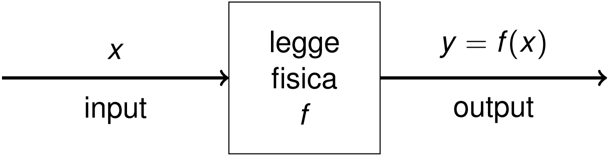
Statistica - 14ª lezione

13 maggio 2021

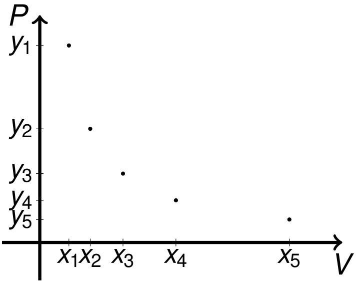
Programma

- Statistica descrittiva
  (riassumere molti dati attraverso poche caratteristiche essenziali)
- Probabilità
  (costruire un modello che preveda il risultato di un esperimento)
- Inferenza statistica
  (tarare i parametri del modello in base ai risultati dell'esperimento)
- Regressione lineare
  (riconoscere relazioni tra dati di tipo diverso)

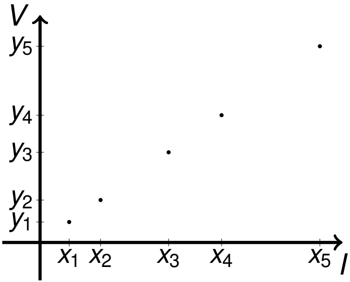
Modelli empirici



Idealmente:



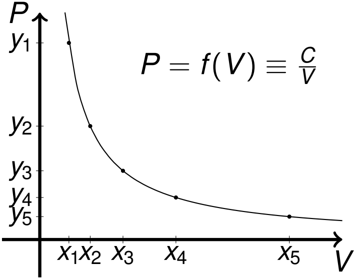
Gas perfetto



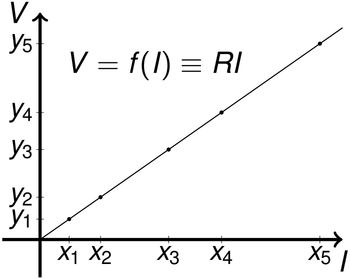
Corrente-tensione

Modelli empirici

Idealmente:

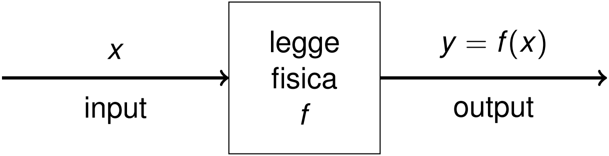


Gas perfetto

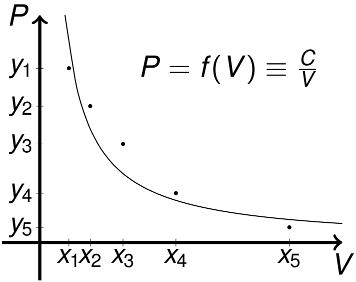


Corrente-tensione

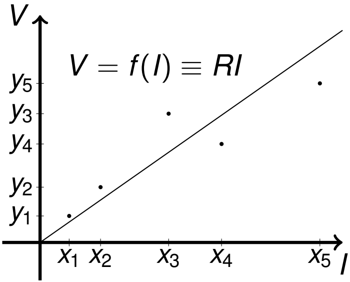
Modelli empirici



Realisticamente:



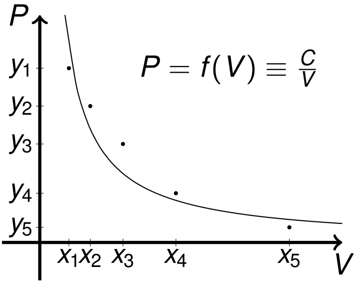
Gas perfetto



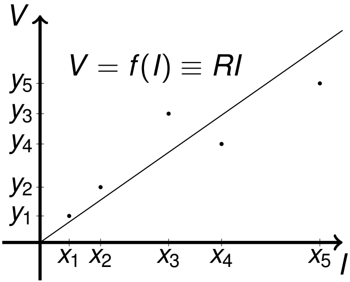
Corrente-tensione

Modelli empirici

Vogliamo trovare f!



