BIOSTATISTICA

6. Modelli statistici: analisi della regressione lineare

Marta Blangiardo, Imperial College, London
Department of Epidemiology and Public Health
m.blangiardo@imperial.ac.uk

MODELLO STATISTICO

Rappresentazione seplificata, analogica e necessaria della realtà

Semplificazione della realtà: il modello di un bacino idrologico, di un aereoplano, del flusso finanziario di un Paese ottenuti riproducendo gli aspetti “essenziali” e eliminando quelli ritenuti “superficiali”.

Analogia della realtà: il modello è una riproduzione della realtà

Rappresentazione necessaria della realtà: anche se è semplificato il modello è necessario per capire la realtà tramite lo studio di relazioni semplici e di maggiore intellegibilità

La specificazione di un modello consiste nell'esplicitare un legame tra i fenomeni di interesse:

$$Y = f(X_1, X_2,\ldots,X_p)$$

Dove Y è la variabile da spiegare, mentre $X_1, X_2,\ldots,X_p$ sono le variabili scelte per spiegare Y tramite la funzione f(.)

Inoltre non è quasi mai plausibile ipotizzare un legame deterministico quindi dobbiamo aggiungere un errore:

$$Y = f(X_1, X_2,\ldots,X_p) + \varepsilon$$

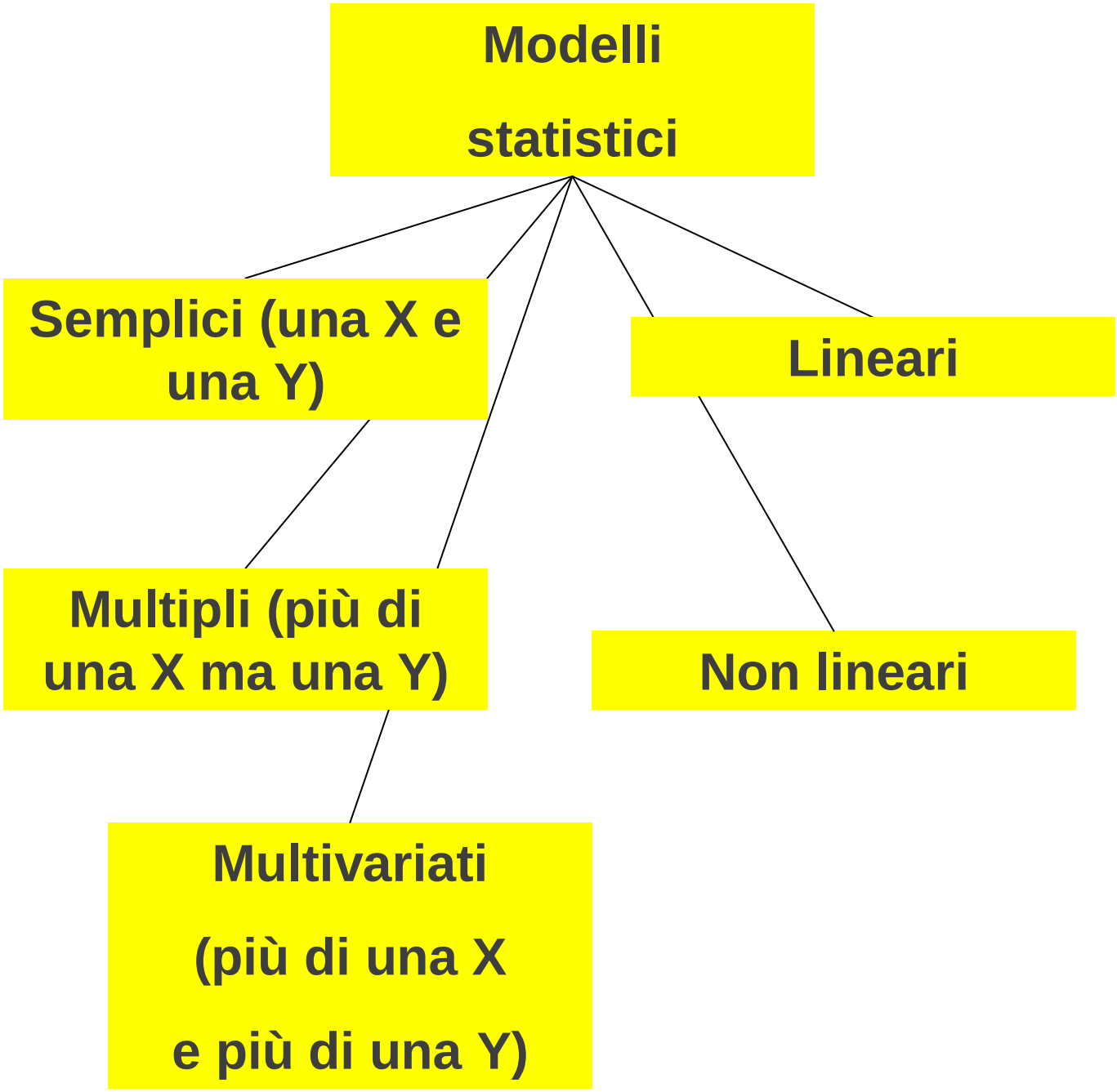
dove ε è una variabile casuale e riassume la nostra ignoranza circa la vera relazione tra Y e X. Per questo motivo la chiameremo *variabile errore*.

In alcuni contesti la specificazione della relazione funzionale risulta immediata dalla natura del problema:

1) Se Y è il peso ed X è l'altezza di una persona adulta la prima relazione da specificare è quella proporzionale (maggiore il peso, maggiore l'altezza e viceversa) $Y=\beta X+ \varepsilon$

2) Se Y è il peso di una mattonella rettangolare per la quale X_1 e X_2 sono rispettivamente la lunghezza e la larghezza, allora una relazione funzionale può essere specificata mediante $Y=\beta X_1X_2+ \varepsilon$

Entrambe le specificazioni evidenziano un parametro β che deve essere determinato per poter utilizzare il modello specificato


Modelli statistici
Semplici (una X e una Y)
Lineari
Multipli (più di una X ma una Y)
Non lineari
Multivariati (più di una X e più di una Y)

Terminologia

$Y = f(X_1, X_2,\ldots,X_p)$

Y: variabile dipendente

$X_1,\ldots,X_p$: variabili esplicative

ε: variabile casuale errore

NOTA: il legame statistico implicato dal modello non è simmetrico. Sono le variabili esplicative a "determinare" la variabile dipendente e NON viceversa.

X: precipitazione giornaliera di un bacino idrografico

Y: livello del fiume che si origina dal bacino

Relazione: X ⇨ Y ma NON Y ⇨ X

X: dose di concime somministrato in un campo di grano

Y: resa di grano in quel terreno

Relazione: X ⇨ Y ma NON Y ⇨ X

Modello di regressione lineare

Il termine REGRESSIONE deriva dall'applicazione svolta dal biologo Galton che nel 1886 esaminò altezze dei figli (Y) in funzione delle altezze dei genitori (X) in Inghilterra e notò una relazione funzionale tra le due variabili: più alti i genitori, più alti i figli e viceversa.

Tuttavia ai genitori che si collocavano agli estremi (molto bassi o molto alti) non corrispondevano figli altrettanto estremi, ovvero Galton osservò che l'altezza dei figli si spostava verso la media e quindi concluse che questo costituiva una regression towards mediocrity e la relazione funzionale fu chiamata "modello di regressione".

Oggi il termine regressione è divenuto significato di "relazione funzionale tra variabili ottenuta con metodi statistici" e la frase "regredire Y su $(X_1,\ldots,X_p)$" significa ricercare una relazione statistica del tipo:

$$Y = f(X_1, X_2,\ldots,X_p) + \varepsilon$$

Il modello di regressione semplice è specificato dalla relazione:

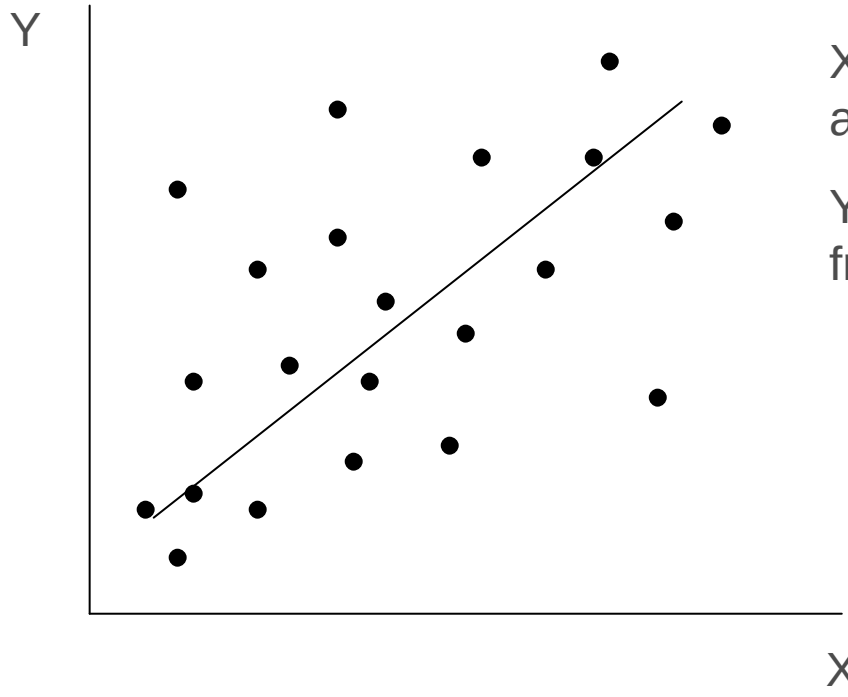
$$y_i = f(x_i;\beta) + \varepsilon_i$$

La funzione $f(x_i;\beta)$ può essere di primo grado, ad esempio:

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Oppure di grado superiore al primo, ad esempio di secondo grado:

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + \varepsilon_i$$



X = velocità di un autoveicolo

Y = spazio di frenata

Modello di regressione lineare semplice

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

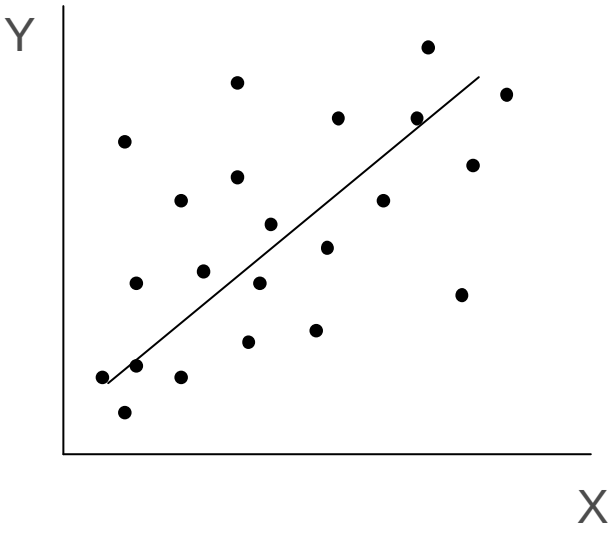
Identifica una retta, nota come la retta di regressione:

β_0: intercetta, il valore di Y_i quando $x_i=0$

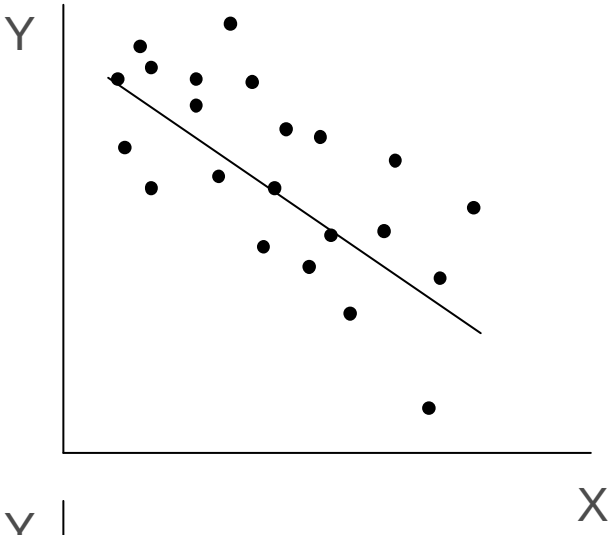
β_1: pendenza, di quanto cambia Y_i quando x_i incrementa di un'unità

ε_i: l'errore che si commette nella spiegazione dellavariabile y_i tramite una funzione lineare di x_i

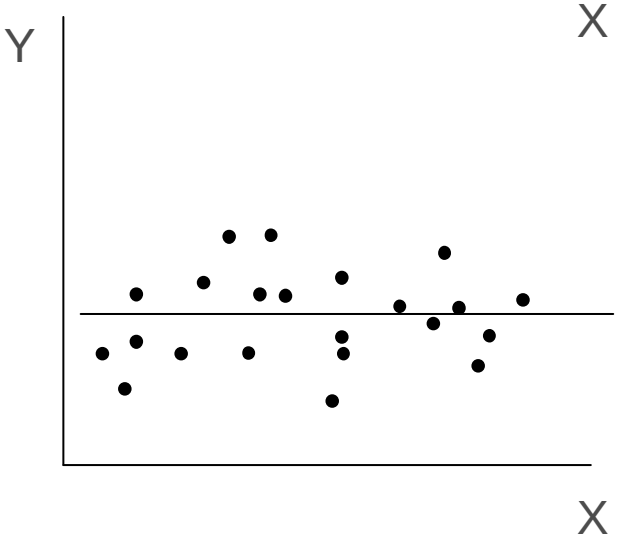
Che relazione c'è tra X e Y?



