

METODI STATISTICI PER LA RICERCA SOCIALE

CAPITOLO 9. REGRESSIONE LINEARE E CORRELAZIONE

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM-88 SOCIOLOGIA E RICERCA SOCIALE

CONCETTI PRELIMINARI

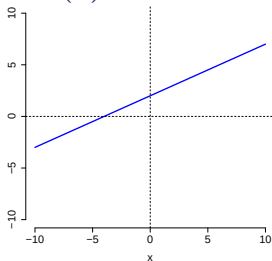
Relazione funzionale

Una variabile Y è funzione di una variabile X se ad ogni valore di X corrisponde uno e un solo valore di Y

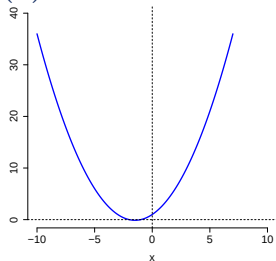
$$X \mapsto Y = f(X)$$

Relazioni funzionali: Esempi

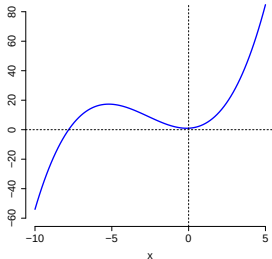
$$f(x) = 2 + 0.5 \cdot x$$



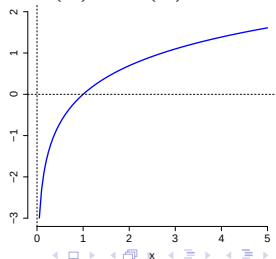
$$f(x) = 1 + 1.5 \cdot x + 0.5 \cdot x^2$$



$$f(x) = 1 + 0.5 \cdot x + 2 \cdot x^2 + 0.25 \cdot x^3$$



$$f(x) = \ln(x) \quad x > 0$$

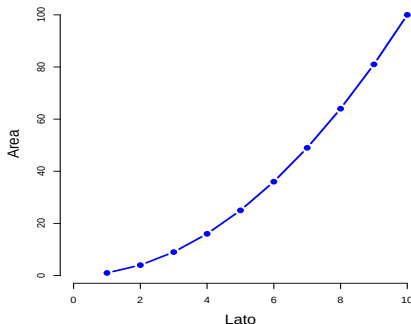


Relazione funzionale quadratica - Area di un quadrato

Se X = Lato di un quadrato e Y = Area del quadrato

$$Y = f(X) = X^2$$

Lato (X)	0	1	2	3	4	5	6	7	8	9	10
Area (Y)	0	1	4	9	16	25	36	49	64	81	100

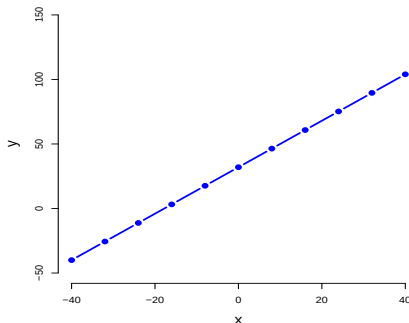


Relazione funzionale lineare (funzione lineare)

Se X = Temperatura in gradi Celsius e Y = Temperatura in gradi Fahrenheit

$$Y = 32 + 1.8 \cdot X$$

X	-40.0	-32.0	-24.0	-16.0	-8.0	0.0	8.0	16.0	24.0	32.0	40.0
Y	-40.0	-25.6	-11.2	3.2	17.6	32.0	46.4	60.8	75.2	89.6	104.0



Funzione lineare

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

dove β_0 e β_1 sono delle quantità costanti

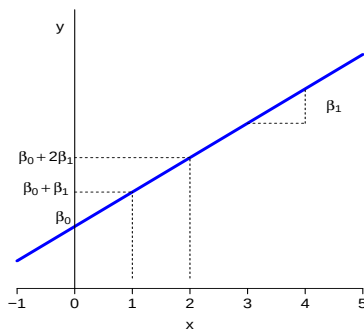
- β_0 è chiamata l'**intercetta** della retta
- L'intercetta è l'altezza alla quale la retta incontra l'asse delle ordinate, ossia il valore di Y per $X = 0$

$$x = 0 \implies y = \beta_0 + \beta_1 \cdot x = \beta_0 + \beta_1 \cdot 0 = \beta_0$$

- β_1 è chiamato **coefficiente angolare** della retta
- Il coefficiente angolare rappresenta la pendenza della retta, ossia la variazione di y per ogni incremento unitario di x (indipendentemente dal valore di x)

$$\beta_1 = \frac{y_{x+1} - y_x}{(x+1) - x} = [\beta_0 + \beta_1 \cdot (x+1)] - [\beta_0 + \beta_1 \cdot x]$$

Funzione lineare



- Intercetta: β_0

$$x = 0 \implies y = \beta_0 + \beta_1 \cdot x = \beta_0$$

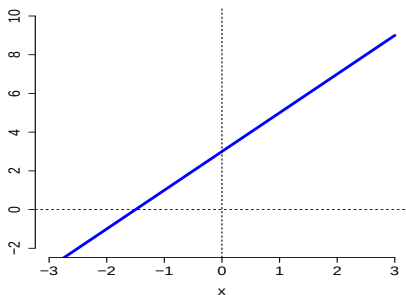
- Coefficiente angolare: β_1

$$\beta_1 = \frac{y_{x+1} - y_x}{(x+1) - x} = [\beta_0 + \beta_1 \cdot (x+1)] - [\beta_0 + \beta_1 \cdot x]$$

Relazione lineare positiva

La relazione tra due variabili è lineare positiva se può essere rappresentata mediante una retta con coefficiente angolare positivo: $\beta_1 > 0$

- Y cresce al crescere di X (la retta ha un andamento dal basso verso l'alto)
- Valori elevati (piccoli) di Y si verificano in corrispondenza di valori elevati (piccoli) valori X (concordanza)

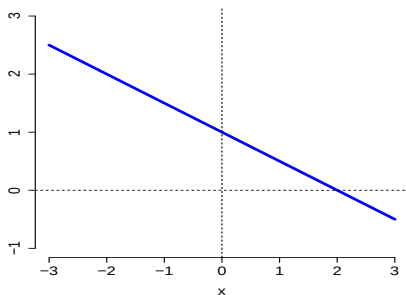


$$y = 3 + 2 \cdot x$$

Relazione lineare negativa

La relazione tra due variabili è lineare negativa se può essere rappresentata mediante una retta con coefficiente angolare negativo: $\beta_1 < 0$

- Y decresce al crescere di X (la retta ha un andamento dall'alto verso il basso)
- Valori elevati (piccoli) di Y si verificano in corrispondenza di valori piccoli (elevati) X (discordanza)

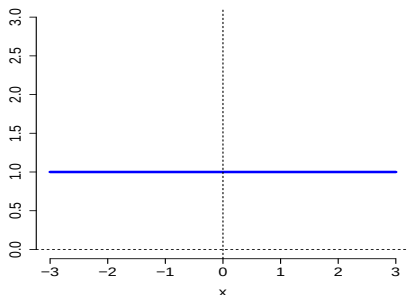


$$y = 1 - 0.5 \cdot x$$

Assenza di relazione lineare

Coefficiente angolare della retta nullo: $\beta_1 = 0$

- Y rimane costante al variare di X (la retta è parallela all'asse delle ascisse)



$$Y = 3 + 0 \cdot X$$

CORRELAZIONE

Associazione tra caratteri qualitativi ordinati/quantitativi

Nello studio dell'associazione tra caratteri qualitativi ordinabili e nello studio dell'associazione tra caratteri quantitativi è possibile valutare sia l'intensità sia il verso dell'associazione

- Poiché le modalità dei due caratteri sono ordinate, possono esistere tra loro due diversi tipi di relazione
 - ✓ Relazione diretta (positiva) o Concordanza
 - ✓ Relazione inversa (negativa) o Discordanza
- Un indice di associazione per caratteri ordinati dovrebbe assumere sia valori negativi che positivi
 - ✓ Valori positivi nel caso in cui esista concordanza tra i due caratteri
 - ✓ Valori negativi nel caso in cui esista discordanza tra i due caratteri

Concordanza e Discordanza: Caratteri qualitativi ordinali

Siano X e Y due caratteri qualitativi ordinali

- Esiste concordanza tra X e Y se le modalità di ordine elevato di X si associano più frequentemente a modalità di ordine elevato di Y , mentre modalità di ordine basso di X si associano più frequentemente a modalità di ordine basso di Y (e viceversa)
- Esiste discordanza tra X e Y se le modalità di ordine elevato di X si associano più frequentemente a modalità di ordine basso di Y , mentre modalità di ordine basso di X si associano più frequentemente a modalità di ordine elevato di Y (e viceversa)
- Misure di associazione tra caratteri qualitativi ordinali dipendono in genere dal numero di coppie di modalità concordanti e discordanti

Concordanza e Discordanza: Caratteri quantitativi

Siano X e Y due caratteri quantitativi

- Esiste concordanza tra X e Y se
 - ✓ al crescere di X cresce Y e al decrescere di X decresce Y
ovvero
 - ✓ al crescere di Y cresce X e al decrescere di Y decresce X
- Esiste discordanza tra X e Y se
 - ✓ al crescere di X decresce Y e al decrescere di X cresce Y
ovvero
 - ✓ al crescere di Y decresce X e al decrescere di Y cresce X

Associazione tra caratteri quantitativi

Istruzione e qualità della vita
(Soggetti di oltre 50 anni)

Unità	Edu	QoL
u_i	x_i	y_i
1	13	74
2	5	66
3	8	62
4	18	78
5	10	77
6	21	84
7	15	80
8	25	85
9	13	74
10	20	85
11	13	64
12	16	83
13	18	76

Edu = Anni di istruzione

QoL = Qualità della vita (Scala: 0 – 100)
(QoL = Quality of life)

Attività fisica e body mass index
(Maschi di 35 – 50 anni)

Unità	PA	BMI
u_i	x_i	y_i
1	4	23.3
2	2	23.6
3	1	23.7
4	5	22.9
5	1	23.5
6	4	23.1
7	3	23.4
8	5	22.9
9	2	23.1
10	5	22.6
11	1	24.5
12	2	23.3
13	4	23.0

PA = Ore di attività fisica settimanali
(PA=Physical activity)

BMI = Body Mass Index = $(\text{Peso} / \text{Altezza}^2)$

Diagramma a dispersione (Scatter plot)

- Per mettere in evidenza la possibile relazione tra due caratteri quantitativi si può rappresentare l'insieme dei dati attraverso un diagramma a dispersione
- Siano X e Y due variabili quantitative di cui si osservano le realizzazioni su n unità statistiche:

$$(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i), \dots, (x_n, y_n)$$

- Le coppie ordinate di modalità dei due caratteri quantitativi osservati per ogni unità del collettivo vengono rappresentate come punti di un piano cartesiano i cui assi ortogonali corrispondono ai caratteri

Diagramma a dispersione - Esempio

X = Anni di istruzione

Y = Qualità della vita

u_1	u_2	...	u_{10}	...	u_{13}
(x_1, y_1)	(x_2, y_2)	...	(x_{10}, y_{10})	...	(x_{13}, y_{13})
(13, 74)	(5, 66)	...	(20, 85)	...	(18, 76)

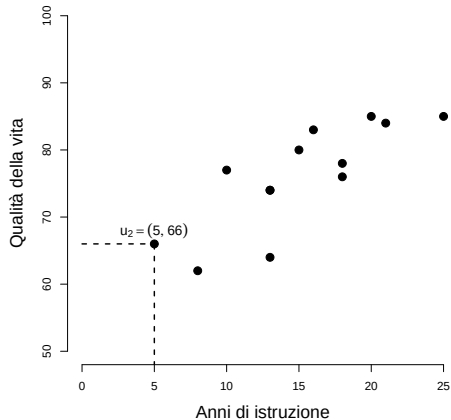
X = Ore di attività fisica (settimanali)

Y = BMI

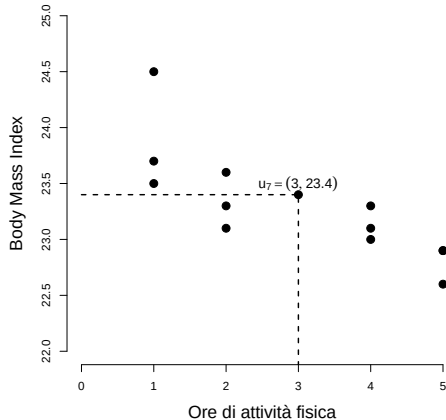
u_1	u_2	...	u_{10}	...	u_{13}
(x_1, y_1)	(x_2, y_2)	...	(x_{10}, y_{10})	...	(x_{13}, y_{13})
(4, 23.3)	(2, 23.6)	...	(5, 22.6)	...	(4, 23.0)