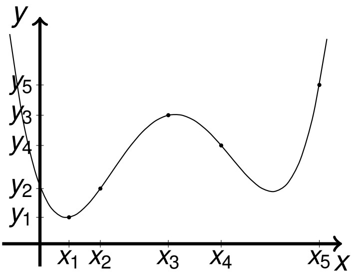
Gas perfetto



Corrente-tensione

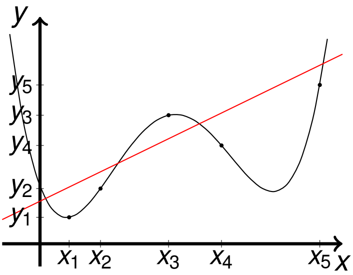
Modello lineare

Per n punti passa sempre un polinomio di grado $n-1$:



Modello lineare

Per n punti passa sempre un polinomio di grado $n-1$:



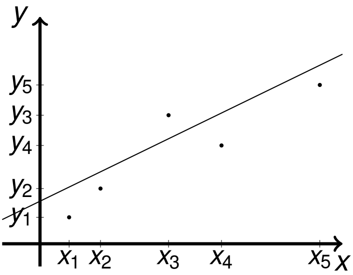
Ma noi vogliamo la funzione più semplice possibile: la retta

$$y = a + bx$$

parametrizzata solo da (a, b).

Modello lineare

Per n punti passa sempre un polinomio di grado $n-1$:



Ma noi vogliamo la funzione più semplice possibile: la retta

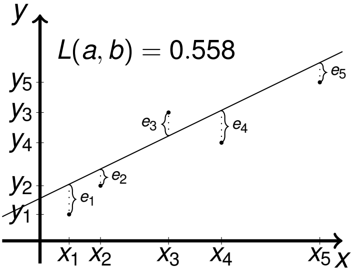
$$y = a + bx$$

parametrizzata solo da (a, b).

Come si capisce se la retta di parametri (a, b) approssima bene gli n punti?

Modello lineare

Per n punti passa sempre un polinomio di grado $n-1$:



Ma noi vogliamo la funzione più semplice possibile: la retta

$$y = a + bx$$

parametrizzata solo da (a, b).

Come si capisce se la retta di parametri (a, b) approssima bene gli n punti?

Dobbiamo introdurre gli errori (o *residui*)

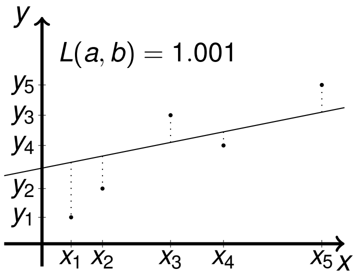
$$e_i := y_i - (a + bx_i)$$

e un funzionale di errore

$$L(a,b) := \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} e_i^2$$

Modello lineare

Per n punti passa sempre un polinomio di grado $n-1$:



Ma noi vogliamo la funzione più semplice possibile: la retta

$$y = a + bx$$

parametrizzata solo da (a, b).

Come si capisce se la retta di parametri (a, b) approssima bene gli n punti?

Dobbiamo introdurre gli errori (o *residui*)

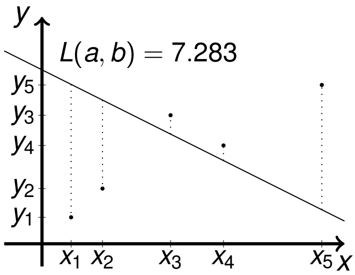
$$e_i := y_i - (a + bx_i)$$

e un funzionale di errore

$$L(a,b) := \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} e_i^2$$

Modello lineare

Per n punti passa sempre un polinomio di grado $n-1$:



Ma noi vogliamo la funzione più semplice possibile: la retta

$$y = a + bx$$

parametrizzata solo da (a, b).

Come si capisce se la retta di parametri (a, b) approssima bene gli n punti?

Dobbiamo introdurre gli errori (o *residui*)

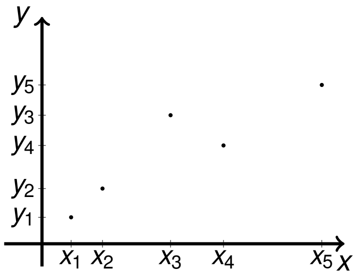
$$e_i := y_i - (a + bx_i)$$

e un funzionale di errore

$$L(a,b) := \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} e_i^2$$

Modello lineare

Per n punti passa sempre un polinomio di grado $n-1$:



Ma noi vogliamo la funzione più semplice possibile: la retta

$$y = a + bx$$

parametrizzata solo da (a, b).

Come si capisce se la retta di parametri (a, b) approssima bene gli n punti?

Dobbiamo introdurre gli errori (o *residui*)

$$e_i := y_i - (a + bx_i)$$

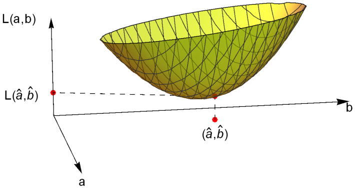
e un funzionale di errore

$$L(a, b) := \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} e_i^2$$

Perché si mette il quadrato?