Covariano positivamente



Covariano negativamente



Non covariano

La covarianza misura l'attitudine a covariare di due caratteri

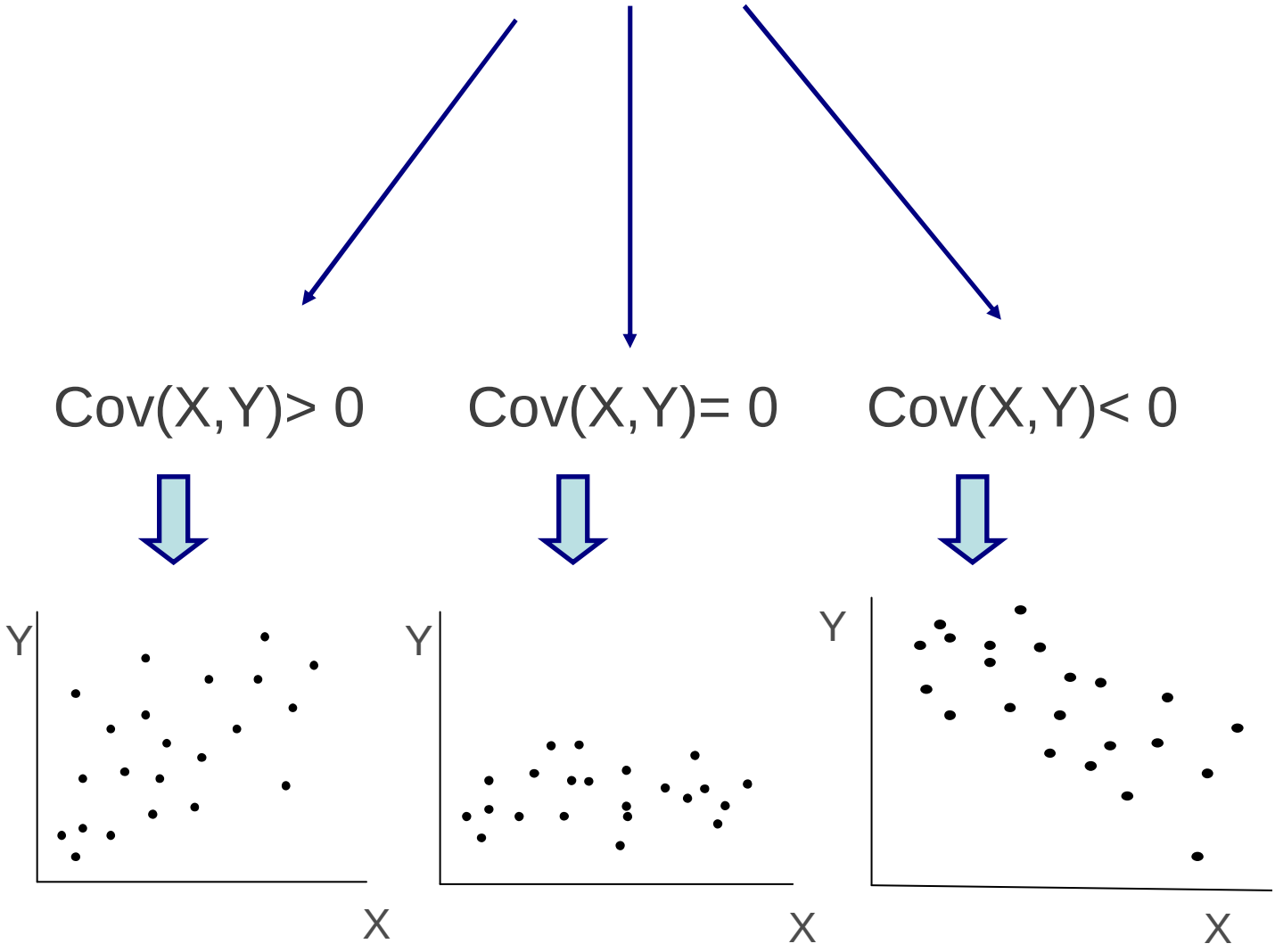
X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
10	14	-5	-4	20
15	17	0	-1	0
20	19	5	1	5
14	16	-1	-2	2
12	15	-3	-3	9
16	21	1	3	3
18	24	3	6	18

$\bar{x} = 15$ $\bar{y} = 18$

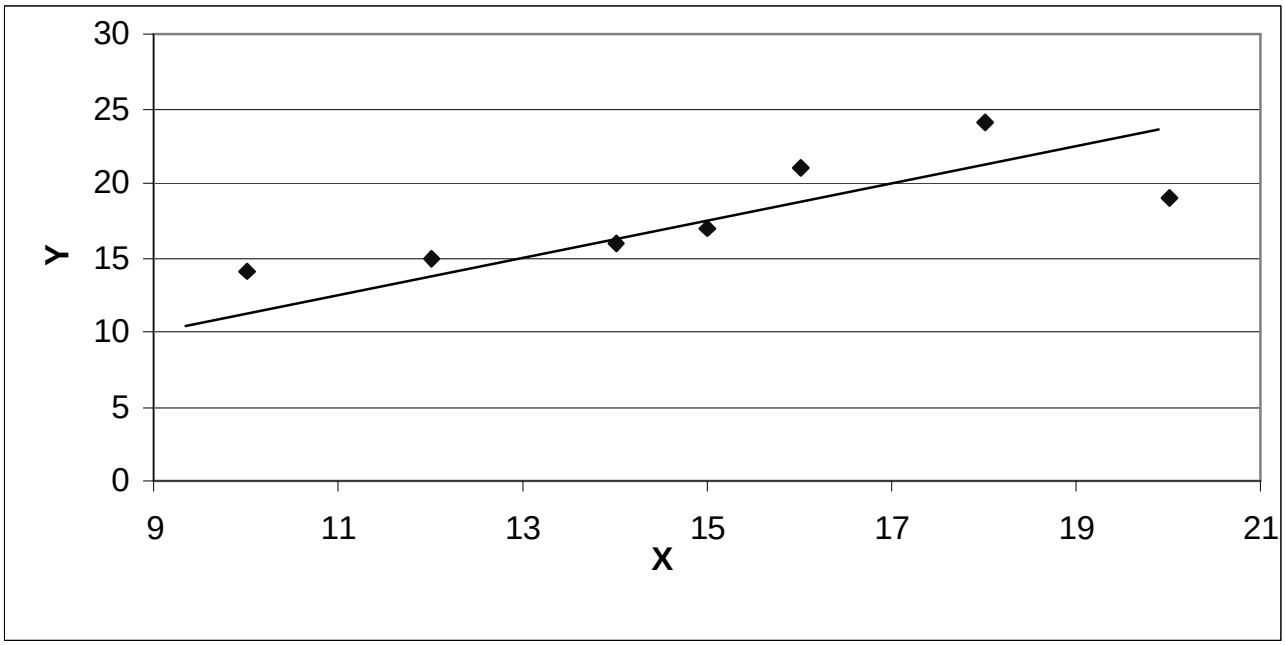
$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{n-1}$$

$$Cov(X,Y) = \frac{20+0+5+2+9+3+18}{7-1} = 9.5$$

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})}{n-1}$$



X	Y
10	14
15	17
20	19
14	16
12	15
16	21
18	24



$$Cov(X,Y)=9.5 > 0$$

E’ utile costruire una misura STANDARDIZZATA che esprima quanto I due caratteri covariano

COEFFICIENTE DI CORRELAZIONE

COVARIANZA

$$\rho = \frac{Cov(X,Y)}{sd(X) \cdot sd(Y)}$$

Deviazione standard

Deviazione standard

-1 0 1

Cov(X,Y)= - sd(X) · sd(Y)

PERFETTA

Correlazione NEGATIVA

Cov(X,Y)=0

NON c’è

correlazione

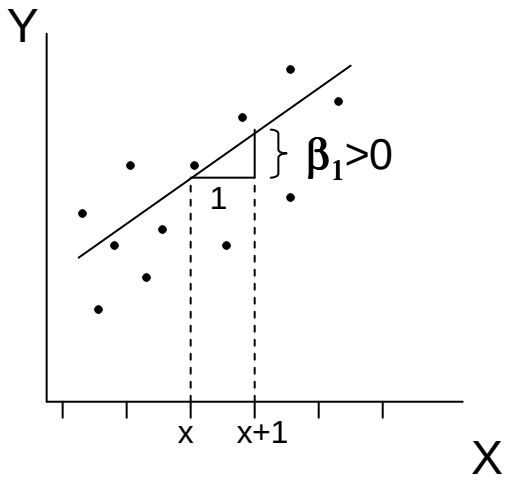
Cov(X,Y)= sd(X) · sd(Y)

PERFETTA

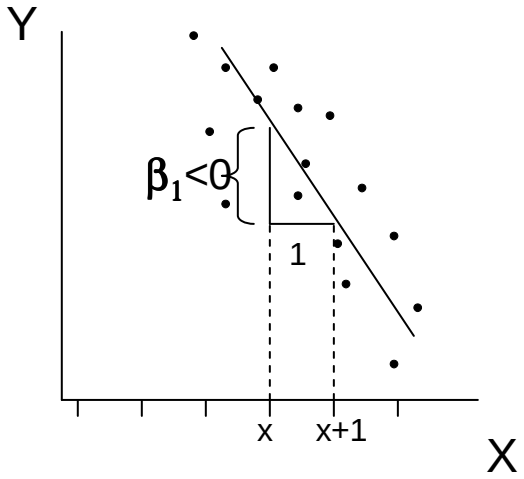
Correlazione POSITIVA

Modello di regressione lineare semplice

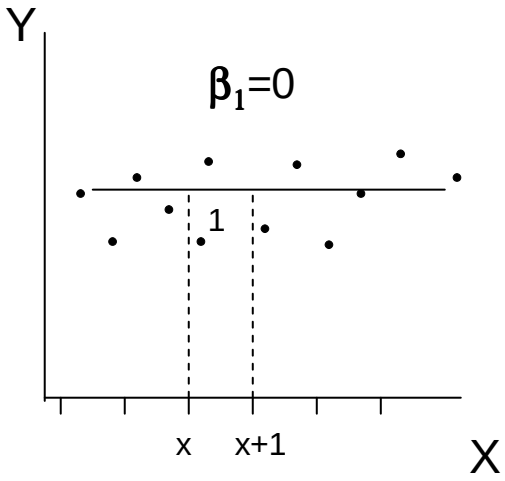
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



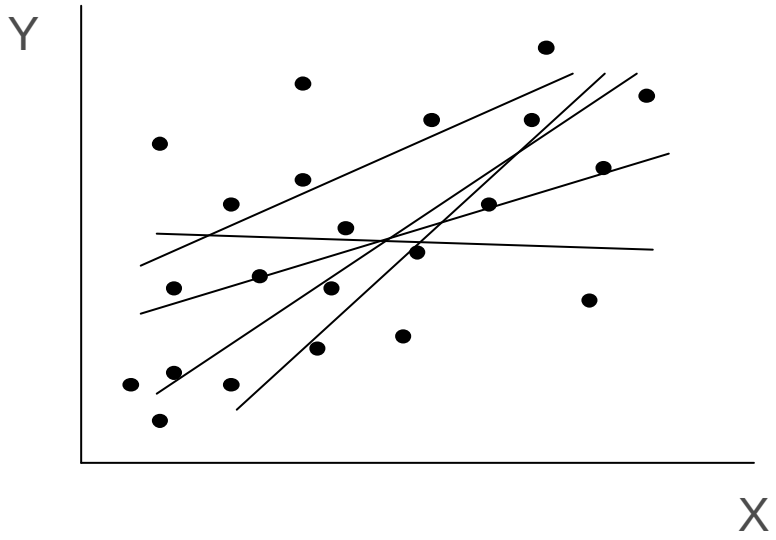
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$



