Regressione Lineare
parte 2

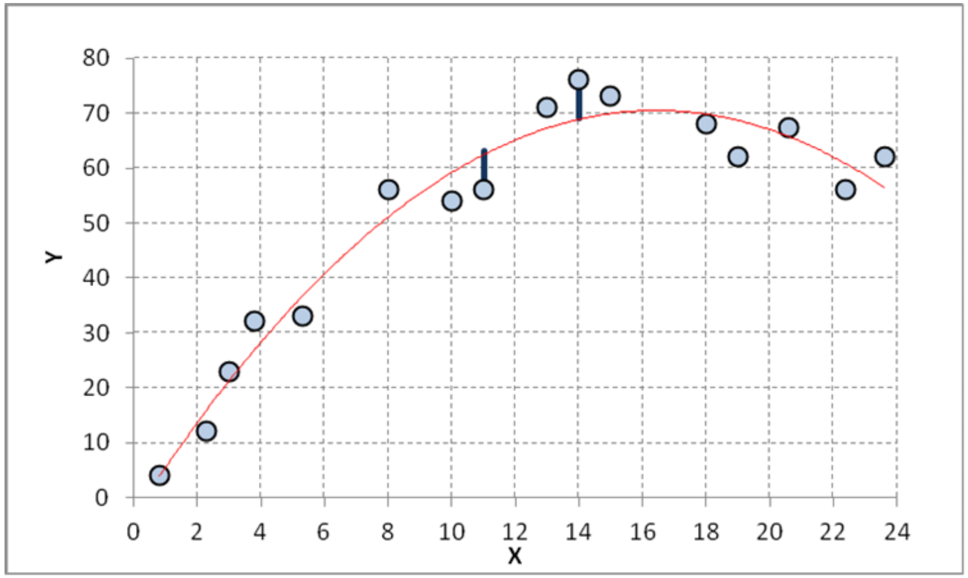
Corso di
Misure Meccaniche e Termiche

David Vetturi

Varianza residua:

Il modello generato attraverso il metodo dei minimi quadrati descrive la variabilità della grandezza Y che può essere pensata essere funzione diretta di X.

La quantità ε^2 indica quanto il modello non è in grado di spiegare la variabilità delle Y complessivamente. Tale "errore" può essere mediato fra tutti gli m punti osservati e dunque

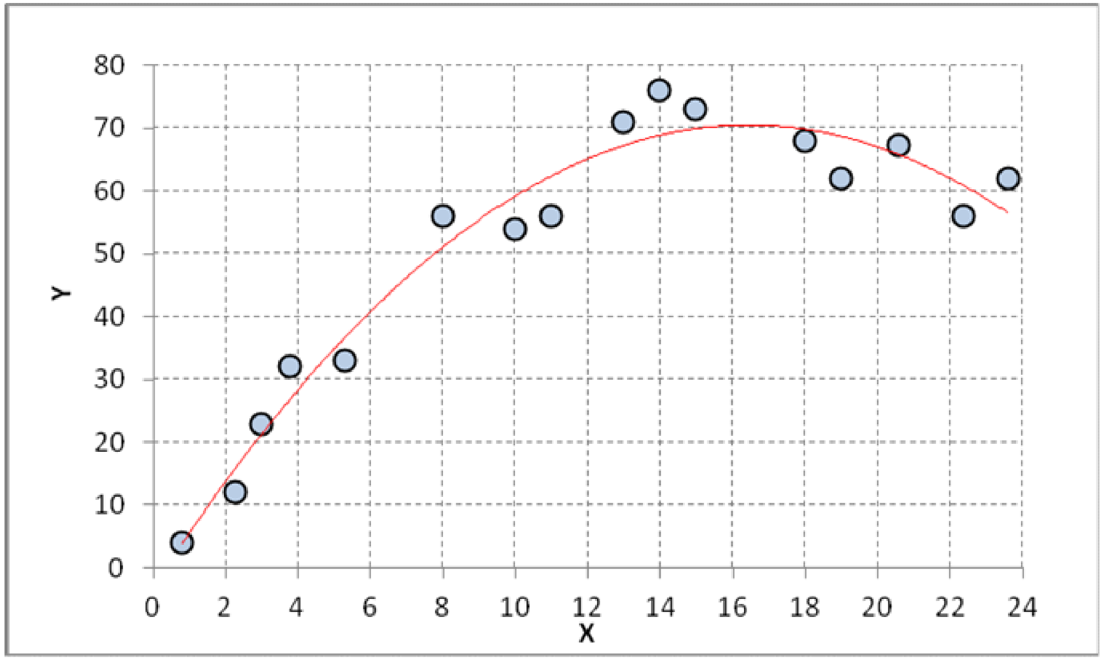


$$\varepsilon^2 = \sum_{k=1}^{m} \varepsilon_k^2 = \sum_{k=1}^{m} \left(y_k - y(x_k)\right)^2$$

$$\sigma_0^{\ 2} = \frac{\sum_{k=1}^{m} \varepsilon_k^2}{m-n} = \frac{\sum_{k=1}^{m} \left(y_k - y(x_k)\right)^2}{m-n}$$

Regressione lineare

Il metodo della stima ai Minimi Quadrati presentato permette di valutare la ennupla di parametri (α_1, α_2 … α_n) che identifica il modello funzionale di relazione fra Y e X all'interno delle funzioni di uno spazio vettoriale generato da una base di funzioni (φ_1, φ_2 … φ_n)



Si può pensare cha la serie di m punti (x_k,y_k) sia estratta dalla distribuzione congiunta $f_{xy}(x,y)$ e che ne rappresenti un campione.