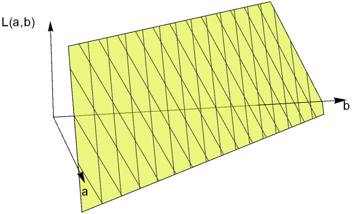
Modello lineare

Col quadrato:



- $L(a, b) \geq 0 \ \forall (a, b)$
- $L(\infty) = +\infty$
- $\exists!$ punto di minimo $(\hat{a}, \hat{b})$

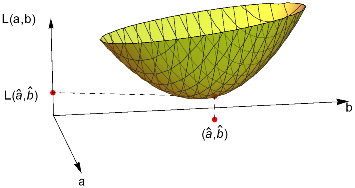
Senza quadrato:



- i residui si compensano
- $\nexists$ minimo

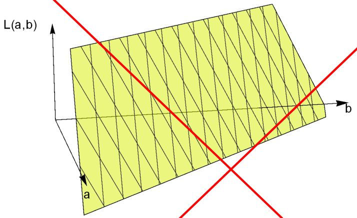
Modello lineare

Col quadrato:



- $L(a, b) \geq 0 \; \forall (a, b)$
- $L(\infty) = +\infty$
- $\exists!$ punto di minimo $(\hat{a}, \hat{b})$

~~Senza quadrato:~~



- ~~i residui si compensano~~
- ~~$\nexists$ minimo~~

$y = \hat{a} + \hat{b}x$ è la *retta dei minimi quadrati* (LSL = *least square line*)

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$\frac{\partial L}{\partial a} = \frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$\begin{aligned}\frac{\partial L}{\partial a} &= \frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \\ &= -2\sum_{i=1}^{n}(y_i - a - bx_i)\end{aligned}$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$\begin{aligned}
\frac{\partial L}{\partial a} &= \frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\
&= -2 \sum_{i=1}^{n} (y_i - a - bx_i) \\
&= -2(n\bar{y} - na - nb\bar{x})
\end{aligned}$$

dove

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i .$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$\begin{aligned}
\frac{\partial L}{\partial a} &= \frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\
&= -2 \sum_{i=1}^{n} (y_i - a - bx_i) \\
&= -2(n\bar{y} - na - nb\bar{x}) \\
&\equiv 0
\end{aligned}$$

dove

$$\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i .$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$
\begin{aligned}
\frac{\partial L}{\partial a} &= \frac{\partial}{\partial a} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\
&= -2 \sum_{i=1}^{n} (y_i - a - bx_i) \\
&= -2(n\bar{y} - na - nb\bar{x}) \\
&\equiv 0 \\
&\qquad \Rightarrow \quad a = \bar{y} - b\bar{x}
\end{aligned}
$$

dove

$$
\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y} := \frac{1}{n} \sum_{i=1}^{n} y_i .
$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$\frac{\partial L}{\partial b} = \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$\begin{aligned}\frac{\partial L}{\partial b} &= \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\ &= -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i)\end{aligned}$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\
&= -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) \\
&= -2 \left(\sum_{i=1}^{n} x_i y_i - na\bar{x} - b \sum_{i=1}^{n} x_i^2 \right)
\end{aligned}$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$
\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\
&= -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) \\
&= -2 \left(\sum_{i=1}^{n} x_i y_i - na\bar{x} - b \sum_{i=1}^{n} x_i^2 \right) \\
&= -2 \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y} - b\bar{x}) - b \sum_{i=1}^{n} x_i^2 \right) \qquad \text{con } a = \bar{y} - b\bar{x}
\end{aligned}
$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$
\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\
&= -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) \\
&= -2 \left(\sum_{i=1}^{n} x_i y_i - na\bar{x} - b \sum_{i=1}^{n} x_i^2 \right) \\
&= -2 \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y} - b\bar{x}) - b \sum_{i=1}^{n} x_i^2 \right) \qquad \text{con } a = \bar{y} - b\bar{x} \\
&\equiv 0
\end{aligned}
$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$
\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\
&= -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) \\
&= -2 \left(\sum_{i=1}^{n} x_i y_i - na\bar{x} - b \sum_{i=1}^{n} x_i^2 \right) \\
&= -2 \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y} - b\bar{x}) - b \sum_{i=1}^{n} x_i^2 \right) \qquad \text{con } a = \bar{y} - b\bar{x} \\
&\equiv 0 \\
&\qquad \Rightarrow \quad b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_j^2 - n\bar{x}^2}
\end{aligned}
$$

Retta dei minimi quadrati

Cerchiamo $(\hat{a}, \hat{b})$ ponendo $\nabla L(\hat{a}, \hat{b}) \equiv 0$:

$$\begin{aligned}
\frac{\partial L}{\partial b} &= \frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\
&= -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) \\
&= -2 \left(\sum_{i=1}^{n} x_i y_i - na\bar{x} - b \sum_{i=1}^{n} x_i^2 \right) \\
&= -2 \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y} - b\bar{x}) - b \sum_{i=1}^{n} x_i^2 \right) \qquad \text{con } a = \bar{y} - b\bar{x} \\
&\equiv 0
\end{aligned}$$

$$\Rightarrow \quad b = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_j^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_j - \bar{x})^2}$$