$y_i = \beta_0 - \beta_1 x_i + \varepsilon_i$



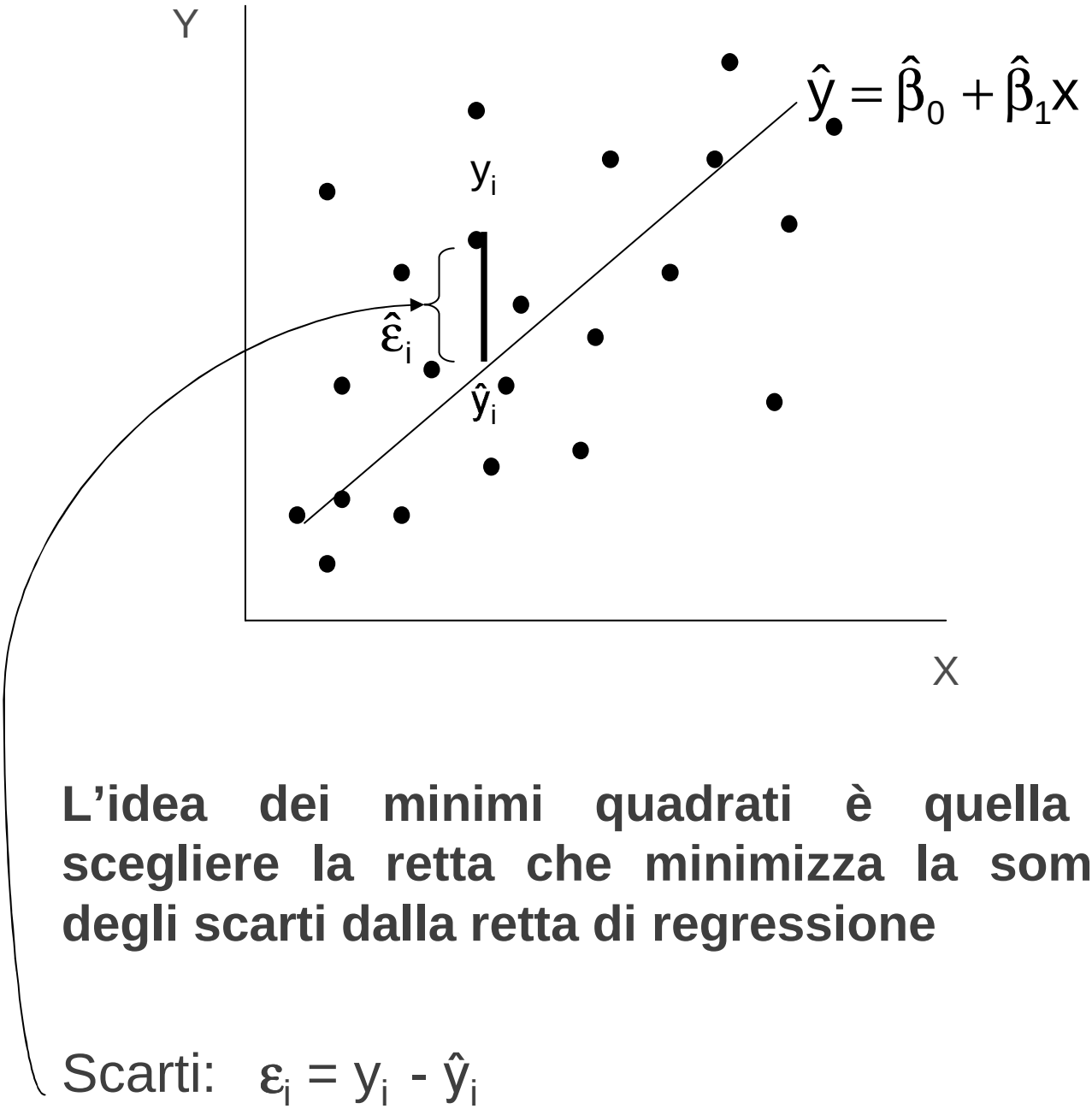
$y_i = \beta_0 + \varepsilon_i$



Per un insieme di punti possono passare infinite rette!

Come scegliere la retta “migliore”?

Metodo dei Minimi Quadrati



L'idea dei minimi quadrati è quella di scegliere la retta che minimizza la somma degli scarti dalla retta di regressione

Scarti: $\varepsilon_i = y_i - \hat{y}_i$

$RSS = \Sigma_i \varepsilon_i^2 = \Sigma_i (y_i - \hat{y}_i)^2 = \Sigma_i (y_i - \beta_0 - \beta_1 x_i)^2$

Si può dimostrare che i parametri che minimizzano la somma degli scarti dalla media al quadrato sono i seguenti:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(X,Y)}{Var(X)}$$

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

dove

$$\bar{x} = \frac{1}{n}\sum x_i;\ \bar{y} = \frac{1}{n}\sum y_i;$$

$$Var(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2;\quad Cov(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Coefficiente di correlazione $\Leftrightarrow$? $\Leftrightarrow$ β_1

$$\rho = \frac{Cov(X,Y)}{sd(X) \cdot sd(Y)} \qquad \hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)}$$

1) Ricavo ρ da $\hat{\beta}_1$

$$\rho = \hat{\beta}_1 \frac{sd(X)}{sd(Y)}$$

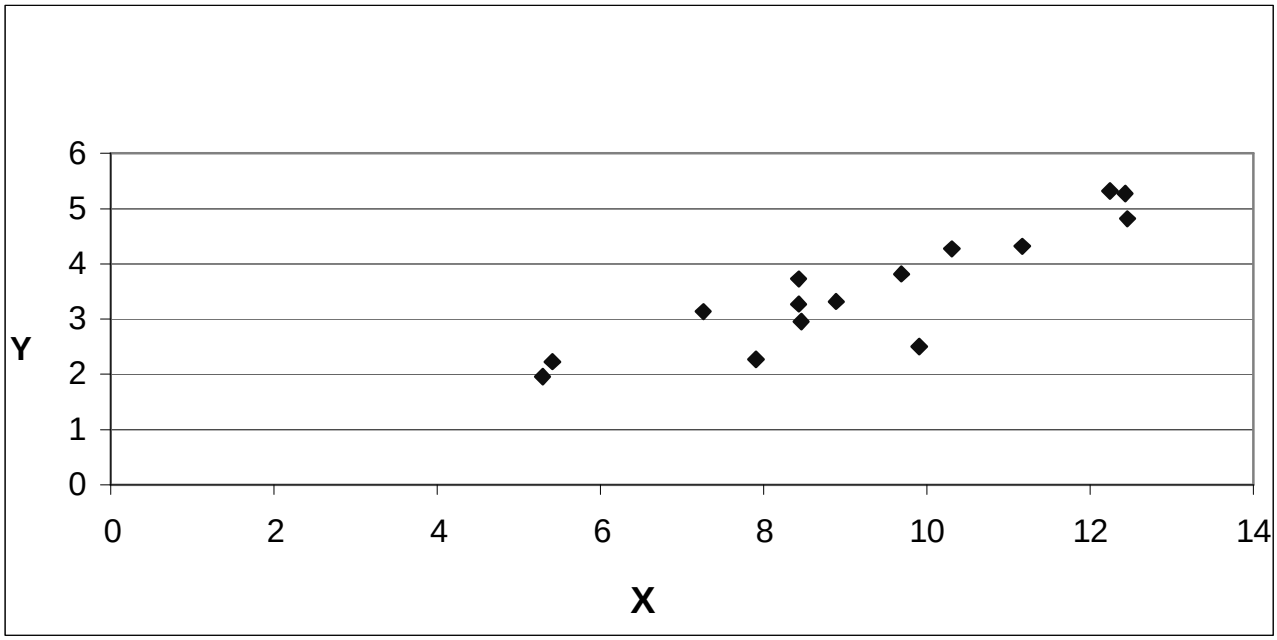
2) Ricavo $\hat{\beta}_1$ da ρ

$$\hat{\beta}_1 = \rho \frac{sd(Y)}{sd(X)}$$

Dalla popolazione di camelie estraiamo un campione di 15 foglie della varietà cordiforme sui quali misuriamo la variabile X (peso vivo) e Y (peso secco). Otteniamo i seguenti valori:

X	Y
9.705	3.816
7.267	3.130
8.459	2.955
12.476	4.809
10.296	4.269
8.424	3.291
7.910	2.274
8.879	3.308
11.160	4.340
5.295	1.948
8.421	3.715
12.232	5.340
5.422	2.212
9.900	2.512
12.441	5.277

Trovare la retta di regressione dei minimi quadrati che spiega Y in funzione di X



Dal campione si calcolano le seguenti quantità

$\overline{x} = 9.2191$ $\overline{y} = 3.5464$

$s^2_x = 5.2140$ $s^2_y = 1.1949$

$n=15$

Per ottenere i parametri della retta di regressione si devono usare le formule seguenti:

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$Cov(X,Y) = \frac{\left[\begin{array}{l}(9.705 - 9.2191) \cdot (3.816 - 3.5464) + \ldots \\ + (12.441 - 9.2191) \,.\, (5.277 - 3.5464)\end{array}\right]}{15\text{-}1}$$

$$Cov(X,Y) = 2.2324$$

$$Var(X) = 5.2140$$

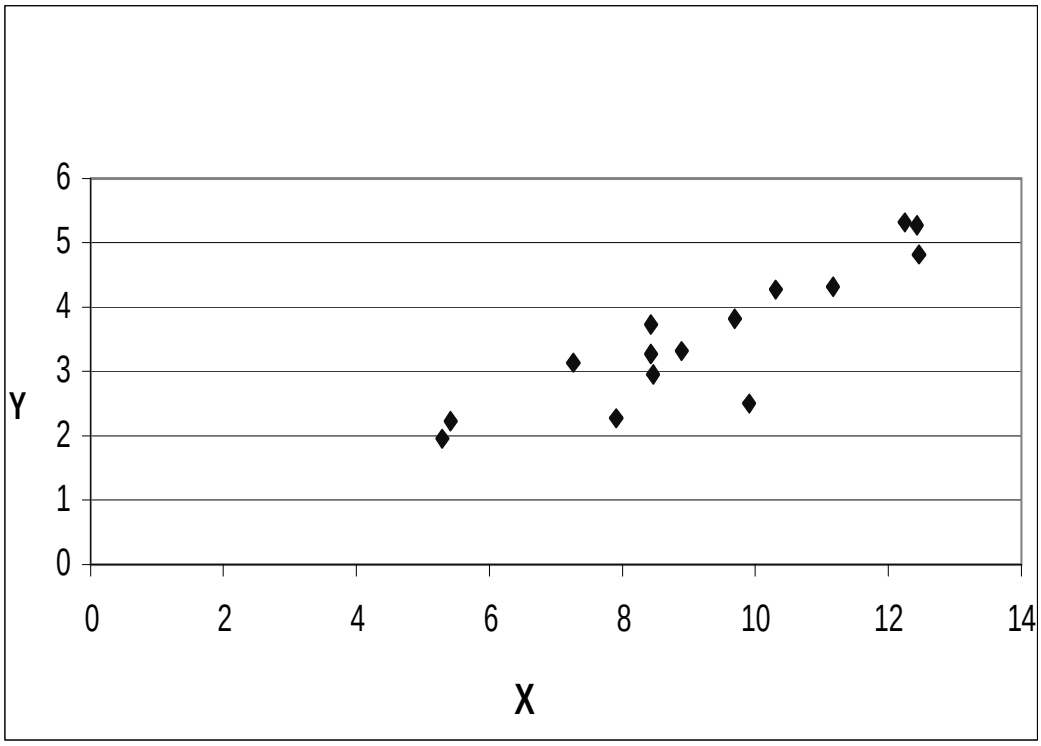
$$\hat{\beta}_1 = 2.2324 / 5.2140 = 0.4282$$

$$\hat{\beta}_0 = 3.5464 - 0.4282 \cdot 9.2191 = -0.4009$$

La retta di regressione che minimizza i quadrati degli scarti dalla media è la seguente:

