

## ***Dipendenza in media***

- La *misurazione della connessione* tra due caratteri si basa solo sulle frequenze congiunte senza tenere conto delle modalità dei due caratteri.
- Quando uno dei due caratteri (tipicamente  $Y$ ) è *quantitativo*, è possibile confrontare le distribuzioni condizionate di  $Y$  tramite le medie condizionate.
- Per l’ $i$ -esima modalità di  $X$  ( $x_i$ ), la **media condizionata di  $Y$**  è data da

$$\bar{y}_i = \frac{1}{n_{i\circ}} \sum_{j=1}^c y_j n_{ij} = \sum_{j=1}^c y_j f_{j|i}$$

- Nel caso in cui il carattere  $Y$  è in classi, occorre utilizzare i valori centrali delle classi al posto delle modalità.
- Il carattere  $Y$  si dice ***indipendente in media*** da  $X$  quando  $X$  non influenza la media di  $Y$ . In termini matematici:

$$\bar{y}_1 = \bar{y}_2 = \dots = \bar{y}_r = \bar{y},$$

dove  $\bar{y}$  è la *media marginale* di  $Y$  data da

$$\bar{y} = \frac{1}{N} \sum_{j=1}^c y_j n_{\circ j}.$$

- Altrimenti,  $Y$  è ***dipendente in media*** da  $X$ .

## *Esempio*

- La media della distribuzione condizionata del *reddito* quando il *titolo di studio* è la *licenza media* viene calcolata come

Reddito	Frequenze ( $n_{ij}$ )	Valori centrali ( $y_j$ )	$y_j n_{ij}$
0-10	88	5	440
10-30	143	20	2.860
30-100	120	65	7.800
<b>Totale</b>	351	-	11.100

$$\bar{y}_1 = \frac{11.100}{351} = 31,62$$

- Si ha anche

$$\bar{y}_2 = 38,53 \quad (\text{titolo di studio} = \text{diploma})$$

$$\bar{y}_3 = 48,83 \quad (\text{titolo di studio} = \text{laurea})$$

e quindi il reddito *dipende in media* dal titolo di studio.

- La media della distribuzione marginale del *reddito* viene calcolata come

Reddito	Frequenze ( $n_{\circ j}$ )	Valori centrali ( $y_j$ )	$y_j n_{\circ j}$
0-10	100	5	500
10-30	200	20	4.000
30-100	200	65	13.000
<b>Totale</b>	500	-	17.500

$$\bar{y} = \frac{17.500}{500} = 35$$

## *Spezzata di regressione*

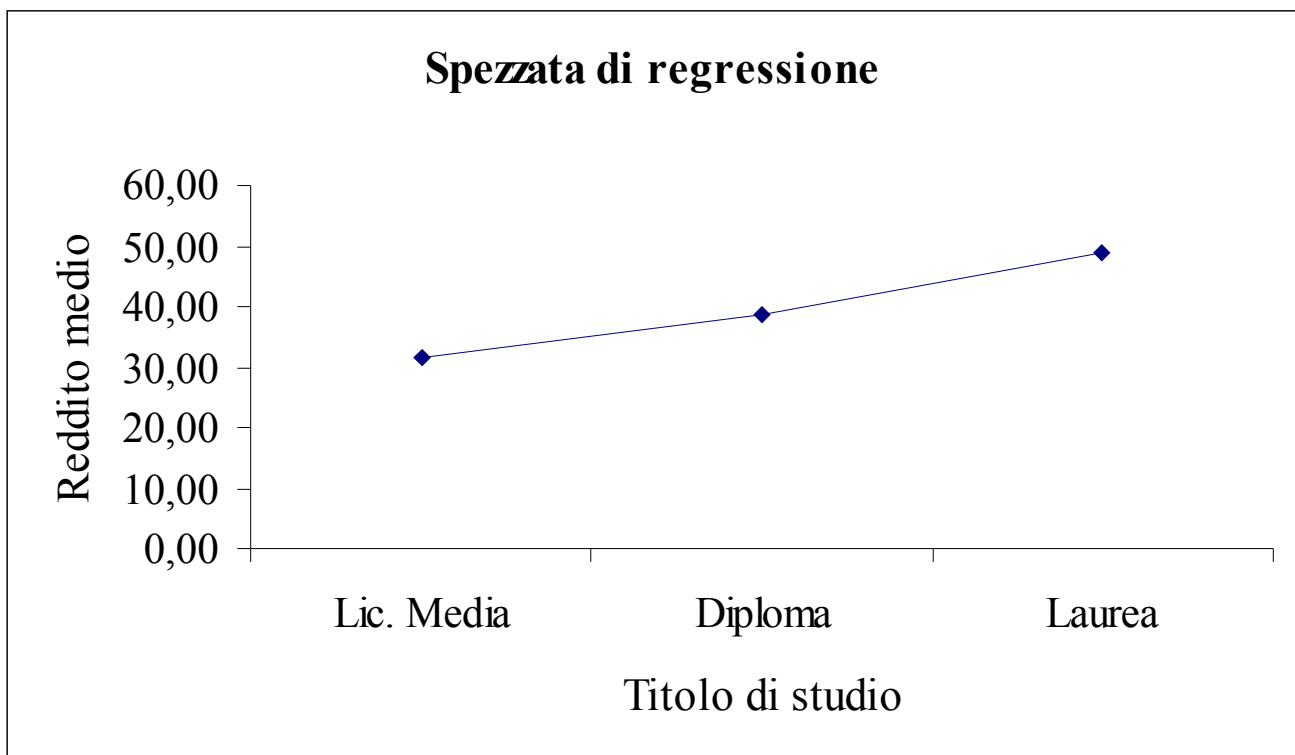
- La *spezzata di regressione* è un grafico che consiste nel rappresentare, e congiungere con dei segmenti, i punti di coordinate