Sintesi dei dati

Dai punti $(x_1, y_1), \dots, (x_n, y_n)$ distilliamo le 5 quantità

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s_{xx} := \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_{xy} := \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad s_{yy} := \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Sintesi dei dati

Dai punti $(x_1, y_1), \ldots, (x_n, y_n)$ distilliamo le 5 quantità

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s_{xx} := \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_{xy} := \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad s_{yy} := \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Tutte le altre quantità della LSL sono funzioni di queste. P. es.:

$$\hat{a} = \bar{y} - \frac{s_{xy}}{s_{xx}}\,\bar{x} \qquad \hat{b} = \frac{s_{xy}}{s_{xx}}$$ parametri della LSL

Sintesi dei dati

Dai punti $(x_1, y_1), \ldots, (x_n, y_n)$ distilliamo le 5 quantità

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s_{xx} := \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_{xy} := \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad s_{yy} := \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Tutte le altre quantità della LSL sono funzioni di queste. P. es.:

$$\hat{a} = \bar{y} - \frac{s_{xy}}{s_{xx}}\,\bar{x} \qquad \hat{b} = \frac{s_{xy}}{s_{xx}} \qquad \text{parametri della LSL}$$

$$\hat{y}_i := \hat{a} + \hat{b}x_i = \bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x}) \qquad \text{output della LSL}$$

Sintesi dei dati

Dai punti $(x_1, y_1), \ldots, (x_n, y_n)$ distilliamo le 5 quantità

$$\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s_{xx} := \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_{xy} := \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad s_{yy} := \sum_{i=1}^{n}(y_i - \bar{y})^2$$

Tutte le altre quantità della LSL sono funzioni di queste. P. es.:

$$\hat{a} = \bar{y} - \frac{s_{xy}}{s_{xx}}\,\bar{x} \qquad \hat{b} = \frac{s_{xy}}{s_{xx}} \qquad \text{parametri della LSL}$$

$$\hat{y}_i := \hat{a} + \hat{b}x_i = \bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x}) \qquad \text{output della LSL}$$

$$\hat{e}_i := y_i - \hat{y}_i = (y_i - \bar{y}) - \frac{s_{xy}}{s_{xx}}(x_i - \bar{x}) \qquad \text{residui della LSL}$$

Varianza totale

$$ss_t := s_{yy} = \sum_i (y_i - \bar{y})^2 \qquad \textit{total sum of squares}$$

Varianza totale

$$ss_t := s_{yy} = \sum_i (y_i - \bar{y})^2 \qquad \textit{total sum of squares}$$

- $ss_t \geq 0$

Varianza totale

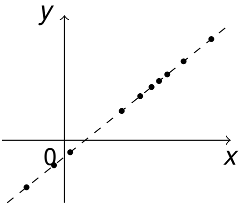
$$ss_t := s_{yy} = \sum_i (y_i - \bar{y})^2 \qquad \textit{total sum of squares}$$

- $ss_t \geq 0$
- ss_t non dipende dall'aver scelto il modello lineare

Varianza totale

$$ss_t := s_{yy} = \sum_i (y_i - \bar{y})^2 \qquad \textit{total sum of squares}$$

- $ss_t \geq 0$
- ss_t non dipende dall'aver scelto il modello lineare
- $ss_t \neq 0$ anche se tutti i punti stanno esattamente su una retta:

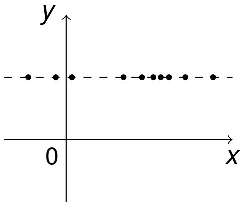


$ss_t = 11.98$

Varianza totale

$$ss_t := s_{yy} = \sum_i (y_i - \bar{y})^2 \qquad \textit{total sum of squares}$$

- $ss_t \geq 0$
- ss_t non dipende dall'aver scelto il modello lineare
- $ss_t \neq 0$ anche se tutti i punti stanno esattamente su una retta:



$ss_t = 0$

Varianza totale

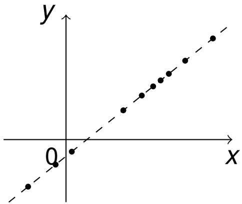
$$ss_t := s_{yy} = \sum_i (y_i - \bar{y})^2 \qquad \textit{total sum of squares}$$

- $ss_t \geq 0$
- ss_t non dipende dall'aver scelto il modello lineare
- $ss_t \neq 0$ anche se tutti i punti stanno esattamente su una retta
- $ss_t = ss_r + ss_e$, dove:
  - ss_r è intrinseca al modello lineare (varianza *spiegata*)
  - ss_e dipende dalla dispersione intorno alla LSL (varianza *residua*)

Varianza totale

$$ss_t := s_{yy} = \sum_i (y_i - \bar{y})^2 \qquad \textit{total sum of squares}$$