Diagramma a dispersione - Esempi

Istruzione e Qualità della vita



Attività fisica e BMI



Variabili scarto

Si considerino le variabili centrate rispetto alla loro media:

$$X' = X - \bar{X} \quad \text{e} \quad Y' = Y - \bar{Y}$$

dove

$$\bar{X} = \frac{x_1 + \dots + x_n}{n} \quad \text{e} \quad \bar{Y} = \frac{y_1 + \dots + y_n}{n}$$

u_i	X	$X' = X - \bar{X}$	Y	$Y' = Y - \bar{Y}$
u_1	x_1	$x_1 - \bar{X}$	y_1	$y_1 - \bar{Y}$
u_2	x_2	$x_2 - \bar{X}$	y_2	$y_2 - \bar{Y}$
...	...	...	...	...
u_i	x_i	$x_i - \bar{X}$	y_i	$y_i - \bar{Y}$
...	...	...	...	...
u_n	x_n	$x_n - \bar{X}$	y_n	$y_n - \bar{Y}$

Concordanza e discordanza

- Due caratteri quantitativi presentano **concordanza** se la maggior parte degli scostamenti sono concordi
 - X' e Y' entrambi positivi (o negativi)
- Due caratteri quantitativi presentano **discordanza** se la maggior parte degli scostamenti sono discordi
 - X' positivo e Y' negativo (o viceversa)
- Rappresentare il diagramma a dispersione dei valori scartati dalla media

Variabili scarto: Esempio – Istruzione e qualità della vita

Unità	Edu	QoL	$\text{Edu} - \overline{\text{Edu}}$	$\text{QoL} - \overline{\text{QoL}}$
u_i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$
1	13	74	-2	-2
2	5	66	-10	-10
3	8	62	-7	-14
4	18	78	3	2
5	10	77	-5	1
6	21	84	6	8
7	15	80	0	4
8	25	85	10	9
9	13	74	-2	-2
10	20	85	5	9
11	13	64	-2	-12
12	16	83	1	7
13	18	76	3	0
Totale	195	988	0	0

$$\overline{\text{Edu}} = \frac{195}{13} = 15$$

$$\overline{\text{QoL}} = \frac{988}{13} = 76$$

Variabili scarto: Esempio – Attività fisica e BMI

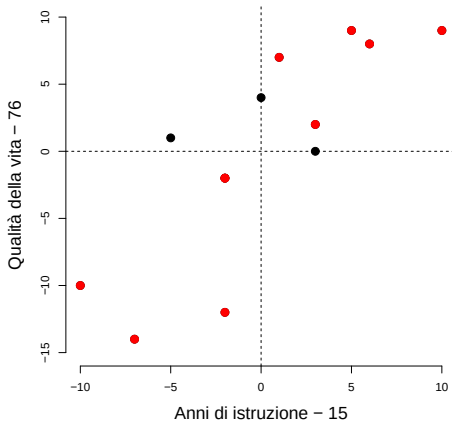
Unità	PA	BMI	$PA - \overline{PA}$	$BMI - \overline{BMI}$
u_i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$
1	4	23.3	1	0.0
2	2	23.6	-1	0.3
3	1	23.7	-2	0.4
4	5	22.9	2	-0.4
5	1	23.5	-2	0.2
6	4	23.1	1	-0.2
7	3	23.4	0	0.1
8	5	22.9	2	-0.4
9	2	23.1	-1	-0.2
10	5	22.6	2	-0.7
11	1	24.5	-2	1.2
12	2	23.3	-1	0.0
13	4	23.0	1	-0.3
Totale	39	302.9	0	0

$$\overline{PA} = \frac{39}{13} = 3$$

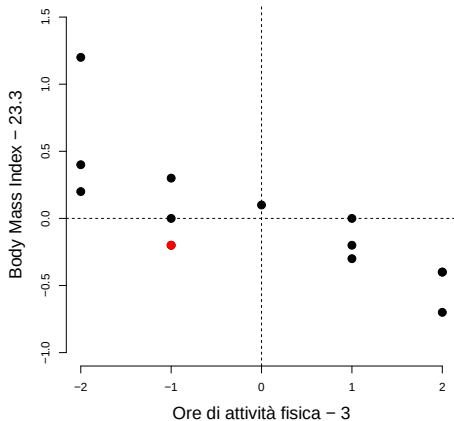
$$\overline{BMI} = \frac{302.9}{13} = 23.3$$

Diagramma a dispersione delle variabili scarto - Esempi

Istruzione e Qualità della vita



Attività fisica e BMI



Covarianza

- Nella popolazione la covarianza è il valore atteso (media) del prodotto degli scarti dalla media:

$$\sigma_{XY} = \mathbb{E}[(X - \mu_X) \cdot (Y - \mu_Y)]$$

Potete pensare che ...

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N [(x_i - \mu_X) \cdot (y_i - \mu_Y)]$$

con

$$\mu_X = \frac{1}{N} \sum_{i=1}^N x_i \quad e \quad \mu_Y = \frac{1}{N} \sum_{i=1}^N y_i$$

dove N è il numero di unità statistiche nella popolazione

- La covarianza misura la concordanza o la discordanza tra due caratteri quantitativi

Covarianza: Proprietà

- È un indice simmetrico: $\sigma_{XY} = \sigma_{YX}$
- Valori della covarianza

$$-\sigma_X \cdot \sigma_Y \leq \sigma_{XY} \leq \sigma_X \cdot \sigma_Y$$

dove σ_X e σ_Y sono le deviazioni standard di X e Y (nella popolazione)

- $\sigma_{XY} > 0 \iff X$ e Y sono concordi
- $\sigma_{XY} < 0 \iff X$ e Y sono discordi
- $\sigma_{XY} = 0 \iff X$ e Y incorrelate (assenza di relazione lineare)
- Se X e Y sono statisticamente indipendenti allora $\sigma_{XY} = 0$. Non è vero il viceversa

Relazione non lineare

u_i	x_i	y_i	$(x_i - \mu_X) \cdot (y_i - \mu_Y)$
1	-2	4	-3.0
2	-1	1	1.5
3	1	1	-1.5
4	2	4	3.0
Totale	0	10	0.0

$$\mu_X = \frac{-2 + (-1) + 1 + 2}{4} = 0 \quad \text{e} \quad \mu_Y = \frac{4 + 1 + 1 + 4}{4} = 2.5$$

$$\sigma_{XY} = \frac{-3.0 + 1.5 - 1.5 + 3.0}{4} = 0$$

N.B Tra X e Y esiste una relazione non lineare $Y = X^2$: Le due variabili non sono statisticamente indipendenti

Covarianza campionaria

- Campione casuale di n unità: $(X_1, Y_1), \dots, (X_n, Y_n)$
- Stimatore della covarianza: Somma dei prodotti degli scostamenti dalle rispettive medie campionarie, $\bar{X}$ e $\bar{Y}$ diviso $n - 1$:

$$\begin{aligned}
 S_{XY} &= \frac{(X_1 - \bar{X}) \cdot (Y_1 - \bar{Y}) + \dots + (X_n - \bar{X}) \cdot (Y_n - \bar{Y})}{n - 1} \\
 &= \frac{\sum_{i=1}^n [(X_i - \bar{X}) \cdot (Y_i - \bar{Y})]}{n - 1}
 \end{aligned}$$

- Si può dimostrare che

$$S_{XY} = \frac{(X_1 \cdot Y_1 + \dots + X_n \cdot Y_n) - n \cdot \bar{X} \cdot \bar{Y}}{n - 1} = \frac{\sum_{i=1}^n X_i \cdot Y_i - n \cdot \bar{X} \cdot \bar{Y}}{n - 1}$$

Proprietà della covarianza campionaria

- È un indice simmetrico: $S_{XY} = S_{YX}$
- Valori della covarianza campionaria

$$-S_X \cdot S_Y \leq S_{XY} \leq S_X \cdot S_Y$$

dove S_X e S_Y sono le deviazioni standard campionarie (corrette) di X e Y :

$$S_X = \sqrt{S_X^2} = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} \quad \text{e} \quad S_Y = \sqrt{S_Y^2} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}}$$

- s_{XY} = Valore osservato della covarianza campionaria su un particolare campione

Istruzione e qualità della vita: Covarianza campionaria tra Edu e QoL

Unità	Edu	QoL	Edu · QoL	$Edu - \overline{Edu}$	$QoL - \overline{QoL}$	$(Edu - \overline{Edu}) \cdot (QoL - \overline{QoL})$
u_i	x_i	y_i	$x_i \cdot y_i$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	13	74	962	-2	-2	4
2	5	66	330	-10	-10	100
3	8	62	496	-7	-14	98
4	18	78	1404	3	2	6
5	10	77	770	-5	1	-5
6	21	84	1764	6	8	48
7	15	80	1200	0	4	0
8	25	85	2125	10	9	90
9	13	74	962	-2	-2	4
10	20	85	1700	5	9	45
11	13	64	832	-2	-12	24
12	16	83	1328	1	7	7
13	18	76	1368	3	0	0
Totale	195	988	15241	0	0	421

$$\overline{Edu} = \frac{195}{13} = 15$$

$$\overline{QoL} = \frac{988}{13} = 76$$

Istruzione e qualità della vita: Covarianza campionaria tra Edu e QoL

$$\begin{aligned}
 S_{Edu, QoL} &= \frac{\sum_{i=1}^n [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{n - 1} \\
 &= \frac{(13 - 15) \cdot (74 - 76) + \dots + (18 - 15) \cdot (76 - 76)}{13 - 1} \\
 &= \frac{4 + \dots + 0}{12} = \frac{421}{12} = 35.083
 \end{aligned}$$

oppure

$$\begin{aligned}
 S_{Edu, QoL} &= \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{n - 1} \\
 &= \frac{(13 \cdot 74 + \dots + 18 \cdot 76) - 13 \cdot 15 \cdot 76}{13 - 1} \\
 &= \frac{15241 - 14820}{12} = \frac{421}{12} = 35.083
 \end{aligned}$$

Il coefficiente di correlazione lineare

- La covarianza dipende dall'unità di misura in cui sono espresse le variabili e dal loro ordine di grandezza rispetto alla media
- Si considera un indice relativo: il **coefficiente di correlazione lineare di Bravais e Pearson**

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

- Il coefficiente di correlazione è uguale alla covarianza tra variabili standardizzate

$$\rho_{X,Y} = \sigma_{Z_X,Z_Y}$$

dove

$$Z_X = \frac{X - \mu_X}{\sigma_X} \quad \text{e} \quad Z_Y = \frac{Y - \mu_Y}{\sigma_Y}$$

Proprietà del coefficiente di correlazione lineare

- Il coefficiente di correlazione e la covarianza hanno lo stesso segno
- Il coefficiente di correlazione non dipende dall'unità di misura delle variabili
- Il coefficiente di correlazione è un indice simmetrico (come la covarianza)

$$\rho_{XY} = \rho_{YX}$$

- $\rho_{XY} = 0 \iff \sigma_{XY} = 0$
- L'indipendenza implica l'incorrelazione, ma non è vero il viceversa:

$$X \text{ e } Y \text{ indipendenti} \Rightarrow \rho_{XY} = 0$$

$$X \text{ e } Y \text{ indipendenti} \nLeftarrow \rho_{XY} = 0$$

- $\sigma_{X,X} = \sigma_X^2$ e $\rho_{X,X} = 1$

Il coefficiente di correlazione lineare: Interpretazione

- Il coefficiente di correlazione lineare assume valori compresi tra -1 e 1:

$$-1 \leq \rho_{XY} \leq 1$$

- $\rho_{XY} = 1 \iff$ Esiste un'associazione lineare perfetta tra X e Y e i due caratteri sono concordi:

$$Y = \alpha + \beta \cdot X \quad \beta > 0 \quad \text{e} \quad X = \tilde{\alpha} + \tilde{\beta} \cdot Y \quad \tilde{\beta} > 0$$