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

$$\hat{y} = -0.4009 + 0.4282 \cdot x$$



Per disegnarla mi servono due punti

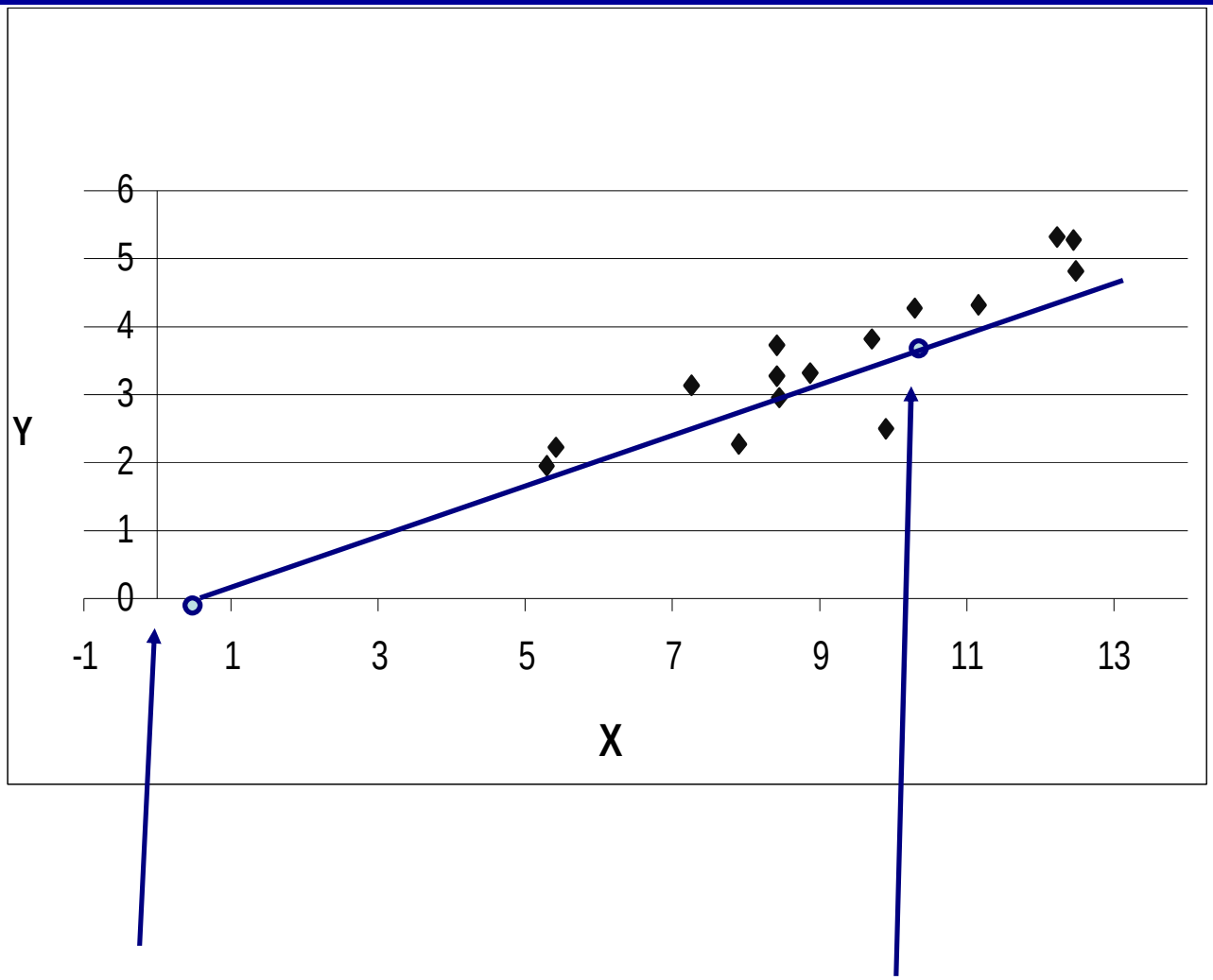
1) Quando y = 0 ⇒ $0 = -0.4009 + 0.4282 \cdot x$

$x = 0.4009 / 0.4282 = 0.9363$

2) Quando x = 10 ⇒ $y = -0.4009 + 0.4282 \cdot 10$

$y = -0.4009 + 4.282$

$= 3.8807$



Punto 1 :

x=-0.4009, y=0

Punto 2 :

x=10, y=3.8807

Come interpretare i due coefficienti del modello di regressione?

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

Intercetta:

- **valore che assume y quando x=0**
- **punto nel quale la retta incorcia l'asse delle y**

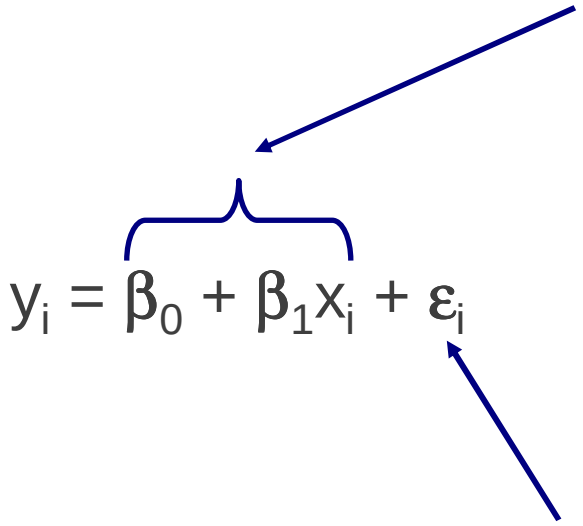
Pendenza:

- **indica di quanto varia y al variare di un'unità di x**
- **Il segno indica il verso dell'inclinazione**

Assunzioni del modello di regressione

Nel ipotizzare un modello di regressione stiamo assumendo che:

1. I dati sperimentali siano un campione casuale estratto da una popolazione di unità x,y per i quali vige la relazione

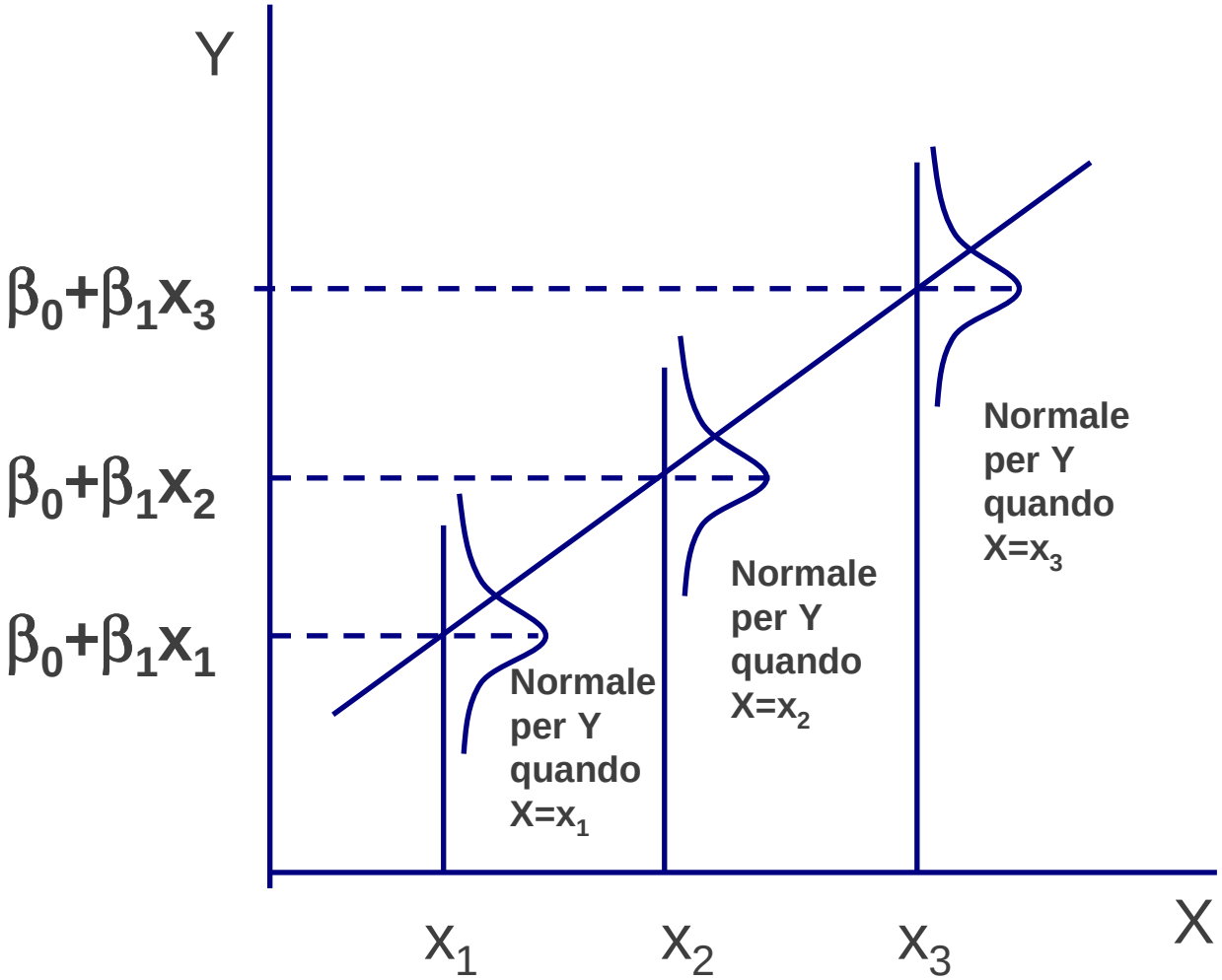


$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

tenendo conto di eventuali cause accidentali

Assunzioni del modello di regressione

2. Fissato un valore di X abbiamo una popolazione di valori di Y distribuiti normalmente con media situata sulla retta di regressione



$$Y|X=x_1 \sim N(\beta_0+\beta_1x_1,\sigma^2)$$
$$Y|X=x_2 \sim N(\beta_0+\beta_1x_2,\sigma^2)$$
$$Y|X=x_3 \sim N(\beta_0+\beta_1x_3,\sigma^2)$$

Assunzioni del modello di regressione

$$Y|X=x_i \sim N(\beta_0+\beta_1 x_i, \sigma^2)$$

3. La varianza rimane la stessa indipendentemente da X

Omoschedasticità

$Var(y_i) = \sigma^2$

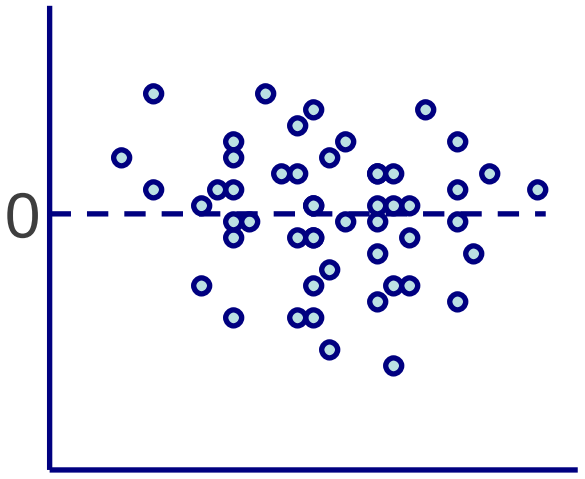
Assunzioni del modello di regressione

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

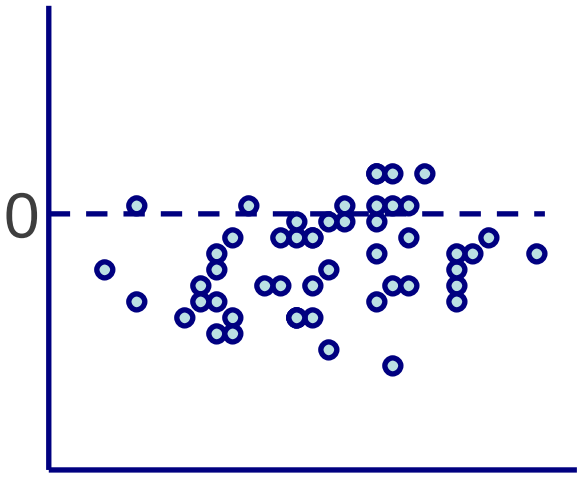
Scarti (Residui):

$$\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$$

$\varepsilon_i \sim N(0, \sigma^2)$ Stessa variabilità di Y

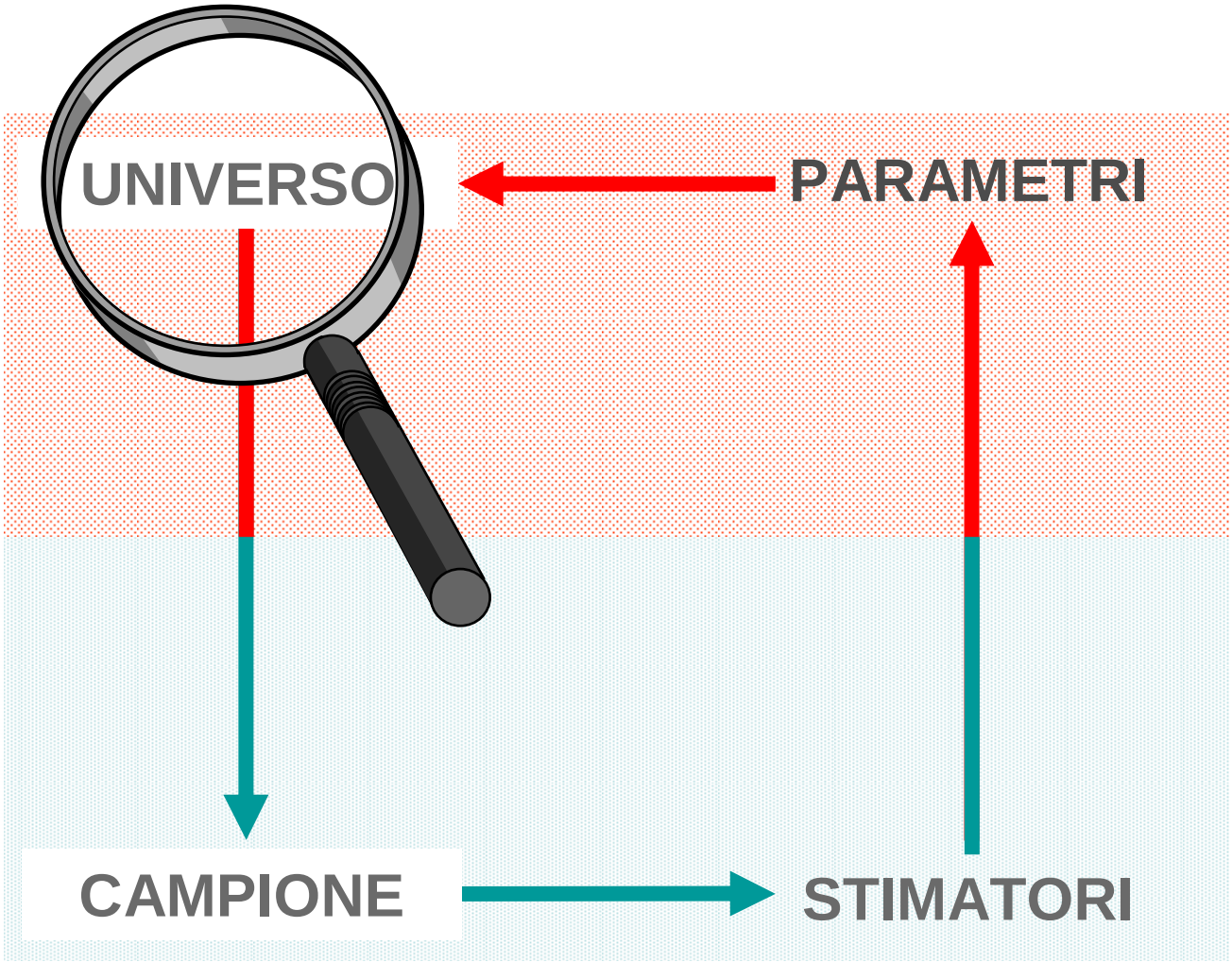


Assunzione rispettata



Assunzione violata

Siamo interessati a valutare l'esistenza di una relazione tra peso vivo e peso secco nella popolazione delle camelie tramite un modello di regressione.