- $ss_t \geq 0$
- ss_t non dipende dall'aver scelto il modello lineare
- $ss_t \neq 0$ anche se tutti i punti stanno esattamente su una retta
- $ss_t = ss_r + ss_e$, dove:
  - ss_r è intrinseca al modello lineare (varianza *spiegata*)
  - ss_e dipende dalla dispersione intorno alla LSL (varianza *residua*)

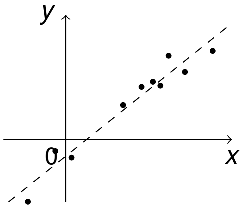


$ss_r = 11.98$
$ss_e = 0$

Varianza totale

$$ss_t := s_{yy} = \sum_i (y_i - \bar{y})^2 \qquad \textit{total sum of squares}$$

- $ss_t \geq 0$
- ss_t non dipende dall'aver scelto il modello lineare
- $ss_t \neq 0$ anche se tutti i punti stanno esattamente su una retta
- $ss_t = ss_r + ss_e$, dove:
  - ss_r è intrinseca al modello lineare (varianza *spiegata*)
  - ss_e dipende dalla dispersione intorno alla LSL (varianza *residua*)



$ss_r = 11.98$
$ss_e = 0.70$

Varianza residua

$$ss_e := \min_{(a,b)} L(a,b)$$

Varianza residua

$$ss_e \;:=\; \min_{(a,b)} L(a,b) \;=\; L(\hat{a},\hat{b})$$

Varianza residua

$$ss_e := \min_{(a,b)} L(a,b) = L(\hat{a},\hat{b}) = \sum_i (y_i - \hat{a} - \hat{b}x_i)^2$$

Varianza residua

$$
\begin{aligned}
ss_e &:= \min_{(a,b)} L(a,b) = L(\hat{a},\hat{b}) = \sum_i (y_i - \hat{a} - \hat{b}x_i)^2 \\
&= \sum_i (y_i - \hat{y}_i)^2
\end{aligned}
$$

Varianza residua

$$
\begin{aligned}
ss_e \;&:=\; \min_{(a,b)} L(a,b) \;=\; L(\hat{a},\hat{b}) \;=\; \sum_i (y_i - \hat{a} - \hat{b}x_i)^2 \\
&=\; \sum_i (y_i - \hat{y}_i)^2 \;=\; \sum_i \hat{e}_i^2
\end{aligned}
$$

Varianza residua

$$
\begin{aligned}
ss_e \;:=\; \min_{(a,b)} L(a,b) \;&=\; L(\hat{a},\hat{b}) \;=\; \sum_i (y_i - \hat{a} - \hat{b}x_i)^2 \\
&=\; \sum_i (y_i - \hat{y}_i)^2 \;=\; \sum_i \hat{e}_i^2 \qquad \textit{error sum of squares}
\end{aligned}
$$

Varianza residua

$$
\begin{aligned}
ss_e &:= \min_{(a,b)} L(a,b) = L(\hat{a},\hat{b}) = \sum_i (y_i - \hat{a} - \hat{b}x_i)^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{e}_i^2 \qquad \textit{error sum of squares}
\end{aligned}
$$

- $ss_e \geq 0$

Varianza residua

$$
\begin{aligned}
ss_e &:= \min_{(a,b)} L(a,b) = L(\hat{a},\hat{b}) = \sum_i (y_i - \hat{a} - \hat{b}x_i)^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{e}_i^2 \qquad \textit{error sum of squares}
\end{aligned}
$$

- $ss_e \geq 0$
- $ss_e = 0$ se e solo se tutti i punti stanno esattamente su una retta

Varianza residua

$$ss_e := \min_{(a,b)} L(a,b) = L(\hat{a},\hat{b}) = \sum_i (y_i - \hat{a} - \hat{b}x_i)^2$$

$$= \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{e}_i^2 \qquad \textit{error sum of squares}$$

- $ss_e \geq 0$
- $ss_e = 0$ se e solo se tutti i punti stanno esattamente su una retta:

$\Rightarrow$: $\quad ss_e = 0 \quad \Rightarrow \quad (y_i - \hat{a} - \hat{b}x_i)^2 = 0$ per ogni i

$\qquad\qquad\qquad \Rightarrow \quad y_i = \hat{a} + \hat{b}x_i$ per ogni i

Varianza residua

$$ss_e := \min_{(a,b)} L(a,b) = L(\hat{a},\hat{b}) = \sum_i (y_i - \hat{a} - \hat{b}x_i)^2$$

$$= \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{e}_i^2 \qquad \textit{error sum of squares}$$