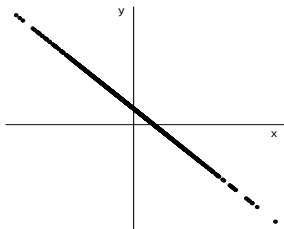
- $\rho_{XY} = -1 \iff$ Esiste un'associazione lineare perfetta tra X e Y e i due caratteri sono discordi:

$$Y = \alpha + \beta \cdot X \quad \beta < 0 \quad \text{e} \quad X = \tilde{\alpha} + \tilde{\beta} \cdot Y \quad \tilde{\beta} < 0$$

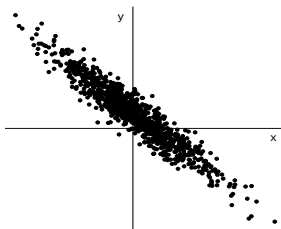
- $\rho_{XY} = 0 \iff$ Assenza di associazione lineare tra X e Y

Il coefficiente di correlazione lineare

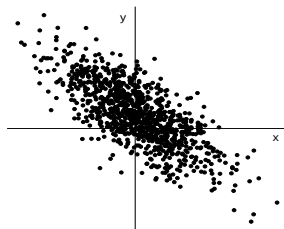
$$\rho_{XY} = -1$$



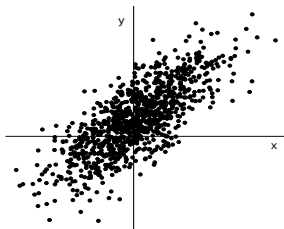
$$\rho_{XY} = -0.95$$



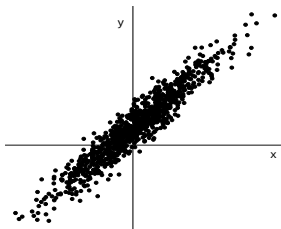
$$\rho_{XY} = -0.75$$



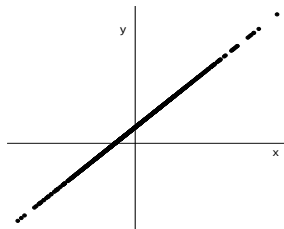
$$\rho_{XY} = 0.75$$



$$\rho_{XY} = 0.95$$

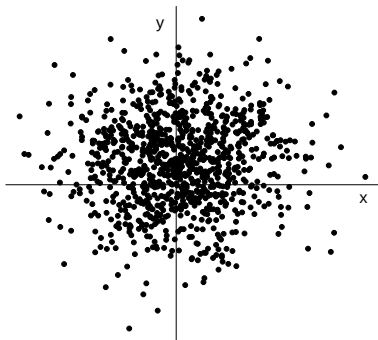


$$\rho_{XY} = 1$$

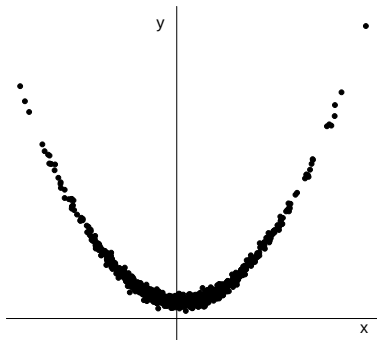


Il coefficiente di correlazione lineare

$$\rho_{XY} = 0$$



$$\rho_{XY} = 0$$



Il coefficiente di correlazione lineare campionario

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y}$$

$$S_{XY} = \frac{\sum_{i=1}^n [(X_i - \bar{X}) \cdot (Y_i - \bar{Y})]}{n-1} = \frac{\sum_{i=1}^n X_i \cdot Y_i - n \cdot \bar{X} \cdot \bar{Y}}{n-1}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^n X_i^2 - n \cdot \bar{X}^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^n Y_i^2 - n \cdot \bar{Y}^2}{n-1}}$$

Il coefficiente di correlazione lineare campionario

- Il coefficiente di correlazione campionario e la covarianza campionaria hanno lo stesso segno
- Il coefficiente di correlazione campionario non dipende dall'unità di misura delle variabili
- Il coefficiente di correlazione campionario è un indice simmetrico (come la covarianza campionaria)

$$r_{XY} = r_{YX}$$

- $r_{XY} = 0$ se e solo se $s_{XY} = 0$
- $s_{X,X} = s_X^2$ e $r_{X,X} = 1$
- Il coefficiente di correlazione campionario assume valori compresi tra -1 e 1:

$$-1 \leq r_{XY} \leq 1$$

Il coefficiente di correlazione lineare campionario

Valore osservato del coefficiente di correlazione campionario (su un particolare campione):

$$r_{XY} = \frac{s_{XY}}{s_X \cdot s_Y}$$

dove s_{XY} , s_X e s_Y sono i valori osservati della covarianza campionaria e delle deviazioni standard campionarie

Coefficiente di correlazione campionario tra Edu e QoL

Unità u_i	Edu x_i	QoL y_i	$(\text{Edu} - \overline{\text{Edu}})^2$ $(x_i - \bar{x})^2$	$(\text{QoL} - \overline{\text{QoL}})^2$ $(y_i - \bar{y})^2$	$(\text{Edu} - \overline{\text{Edu}}) \cdot (\text{QoL} - \overline{\text{QoL}})$ $(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	13	74	4	4	4
2	5	66	100	100	100
3	8	62	49	196	98
4	18	78	9	4	6
5	10	77	25	1	-5
6	21	84	36	64	48
7	15	80	0	16	0
8	25	85	100	81	90
9	13	74	4	4	4
10	20	85	25	81	45
11	13	64	4	144	24
12	16	83	1	49	7
13	18	76	9	0	0
Totale	195	988	366	744	421

$$\overline{\text{Edu}} = \frac{195}{13} = 15$$

$$\overline{\text{QoL}} = \frac{988}{13} = 76$$

Coefficiente di correlazione campionario tra Edu e QoL

$$\begin{aligned}
 s_{Edu, QoL} &= \frac{\sum_{i=1}^n [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{n - 1} \\
 &= \frac{(13 - 15) \cdot (74 - 76) + \dots + (18 - 15) \cdot (76 - 76)}{13 - 1} \\
 &= \frac{4 + \dots + 0}{12} = \frac{421}{12} = 35.083
 \end{aligned}$$

$$\begin{aligned}
 s_{Edu}^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \\
 &= \frac{(13 - 15)^2 + \dots + (18 - 15)^2}{13 - 1} \\
 &= \frac{4 + \dots + 9}{12} = \frac{366}{12} = 30.5
 \end{aligned}$$

$$\begin{aligned}
 s_{QoL}^2 &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1} \\
 &= \frac{(74 - 76)^2 + \dots + (76 - 76)^2}{15 - 1} \\
 &= \frac{4 + \dots + 0}{12} = \frac{744}{12} = 62
 \end{aligned}$$

$$r_{XY} = \frac{35.083}{\sqrt{30.5} \sqrt{62}} = 0.807$$

Inferenza per il coefficiente di correlazione

- Popolazioni: X e Y hanno una distribuzione congiunta Normale e quindi

$$X \sim N(\mu_X, \sigma_X^2) \quad e \quad Y \sim N(\mu_Y, \sigma_Y^2)$$

- Parametro = Coefficiente di correlazione tra X e Y nella popolazione: ρ_{XY}
- Sotto l'ipotesi di normalità: $\rho_{XY} = 0 \iff X$ e Y sono indipendenti

- Ipotesi

$$H_0 : \rho_{XY} = 0 \quad \text{versus} \quad H_a : \rho_{XY} \neq 0$$

- Statistica test

$$T = \frac{r_{XY}}{\sqrt{(1 - r_{XY}^2)/(n - 2)}}$$

- Sotto l'ipotesi nulla, $T \sim t_{n-2}$
- Regione di rifiuto al livello di significatività α

$$RC_\alpha = t \leq -t_{\alpha/2, n-2} \quad \text{oppure} \quad t \geq t_{\alpha/2, n-2}$$

- P -valore

$$p\text{-value} = Pr(T \leq -|t_{oss}| \text{ oppure } T \geq |t_{oss}|; H_0) = 2 \cdot [1 - Pr(T \leq |t_{oss}|; H_0)]$$

Inferenza per il coefficiente di correlazione tra Edu e QoL

- $\alpha = 0.05 \implies n = 13 \implies t_{0.025, 13-2} = t_{0.025, 11} = 2.20$
- Regione di rifiuto: $RC_{0.05} = T \leq -2.20$ oppure $T \geq 2.20$
- Coefficiente di correlazione campionario: $r = 0.807$
- Valore osservato della statistica test

$$t_{oss} = \frac{0.807}{\sqrt{(1 - (0.807)^2)/(13 - 2)}} = \frac{0.807}{\sqrt{(1 - 0.651)/11}} = \frac{0.807}{0.178} = 4.529$$

- Decisione: $t_{oss} = 4.529 > 2.2 \iff t_{oss} = 4.529 \in RC_{0.05} \implies$ I dati mostrano evidenza contro H_0 al livello di significatività del 5%.
- $p - value = 0.00086$

Esempio: Attività fisica e BMI

Unità	PA	BMI	PA ²	BMI ²	PA · BMI
u_i	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$
1	4	23.3	16	542.89	93.2
2	2	23.6	4	556.96	47.2
3	1	23.7	1	561.69	23.7
4	5	22.9	25	524.41	114.5
5	1	23.5	1	552.25	23.5
6	4	23.1	16	533.61	92.4
7	3	23.4	9	547.56	70.2
8	5	22.9	25	524.41	114.5
9	2	23.1	4	533.61	46.2
10	5	22.6	25	510.76	113.0
11	1	24.5	1	600.25	24.5
12	2	23.3	4	542.89	46.6
13	4	23.0	16	529.00	92.0
Totale	39	302.9	147	7060.29	901.5

$$\overline{PA} = \frac{39}{13} = 3$$

$$\overline{BMI} = \frac{302.9}{13} = 23.3$$

Covarianza campionaria e coefficiente di correlazione campionario

$$s_{PA,BMI} = \frac{\sum_{i=1}^n x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{n - 1} = \frac{901.5 - 13 \cdot 3 \cdot 23.3}{13 - 1} = \frac{-7.2}{12} = -0.6$$

$$\begin{aligned} s_{PA}^2 &= \frac{\sum_{i=1}^n x_i^2 - n \cdot \bar{x}^2}{n - 1} \\ &= \frac{147 - 13 \cdot (3)^2}{13 - 1} \\ &= \frac{30}{12} = 2.5 \end{aligned}$$

$$\begin{aligned} s_{BMI}^2 &= \frac{\sum_{i=1}^n y_i^2 - n \cdot \bar{y}^2}{n - 1} \\ &= \frac{7060.29 - 13 \cdot (23.3)^2}{13 - 1} \\ &= \frac{2.72}{12} = 0.227 \end{aligned}$$

$$r_{XY} = \frac{-0.6}{\sqrt{2.5} \cdot \sqrt{0.227}} = -0.797$$

Inferenza per il coefficiente di correlazione tra PA e BMI

- $\alpha = 0.05 \implies n = 13 \implies t_{0.025, 13-2} = t_{0.025, 11} = 2.20$
- Regione di rifiuto: $RC_{0.05} = T \leq -2.20$ oppure $T \geq 2.20$
- Coefficiente di correlazione campionario: $r = -0.797$
- Valore osservato della statistica test

$$t_{oss} = \frac{-0.797}{\sqrt{(1 - (-0.797)^2)/(13 - 2)}} = \frac{-0.797}{\sqrt{(1 - 0.635)/11}} = \frac{0.762}{0.182} = -4.377$$

- Decisione: $t_{oss} = -4.377 < -2.2 \iff t_{oss} = -4.377 \in RC_{0.05} \implies$ I dati mostrano evidenza contro H_0 al livello di significatività del 5%.
- $p - value = 0.001104$

REGRESSIONE LINEARE

Regressione lineare: Variabili risposta e variabili esplicative

Date due variabili quantitative X e Y , supponiamo di essere interessati a studiare come la variabile Y varia in risposta a variazioni della variabile X

Y = Variabile risposta (Variabile dipendente)