$$(x_i, \bar{y}_i), \quad i = 1, \dots, r.$$

- La spezzata di regressione permette di intuire come varia la media di  $Y$  al variare della modalità di  $X$ .

## *Esempio*

- Per la distribuzione doppia dei caratteri *titolo di studio* e *reddito*



## ***Misura della dipendenza in media***

- Per verificare se c’è o meno *dipendenza in media* di  $Y$  da  $X$  si può utilizzare la ***devianza spiegata*** che è definita come

$$D_S = \sum_{i=1}^r (\bar{y}_i - \bar{y})^2 n_{i\circ}$$

- La devianza spiegata è sempre *non negativa* ( $D_S \geq 0$ ); in particolare:

➤  $D_S = 0 \Rightarrow$  *indipendenza in media* di  $Y$  da  $X$ .

➤  $D_S > 0 \Rightarrow$  *dipendenza in media* di  $Y$  da  $X$ .

- Il massimo della devianza spiegata è dato dalla ***devianza totale*** di  $Y$ ,

$$D_Y = \sum_{j=1}^c (y_j - \bar{y})^2 n_{\circ j}$$

- La differenza tra la devianza totale e quella spiegata è chiamata ***devianza residua*** che è sempre non negativa

$$D_R = D_Y - D_S = \sum_{i=1}^r \sum_{j=1}^c (y_{ij} - \bar{y}_i)^2 n_{ij}$$

- E’ quindi possibile definire un *indice relativo* per misurare la dipendenza in media, chiamato ***rapporto di correlazione***, come

$$\eta^2 = \frac{D_S}{D_Y} = 1 - \frac{D_R}{D_Y}$$

- Per l’interpretazione del valore assunto da  $\eta^2$  si consideri che:
  - $\eta^2 = 0 \Rightarrow$  *indipendenza in media* ( $D_S = 0, D_R = D_Y$ ).
  - $\eta^2 > 0 \Rightarrow$  *dipendenza in media* ( $D_S > 0$ ).
  - $\eta^2 = 1 \Rightarrow$  *massima dipendenza in media* ( $D_S = D_Y, D_R = 0$ ).
- La *dipendenza in media* ***implica*** la *dipendenza statistica* e quindi se

$$\eta^2 > 0 \quad \text{allora} \quad \chi^2 > 0.$$

- L’*indipendenza in media* ***non implica*** l’*indipendenza statistica* e quindi può accadere che

$$\eta^2 = 0 \quad \text{e} \quad \chi^2 > 0.$$

## *Esempio*

- Per la distribuzione doppia dei caratteri *titolo di studio* e *reddito*, la *devianza spiegata* viene calcolata nel modo seguente

<i>i</i>	$\bar{y}_i$	$n_{i\circ}$	$(\bar{y}_i - \bar{y})^2 n_{i\circ}$
1	31,62	351	4009,96
2	38,53	85	1059,18
3	48,83	64	12241,21
<b>Totale</b>	–	500	17310,35

$$D_S = 17310,65$$

che è maggiore di 0 in quanto c’è dipendenza in media.

- La devianza totale del *reddito* è  $D_Y = 315.000$  in quanto

Reddito	Frequenze ( $n_{\circ j}$ )	Valori centrali ( $y_j$ )	$(y_j - \bar{y})^2 n_{\circ j}$
0–10	100	5	90.000
10–30	200	20	45.000
30–100	200	65	180.000
<b>Totale</b>	500	–	315.000

- Il *rapporto di correlazione* è quindi pari a

$$\eta^2 = \frac{17.297,36}{315.000} = 0,0549$$

che indica una modesta dipendenza in media del reddito dal titolo di studio.

