La Retta di Regressione

M.M.I.B. 2019/20
corso di laurea in Scienze Biologiche
Sapienza Università di Roma

Riflessioni per cominciare

Quale modello per i dati osservati negli esempi precedenti?

Riflessioni per cominciare

Quale modello per i dati osservati negli esempi precedenti?

È possibile descrivere la relazione tra queste coppie di dati viste nei diversi esempi facendo uso di un modello statistico?

Riflessioni per cominciare

Quale modello per i dati osservati negli esempi precedenti?

È possibile descrivere la relazione tra queste coppie di dati viste nei diversi esempi facendo uso di un modello statistico?

Se lo scatter plot evidenzia che i punti sono disposti attorno a una retta crescente o decrescente, si parla di correlazione lineare,

Riflessioni per cominciare

Quale modello per i dati osservati negli esempi precedenti?

È possibile descrivere la relazione tra queste coppie di dati viste nei diversi esempi facendo uso di un modello statistico?

Se lo scatter plot evidenzia che i punti sono disposti attorno a una retta crescente o decrescente, si parla di correlazione lineare,

In tal caso si può tracciare una **retta di regressione** (anche **retta interpolante**) a partire dai dati.

Riflessioni per cominciare

Quale modello per i dati osservati negli esempi precedenti?

È possibile descrivere la relazione tra queste coppie di dati viste nei diversi esempi facendo uso di un modello statistico?

Se lo scatter plot evidenzia che i punti sono disposti attorno a una retta crescente o decrescente, si parla di correlazione lineare,

In tal caso si può tracciare una **retta di regressione** (anche **retta interpolante**) a partire dai dati.

Perciò si può pensare a un modello lineare. E si potrà usare il modello anche per fare previsioni sulla variazione della y al variare di x (**estrapolazione**).

La retta di regressione

Un modello di regressione lineare è un modello statistico che descrive una relazione lineare tra due variabili quantitative,

La retta di regressione

Un modello di regressione lineare è un modello statistico che descrive una relazione lineare tra due variabili quantitative,

descrive come cambia (linearmente) una variabile di risposta y quando cambia la variabile esplicativa x,

La retta di regressione

Un modello di regressione lineare è un modello statistico che descrive una relazione lineare tra due variabili quantitative,

descrive come cambia (linearmente) una variabile di risposta y quando cambia la variabile esplicativa x,

spesso viene usata per prevedere nuovi valori di y da nuovi valori di x,

La retta di regressione

Un modello di regressione lineare è un modello statistico che descrive una relazione lineare tra due variabili quantitative,

descrive come cambia (linearmente) una variabile di risposta y quando cambia la variabile esplicativa x,

spesso viene usata per prevedere nuovi valori di y da nuovi valori di x,

la retta di regressione lineare è descritta da una relazione

$$y = bx + a + e$$

dove e è un errore casuale che si suppone essere una variabile aleatoria a media nulla.

Back to the future!