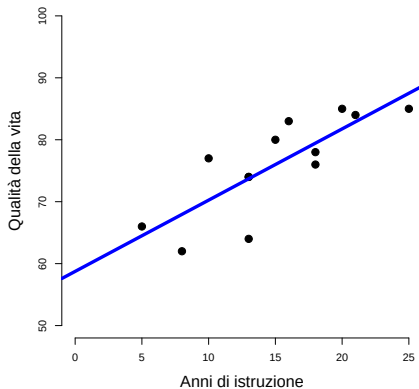
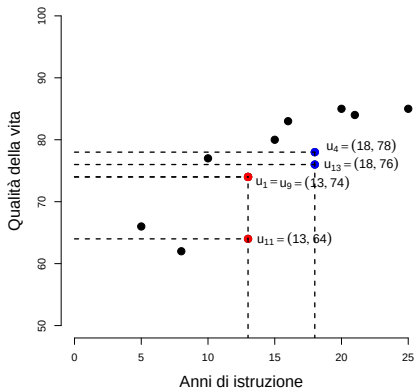
X = Variabile esplicativa (Variabile indipendente)

- Le variabili sono trattate in modo **asimmetrico**
- La relazione tra le due variabili può essere rappresentata attraverso un modello di regressione
 - ✓ Individuare un modello matematico (funzione) che descriva in modo dettagliato la relazione che emerge dai dati

Relazione statistica

- Negli studi empirici la relazione che può essere osservata tra le variabili non è mai una relazione matematica esatta: A un dato valore della variabile esplicativa X possono corrispondere diversi valori di Y
- La realtà è complessa: I modelli sono semplici approssimazioni delle relazioni che esistono tra le variabili nella popolazione
- Una funzione lineare rappresenta il modello matematico più semplice per descrivere la relazione fra due variabili quantitative

Diagramma a dispersione - Esempio: Anni di istruzione e qualità della vita



Il modello di regressione lineare semplice

Le osservazioni $y_1, \dots, y_i, \dots, y_n$ sono realizzazioni di variabili aleatorie $Y_1, \dots, Y_i, \dots, Y_n$ Normali indipendenti aventi media che è funzione lineare della variabile esplicativa X , e varianza costante indipendentemente dal valore della variabile esplicativa X :

Per $i = 1, \dots, n$ $Y_i | X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ indipendenti

Quindi, per ogni $i = 1, \dots, n$

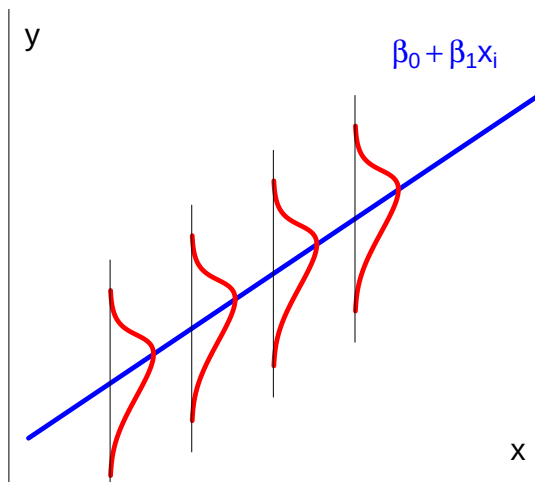
$$\mathbb{E}(Y_i | X_i = x_i) = \beta_0 + \beta_1 \cdot x_i$$

e

$$\text{Var}(Y_i | X_i = x_i) = \sigma^2$$

La condizione di varianza costante è detta **ipotesi di omoschedasticità**

Modello di regressione



Il modello di regressione lineare semplice: Assunzioni

Assunzione 1. Per ogni unità $i = 1, \dots, n$

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

Assunzione 2. Gli errori ϵ_i , $i = 1, \dots, n$, sono variabili aleatorie indipendenti aventi distribuzione Normale di media nulla e varianza costante indipendentemente dal valore della variabile esplicativa X (omoschedasticità)

Per $i = 1, \dots, n$ $\epsilon_i \sim N(0, \sigma^2)$ indipendenti

- ✓ Ciascuna osservazione ha il proprio valore dell'errore, ϵ_i : Il termine di errore ϵ_i rappresenta la deviazione di Y_i dalla media $\beta_0 + \beta_1 \cdot x_i$
- ✓ Il termine di errore ϵ_i rappresenta l'errore che si commette nel prevedere l'osservazione individuale usando la media di Y in corrispondenza di un valore fissato di X , $\beta_0 + \beta_1 \cdot x_i$
- ✓ Per ogni valore x_i della variabile esplicativa, X , la variabilità nei valori di Y corrisponde alla variabilità nei valori ϵ_i
- ✓ Gli n valori di ϵ non sono noti (sono variabili aleatorie), come non sono i noti i parametri β_0 , β_1 e σ^2 e quindi la vera media $\beta_0 + \beta_1 \cdot x_i$

Interpretazione dei coefficienti di regressione

Nel modello di regressione lineare semplice

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, n$$

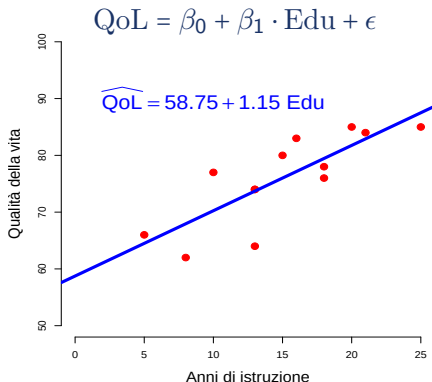
- Il parametro β_0 (intercetta della retta di regressione) rappresenta il valore atteso della variabile risposta Y per $X = 0$

$$\mathbb{E}(Y_i | X_i = 0) = \beta_0 + \beta_1 \cdot 0 = \beta_0$$

- Il parametro β_1 (coefficiente angolare della retta) descrive come la variabile risposta Y varia **in media** per ogni incremento unitario della variabile esplicativa X

$$\begin{aligned} & \frac{\mathbb{E}(Y_i | X_i = x_i + 1) - \mathbb{E}(Y_i | X_i = x_i)}{(x_i + 1) - x_i} \\ &= [\beta_0 + \beta_1 \cdot (x_i + 1)] - [\beta_0 + \beta_1 \cdot x_i] = \beta_1 \end{aligned}$$

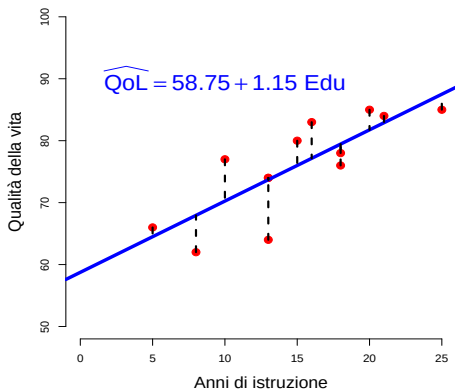
Edu e QoL ($r_{Edu, QoL} = 0.807$)



- $\hat{\beta}_0 = 58.75$: Qualità della vita media per soggetti con zero anni di istruzione
- $\hat{\beta}_1 = 1.15$: La qualità della vita aumenta in media di 1.15 punti per ogni incremento unitario del numero di anni di istruzione (ossia per ogni anno in più di istruzione)

Stima puntuale dei coefficienti di regressione

- Individuare una retta che per ogni valore di x restituisca un valore di y prossimo ai valori osservati
- La retta deve essere il più vicina possibile ai punti osservati nella direzione verticale



Il metodo dei minimi quadrati

Il metodo dei minimi quadrati consiste nel ricercare quei valori di β_0 e β_1 che rendono minima la funzione

$$G(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Stime dei minimi quadrati dei coefficienti di regressione

$$\widehat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Si noti che

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\left(\sum_{i=1}^n x_i \cdot y_i \right) - n \cdot \bar{x} \cdot \bar{y}}{\left(\sum_{i=1}^n x_i^2 \right) - n \cdot (\bar{x})^2}$$

Edu e QoL: Stima dei coefficienti di regressione

Unità u_i	Edu x_i	QoL y_i	$(\text{Edu} - \overline{\text{Edu}})^2$ $(x_i - \bar{x})^2$	$(\text{Edu} - \overline{\text{Edu}}) \cdot (\text{QoL} - \overline{\text{QoL}})$ $(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	13	74	4	4
2	5	66	100	100
3	8	62	49	98
4	18	78	9	6
5	10	77	25	-5
6	21	84	36	48
7	15	80	0	0
8	25	85	100	90
9	13	74	4	4
10	20	85	25	45
11	13	64	4	24
12	16	83	1	7
13	18	76	9	0
Totale	195	988	366	421

Edu e QoL: Stima dei coefficienti di regressione

- Medie campionarie

$$\overline{\text{Edu}} = \frac{195}{13} = 15 \qquad \overline{\text{QoL}} = \frac{988}{13} = 76$$

- Covarianza campionaria tra Edu e QoL e varianza campionaria (corretta) di Edu

$$s_{\text{Edu}, \text{QoL}} = \frac{421}{12} = 35.083 \qquad s_{\text{Edu}}^2 = \frac{366}{12} = 30.5$$

- Stime dei coefficienti di regressione: $\mathbb{E}[\text{QoL} \mid \text{Edu}] = \beta_0 + \beta_1 \cdot \text{Edu}$

$$\widehat{\beta}_1 = \frac{s_{\text{Edu}, \text{QoL}}}{s_{\text{Edu}}^2} = \frac{35.083}{30.5} = 1.15$$

$$\widehat{\beta}_0 = 76 - 1.15 \cdot 15 = 76 - 17.25 = 58.75$$

Valori Teorici e Residui di Regressione

- Valori Teorici

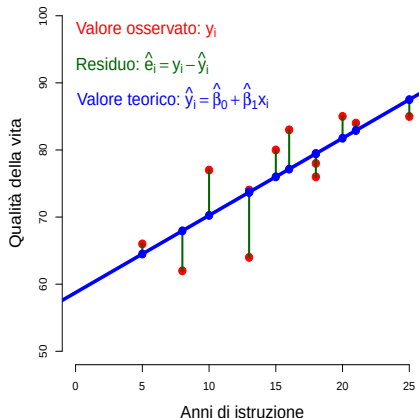
$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot x_i \quad i = 1, \dots, n$$

- I valori teorici sono i valori di Y forniti dalla retta di regressione stimata in corrispondenza dei valori osservati x_i , $i = 1, \dots, n$, ossia le stime delle medie di Y in ogni sottopopolazione definita dai valori osservati della X
- La funzione di regressione stimata fornisce una stima del valor atteso di Y al variare di X
- Residui di regressione

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 \cdot x_i \quad i = 1, \dots, n$$

Retta di regressione

La retta dei minimi quadrati di Y su X è quella che rende minima la somma dei quadrati delle distanze verticali (ossia nelle ordinate) tra il punto di coordinate (x_i, y_i) - valori osservati - e il punto di coordinate $(x_i, \hat{y}_i)$



Edu e QoL: Valori Teorici e residui di regressione

Unità	Edu	QoL	$\widehat{QoL}_i$	$(QoL_i - \widehat{QoL}_i)$
u_i	x_i	y_i	$\hat{y}_i$	$\hat{e}_i = (y_i - \hat{y}_i)$
1	13	74	73.70	0.30
2	5	66	64.50	1.50
3	8	62	67.95	-5.95
4	18	78	79.45	-1.45
5	10	77	70.25	6.75
6	21	84	82.90	1.10
7	15	80	76.00	4.00
8	25	85	87.50	-2.50
9	13	74	73.70	0.30
10	20	85	81.75	3.25
11	13	64	73.70	-9.70
12	16	83	77.15	5.85
13	18	76	79.45	-3.45

$$\begin{aligned}\widehat{QoL}_4 &= \hat{\beta}_0 + \hat{\beta}_1 \cdot Edu_4 \\ &= 58.75 + 1.15 \cdot 18 \\ &= 79.45\end{aligned}$$

$$\begin{aligned}\hat{e}_{11} &= QoL_{11} - \widehat{QoL}_{11} \\ &= 64 - 73.70 \\ &= 64 - (58.75 + 1.15 \cdot 13) \\ &= -9.7\end{aligned}$$

Alcune proprietà della retta di regressione

$$\begin{aligned}\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i & \text{e} & & \hat{e}_i &= y_i - \hat{y}_i \\ \hat{\beta}_0 &= (\bar{y} - \hat{\beta}_1 \cdot \bar{x}) & & & \hat{\beta}_1 &= \frac{s_{XY}}{s_X^2}\end{aligned}$$

- La retta di regressione stimata passa per il punto di coordinate $(\bar{x}, \bar{y})$

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot \bar{x} = (\bar{y} - \hat{\beta}_1 \cdot \bar{x}) + \hat{\beta}_1 \cdot \bar{x} = \bar{y}$$