- $ss_e \geq 0$
- $ss_e = 0$ se e solo se tutti i punti stanno esattamente su una retta:

$$\Rightarrow: \quad ss_e = 0 \quad \Rightarrow \quad (y_i - \hat{a} - \hat{b}x_i)^2 = 0 \text{ per ogni } i$$

$$\Rightarrow \quad y_i = \hat{a} + \hat{b}x_i \text{ per ogni } i$$

$$\Leftarrow: \quad y_i = a + bx_i \text{ per ogni } i \quad \Rightarrow \quad L(a,b) = 0$$

$$\Rightarrow \quad ss_e = \min_{(a,b)} L(a,b) = 0$$

Varianza residua

$$ss_e := \min_{(a,b)} L(a,b) = L(\hat{a},\hat{b}) = \sum_i (y_i - \hat{a} - \hat{b}x_i)^2$$

$$= \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{e}_i^2 \qquad \textit{error sum of squares}$$

- $ss_e \geq 0$
- $ss_e = 0$ se e solo se tutti i punti stanno esattamente su una retta:

$$\Rightarrow: \quad ss_e = 0 \quad \Rightarrow \quad (y_i - \hat{a} - \hat{b}x_i)^2 = 0 \text{ per ogni } i$$

$$\Rightarrow \quad y_i = \hat{a} + \hat{b}x_i \text{ per ogni } i$$

$$\Leftarrow: \quad y_i = a + bx_i \text{ per ogni } i \quad \Rightarrow \quad L(a,b) = 0$$

$$\Rightarrow \quad ss_e = \min_{(a,b)} L(a,b) = 0$$

- Se $ss_e = 0$, tutti i punti stanno sulla LSL.

Varianza spiegata

$$ss_r \;:=\; \sum_i (\bar{y} - \hat{y}_i)^2$$

Varianza spiegata

$$ss_r \ := \ \sum_i (\bar{y} - \hat{y}_i)^2 \ = \ \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right) \right]^2$$

Varianza spiegata

$$
\begin{aligned}
ss_r \;&:=\; \sum_i (\bar{y} - \hat{y}_i)^2 \;=\; \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}\,(x_i - \bar{x})\right)\right]^2 \\
&=\; \left(\frac{s_{xy}}{s_{xx}}\right)^2 \sum_i (x_i - \bar{x})^2
\end{aligned}
$$

Varianza spiegata

$$
\begin{aligned}
ss_r \;&:=\; \sum_i (\bar{y} - \hat{y}_i)^2 \;=\; \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}\,(x_i - \bar{x})\right)\right]^2 \\
&=\; \left(\frac{s_{xy}}{s_{xx}}\right)^2 \sum_i (x_i - \bar{x})^2 \;=\; \frac{s_{xy}^2}{s_{xx}}
\end{aligned}
$$

Varianza spiegata

$$
\begin{aligned}
ss_r &:= \sum_i (\bar{y} - \hat{y}_i)^2 = \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right) \right]^2 \\
&= \left(\frac{s_{xy}}{s_{xx}} \right)^2 \sum_i (x_i - \bar{x})^2 = \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}
\end{aligned}
$$

Varianza spiegata

$$
\begin{aligned}
ss_r \; &:= \; \sum_i (\bar{y} - \hat{y}_i)^2 \; = \; \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right) \right]^2 \\
&= \left(\frac{s_{xy}}{s_{xx}} \right)^2 \sum_i (x_i - \bar{x})^2 \; = \; \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}
\end{aligned}
$$

- $ss_r \geq 0$

Varianza spiegata

$$
\begin{aligned}
ss_r \;&:=\; \sum_i (\bar{y} - \hat{y}_i)^2 \;=\; \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}\,(x_i - \bar{x})\right)\right]^2 \\
&= \left(\frac{s_{xy}}{s_{xx}}\right)^2 \sum_i (x_i - \bar{x})^2 \;=\; \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}
\end{aligned}
$$

- $ss_r \geq 0$
- $ss_t = ss_r + ss_e$

Varianza spiegata

$$
\begin{aligned}
ss_r \;&:=\; \sum_i (\bar{y} - \hat{y}_i)^2 \;=\; \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}\,(x_i - \bar{x})\right)\right]^2 \\
&= \left(\frac{s_{xy}}{s_{xx}}\right)^2 \sum_i (x_i - \bar{x})^2 \;=\; \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}
\end{aligned}
$$

- $ss_r \geq 0$
- $ss_t = ss_r + ss_e$, perché

  $$ss_e \;:=\; \sum (y_i - \hat{y}_i)^2$$

Varianza spiegata

$$ss_r \ := \ \sum_i (\bar{y} - \hat{y}_i)^2 \ = \ \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right) \right]^2$$

$$= \left(\frac{s_{xy}}{s_{xx}} \right)^2 \sum_i (x_i - \bar{x})^2 \ = \ \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}$$

- $ss_r \geq 0$
- $ss_t = ss_r + ss_e$, perché

$$ss_e \ := \ \sum (y_i - \hat{y}_i)^2 \ = \ \sum \left[y_i - \left(\bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right) \right]^2$$