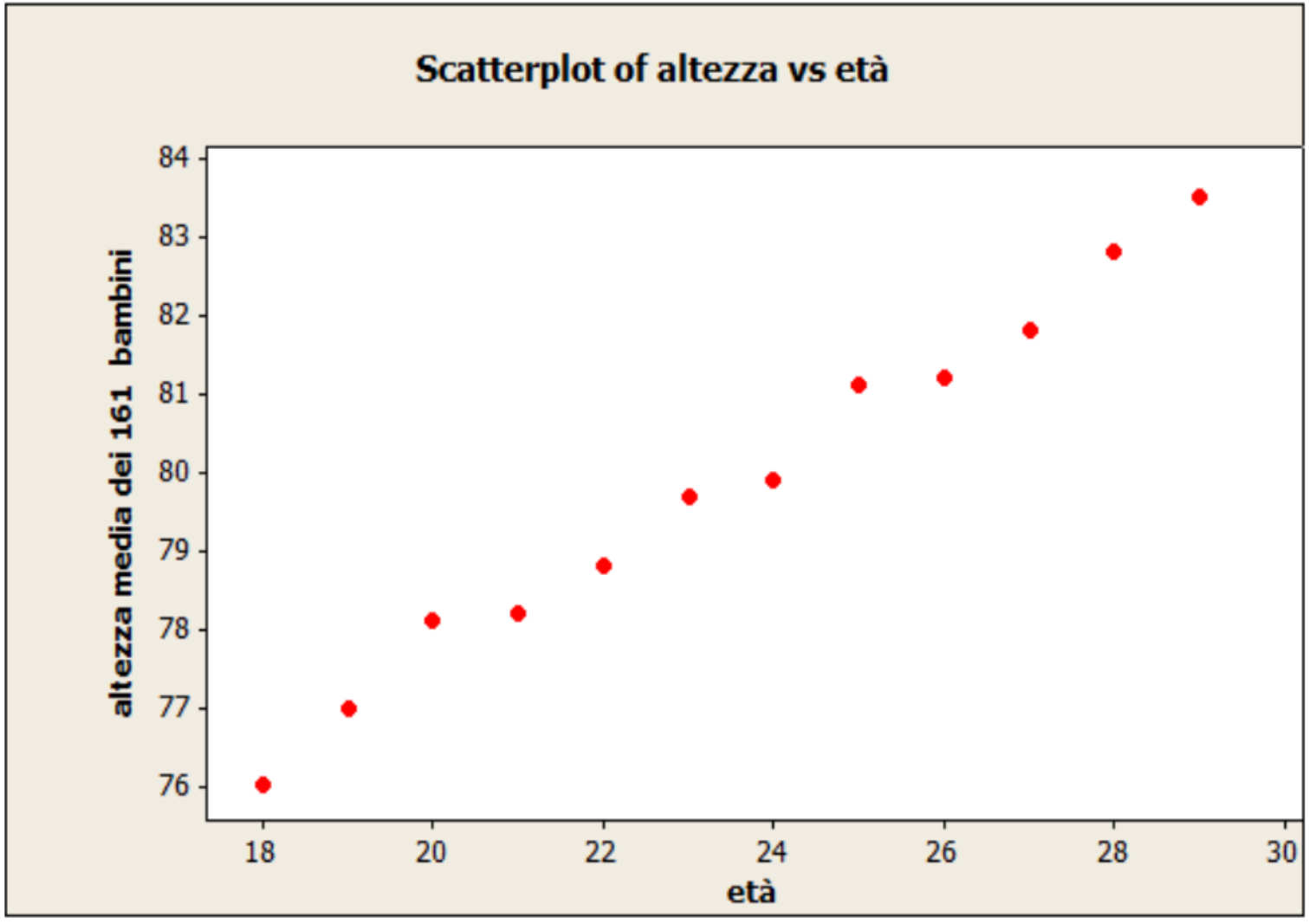
Consideriamo nuovamente l'esempio 2, cioè i dati ralativi ad un gruppo di 161 bambini del villaggio di Kalama (Egitto), per capire se possiamo proporre un legame lineare tra l'altezza media (in centimetri) del gruppo di bambini e la loro età in mesi, dati riproposti nella seguente tabella

Back to the future!

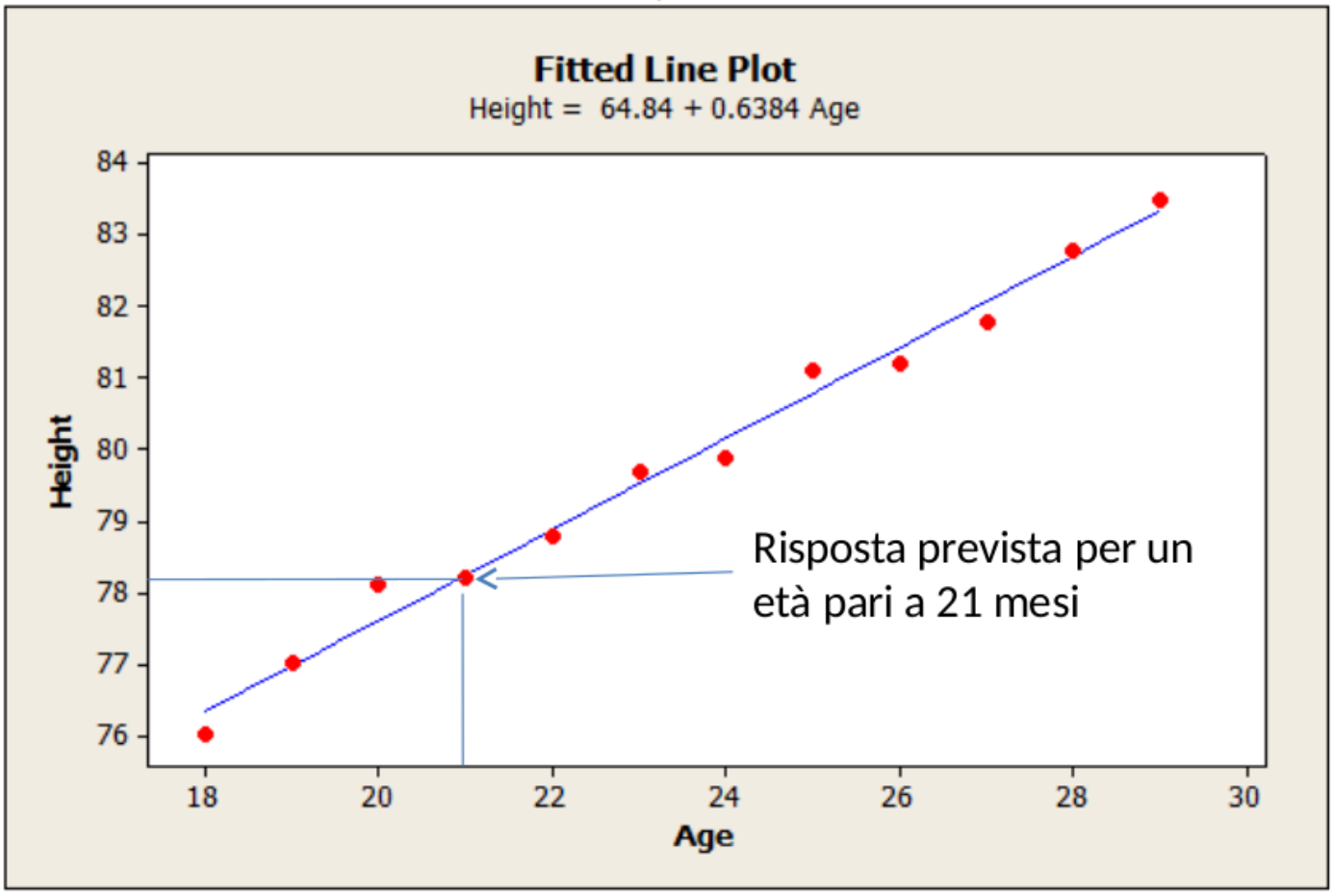
Consideriamo nuovamente l'esempio 2, cioè i dati ralativi ad un gruppo di 161 bambini del villaggio di Kalama (Egitto), per capire se possiamo proporre un legame lineare tra l'altezza media (in centimetri) del gruppo di bambini e la loro età in mesi, dati riproposti nella seguente tabella