- La somma (media) dei valori teorici è uguale alla somma (media) dei valori osservati

$$\sum_{i=1}^n \hat{y}_i = \sum_{i=1}^n y_i \iff \bar{\hat{y}} = \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y}$$

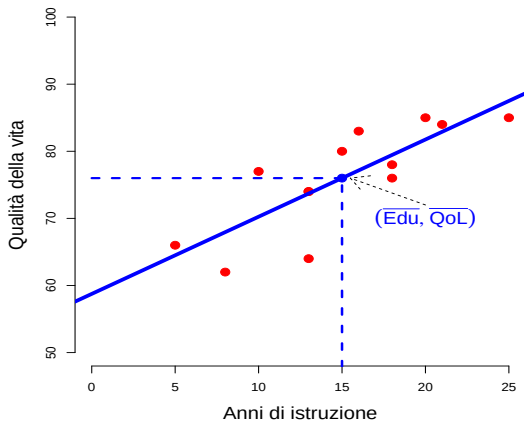
- La somma dei residui è uguale a zero

$$\hat{e}_1 + \dots + \hat{e}_n = \sum_{i=1}^n \hat{e}_i = 0$$

Esempio: Edu e QoL

$$\widehat{\text{QoL}} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot \text{Edu} = 58.75 - 1.15 \cdot \text{Edu}$$

$$\widehat{\beta}_0 + \widehat{\beta}_1 \cdot \overline{\text{Edu}} = 58.75 - 1.15 \cdot 15 = 76 = \overline{\text{QoL}}$$



Edu e QoL: Valori teorici e residui di regressione

Unità	Edu	QoL	$\widehat{QoL}_i$	$(QoL_i - \widehat{QoL}_i)$
u_i	x_i	y_i	$\hat{y}_i$	$\hat{e}_i = (y_i - \hat{y}_i)$
1	13	74	73.70	0.30
2	5	66	64.50	1.50
3	8	62	67.95	-5.95
4	18	78	79.45	-1.45
5	10	77	70.25	6.75
6	21	84	82.90	1.10
7	15	80	76.00	4.00
8	25	85	87.50	-2.50
9	13	74	73.70	0.30
10	20	85	81.75	3.25
11	13	64	73.70	-9.70
12	16	83	77.15	5.85
13	18	76	79.45	-3.45
Totale	195	988	988	0
Media	15	76	76	

Coefficiente di regressione e coefficiente di correlazione

Nella popolazione ...

$$\beta_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X}$$

- L'inclinazione (positiva o negativa) della retta di regressione è determinata dal coefficiente di correlazione lineare
- Il coefficiente di regressione β_1 e il coefficiente di correlazione ρ_{XY} hanno lo stesso segno
- Se $\rho_{XY} = 0$, allora $\beta_1 = 0 \iff$ Retta di regressione parallela all'asse delle ascisse $\iff$ Assenza di dipendenza lineare tra Y e X
- Se $\rho_{XY} = \pm 1$, allora $\beta_1 = \pm \frac{\sigma_Y}{\sigma_X}$
 - ✓ Punti perfettamente allineati lungo la retta di regressione
 - ✓ Ad ogni variazione di $1 \cdot \sigma_X$ in X corrisponde una variazione di $1 \cdot \sigma_Y$ in Y

Coefficiente di regressione e coefficiente di correlazione campionario

Nel campione ...

$$\widehat{\beta}_1 = \frac{s_{XY}}{s_X^2} = r_{XY} \frac{s_Y}{s_X}$$

Infatti

$$\widehat{\beta}_1 = \frac{s_{XY}}{s_X^2} = \frac{s_{XY}}{\sqrt{s_X^2} \sqrt{s_X^2}} \cdot \frac{\sqrt{s_Y^2}}{\sqrt{s_Y^2}} = \frac{s_{XY}}{\sqrt{s_Y^2} \sqrt{s_X^2}} \cdot \frac{\sqrt{s_Y^2}}{\sqrt{s_X^2}} = r_{XY} \frac{s_Y}{s_X}$$

ossia

$$\widehat{\beta}_1 = r_{XY} \frac{s_Y}{s_X} \qquad r_{XY} = \widehat{\beta}_1 \frac{s_X}{s_Y}$$

Coefficiente di regressione e coefficiente di correlazione campionario

- L'inclinazione (positiva o negativa) della retta di regressione (stimata) è determinata dal coefficiente di correlazione lineare campionario
- Il coefficiente di regressione $\widehat{\beta}_1$ e il coefficiente di correlazione r_{XY} hanno lo stesso segno
- Se $r_{XY} = 0$, allora $\widehat{\beta}_1 = 0 \iff$ Retta di regressione (stimata) parallela all'asse delle ascisse
- Se $r_{XY} = \pm 1$, allora $\widehat{\beta}_1 = \pm \frac{s_Y}{s_X}$
 - ✓ Punti perfettamente allineati lungo la retta di regressione
 - ✓ Ad ogni variazione di $1 \cdot s_X$ in X corrisponde una variazione di $1 \cdot s_Y$ in Y
- $s_X \cdot \widehat{\beta}_1 = r_{XY} \cdot s_Y$: Un incremento in x di s_X unità determina una variazione attesa in Y di r_{XY} deviazioni standard di Y
- $s_X = s_Y \implies \widehat{\beta}_1 = r_{XY}$

La correlazione implica la regressione verso la media

$$\beta_1 = \rho_{XY} \frac{\sigma_Y}{\sigma_X} \quad \text{quindi} \quad \beta_1 \sigma_X = \rho_{X,Y} \sigma_Y$$

- Un incremento in X di σ_X unità determina una variazione attesa in Y di $\sigma_X \beta_1$ unità
- Poiché $\beta_1 \sigma_X = \rho_{X,Y} \sigma_Y$ un incremento in X di σ_X unità corrisponde a una variazione attesa in Y di $\rho_{X,Y}$ deviazioni standard di Y
- Più grande è in valore assoluto il coefficiente di correlazione, più forte è l'associazione nel senso che una variazione di una deviazione standard in X corrisponde a una più grande proporzione di variazione di deviazioni standard in Y

La correlazione implica la regressione verso la media: Esempio

- Lo scienziato inglese Francis Galton (16/02/1822 - 17/01/1911) pone le basi del concetto di regressione e correlazione dopo il 1880
- Galton notò che la correlazione tra altezza dei genitori (media delle altezze del padre e della madre) e l'altezza dei figli è circa $r_{X,Y} = 0.5$
- Dalla relazione tra regressione e correlazione si ha che una variazione di una deviazione standard nell'altezza dei genitori (s_X) corrisponde a una variazione di circa la metà della deviazione standard nell'altezza dei figli ($r_{X,Y} \cdot s_Y = 0.5s_Y$)
 - ✓ Per genitori di altezza media, l'altezza media del figlio è pari alla media
 - ✓ Se l'altezza dei genitori è una deviazione standard sopra la media delle altezze dei genitori, l'altezza attesa del figlio è pari a mezza deviazione standard delle altezze dei figli sopra la media delle altezze dei figli
 - ✓ Se l'altezza dei genitori è due deviazioni standard sotto la media delle altezze dei genitori, l'altezza attesa del figlio è pari a una deviazione standard delle altezze dei figli sotto la media delle altezze dei figli



Esempio: Edu e QoL

Unità u_i	Edu x_i	QoL y_i	$(\text{Edu} - \overline{\text{Edu}})^2$ $(x_i - \bar{x})^2$	$(\text{QoL} - \overline{\text{QoL}})^2$ $(y_i - \bar{y})^2$	$(\text{Edu} - \overline{\text{Edu}}) \cdot (\text{QoL} - \overline{\text{QoL}})$ $(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	13	74	4	4	4
2	5	66	100	100	100
3	8	62	49	196	98
4	18	78	9	4	6
5	10	77	25	1	-5
6	21	84	36	64	48
7	15	80	0	16	0
8	25	85	100	81	90
9	13	74	4	4	4
10	20	85	25	81	45
11	13	64	4	144	24
12	16	83	1	49	7
13	18	76	9	0	0
Totale	195	988	366	744	421

Esempio: Edu e QoL

$$\widehat{QoL} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot Edu$$

$$s_{Edu, QoL} = 35.083 \quad s_{Edu}^2 = 30.5 \quad s_{Edu} = 5.523 \quad s_{QoL}^2 = 62 \quad s_{QoL} = 7.874$$

$$r_{Edu, QoL} = \frac{35.083}{5.523 \cdot 7.874} = 0.807$$



$$\widehat{\beta}_1 = \frac{s_{Edu, QoL}}{s_{Edu}^2} = \frac{35.083}{30.5} = 1.15$$



$$\widehat{\beta}_1 = r_{Edu, QoL} \frac{s_{QoL}}{s_{Edu}} = 0.807 \frac{7.874}{5.523} = 1.15$$

Un incremento di $s_{Edu} = 5.523$ nel numero di anni di istruzione determina un *incremento* atteso nella qualità della vita di $r_{Edu, QoL} = 0.807$ deviazioni standard della qualità della vita (ossia un incremento di $0.807 \cdot 7.874 = 6.354$ punti)

Scomposizione della somma dei quadrati

$$\begin{array}{ccccc} \text{Somma dei} & & \text{Somma dei} & & \text{Somma dei} \\ \text{Quadrati} & = & \text{Quadrati della} & + & \text{Quadrati degli} \\ \text{Totale} & & \text{Regressione} & & \text{Residui} \end{array}$$



$$\begin{array}{ccccc} \text{Total} & & \text{Regression} & & \text{Sum of} \\ \text{Sum of Squares} & = & \text{Sum of Squares} & + & \text{Squared Errors} \\ \text{(TSS)} & = & \text{(RSS)} & + & \text{(SSE)} \end{array}$$



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

Scomposizione della somma dei quadrati

- Somma dei Quadrati Totale $\iff$ Devianza Totale

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

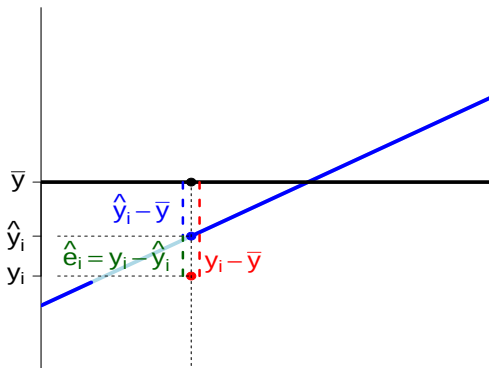
- Somma dei Quadrati di Regressione $\iff$ Devianza di Regressione (Devianza Spiegata): Devianza dei valori teorici

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

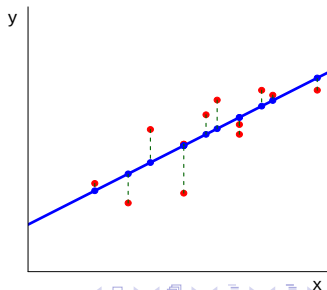
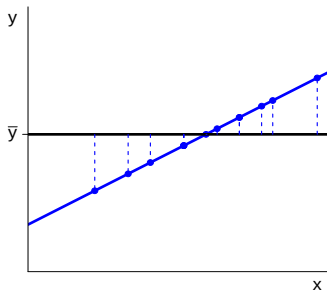
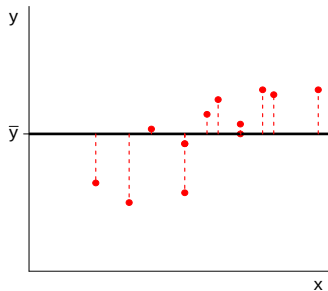
- Somma dei Quadrati dei Residui $\iff$ Devianza Residua

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (\hat{e}_i)^2$$

Scomposizione della somma dei quadrati



Scomposizione della somma dei quadrati



Interpretazione

$TSS = 0 \iff y_i = \bar{y}$ per ogni $i = 1, \dots, n$

SSE = Grado di incertezza residua

- $SSE = 0 \iff$ la relazione lineare prevede perfettamente le osservazioni
- Maggiore è la dispersione dei dati intorno alla retta maggiore è il valore di SSE

RSS = Parte della devianza totale (TSS) riprodotta dai valori teorici attraverso la retta di regressione

- $RSS = 0 \iff \hat{y}_i = \bar{y}$ per ogni $i = 1, \dots, n$
- $RSS = TSS \iff \hat{y}_i = y_i$ per ogni $i = 1, \dots, n$