Dalla popolazione di camelie estraiamo un campione di 15 foglie della varietà cordiforme sui quali misuriamo il peso vivo e il peso secco.

La retta di regressione dei minimi quadrati è la seguente:

$$y = -0.4009 + 0.4282 \cdot x$$

Come valutiamo se la relazione tra le due variabili è significativa o no?

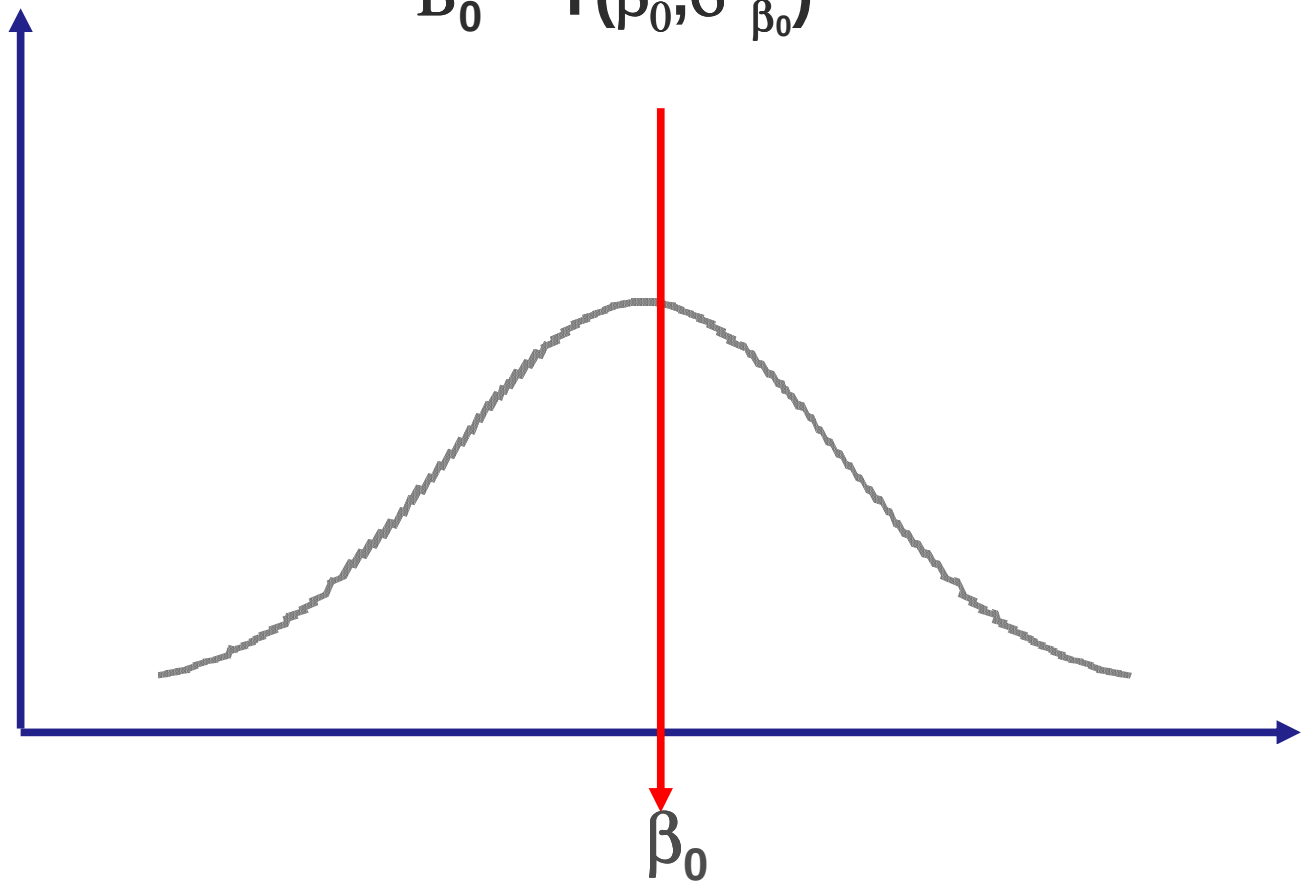
CAMPIONE → **STIMATORI**

Facciamo INFERENZA sui parametri della retta di regressione.

POPOLAZIONE BERSAGLIO

Tutti i possibili campioni

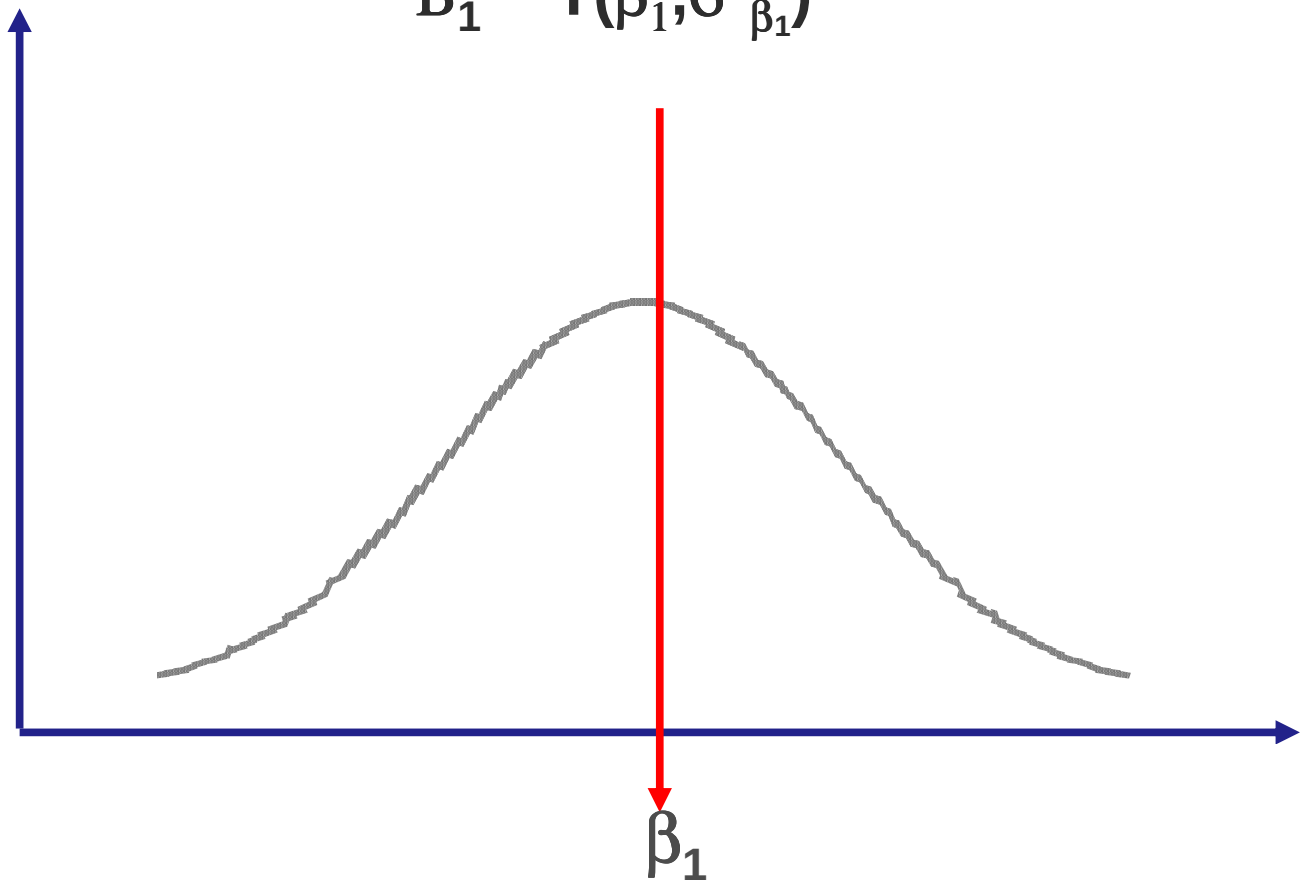
$$\hat{B}_0 \sim T(\beta_0, \sigma^2_{\beta_0})$$



POPOLAZIONE BERSAGLIO

Tutti i possibili campioni

$$\hat{B}_1 \sim T(\beta_1, \sigma^2_{\beta_1})$$



Usiamo $\hat{\beta}_0$ e $\hat{\beta}_1$ per stimare i veri valori dei parametri β_0 e β_1.

$$\hat{\beta}_0 \longrightarrow T(\beta_0, \sigma^2_{\beta_0})$$

Ipotesi nulla:

$H_0: \beta_0 = 0$

La retta di regressione passa per il punto di coordinate (0,0)

Test del T di Student

Dal campione:

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

Stima campionaria

$$se(\beta_0) = s \sqrt{\frac{1}{n} + \frac{\bar{(x)}^2}{Dev(x)}}$$

Deviazione standard campionaria

Usiamo $\hat{\beta}_0$ e $\hat{\beta}_1$ per stimare i veri valori dei parametri β_0 e β_1.

$$\hat{\beta}_1 \longrightarrow T(\beta_1, \sigma^2_{\beta_1})$$

Ipotesi nulla:

$H_1: \beta_1 = 0$

La retta di regressione ha pendeza 0

↓

Test del T di Student

Dal campione:

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$$

Stima campionaria

$$se(\beta_1) = \frac{s}{\sqrt{Dev(x)}}$$

Deviazione standard campionaria

L'errore standard di entrambi i parametri è funzione di s

$$s = \sqrt{\frac{(n-1)\, s^2_y\, (1 - \rho^2_{xy})}{n-2}}$$

I valori empirici per il test T di student sono

β_1

$$t_g = \frac{\hat{\beta}_1 - 0}{es(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - 0}{\frac{s}{\sqrt{Dev(x)}}}$$

n-2

β_0

$$t_g = \frac{\hat{\beta}_0 - 0}{es(\hat{\beta}_0)} = \frac{\hat{\beta}_0 - 0}{s\sqrt{\frac{1}{n} + \frac{\overline{(x)}^2}{Dev(x)}}}$$

n-2

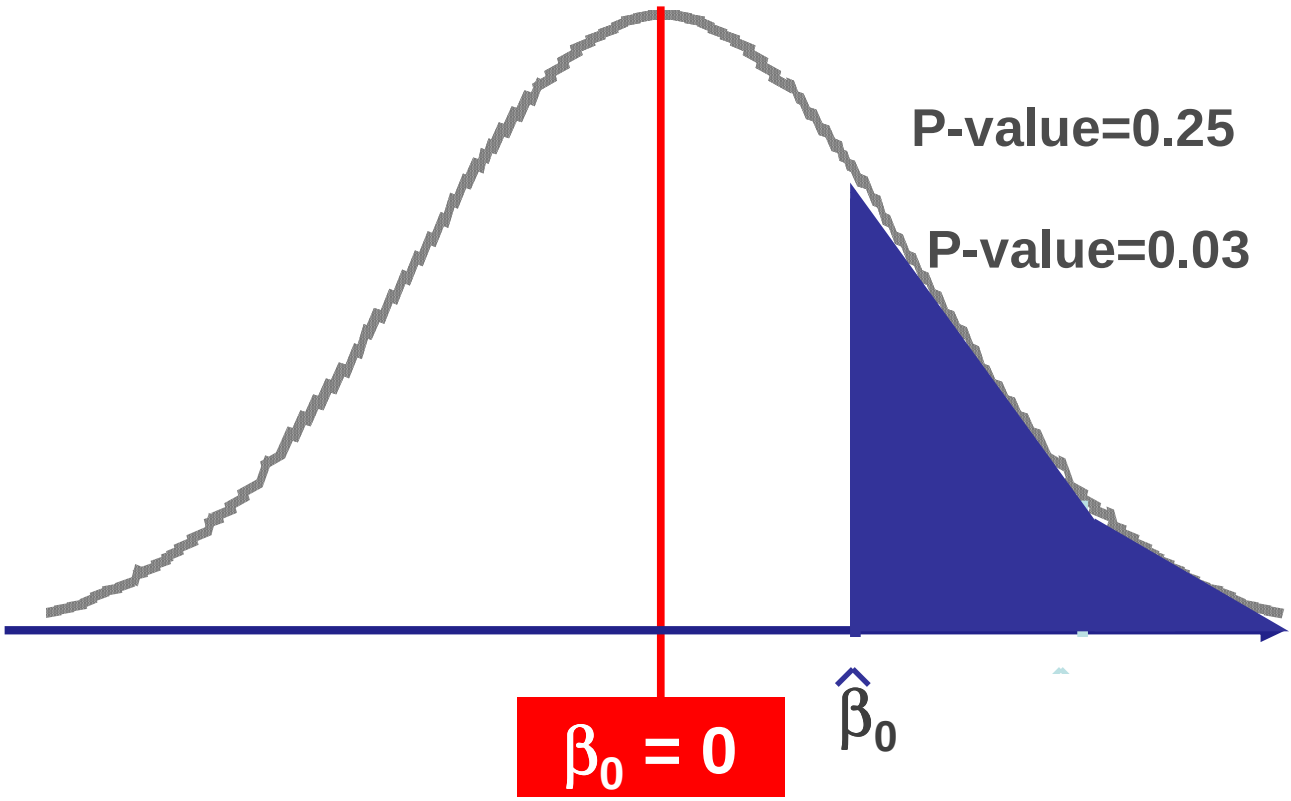
P-Value: probabilità di ottenere un risultato campionario altrettanto o più estremo di quello osservato, se H_0 è vera

$$\text{P-value} = \Pr(\hat{B}_0 > \hat{\beta}_0 \text{ sotto } H_0)$$

Più piccolo è il valore del p-value,

1) più "estremo" è il valore osservato

2) Più bassa l'evidenza che i dati siano coerenti con la distribuzione sotto l'ipotesi nulla