mesi	18	19	20	21	22	23
cm	76.01	77.00	78.10	78.20	78.80	79.70
mesi	24	25	26	27	28	29
cm	79.90	81.10	81.20	81.80	82.20	83.50

scatter plot



scatter plot



Significato della retta

La retta che interpola i punti sperimentali rappresenta una stima della retta di regressione e permette di prevedere, data l'età di un bambino, quale dovrebbe essere in media la sua altezza,

Significato della retta

La retta che interpola i punti sperimentali rappresenta una stima della retta di regressione e permette di prevedere, data l'età di un bambino, quale dovrebbe essere in media la sua altezza,

se osserviamo altezze molto lontane dalla retta, abbiamo a che fare con dei casi anomali.

Significato della retta

La retta che interpola i punti sperimentali rappresenta una stima della retta di regressione e permette di prevedere, data l'età di un bambino, quale dovrebbe essere in media la sua altezza,

se osserviamo altezze molto lontane dalla retta, abbiamo a che fare con dei casi anomali.

Si noti che nel contesto della regressione occorre scegliere quale variabile è esplicativa.

Metodo dei minimi quadrati

Poiché la relazione espressa dal modello di regressione lineare semplice consiste nell'equazione di una retta, si tratterà di trovare la retta che meglio approssima i punti osservati.

Metodo dei minimi quadrati

Poiché la relazione espressa dal modello di regressione lineare semplice consiste nell'equazione di una retta, si tratterà di trovare la retta che meglio approssima i punti osservati.

Secondo il metodo dei minimi quadrati si sceglie la "migliore" retta di regressione minimizzando la somma dei quadrati delle distanze verticali tra i punti osservati e la retta stessa, ovvero la somma dei quadrati dei **residui** (anche detti errori di previsione).

Metodo dei minimi quadrati

Poiché la relazione espressa dal modello di regressione lineare semplice consiste nell'equazione di una retta, si tratterà di trovare la retta che meglio approssima i punti osservati.

Secondo il metodo dei minimi quadrati si sceglie la "migliore" retta di regressione minimizzando la somma dei quadrati delle distanze verticali tra i punti osservati e la retta stessa, ovvero la somma dei quadrati dei **residui** (anche detti errori di previsione).

Adrien Marie Legendre (1752-1833, Parigi).