Scomposizione della somma dei quadrati: Esempio Edu e QoL

Unità u_i	QoL y_i	$\widehat{QoL}_i$ $\hat{y}_i$	$(QoL_i - \overline{QoL}_i)$ $(y_i - \bar{y})^2$	$(\widehat{QoL}_i - \overline{QoL}_i)$ $(\hat{y}_i - \bar{y})^2$	$(QoL_i - \widehat{QoL}_i)^2$ $\hat{e}_i^2$
1	74	73.70	4	5.29	0.09
2	66	64.50	100	132.31	2.26
3	62	67.95	196	64.83	35.38
4	78	79.45	4	11.91	2.10
5	77	70.25	1	33.08	45.58
6	84	82.90	64	47.63	1.21
7	80	76.00	16	0.00	16.00
8	85	87.50	81	132.31	6.26
9	74	73.70	4	5.29	0.09
10	85	81.75	81	33.08	10.55
11	64	73.70	144	5.29	94.08
12	83	77.15	49	1.32	34.22
13	76	79.45	0	11.91	11.91
Totale	988	988	744	484.27	259.73

$$TSS = 744 = 484.27 + 259.73 = RSS + SSE$$

Indice di determinazione lineare

$$R^2 = \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{SSE}{TSS}$$

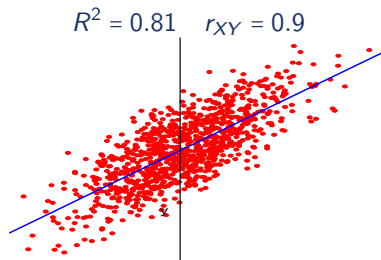
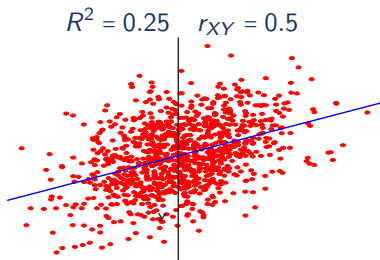
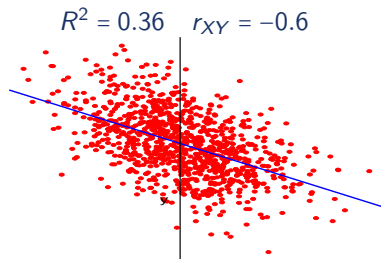
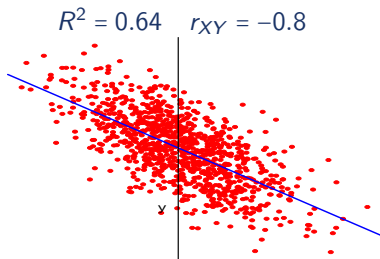
- Indica la proporzione di variabilità totale di Y spiegata dalla variabile esplicativa X attraverso il modello di regressione
- Riduzione proporzionale dell'errore che si ha impiegando la retta di regressione invece che semplicemente la media per prevedere Y
- R^2 non dipende dall'unità di misura delle variabili ed è un indice simmetrico
- $0 \leq R^2 \leq 1$: Misura della forza dell'associazione lineare
- $R^2 = 0$ ($RSS = 0$ e $SSE = TSS$): assenza di relazione statistica lineare

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \hat{\beta}_0 = \bar{y} \quad \text{per ogni } i$$

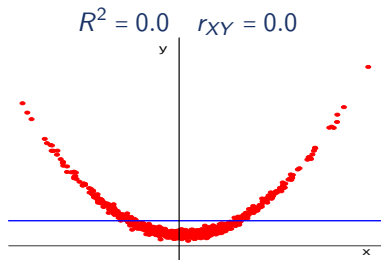
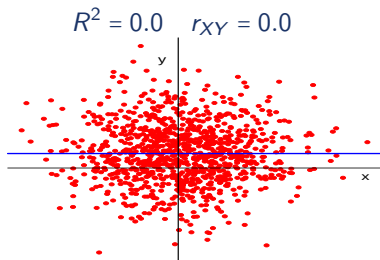
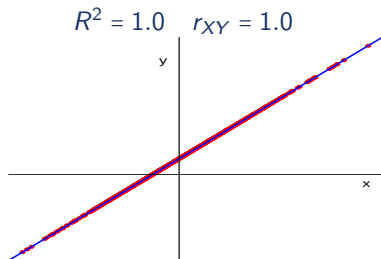
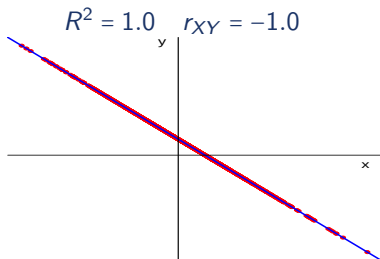
- $R^2 = 1$ ($RSS = TSS$ e $SSE = 0$): perfetta dipendenza lineare (tutti i punti campionari si trovano sulla retta di regressione)
- Si può dimostrare che

$$R^2 = r_{XY}^2 = \left(\frac{s_{XY}}{s_X s_Y} \right)^2$$

Indice di determinazione lineare e correlazione



Indice di determinazione lineare e correlazione



Indice di determinazione lineare tra Edu e QoL

$$\widehat{QoL} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot Edu = 58.75 + 1.15 \cdot Edu$$

$$TSS = 744 \quad RSS = 484.265 \quad SSE = 259.735 \quad r_{Edu, QoL} = 0.807$$

Quindi

$$R^2 = \frac{RSS}{TSS} = \frac{484.265}{744} = 0.651$$



$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{259.735}{744} = 1 - 0.349 = 0.651$$



$$R^2 = (r_{Edu, QoL})^2 = (0.807)^2 = 0.651$$

Il 65.1% della variabilità totale della qualità della vita è spiegata dagli anni di istruzione tramite il modello di regressione lineare

Interpolazione e estrapolazione

- Utilizzare la retta di regressione per calcolare il valore (medio) della variabile risposta Y per un dato valore della variabile esplicativa x non osservato, ma interno ai dati

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x^*$$

$$x^* \neq x_i \text{ per } i = 1, \dots, n \quad \text{e} \quad \min_{i=1, \dots, n} x_i < x^* < \max_{i=1, \dots, n} x_i$$

- Utilizzare la retta di regressione per prevedere il valore (medio) della variabile risposta Y per un dato valore della variabile esplicativa x esterno ai dati osservati

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x^* \quad x^* \notin \left[\min_{i=1, \dots, n} x_i; \max_{i=1, \dots, n} x_i \right]$$

L'accuratezza dell'estrapolazione dipende da:

- ✓ quanto la retta si adatta bene ai dati;
- ✓ quanto il valore di x^* è lontano dai valori osservati

Edu e QoL: Interpolazione e estrapolazione

- Retta dei minimi quadrati

$$\widehat{\text{QoL}} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot \text{Edu} = 58.75 + 1.15 \cdot \text{Edu}$$

- Range (osservato) della variabile Edu

$$\min_{i=1,\dots,13} \text{Edu}_i = 5 < x^* < \max_{i=1,\dots,13} \text{Edu}_i = 25$$

- Interpolazione: $\text{Edu}^* = 12$

$$\widehat{\text{QoL}}_{\text{Edu}=12} = 58.75 + 1.15 \cdot 12 = 72.55$$

- Estrapolazione

$$\begin{aligned} \widehat{\text{QoL}}_{\text{Edu}=1} &= 58.75 + 1.15 \cdot 1 = 59.90 \\ \widehat{\text{QoL}}_{\text{Edu}=4} &= 58.75 + 1.15 \cdot 4 = 63.35 \\ \widehat{\text{QoL}}_{\text{Edu}=26} &= 58.75 + 1.15 \cdot 26 = 88.65 \\ \widehat{\text{QoL}}_{\text{Edu}=35} &= 58.75 + 1.15 \cdot 35 = 99.00 \end{aligned}$$

Stimatori dei minimi quadrati

Nel modello di regressione lineare semplice le caratteristiche della popolazione che interessa studiare sono i coefficienti β_0 e β_1

- I valori della variabile esplicativa sono fissati
- Indichiamo con $\widehat{B}_0$ e $\widehat{B}_1$ gli stimatori dei minimi quadrati

$$\widehat{B}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\widehat{B}_0 = \bar{Y} - \widehat{B}_1 \bar{x}$$

- Il valori assunti da $\widehat{B}_0$ e $\widehat{B}_1$ in corrispondenza di un particolare campione sono detto stime dei minimi quadrati dei coefficienti di regressione β_0 e β_1 :
 $\widehat{\beta}_0$ e $\widehat{\beta}_1$ = stime dei minimi quadrati dei coefficienti di regressione β_0 e β_1
- Gli stimatori dei minimi quadrati hanno buone proprietà: Sono stimatori non distorti e hanno varianza minima nella classe degli stimatori corretti di un certo tipo

Varianze e errori standard degli stimatori dei minimi quadrati

- $\sigma^2 = \text{var}(\epsilon_i)$: Varianza degli errori di regressione
- Varianze

$$\text{Var}(\hat{B}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\text{Var}(\hat{B}_0) = \sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

- Errori standard

$$\sigma(\hat{B}_0) = \sqrt{\text{Var}(\hat{B}_0)} \quad \text{e} \quad \sigma(\hat{B}_1) = \sqrt{\text{Var}(\hat{B}_1)}$$

- Gli errori standard rappresentano una misura della precisione degli stimatori dei coefficienti di regressione

Stima della varianza degli errori

$$s^2 = \frac{SSE}{n-2} = \frac{\widehat{e}_1^2 + \dots + \widehat{e}_n^2}{n-2} = \frac{\sum_{i=1}^n \widehat{e}_i^2}{n-2}$$

- È uno stimatore non distorto della varianza degli errori di regressione σ^2
- Errore standard di regressione

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n \widehat{e}_i^2}{n-2}}$$

- s è una misura della variabilità degli scostamenti dei valori osservati dai valori teorici: variabilità campionaria intorno alla retta dei minimi quadrati
- Esempio: Stima della varianza degli errori: QoL e Edu

$$\widehat{QoL} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot Edu = 58.75 + 1.15 \cdot Edu \quad \text{e} \quad SSE = 259.735$$

Quindi

$$s^2 = \frac{SSE}{n-2} = \frac{259.735}{13-2} = 23.61 \quad \text{e} \quad s = \sqrt{s^2} = \sqrt{23.61} = 4.86$$

Stimatori delle varianze e degli errori standard degli stimatori dei minimi quadrati

- Stimatori delle varianze

$$\widehat{Var}(\widehat{B}_1) = \frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\widehat{Var}(\widehat{B}_0) = s^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

- Stimatori degli errori standard

$$\widehat{e.s.}(\widehat{B}_1) = \sqrt{\widehat{Var}(\widehat{B}_1)} = \sqrt{\frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\widehat{e.s.}(\widehat{B}_0) = \sqrt{\widehat{Var}(\widehat{B}_0)} = \sqrt{s^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

Stima delle varianze e degli errori standard degli stimatori dei minimi quadrati - Esempio: Edu e QoL

$$\overline{\text{Edu}} = 15 \quad \sum_{i=1}^n (\text{Edu}_i - \overline{\text{Edu}})^2 = 366 \quad \text{e} \quad s^2 = \frac{259.735}{11} = 23.61$$

Quindi

$$\widehat{\text{Var}}(\widehat{B}_1) = \frac{s^2}{\sum_{i=1}^n (\text{Edu}_i - \overline{\text{Edu}})^2} = \frac{23.61}{366} = 0.065$$

$$\widehat{\text{e.s.}}(\widehat{B}_1) = \sqrt{\widehat{\text{Var}}(\widehat{B}_1)} = \sqrt{0.065} = 0.254$$

$$\begin{aligned} \widehat{\text{Var}}(\widehat{B}_0) &= s^2 \cdot \left[\frac{1}{n} + \frac{\overline{\text{Edu}}^2}{\sum_{i=1}^n (\text{Edu}_i - \overline{\text{Edu}})^2} \right] = 23.61 \cdot \left[\frac{1}{13} + \frac{(15)^2}{366} \right] \\ &= 23.61 \cdot 0.692 = 16.332 \end{aligned}$$

$$\widehat{\text{e.s.}}(\widehat{B}_0) = \sqrt{\widehat{\text{Var}}(\widehat{B}_0)} = \sqrt{16.332} = 4.041$$