## ***Regressione***

- Concetto utilizzato quando *entrambi i caratteri* ( $X$  e  $Y$ ) *sono quantitativi*.
- In questo caso si preferisce utilizzare direttamente la distribuzione doppia unitaria

$X$	$Y$
$x_1$	$y_1$
$x_2$	$y_2$
$\vdots$	$\vdots$
$x_N$	$y_N$

- Solitamente si intende che la variabile  $Y$  dipenda da  $X$ . Quindi  $X$  è chiamata ***variabile indipendente*** (o ***esplicativa***) mentre  $Y$  è chiamata ***variabile dipendente*** (o ***risposta***).
- L’obiettivo è capire come  $X$  influenza  $Y$  e approssimare tale relazione tramite una semplice funzione matematica

$$y = f(x).$$

- Un’analisi preliminare può essere effettuata tramite un ***grafico a dispersione*** che consiste nel rappresentare punti di coordinate

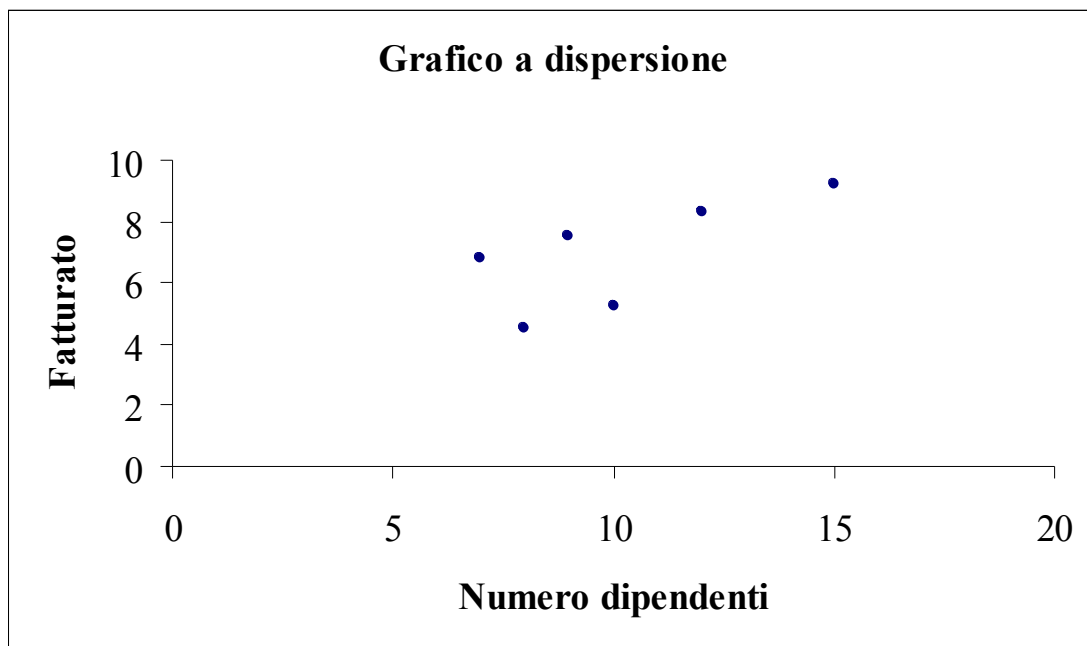
$$(x_i, y_i), \quad i = 1, \dots, N.$$

## *Esempio*

- Per un collettivo di  $N = 6$  imprese sono state rilevate le modalità dei caratteri *numero di dipendenti* ( $X$ ) e *fatturato* ( $Y$ ).

$X$	$Y$
15	9,2
10	5,2
8	4,5
7	6,8
12	8,3
9	7,5

- Sulla base di questi dati si vuole stabilire come il numero di dipendenti *influenza* il fatturato di un’impresa. Dal *grafico a dispersione* si nota che al crescere di  $X$  cresce  $Y$ .