PROBLEMA: l'ipotesi è bidirezionale

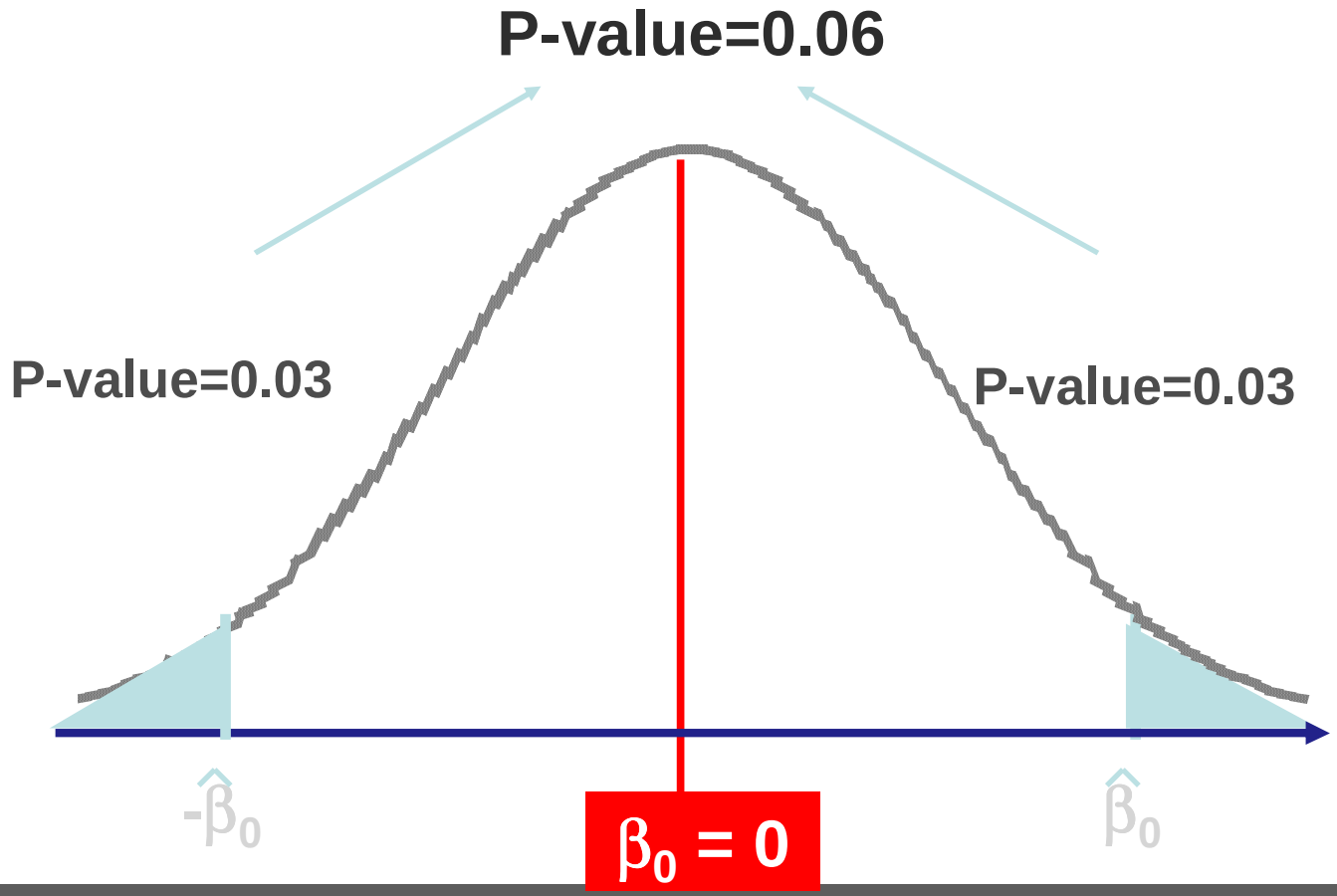
$H_0: \beta_0 = 0$ vs $H_1: \beta_0 \neq 0$

Unidirezionale

P-value = Pr ($\hat{B}_0 > \hat{\beta}_0$ sotto H_0)

Bidirezionale

2*P-value



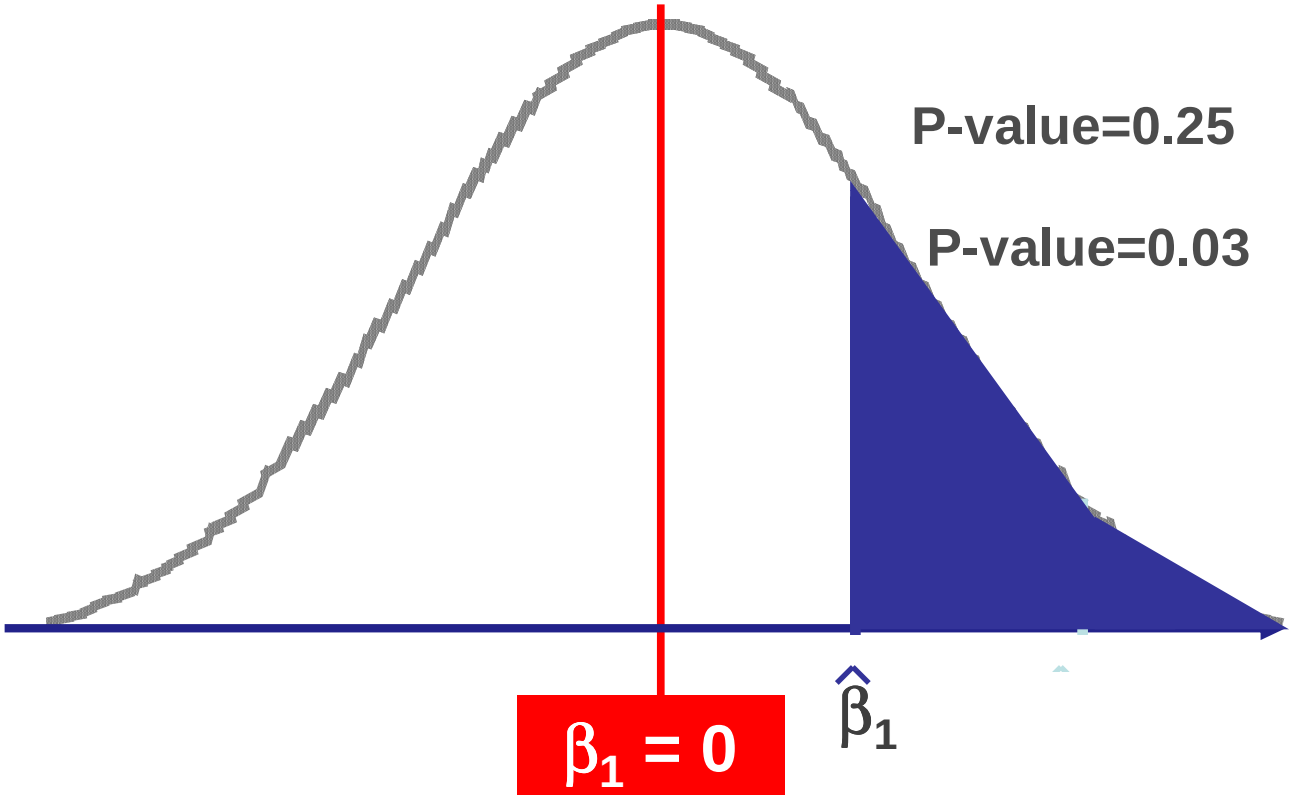
P-Value: probabilità di ottenere un risultato campionario altrettanto o più estremo di quello osservato, se H_0 è vera

$$\text{P-value} = \Pr\left(\hat{B}_1 > \hat{\beta}_1 \text{ sotto } H_0\right)$$

Più piccolo è il valore del p-value,

1) più "estremo" è il valore osservato

2) Più bassa l'evidenza che i dati siano coerenti con la distribuzione sotto l'ipotesi nulla



PROBLEMA: l'ipotesi è bidirezionale

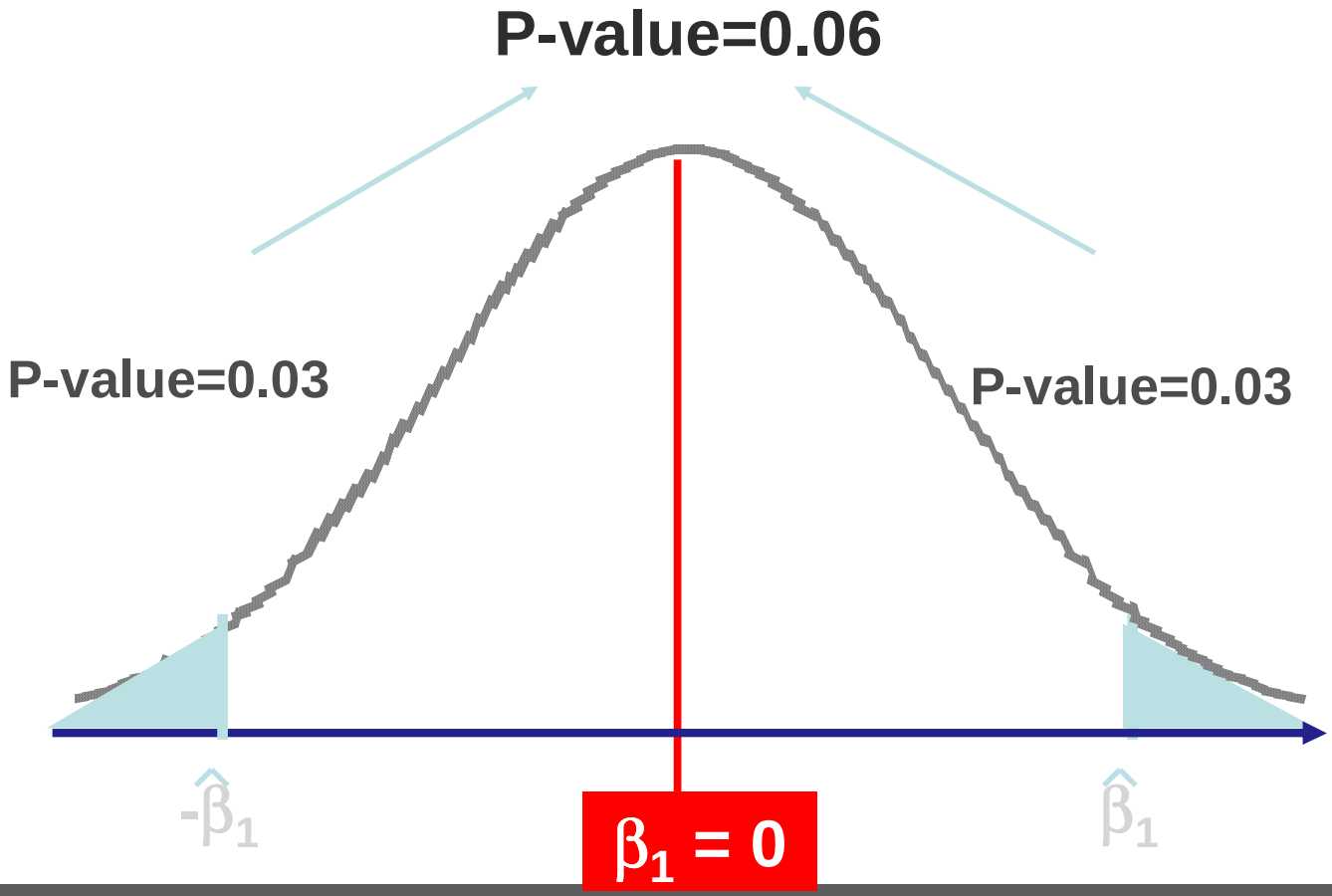
$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

Unidirezionale

P-value = Pr ($\hat{B}_1 > \hat{\beta}_1$ sotto H_0)

Bidirezionale

2*P-value



Siamo interessati a valutare l'esistenza di una relazione tra peso vivo e peso secco nella popolazione delle camelie tramite un modello di regressione.