Normalità degli stimatori dei minimi quadrati

Sotto l'assunzione di normalità degli errori, gli stimatori dei minimi quadrati $\widehat{B}_0$ e $\widehat{B}_1$ hanno distribuzione Normale:

$$\begin{aligned}\widehat{B}_0 &\sim N\left(\beta_0, \sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]\right) \\ \widehat{B}_1 &\sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\end{aligned}$$

Quindi

$$\begin{aligned}\frac{\widehat{B}_0 - \beta_0}{e.s.(\widehat{B}_0)} &\sim N(0, 1) & e.s.(\widehat{B}_0) &= \sqrt{\sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]} \\ \frac{\widehat{B}_1 - \beta_1}{e.s.(\widehat{B}_1)} &\sim N(0, 1) & e.s.(\widehat{B}_1) &= \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\end{aligned}$$

Intervalli di confidenza per i coefficienti di regressione

- Si noti che ...

$$\frac{\widehat{B}_0 - \beta_0}{\widehat{e.s.}(\widehat{B}_0)} \sim t_{n-2} \quad \widehat{e.s.}(\widehat{B}_0) = \sqrt{s^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

$$\frac{\widehat{B}_1 - \beta_1}{\widehat{e.s.}(\widehat{B}_1)} \sim t_{n-2} \quad \widehat{e.s.}(\widehat{B}_1) = \sqrt{\frac{s^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

dove

$$s^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n \widehat{e}_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{n-2}$$

- Fissato il livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\beta_0) = \widehat{\beta}_0 \pm t_{n-2}(\alpha/2) \cdot \widehat{e.s.}(\widehat{B}_0)$$

$$IC_{1-\alpha}(\beta_1) = \widehat{\beta}_1 \pm t_{n-2}(\alpha/2) \cdot \widehat{e.s.}(\widehat{B}_1)$$

dove $t_{n-2}(\alpha/2)$ è tale che

$$Pr(-t_{\alpha/2, n-2} < t_{n-2} < t_{\alpha/2, n-2}) = 1 - \alpha$$

Intervalli di confidenza per i coefficienti di regressione

Esempio: Edu e QoL

Variabile	Stima	Errore Standard
Costante	58.75	4.041
QoL	1.15	0.254
Errore standard dei residui: $s = 4.859$ con 11 gradi di libertà		

Fissato $1 - \alpha = 0.95$, si ha $t_{11}(0.05/2) = t_{11}(0.025) = 2.20$. Quindi,

- Intervallo di confidenza per β_0

$$IC_{0.95}(\beta_0) = 58.75 \pm 2.2 \cdot 4.041 = 58.75 \pm 8.895 = (49.851; 67.641)$$

Possiamo essere fiduciosi al 95% che il coefficiente β_0 sia compreso tra 49.851 e 67.641

- Intervallo di confidenza per β_1

$$IC_{0.95}(\beta_1) = 1.15 \pm 2.2 \cdot 0.254 = 1.15 \pm 0.559 = (0.591; 1.709)$$

Possiamo essere fiduciosi al 95% che il coefficiente β_1 sia compreso tra 0.591 e 1.709: La qualità della vita media aumenta almeno 0.591 punti e al massimo di 1.709 punti per ogni anno in più di istruzione

Verifica di ipotesi per β_1

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

- Statistica Test

$$T = \frac{\widehat{B}_1 - 0}{\widehat{e.s.}(\widehat{B}_1)} \sim t_{n-2} \text{ Sotto } H_0$$

- Regione di rifiuto al livello di significatività α

$$RC_\alpha : T \leq -t_{(n-2),\alpha/2} \text{ o } T \geq t_{(n-2),\alpha/2}$$

- Utilizzo del p -valore $p\text{-value} = 2 \cdot [1 - Pr(t_{n-2} \leq |t_{oss}|; H_0)]$
- Se NON è rifiutata l'ipotesi nulla $H_0 : \beta_1 = 0$, allora si conclude che il coefficiente di regressione β_1 non è statisticamente diverso da zero al livello di significatività α : i dati suggeriscono che non esiste una relazione lineare significativa tra le due variabili
- Sotto le ipotesi del modello di regressione, se H_0 è vera allora le variabili sono statisticamente indipendenti

Verifica di ipotesi per β_1 : Esempio Edu e QoL

Variabile	Stima	Errore Standard
Costante	58.75	4.041
QoL	1.15	0.254
Errore standard dei residui: $s = 4.859$ con 11 gradi di libertà		

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

Fissato $\alpha = 0.05$, si ha $t_{11}(0.05/2) = t_{11}(0.025) = 2.20$.

- Regione critica: $RC_{0.05} = T \leq -2.2$ o $T \geq 2.2$
- Valore osservato della statistica test

$$t_{oss} = \frac{1.15}{0.254} = 4.529$$

$$t_{oss} = 4.529 > 2.2 \Leftrightarrow t_{oss} \in RC_{0.05} \Rightarrow \text{si rifiuta } H_0 : \beta_1 = 0$$

- P -valore

$$p\text{-value} = 2 \cdot [1 - Pr(t_{11} \leq |4.529|; H_0)] = 0.00086$$

Verifica di ipotesi per β_1 versus ρ

$$H_0: \beta_1 = 0 \quad \text{vs} \quad H_a: \beta_1 \neq 0 \quad \Longleftrightarrow \quad H_0: \rho_{XY} = 0 \quad \text{vs} \quad H_a: \rho_{XY} \neq 0$$

- Statistica Test

$$T = \frac{\widehat{\beta}_1}{\widehat{e.s.}(\widehat{\beta}_1)} \quad \Longleftrightarrow \quad T = \frac{r_{XY}}{\sqrt{(1 - r_{XY}^2)/(n - 2)}}$$

Entrambe le statistiche test hanno una distribuzione sotto l'ipotesi nulla t_{n-2}

- Le regioni critiche coincidono
- Il valore osservato della statistica test coincide

$$\frac{\widehat{\beta}_1}{\widehat{e.s.}(\widehat{\beta}_1)} = \frac{r_{XY}}{\sqrt{(1 - r_{XY}^2)/(n - 2)}}$$

- Il p -valore coincide

Verifica di ipotesi per β_1 versus ρ : Esempio Edu e QoL

- $H_0 : \rho_{X,Y} = 0$ versus $H_a : \rho_{X,Y} \neq 0$
- Statistica test:

$$T = \frac{r_{X,Y}}{(1 - r_{X,Y}^2)/(n-2)} \sim t_{n-2} \text{ sotto } H_0$$

- $n = 13$ quindi per $\alpha = 0.05$ si ha $t_{0.025,13-2} = t_{0.025,11} = 2.2$
- Regione di rifiuto: $RC_{0.05} = T \leq -2.2$ oppure $T \geq 2.2$
- Coefficiente di correlazione campionario: $r = 0.807$
- Valore osservato della statistica test

$$t_{oss} = \frac{0.807}{\sqrt{(1 - (0.807)^2)/(13-2)}} = \frac{0.807}{\sqrt{(1 - 0.651)/11}} = \frac{0.807}{0.178} = 4.529$$

- Decisione: $t_{oss} = 4.529 > 2.2 \iff t_{oss} = 4.529 \in RC_{0.05} \implies$ I dati mostrano evidenza contro H_0 al livello di significatività del 5%.
- $p\text{-value} = 0.00086$
- Confrontare con i risultati sulla verifica di ipotesi per β_1

Verifica di ipotesi per β_0

$$H_0 : \beta_0 = 0 \quad \text{versus} \quad H_a : \beta_0 \neq 0$$

- Statistica Test

$$T = \frac{\widehat{B}_0}{\widehat{e.s.}(\widehat{B}_0)} \sim t_{n-2} \text{ Sotto } H_0$$

- Regione di rifiuto al livello di significatività α

$$RC_\alpha T \leq -t_{(n-2), \alpha/2} \text{ o } T \geq t_{(n-2), \alpha/2}$$

- Utilizzo del $p - value = 2 \cdot [1 - Pr(t_{n-2} \leq |t_{oss}|; H_0)]$
- Se NON è rifiutata l'ipotesi nulla $H_0 : \beta_0 = 0$, allora si conclude che
 - ✓ il coefficiente di regressione β_0 non è statisticamente diverso da zero al livello di significatività α
 - ✓ i dati suggeriscono che la relazione tra le variabili può essere descritta da una retta passante per l'origine (relazione di proporzionalità tra variabile risposta e variabile esplicativa): In questo caso lo stimatore dei minimi quadrati ha un'espressione diversa

Verifica di ipotesi per β_0 : Esempio Edu e QoL

Variabile	Stima	Errore Standard
Costante	58.75	4.041
QoL	1.15	0.254
Errore standard dei residui: $s = 4.859$ con 11 gradi di libertà		

$$H_0 : \beta_0 = 0 \quad \text{versus} \quad H_a : \beta_0 \neq 0$$

Fissato $\alpha = 0.05$, si ha $t_{11}(0.05/2) = t_{11}(0.025) = 2.2$.

- Regione critica: $RC_{0.05} = T \leq -2.2$ o $T \geq 2.2$
- Valore osservato della statistica test

$$t_{oss} = \frac{58.75}{4.041} = 14.536$$

$$t_{oss} = 14.536 > 2.2 \Leftrightarrow t_{oss} \in RC_{0.05} \Rightarrow \text{si rifiuta } H_0 : \beta_0 = 0$$

- Livello di significatività osservato (p -value)

$$p - \text{value} = 2 \cdot [1 - Pr(t_{11} < |14.536|; H_0)] = 0.000000015$$

Intervalli di confidenza e verifica di ipotesi

- Si considerino le ipotesi

$$H_0 : \beta_0 = 0 \quad \text{versus} \quad H_a : \beta_0 \neq 0 \quad \text{e} \quad H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

Fissato α , il livello di significatività del test, per entrambi i test la regione critica è

$$RC_\alpha = t \leq -t_{n-2}(\alpha/2) \quad \text{oppure} \quad t \geq t_{n-2}(\alpha/2)$$

Calcolati i valori osservati delle statistiche test

$$t_{\text{oss}}^{\beta_0} = \frac{\widehat{\beta}_0}{\widehat{\text{e.s.}}(\widehat{B}_0)} \quad \text{e} \quad t_{\text{oss}}^{\beta_1} = \frac{\widehat{\beta}_1}{\widehat{\text{e.s.}}(\widehat{B}_1)}$$

si rifiuta $H_0 : \beta_0 = 0$ se $t_{\text{oss}}^{\beta_0} \in RC_\alpha$ e si rifiuta $H_0 : \beta_1 = 0$ se $t_{\text{oss}}^{\beta_1} \in RC_\alpha$

- Fissato il livello di confidenza $1 - \alpha$

$$\widehat{\beta}_0 \pm t_{n-2}(\alpha/2) \cdot \widehat{\text{e.s.}}(\widehat{B}_0) \quad \text{e} \quad \widehat{\beta}_1 \pm t_{n-2}(\alpha/2) \cdot \widehat{\text{e.s.}}(\widehat{B}_1)$$

Intervalli di confidenza e Verifica di ipotesi

L'intervallo di confidenza include lo zero?