Metodo dei minimi quadrati

Il residuo è la differenza fra un valore osservato della variabile di risposta e il valore previsto dalla stima della retta di regressione $y = bx + a$,

Metodo dei minimi quadrati

Il residuo è la differenza fra un valore osservato della variabile di risposta e il valore previsto dalla stima della retta di regressione $y = bx + a$, quindi

$$\text{residuo}_i = \text{res}_i = [y_{i,\text{osservato}} - y_{i,\text{previsto}}]$$

Metodo dei minimi quadrati

Il residuo è la differenza fra un valore osservato della variabile di risposta e il valore previsto dalla stima della retta di regressione $y = bx + a$, quindi

$$\text{residuo}_i = \text{res}_i = [y_{i,\text{osservato}} - y_{i,\text{previsto}}]$$

equivalentemente

$$\text{res}_i = [y_i - (bx_i + a)]$$

Metodo dei minimi quadrati

Il residuo è la differenza fra un valore osservato della variabile di risposta e il valore previsto dalla stima della retta di regressione $y = bx + a$, quindi

$$\text{residuo}_i = \text{res}_i = [y_{i,\text{osservato}} - y_{i,\text{previsto}}]$$

equivalentemente

$$\text{res}_i = [y_i - (bx_i + a)]$$

Un residuo può essere positivo o negativo, per determinare la retta dei minimi quadrati si minimizza la **somma dei quadrati dei residui**.

La retta di regressione dei minimi quadrati

Partendo dalle osservazione $\{x_1, \ldots, x_n\}$ e $\{y_1, \ldots, y_n\}$

1. ricaviamo le medie $\overline{x}$ e $\overline{y}$,

La retta di regressione dei minimi quadrati

Partendo dalle osservazione $\{x_1, \ldots, x_n\}$ e $\{y_1, \ldots, y_n\}$

1. ricaviamo le medie $\overline{x}$ e $\overline{y}$,
2. ricaviamo le deviazioni standard delle due variabili s_x e s_y,

La retta di regressione dei minimi quadrati

Partendo dalle osservazione $\{x_1, \ldots, x_n\}$ e $\{y_1, \ldots, y_n\}$

1. ricaviamo le medie $\overline{x}$ e $\overline{y}$,
2. ricaviamo le deviazioni standard delle due variabili s_x e s_y,
3. calcoliamo la loro correlazione r,

La retta di regressione dei minimi quadrati

Partendo dalle osservazione $\{x_1, \ldots, x_n\}$ e $\{y_1, \ldots, y_n\}$

1. ricaviamo le medie $\overline{x}$ e $\overline{y}$,

2. ricaviamo le deviazioni standard delle due variabili s_x e s_y,

3. calcoliamo la loro correlazione r,

allora la retta di regressione dei minimi quadrati ha la seguente espressione

$$y = bx + a \qquad \text{dove} \qquad b = r\frac{s_y}{s_x} \qquad \text{e} \qquad a = \overline{y} - b\overline{x}$$

Esempio 2

Tornando all'esempio 2 abbiamo che

1. $\bar{x} = 23,5$ e $\bar{y} = 79,79$,
2. $s_x = 3,61$ e $s_y = 2,25$,
3. $$r = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s_x} \right] \left[\frac{y_i - \bar{y}}{s_y} \right]$$
$$= \frac{1}{11} \sum_{i=1}^{12} \left[\frac{x_i - 23,5}{3,61} \right] \left[\frac{y_i - 79,79}{2,25} \right] = 0,99$$

Esempio 2

Tornando all'esempio 2 abbiamo che

1. $\bar{x} = 23,5$ e $\bar{y} = 79,79$,
2. $s_x = 3,61$ e $s_y = 2,25$,
3. $$r = \frac{1}{n-1} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s_x} \right] \left[\frac{y_i - \bar{y}}{s_y} \right]$$
$$= \frac{1}{11} \sum_{i=1}^{12} \left[\frac{x_i - 23,5}{3,61} \right] \left[\frac{y_i - 79,79}{2,25} \right] = 0,99$$

quindi abbiamo che la retta di regressione dei minimi quadrati ha l'espressione

$$y = bx + a \qquad \text{con} \qquad b = 0,62 \qquad a = 65,29$$

Sui residui

Perché ci sia un buon adattamento del modello ai dati vogliamo che

Sui residui

Perché ci sia un buon adattamento del modello ai dati vogliamo che

i residui non individuino alcun andamento ulteriormente interpolabile con termini di ordine superiore,

Sui residui

Perché ci sia un buon adattamento del modello ai dati vogliamo che

i residui non individuino alcun andamento ulteriormente interpolabile con termini di ordine superiore,

il segno dei residui sia "casuale", sia cioè, in qualche modo, ripartito equamente tra $+$ e $-$, per escludere errori sistematici,

Sui residui

Perché ci sia un buon adattamento del modello ai dati vogliamo che

i residui non individuino alcun andamento ulteriormente interpolabile con termini di ordine superiore,

il segno dei residui sia "casuale", sia cioè, in qualche modo, ripartito equamente tra $+$ e $-$, per escludere errori sistematici,

rappresentando su di un grafico i punti (x_i, r_i) se la relazione tra x e y è lineare dovremmo osservare dei punti che oscillano casualmente sopra e sotto la quota 0.

Sui residui

Riguardo all'esempio 2 abbiamo