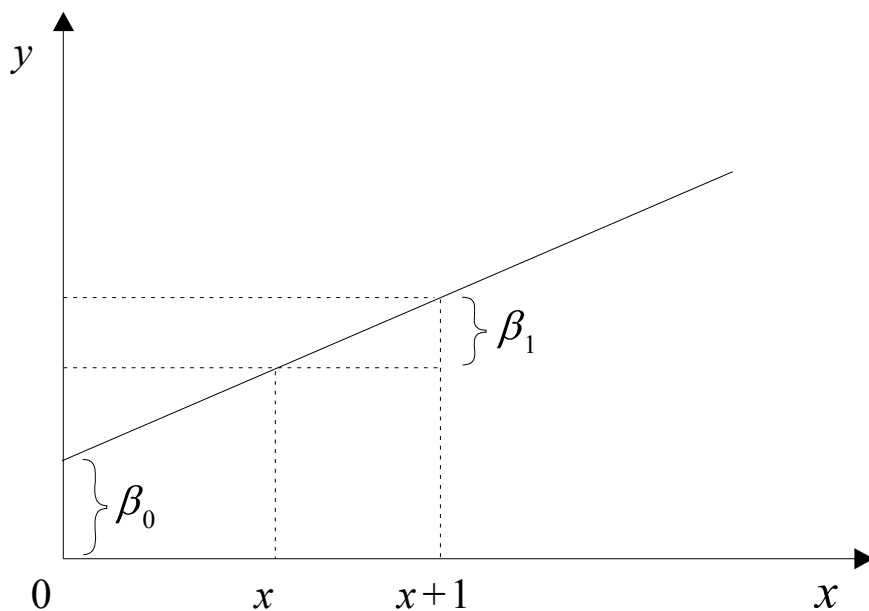


## *Regressione lineare semplice*

- In questo ambito, per approssimare la relazione tra  $Y$  e  $X$  si utilizza una *funzione lineare*

$$y = \beta_0 + \beta_1 x,$$

dove  $\beta_0$  (*intercetta*) e  $\beta_1$  (*coefficiente angolare*) sono chiamati *parametri della retta interpolatrice*.



- I parametri della retta ( $\beta_0$  e  $\beta_1$ ) vanno scelti in modo da approssimare al meglio la relazione tra  $Y$  e  $X$ , ossia *minimizzando l'errore di previsione*. Il metodo normalmente utilizzato a tale fine è conosciuto con il nome di *metodi dei minimi quadrati*.

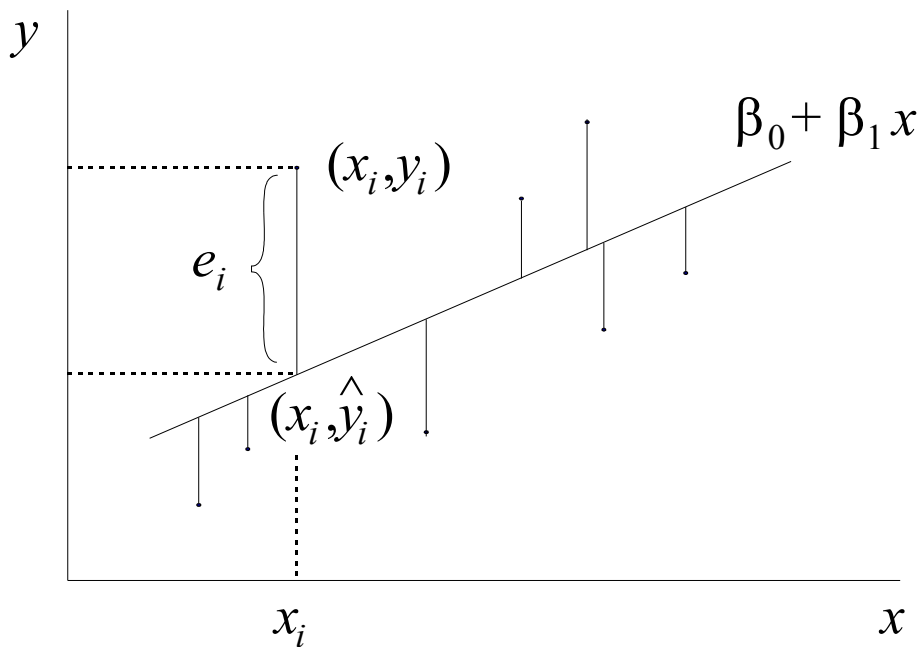
## ***Metodo dei minimi quadrati***

- Per un certo valore dei parametri ( $\beta_0$  e  $\beta_1$ ) si definisce ***valore teorico*** (o ***previsto***) corrispondente alla  $i$ -esima osservazione, la quantità

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

- Il corrispondente ***errore di previsione*** (o ***residuo***) è

$$e_i = y_i - \hat{y}_i$$



- La ***somma dei quadrati dei residui*** è una misura complessiva dell’errore di previsione, definita come

$$S(\beta_0, \beta_1) = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (y_i - \beta_0 - \beta_1 x_i)^2$$