Varianza spiegata

$$
\begin{aligned}
ss_r \; &:= \; \sum_i (\bar{y} - \hat{y}_i)^2 \; = \; \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right) \right]^2 \\
&= \left(\frac{s_{xy}}{s_{xx}} \right)^2 \sum_i (x_i - \bar{x})^2 \; = \; \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}
\end{aligned}
$$

- $ss_r \geq 0$
- $ss_t = ss_r + ss_e$, perché

$$
\begin{aligned}
ss_e \; &:= \; \sum (y_i - \hat{y}_i)^2 \; = \; \sum \left[y_i - \left(\bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right) \right]^2 \\
&= \sum \left[(y_i - \bar{y}) - \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right]^2
\end{aligned}
$$

Varianza spiegata

$$ss_r := \sum_i (\bar{y} - \hat{y}_i)^2 = \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x})\right)\right]^2$$
$$= \left(\frac{s_{xy}}{s_{xx}}\right)^2 \sum_i (x_i - \bar{x})^2 = \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}$$

- $ss_r \geq 0$
- $ss_t = ss_r + ss_e$, perché

$$ss_e := \sum (y_i - \hat{y}_i)^2 = \sum \left[y_i - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x})\right)\right]^2$$
$$= \sum \left[(y_i - \bar{y}) - \frac{s_{xy}}{s_{xx}}(x_i - \bar{x})\right]^2 = s_{yy} - 2\frac{s_{xy}}{s_{xx}} s_{xy} + \left(\frac{s_{xy}}{s_{xx}}\right)^2 s_{xx}$$

Varianza spiegata

$$ss_r := \sum_i (\bar{y} - \hat{y}_i)^2 = \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x})\right)\right]^2$$

$$= \left(\frac{s_{xy}}{s_{xx}}\right)^2 \sum_i (x_i - \bar{x})^2 = \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}$$

- $ss_r \geq 0$
- $ss_t = ss_r + ss_e$, perché

$$ss_e := \sum (y_i - \hat{y}_i)^2 = \sum \left[y_i - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x})\right)\right]^2$$

$$= \sum \left[(y_i - \bar{y}) - \frac{s_{xy}}{s_{xx}}(x_i - \bar{x})\right]^2 = s_{yy} - 2\frac{s_{xy}}{s_{xx}} s_{xy} + \left(\frac{s_{xy}}{s_{xx}}\right)^2 s_{xx}$$

$$= s_{yy} - \frac{s_{xy}^2}{s_{xx}}$$

Varianza spiegata

$$
\begin{aligned}
ss_r \; := \; \sum_i (\bar{y} - \hat{y}_i)^2 \; &= \; \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right) \right]^2 \\
&= \left(\frac{s_{xy}}{s_{xx}} \right)^2 \sum_i (x_i - \bar{x})^2 \; = \; \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}
\end{aligned}
$$

- $ss_r \geq 0$
- $ss_t = ss_r + ss_e$, perché

$$
\begin{aligned}
ss_e \; := \; \sum (y_i - \hat{y}_i)^2 \; &= \; \sum \left[y_i - \left(\bar{y} + \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right) \right]^2 \\
&= \sum \left[(y_i - \bar{y}) - \frac{s_{xy}}{s_{xx}} (x_i - \bar{x}) \right]^2 \; = \; s_{yy} - 2 \frac{s_{xy}}{s_{xx}} s_{xy} + \left(\frac{s_{xy}}{s_{xx}} \right)^2 s_{xx} \\
&= \; s_{yy} - \frac{s_{xy}^2}{s_{xx}} \; = \; ss_t - ss_r
\end{aligned}
$$

Varianza spiegata

$$ss_r := \sum_i (\bar{y} - \hat{y}_i)^2 = \sum_i \left[\bar{y} - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x})\right)\right]^2$$

$$= \left(\frac{s_{xy}}{s_{xx}}\right)^2 \sum_i (x_i - \bar{x})^2 = \frac{s_{xy}^2}{s_{xx}} \qquad \textit{regression sum of squares}$$

- $ss_r \geq 0$
- $ss_t = ss_r + ss_e$, perché

$$ss_e := \sum (y_i - \hat{y}_i)^2 = \sum \left[y_i - \left(\bar{y} + \frac{s_{xy}}{s_{xx}}(x_i - \bar{x})\right)\right]^2$$

$$= \sum \left[(y_i - \bar{y}) - \frac{s_{xy}}{s_{xx}}(x_i - \bar{x})\right]^2 = s_{yy} - 2\frac{s_{xy}}{s_{xx}} s_{xy} + \left(\frac{s_{xy}}{s_{xx}}\right)^2 s_{xx}$$

$$= s_{yy} - \frac{s_{xy}^2}{s_{xx}} = ss_t - ss_r$$

Coefficiente di determinazione

$$r^2 := \frac{ss_r}{ss_t}$$

Coefficiente di determinazione

$$r^2 := \frac{SS_r}{SS_t} = \frac{SS_t - SS_e}{SS_t}$$

Coefficiente di determinazione

$$
\begin{aligned}
r^2 &:= \frac{SS_r}{SS_t} = \frac{SS_t - SS_e}{SS_t} \\
&= 1 - \frac{SS_e}{SS_t}
\end{aligned}
$$