- No $\Rightarrow$ il parametro è significativamente diverso da zero: il valore osservato della statistica test appartiene alla regione critica
- Si $\Rightarrow$ il parametro NON è significativamente diverso da zero: il valore osservato della statistica test NON appartiene alla regione critica

Esempio: Edu e QoL

Variabile	Stima	Errore Standard
Costante	58.75	4.041
QoL	1.15	0.254
Errore standard dei residui: $s = 4.859$ con 11 gradi di libertà		

- Verifica di ipotesi

$$H_0 : \beta_0 = 0 \text{ vs } H_a : \beta_0 \neq 0 \quad \text{e} \quad H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

$$RC_{0.05} = t \leq -2.2 \quad \text{oppure} \quad t \geq 2.2$$

$$t_{oss}^{\beta_0} = 14.536 > 2.2 \Leftrightarrow t_{oss}^{\beta_0} \in RC_{0.05} : \text{ si rifiuta } H_0 : \beta_0 = 0$$

$$t_{oss}^{\beta_1} = 4.529 > 2.2 \Leftrightarrow t_{oss}^{\beta_1} \in RC_{0.05} : \text{ si rifiuta } H_0 : \beta_1 = 0$$

- Intervalli di confidenza al livello di confidenza del 95%

$$IC_{0.95}(\beta_0) = 58.75 \pm 4.041 \cdot 2.2 = (49.851; 67.641)$$

$$IC_{0.95}(\beta_1) = 1.15 \pm 0.254 \cdot 2.2 = (0.591; 1.709)$$

Nessuno dei due intervalli include lo zero

Stimatore della risposta media: $\mathbb{E}[Y_i | X_i = x^*]$

$$\widehat{Y}_{x^*} = \widehat{B}_0 + \widehat{B}_1 x^*$$

- Correttezza $\mathbb{E}(\widehat{Y}_{x^*}) = \beta_0 + \beta_1 x^*$
- Varianza & Errore Standard

$$\text{Var}(\widehat{Y}_{x^*}) = \sigma^2 \cdot \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad \text{e.s.}(\widehat{Y}_{x^*}) = \sqrt{\text{Var}(\widehat{Y}_{x^*})}$$

- Stimatore della varianza e dell'errore standard dello stimatore di $\widehat{Y}_{x^*}$

$$\widehat{\text{Var}}(\widehat{Y}_{x^*}) = s^2 \cdot \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad \widehat{\text{e.s.}}(\widehat{Y}_{x^*}) = \sqrt{\widehat{\text{Var}}(\widehat{Y}_{x^*})}$$

- Intervallo di confidenza al livello di confidenza del $(1 - \alpha)$

$$\begin{aligned} IC_{1-\alpha}(\mathbb{E}[Y_i | X_i = x^*]) &= \widehat{Y}_{x^*} \pm t_{n-2, \alpha/2} \widehat{\text{e.s.}}(\widehat{Y}_{x^*}) \\ &= \widehat{Y}_{x^*} \pm t_{n-2, \alpha/2} \sqrt{s^2 \cdot \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \end{aligned}$$

Esempio: QoL medio

$$\widehat{\text{QoL}}_i = 58.75 + 1.15 \cdot \text{Edu}_i \quad \overline{\text{Edu}} = 15 \quad \text{e} \quad s^2 = 23.61$$

- $\text{Edu}^* = 20$ (Nei dati $\text{Edu}_{10} = 20$)
- Stima della risposta media: $\widehat{\text{QoL}}_{\text{Edu}^*} = 58.75 + 1.15 \cdot 20 = 81.75$
- Errore standard

$$\begin{aligned} \widehat{\text{e.s.}}(\widehat{\text{QoL}}_{\text{Edu}^*}) &= \sqrt{s^2 \cdot \left[\frac{1}{n} + \frac{(\text{Edu}^* - \overline{\text{Edu}})^2}{\sum_{i=1}^n (\text{Edu}_i - \overline{\text{Edu}})^2} \right]} \\ &= \sqrt{23.61 \cdot \left[\frac{1}{13} + \frac{(20 - 15)^2}{366} \right]} = \sqrt{3.43} = 1.85 \end{aligned}$$

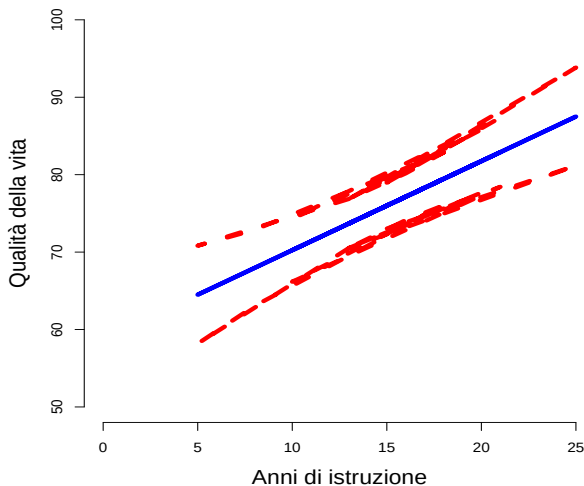
- Intervallo di confidenza al livello di confidenza del $(1 - \alpha) = 0.95$

$$IC_{0.95}(\mathbb{E}[\text{QoL} \mid \text{Edu}^*]) = 81.75 \pm 2.2 \cdot 1.85 = (77.68; 85.83)$$

Edu e QoL

Unità	Edu _i	QoL _i	$\widehat{QoL}_i$	$\widehat{e.s.}(\widehat{QoL}_i)$	$IC_{0.95}(\mathbb{E}[QoL_i Edu_i])$	
u_i	x_i	y_i	$\hat{y}_i$	$\hat{e.s.}(\hat{y}_i)$	$IC_{0.95}(\mathbb{E}[Y_i X_i = x_i])$	
1	13	74	73.70	1.44	70.53	76.87
2	5	66	64.50	2.88	58.17	70.83
3	8	62	67.95	2.23	63.04	72.86
4	18	78	79.45	1.55	76.04	82.86
5	10	77	70.25	1.85	66.17	74.32
6	21	84	82.90	2.03	78.42	87.38
7	15	80	76.00	1.35	73.03	78.97
8	25	85	87.50	2.88	81.17	93.83
9	13	74	73.70	1.44	70.53	76.87
10	20	85	81.75	1.85	77.68	85.83
11	13	64	73.70	1.44	70.53	76.87
12	16	83	77.15	1.37	74.13	80.17
13	18	76	79.45	1.55	76.04	82.86

Bande di confidenza (livello di confidenza $1 - \alpha = 0.95$)



Previsione

Prevedere il valore della variabile risposta per una nuova unità statistica di cui è noto il valore della variabile esplicativa (x^*):

- Previsione

$$\widehat{Y}_{x^*}^P = \widehat{B}_0 + \widehat{B}_1 x^*$$

- Varianza e Errore Standard

$$Var(\widehat{Y}_{x^*}^P) = \sigma^2 \cdot \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$e.s.(\widehat{Y}_{x^*}^P) = \sqrt{Var(\widehat{Y}_{x^*}^P)}$$

- Stimatore della varianza e dell'errore standard dello stimatore di $\widehat{Y}_{x^*}^P$

$$\widehat{Var}(\widehat{Y}_{x^*}^P) = s^2 \cdot \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\widehat{e.s.}(\widehat{Y}_{x^*}^P) = \sqrt{\widehat{Var}(\widehat{Y}_{x^*}^P)}$$

Intervallo di previsione

- Intervallo di previsione al livello di confidenza del $(1 - \alpha)$

$$\begin{aligned}
 IC_{1-\alpha}(Y_{x^*}^P) &= \widehat{Y}_{x^*}^P \pm t_{n-2, \alpha/2} \widehat{e.s.}(\widehat{Y}_{x^*}^P) \\
 &= \widehat{Y}_{x^*}^P \pm t_{n-2, \alpha/2} \sqrt{s^2 \cdot \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}
 \end{aligned}$$

Esempio: Edu e QoL

$$\widehat{\text{QoL}}_i = 58.75 + 1.15 \cdot \text{Edu}_i \quad \overline{\text{Edu}} = 15 \quad \text{e} \quad s^2 = 23.61$$

- $\text{Edu}^* = 20$ (Soggetto che ha un numero di anni di istruzione uguale al soggetto u_{10})
- Previsione: $\widehat{\text{QoL}}_{\text{Edu}^*}^P = 58.75 + 1.15 \cdot 20 = 81.75$
- Errore standard

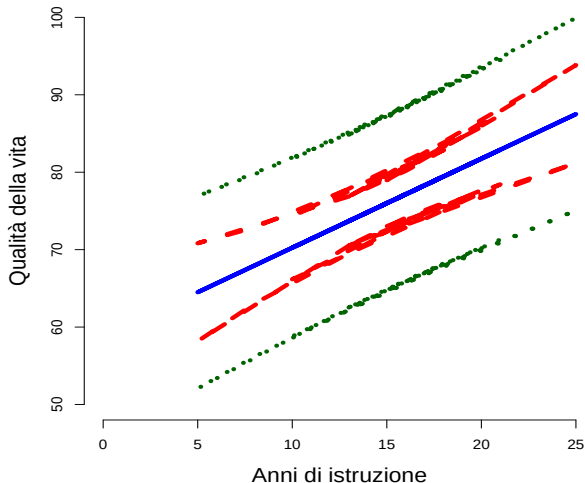
$$\begin{aligned} \widehat{e.s.}(\widehat{\text{QoL}}_{\text{Edu}^*}^P) &= \sqrt{s^2 \cdot \left[1 + \frac{1}{n} + \frac{(\text{Edu}^* - \overline{\text{Edu}})^2}{\sum_{i=1}^n (\text{Edu}_i - \overline{\text{Edu}})^2} \right]} \\ &= \sqrt{23.61 \cdot \left[1 + \frac{1}{13} + \frac{(20 - 15)^2}{366} \right]} = \sqrt{27.04} = 5.20 \end{aligned}$$

- Intervallo di previsione al livello di confidenza del $(1 - \alpha) = 0.95$

$$IC_{0.95}(\text{QoL}_{\text{Edu}^*}^P) = 81.75 \pm 2.2 \cdot 5.2 = (70.31; 93.20)$$

Edu e QoL

Unità	Edu _i	QoL _i	$\widehat{QoL}_i^P$	$\widehat{e.s.}(\widehat{QoL}_i^P)$	$IP_{0.95}(QoL_{Edu_i}^P)$	
u_i	x_i	y_i	$\hat{y}_i^P$	$\hat{e.s.}(\hat{y}_i^P)$	$IP_{0.95}(Y_{x_i}^P)$	
1	13	74	73.70	1.44	62.54	84.85
2	5	66	64.50	2.88	52.07	76.92
3	8	62	67.95	2.23	56.18	79.72
4	18	78	79.45	1.55	68.23	90.68
5	10	77	70.25	1.85	58.80	81.69
6	21	84	82.90	2.03	71.31	94.50
7	15	80	76.00	1.35	64.90	87.10
8	25	85	87.50	2.88	75.08	99.93
9	13	74	73.70	1.44	62.54	84.85
10	20	85	81.75	1.85	70.31	93.20
11	13	64	73.70	1.44	62.54	84.85
12	16	83	77.15	1.37	66.04	88.26
13	18	76	79.45	1.55	68.23	90.68

Bande di confidenza e previsione (livello di confidenza $1 - \alpha = 0.95$)

Esempio: Intervallo di confidenza e di previsione

$$\widehat{\text{QoL}}_i = 58.75 + 1.15 \cdot \text{Edu}_i \quad \overline{\text{Edu}} = 15 \quad \text{e} \quad s^2 = 23.61$$

- $\text{Edu}^* = 22$
- Stima della risposta media/Previsione:

$$\widehat{\text{QoL}}_{\text{Edu}^*} = \widehat{\text{QoL}}_{\text{Edu}^*}^P = 58.75 + 1.15 \cdot 22 = 84.05$$

- Errore standard (della stimatore della risposta media)

$$\begin{aligned} \widehat{\text{e.s.}}(\widehat{\text{QoL}}_{\text{Edu}^*}) &= \sqrt{s^2 \cdot \left[\frac{1}{13} + \frac{(\text{Edu}^* - \overline{\text{Edu}})^2}{\sum_{i=1}^n (\text{Edu}_i - \overline{\text{Edu}})^2} \right]} \\ &= \sqrt{23.61 \cdot \left[\frac{1}{13} + \frac{(22 - 15)^2}{366} \right]} = \sqrt{4.98} = 2.23 \end{aligned}$$

- Intervallo di confidenza al livello di confidenza del $(1 - \alpha) = 0.95$

$$IC_{0.95}(\mathbb{E}[\text{QoL} \mid \text{Edu}^*]) = 84.05 \pm 2.2 \cdot 2.23 = (79.14; 88.96)$$

Esempio: Intervallo di confidenza e di previsione

- Errore standard (della previsione)

$$\begin{aligned}\widehat{e.s.}(\widehat{QoL}_{Edu^*}^P) &= \\ &= \sqrt{s^2 \cdot \left[1 + \frac{1}{13} + \frac{(Edu^* - \overline{Edu})^2}{\sum_{i=1}^n (Edu_i - \overline{Edu})^2} \right]} \\ &= \sqrt{23.61 \cdot \left[1 + \frac{1}{13} + \frac{(22 - 15)^2}{366} \right]} = \sqrt{28.59} = 5.35\end{aligned}$$

- Intervallo di confidenza al livello di confidenza del $(1 - \alpha) = 0.95$

$$IP_{0.95}(QoL_{Edu^*}^P) = 84.05 \pm 2.2 \cdot 5.35 = (72.28; 95.82)$$