X	Y
9.705	3.816
7.267	3.130
8.459	2.955
12.476	4.809
10.296	4.269
8.424	3.291
7.910	2.274
8.879	3.308
11.160	4.340
5.295	1.948
8.421	3.715
12.232	5.340
5.422	2.212
9.900	2.512
12.441	5.277

Campione

n=15

Dal campione otteniamo le seguenti informazioni

Stimatore	Misura di variabilità	
$\bar{y} = \frac{\sum_i y_i}{n} = 3.5464$	Deviazione standard $s_y = \sqrt{\frac{\sum_i (y_i - \bar{y})^2}{n-1}} = 1.1773$	$s = \sqrt{\frac{(n-1)s_y^2(1-\rho_{xy}^2)}{n-2}} = 0.5465$
$\bar{x} = \frac{\sum_i x_i}{n} = 9.2191$	Varianza $s_x^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1} = 5.2140$	
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -0.4009$	Errore standard $es_{\beta_0}^2 = s\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{Dev_x}} = 2.722$	$t_g = \hat{\beta}_0 \sqrt{\frac{1}{es^2_{\hat{\beta}_0}}} = -0.14$
$\hat{\beta}_1 = \frac{Cov(x,y)}{s_x^2} = 0.4282$	Errore standard $es_{\beta_1}^2 = \frac{s}{\sqrt{Dev_x}} = 0.064$	$t_g = \hat{\beta}_1 \sqrt{\frac{1}{es^2_{\hat{\beta}_1}}} = 6.69$

P-value(β_0) = Pr ($\hat{B}_0 > \hat{\beta}_0$ sotto H_0)

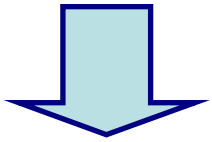
2*P-value(β_0) >2*0.4 che trovo sulle tavole

Non ho sufficiente evidenza per rifiutare H_0

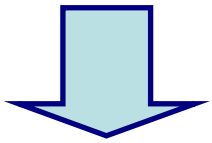
Concludo che β_0 non è significativamente diverso da 0

$$\text{P-value}(\beta_1) = \Pr\left(\hat{B}_1 > \hat{\beta}_1 \text{ sotto } H_0\right)$$

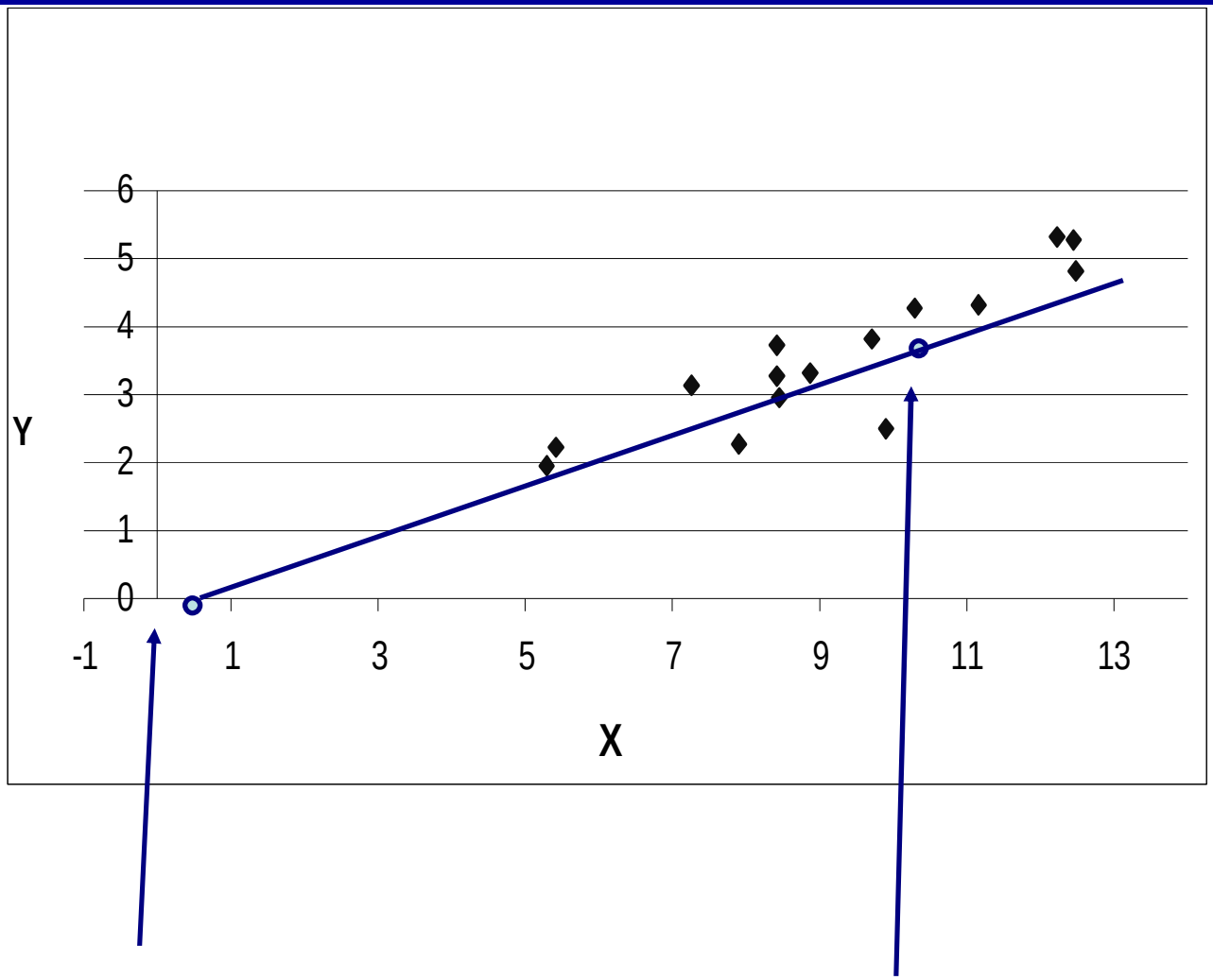
$$2*\text{P-value}(\beta_1) < 2*0.0005$$



Rifiuto H_0



Concludo che β_1 è significativamente diverso da 0



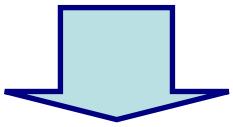
Punto 1 :

x=-0.4009, y=0

Punto 2 :

x=10, y=3.8807

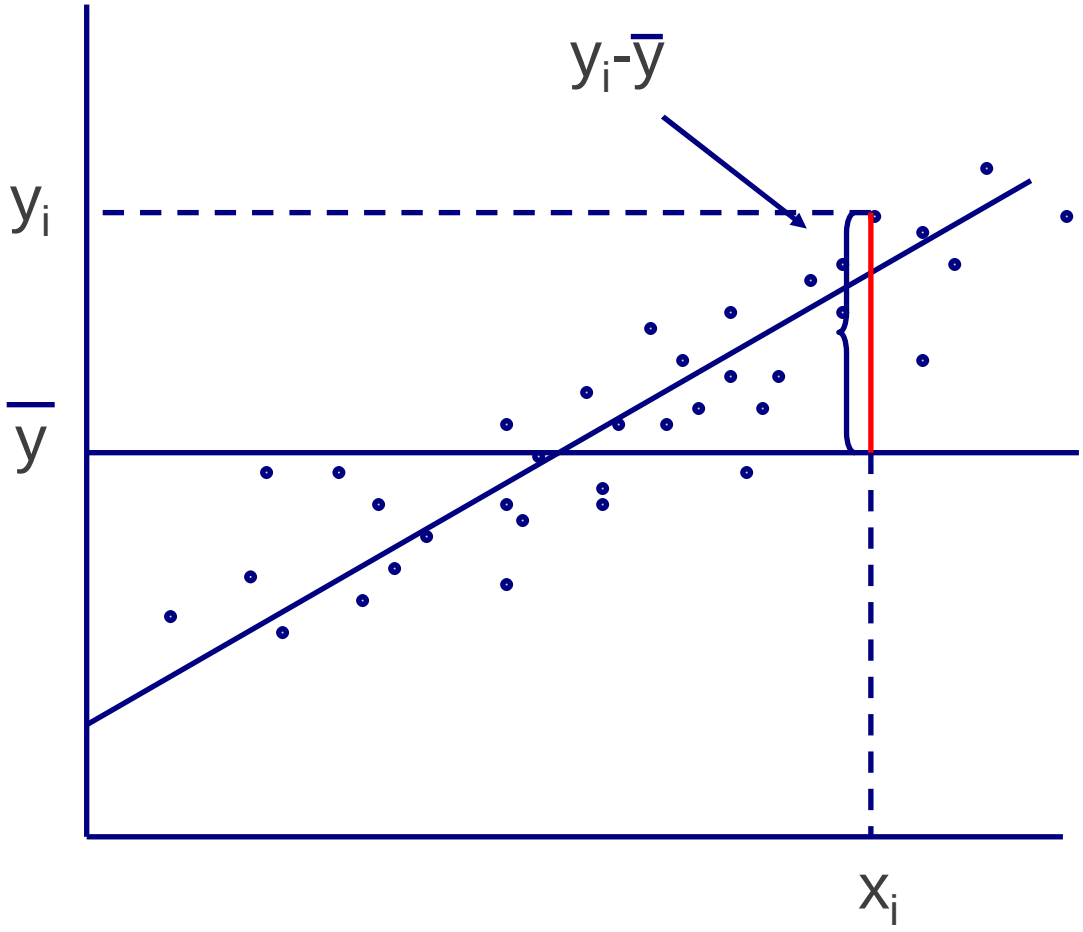
Se concludo che $\beta_1=0$



Il modello di regressione lineare non è adatto ad interpretare la relazione tra X e Y.

C'è un modo per valutare analiticamente la bontà di adattamento del modello?

H_0: **il modello non si adatta ai dati**



Devianza totale:

$$Dev_{TOT} = \Sigma_i (y_i - \bar{y})^2$$

X	Y
9.705	3.816
7.267	3.130
8.459	2.955
12.476	4.809
10.296	4.269
8.424	3.291
7.910	2.274
8.879	3.308
11.160	4.340
5.295	1.948
8.421	3.715
12.232	5.340
5.422	2.212
9.900	2.512
12.441	5.277

Campione

n=15

$\bar{y} = 3.5464$

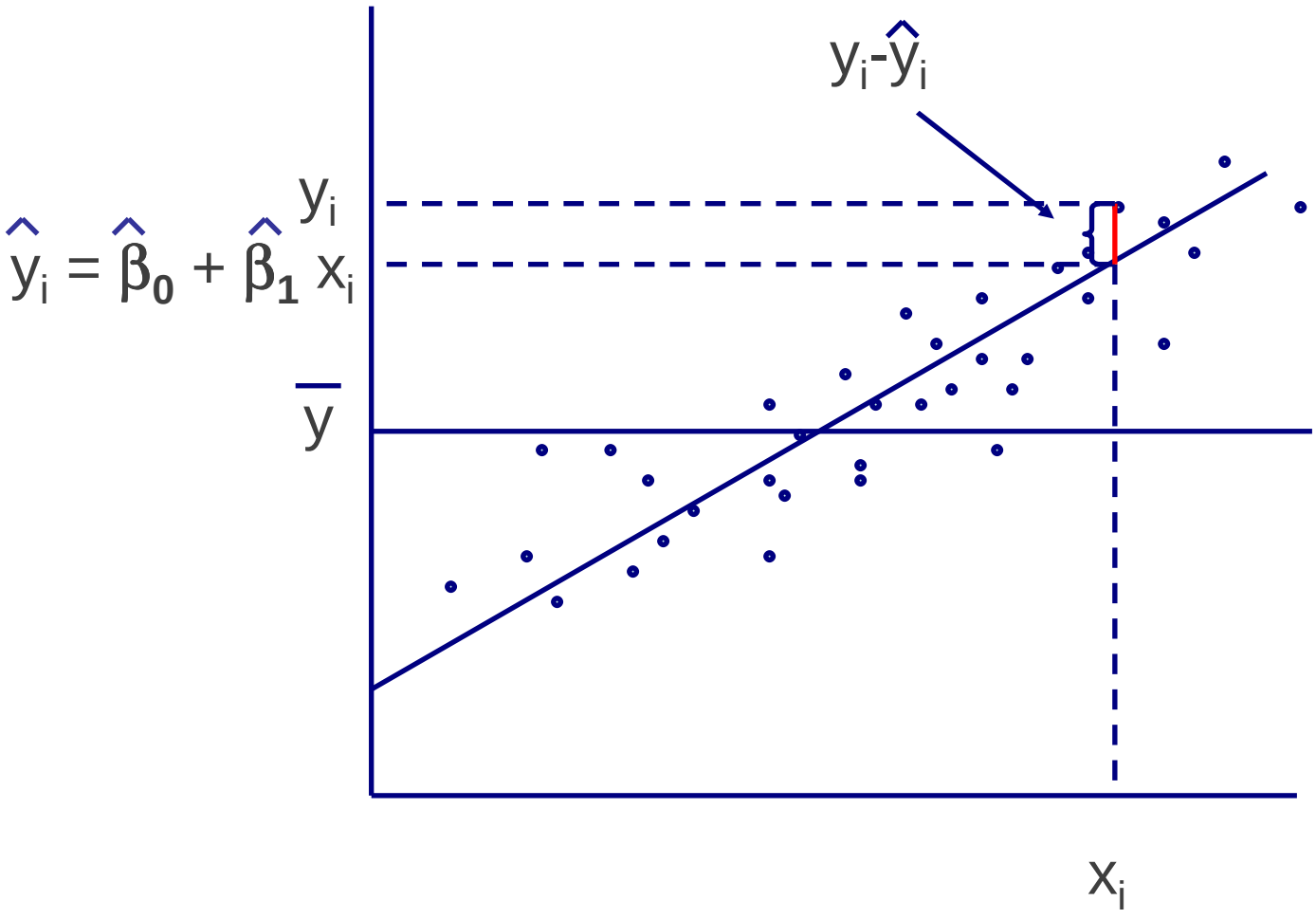
Devianza totale = $\sum_i (y_i - \bar{y})^2$

