, si può calcolare il coefficiente di determinazione a partire dalla devianza dell'errore o dalla devianza di regressione. Partendo dalla devianza dell'errore, si ha:

X	**Y**	$\widehat{Y}$	$y_i - \hat{y}$	$(y_i - \hat{y})$^2
480	180	177.01	2.99	8.9401
500	150	181.61	-31.61	999.1921
380	170	154.01	15.99	255.6801
1100	350	319.61	30.39	923.5521
1100	460	319.61	140.39	19709.3521
230	60	119.51	-59.51	3541.4401
490	240	179.31	60.69	3683.2761
250	90	124.11	-34.11	1163.4921
300	110	135.61	-25.61	655.8721
510	250	183.91	66.09	4367.8881
1300	200	365.61	-165.61	27426.6721
Totale				**62735.36**

Devianza della regressione=Devianza Totale-Devianza dell'errore

Devianza Totale= 137472.73

Devianza dell'errore=62735.36

Devianza della regressione (devianza spiegata)= 137472.73-62735.36=74737.37

$R^2 = \frac{74737.37}{137472.73} = 0.543$ Oppure $R^2 = 1 - \frac{Devianza\ errore}{Devianza\ Totale} = 0.543$

Esercizio 3. Regressione su tabella a doppia entrata

In 115 supermercati sono stati rilevati il prezzo di vendita (X) e il numero delle confezioni vendute (Y) di un certo tipo di prodotto e si ottiene la seguente distribuzione doppia:

X|Y	**[50,100]**	**[102,170]**	**[172,190]**	**Totale**
[0.60, 0.70)	5	13	21	39
[0.70, 0.80)	8	40	2	50
[0.80, 0.90]	20	6	0	26
Totale	33	59	23	115

Valutare se esiste dipendenza lineare tra prezzo di vendita e numero di confezioni vendute.

Svolgimento:

Riportiamo la tabella dati considerando i valori centrali delle classi:

X|Y	**75**	**136**	**181**	**Totale**
0.65	5	13	21	39
0.75	8	40	2	50
0.85	20	6	0	26
Totale	33	59	23	115

vogliamo stimare il modello $Y = \beta_0 + \beta_1 X$ dove Y= numero di confezioni vendute e X=prezzo di vendita.

Le medie condizionate di X e Y risultano:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{h=3} x_i n_{i.} = \frac{1}{115}(0.65 * 39) + (0.75 * 50) + (0.85 * 26) = 0.738$$

$$\mu_y = \frac{1}{n}\sum_{j=1}^{k=3} y_i n_{.j} = \frac{1}{115}(75 * 33) + (136 * 59) + (181 * 23) = 127.495 \cong 128$$

Per calcolare le varianze facciamo riferimento agli scarti dalla media al quadrato:

X|Y	$(\mathbf{75-128})^2$	$(\mathbf{136-128})^2$	$(\mathbf{181-128})^2$	**Totale**
$(\mathbf{0.65-0.738})^2$	5	13	21	39
$(\mathbf{0.75-0.738})^2$	8	40	2	50
$(\mathbf{0.85-0.738})^2$	20	6	0	26
Totale	33	59	23	115

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{h=3}(x_i - \mu_x)^2 n_{i.} = \frac{1}{115}[((0.65-0.738)^2 * 39] + [\,(0.75-0.738)^2 * 5$$

$$+ [\,(0.85-0.738)^2 * 26] = \frac{0.302 + 0.00072 + 0.326}{115} = \frac{0.629}{115} = 0.0055$$

$$\sigma_x = 0.074$$

$$\sigma_y^2 = \frac{1}{n}\sum_{j=1}^{k=3}(y_i - \mu_y)^2 n_{.j} = \frac{1}{115}[((75-128)^2 * 33] + [\,(136-128)^2 * 59] + [\,(181-128)^2 * 23] =$$

$$= \frac{92697 + 3776 + 64607}{115} = \frac{161080}{115} = 1400.69$$

$$\sigma_y = 37.43$$

Per calcolare la covarianza:

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{h}\sum_{j=1}^{k}(x_i - \mu_x)(y_i - \mu_y) * n_{ij} =$$

$$\sigma_{xy} = \frac{1}{115}[(0.65-0.738)(75-128)] * 5 + [(0.65-0.738)(136-128)] * 13$$

$$+ [(0.65-0.738)(181-128)] * 21] + [(0.75-0.738)(75-128)] * 8 + \cdots = -1.714$$

Quindi calcoliamo i coefficienti della retta di regressione:

$$\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{-1.714}{0.0055} = -311.61$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 128 + 311.61(0.738) = 357.97$$

Il valore stimato $\hat{y}_i$ corrispondente ad un dato valore x_i assegnato è pari a:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Per ogni valore di X considerato avremo quindi:

$$\hat{y}_1 = 357.97 - 311.61(0.65) = 155.42$$

$$\hat{y}_2 = 357.97 - 311.61(0.75) = 124.26$$

$$\hat{y}_3 = 357.97 - 311.61(0.85) = 93.10$$

Valutiamo infine la bontà di adattamento del modello ai dati attraverso il coefficiente di determinazione R^2. Siccome abbiamo già calcolato la devianza totale, è sufficiente calcolare la devianza residua. Per agevolare i conti, dal momento che si tratta di dati organizzati in tabella a doppia entrata si procede nel modo seguente:

$y_i \mid \hat{y}_j$	$\hat{y}_1 = 155.42$	$\hat{y}_2 = 124.26$	$\hat{y}_3 = 93.10$	**Totale**
$y_1 = 75$	5	13	21	39
$y_2 = 136$	8	40	2	50
$y_3 = 181$	20	6	0	26
Totale	33	59	23	115

La devianza residua è pari a:

$$\sum_{j=1}^{k}\sum_{i=1}^{h}(y_i - \hat{y}_j)^2 * n_{ij} = (75 - 155.42)^2 * 5 + (75 - 124.26)^2 * 13 + (75 - 93.10)^2 * 21$$
$$+ (136 - 155.42)^2 * 8 + (136 - 124.26)^2 * 40 + \cdots += 115560.94$$

La devianza totale invece è pari a:161080

$$R^2 = 1 - \frac{Devianza\ errore}{Devianza\ Totale} = 1 - \frac{\sum_{j=1}^{k}\sum_{i=1}^{h}(y_i - \hat{y}_j)^2 * n_{ij}}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{115560.94}{161080} = 1 - 0.72 = 0.28$$