Coefficiente di determinazione

$$
\begin{aligned}
r^2 &:= \frac{SS_r}{SS_t} = \frac{SS_t - SS_e}{SS_t} \\
&= 1 - \frac{SS_e}{SS_t} \qquad \textit{coefficiente di determinazione}
\end{aligned}
$$

Coefficiente di determinazione

$$
\begin{aligned}
r^2 &:= \frac{SS_r}{SS_t} = \frac{SS_t - SS_e}{SS_t} \\
&= 1 - \frac{SS_e}{SS_t} \qquad \textit{coefficiente di determinazione}
\end{aligned}
$$

- $0 \leq r^2 \leq 1$

Coefficiente di determinazione

$$r^2 := \frac{ss_r}{ss_t} = \frac{ss_t - ss_e}{ss_t}$$
$$= 1 - \frac{ss_e}{ss_t} \qquad \textit{coefficiente di determinazione}$$

- $0 \leq r^2 \leq 1$
- $r^2 \simeq 1 \quad \Rightarrow \quad$ i punti si dispongono bene su una retta

Coefficiente di determinazione

$$
\begin{aligned}
r^2 \;&:=\; \frac{ss_r}{ss_t} \;=\; \frac{ss_t - ss_e}{ss_t} \\
&=\; 1 - \frac{ss_e}{ss_t} \qquad \textit{coefficiente di determinazione}
\end{aligned}
$$

- $0 \leq r^2 \leq 1$
- $r^2 \simeq 1 \quad \Rightarrow \quad$ i punti si dispongono bene su una retta
- r^2 è la "percentuale di variabilità spiegata dal modello lineare"

Coefficiente di determinazione

$$r^2 := \frac{SS_r}{SS_t} = \frac{SS_t - SS_e}{SS_t}$$
$$= 1 - \frac{SS_e}{SS_t} \qquad \textit{coefficiente di determinazione}$$

- $0 \leq r^2 \leq 1$
- $r^2 \simeq 1 \quad \Rightarrow \quad$ i punti si dispongono bene su una retta
- r^2 è la "percentuale di variabilità spiegata dal modello lineare"
- $r^2 = \dfrac{s_{xy}^2 / s_{xx}}{s_{yy}}$

Coefficiente di determinazione

$$
\begin{aligned}
r^2 &:= \frac{ss_r}{ss_t} = \frac{ss_t - ss_e}{ss_t} \\
&= 1 - \frac{ss_e}{ss_t} \qquad \textit{coefficiente di determinazione}
\end{aligned}
$$

- $0 \leq r^2 \leq 1$
- $r^2 \simeq 1 \quad \Rightarrow \quad$ i punti si dispongono bene su una retta
- r^2 è la “percentuale di variabilità spiegata dal modello lineare”
- $r^2 = \dfrac{s_{xy}^2 / s_{xx}}{s_{yy}} = \dfrac{s_{xy}^2}{s_{xx} s_{yy}}$

Coefficiente di correlazione

$$r := \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}}$$ *coefficiente di correlazione*

Coefficiente di correlazione

$$r := \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} \qquad \textit{coefficiente di correlazione}$$

- $-1 \leq r \leq 1$

Coefficiente di correlazione

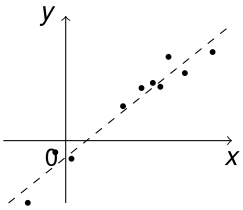
$$r := \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} \qquad \textit{coefficiente di correlazione}$$

- $-1 \leq r \leq 1$
- r^2 è il coefficiente di determinazione

Coefficiente di correlazione

$$r := \frac{s_{xy}}{\sqrt{s_{xx} s_{yy}}} \qquad \textit{coefficiente di correlazione}$$

- $-1 \leq r \leq 1$
- r^2 è il coefficiente di determinazione
- $r > 0.9 \quad \Rightarrow$ i dati approssimano una retta di pendenza positiva

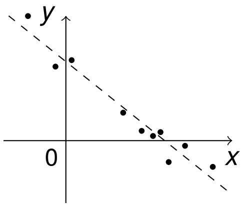


$r = 0.972$
$r^2 = 0.945$

Coefficiente di correlazione

$$r := \frac{s_{xy}}{\sqrt{s_{xx}s_{yy}}} \qquad \textit{coefficiente di correlazione}$$

- $-1 \leq r \leq 1$
- r^2 è il coefficiente di determinazione
- $r > 0.9 \quad \Rightarrow$ i dati approssimano una retta di pendenza positiva
- $r < -0.9 \quad \Rightarrow$ i dati approssimano una retta di pendenza negativa



$$r = -0.972$$
$$r^2 = 0.945$$