$= (3.816-3.5464)^2 + (3.130-3.5464)^2 + \ldots$

$\ldots + (2.512-3.5464)^2 + (5.277-3.5464)^2 =$

$= 16.7289$

Quanta parte della variabilità totale è residua?



Devianza RESIDUA:

$$Dev_R = \Sigma_i (y_i - \hat{y}_i)^2$$

X	Y	$\hat{Y}$
9.705	3.816	3.754
7.267	3.130	2.711
8.459	2.955	3.221
12.476	4.809	4.941
10.296	4.269	4.007
8.424	3.291	3.206
7.910	2.274	2.986
8.879	3.308	3.401
11.160	4.340	4.377
5.295	1.948	1.866
8.421	3.715	3.205
12.232	5.340	4.836
5.422	2.212	1.921
9.900	2.512	3.838
12.441	5.277	4.926

$\Rightarrow \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$-0.4007 + 0.4282 \cdot 9.705 = 3.754$

.

.

.

.

.

.

.

$-0.4007 + 0.4282 \cdot 12.441 = 4.836$

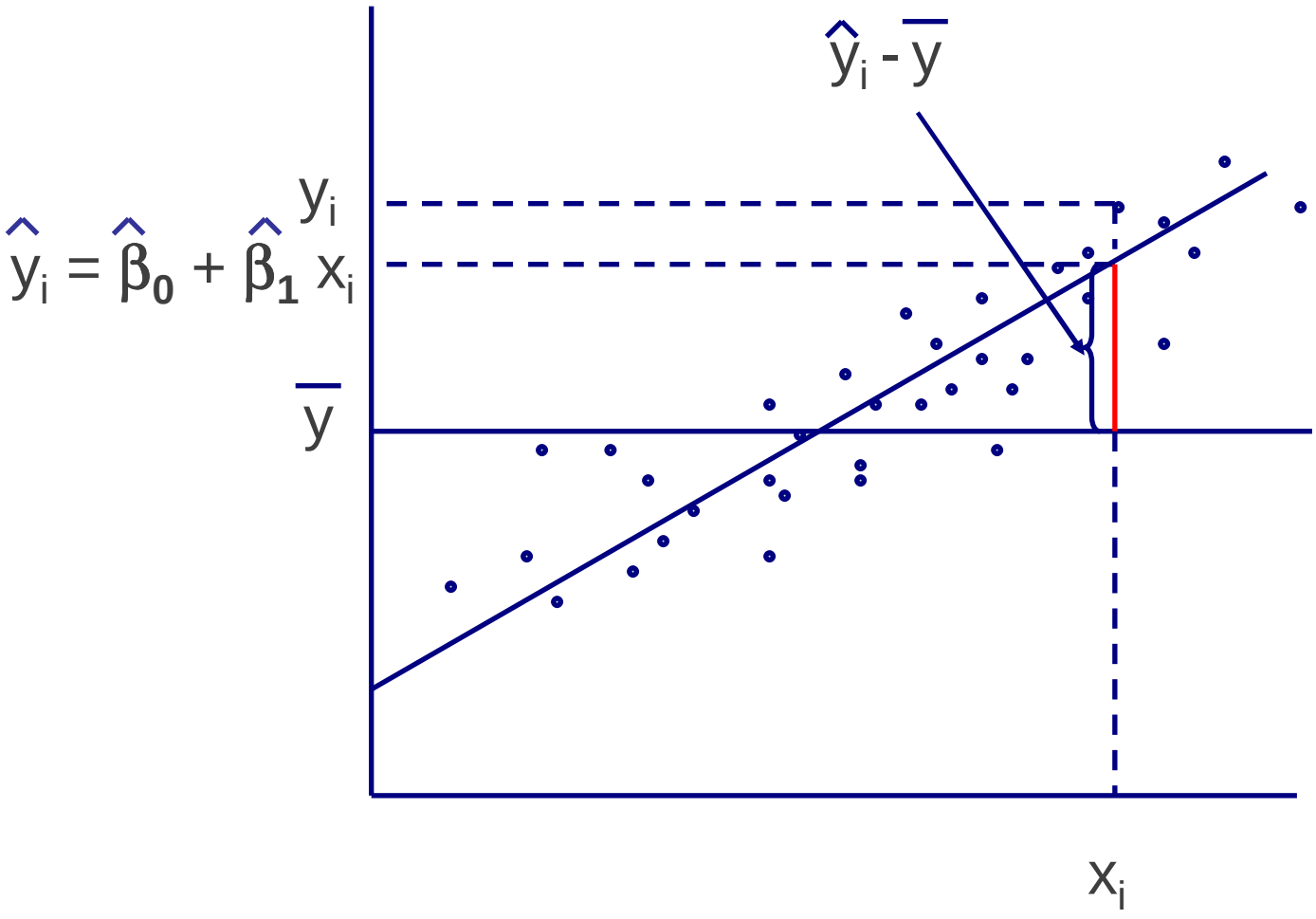
Devianza residua = $\sum_i (y_i - \hat{y}_i)^2$

$= (3.816-3.754)^2 + (3.130-2.711)^2 + \ldots$

$\ldots + (2.512-3.838)^2 + (5.277-4.926)^2 =$

$= 3.3472$

Quanta parte della variabilità totale è spiegata dal modello di regressione?



Devianza SPIEGATA dal modello :

$$Dev_S = \Sigma_i (\hat{y}_i - \bar{y})^2$$

X	Y	$\hat{Y}$
9.705	3.816	3.754
7.267	3.130	2.711
8.459	2.955	3.221
12.476	4.809	4.941
10.296	4.269	4.007
8.424	3.291	3.206
7.910	2.274	2.986
8.879	3.308	3.401
11.160	4.340	4.377
5.295	1.948	1.866
8.421	3.715	3.205
12.232	5.340	4.836
5.422	2.212	1.921
9.900	2.512	3.838
12.441	5.277	4.926

$\Rightarrow \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\bar{y} = 3.5464$

Devianza spiegata = $\sum_i (\hat{y}_i - \bar{y})^2$

$= (3.754-3.5464)^2 + (2.711-2.5464)^2 + \ldots$

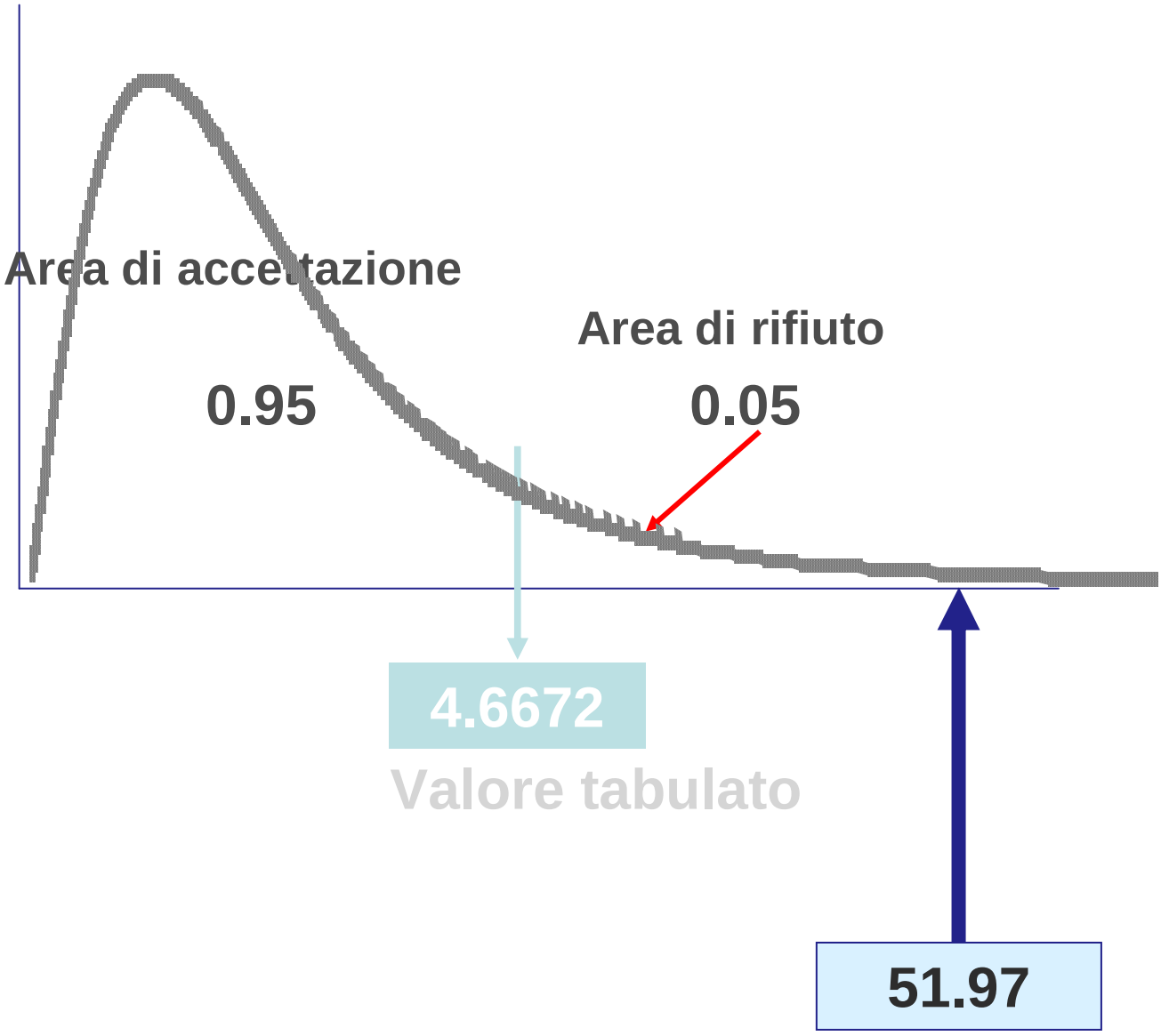
$\ldots + (3.838-3.5464)^2 + (4.926-3.5464)^2 =$

$= 13.3817$

Fonti di variabilità	devianza	Gradi di libertà	varianza
Spiegata dal modello	13.3817 +	k-1 = 1	13.3817
Residua	3.3472 =	n-k = 13	0.2575
Totale	16.7289	n-1=14	

$$F_{1,\,13} = \frac{\text{Varianza spiegata}}{\text{Varianza residua}} = \frac{13.3817}{0.2575} = 51.97$$

Distribuzione $F_{1,13}$



Valore empirico

rifiutiamo H_0 ovvero la relazione tra le due variabili è ben spiegata da un modello di regressione lineare

$p < 0.05$

Esercizio di riepilogo

Si vuole valutare la relazione tra peso alla nascita e settimane di gestazione in UK. A tal fine si estrae un campione di 26 bambini nati a University College Hospital di Londra, della stessa razza e dello stesso genere. I dati sono I seguenti:

X: 42 41 39 40 40 40 39 39 41 42 41 43 43 41 38 37 38 43 35 37 35 38 40 42 39 34

Y: 3.180 2.780 3.630 3.900 3.310 2.896 2.780 3.800 3.900 4.020 4.180 3.460 4.400 3.800 2.990 3.160 2.720 3.560 2.640 2.400 2.320 2.910 3.200 3.800 3.560 2.538

Stimare i parametri della retta di regressione dei minimi quadrati.