Per ogni campione estratto dalla popolazione X-Y il metodo associa una ennupla $(\alpha_1, \alpha_2 \ldots \alpha_n)$ che può essere vista come una variabile casuale a n-dimensioni funzione (attraverso il metodo) della VC doppia X-Y

La quantità σ_0^2 indica la variabilità non spiegata dal modello e permette di valutare la variabilità dei parametri α (variabile casuale a n-dimensioni).

Varianza dei parametri α:

Nel caso in cui le funzioni di base risultano fra loro ortogonali le quantità αi risultano tra loro indipendenti e la loro varianza vale:

$$\sigma_{\alpha_i'}{}^2 = \frac{\sigma_0^2}{\sum_{k=1}^{m} \varphi_i'(x_k)^2}$$

Mentre nel caso generale in cui le funzioni di base non sono fra loro ortogonali si ha che la matrice di covarianza della variabile casuale α a n dimensioni è data da:

$$C_{\alpha\alpha} = \sigma_0{}^2 \cdot A^{-1}$$

dove $C\alpha\alpha$ è la matrice di covarianza e A la matrice del sistema lineare che permette di valutare i parametri α

Varianza del modello:

Analogamente al caso dei parametri α, anche il modello che questi definiscono è funzione del “campione” estratto da X-Y.

La varianza dei parametri può essere dunque propagata sul modello y(x).

Per ogni valore di x si ha:

$$\sigma_{y(x)}{}^{2}=\sum_{i=1}^{n}\varphi_i'(x)^2\cdot\sigma_{\alpha_i'}^2$$

Varianza del modello:

Mentre nel caso generale in cui le funzioni di base non sono fra loro ortogonali si ha:

$$\sigma_{y(x)}{}^2 = [\varphi] \cdot C_{\alpha\alpha} \cdot [\varphi]^T$$

dove

$$[\varphi] = [\varphi_1(x) \quad \varphi_2(x) \quad .. \quad \varphi_n(x)]$$

Varianza della previsione:

Se il modello viene utilizzato per fare previsioni, ovvero se viene utilizzato per valutare il valore di *y* in corrispondenza di una certa ascissa *x* si ha che le cause di variabilità sono due:

- la variabilità del modello $\sigma^2_{y(x)}$

- la variabilità delle Y non spiegata dal modello σ^2_0

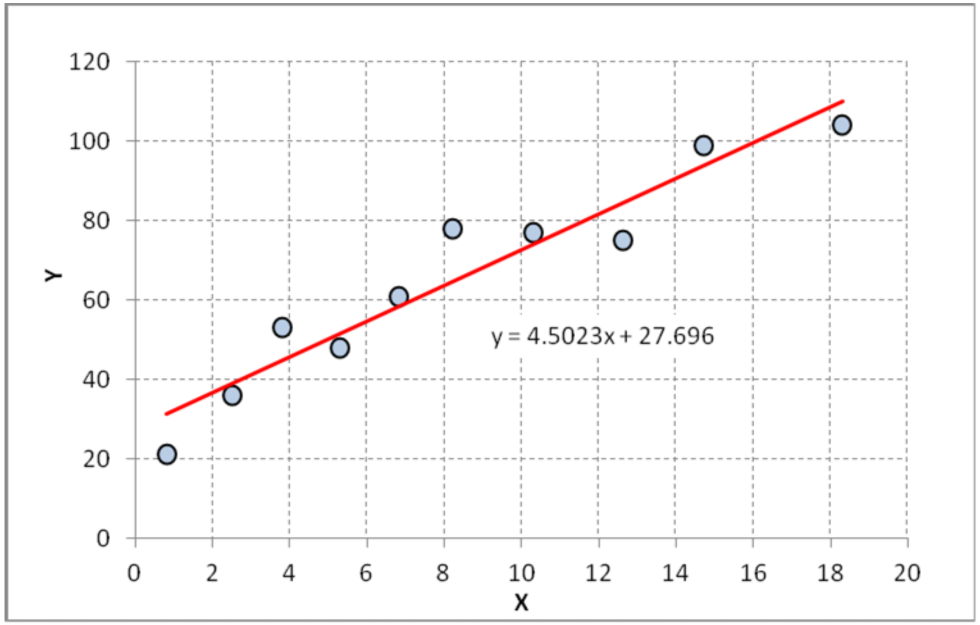
e dunque:

$$\sigma_{y_{prev}}{}^2 = \sigma_{y(x)}{}^2 + \sigma_0^2$$

Retta ai minimi quadrati:
esempio numerico - continuazione

Come mostrato precedentemente è possibile calcolare i parametri della retta ai minimi quadrati a partire dai punti (x_k, y_k) dati. Lo scarto indica quanto il modello non è in grado di spiegare la variabilità delle ordinate.

x	y	y(x)	scarto
0.8	21	31.30	-10.30
2.5	36	38.95	-2.95
3.8	53	44.80	8.20
5.3	48	51.56	-3.56
6.8	61	58.31	2.69
8.2	78	64.61	13.39
10.3	77	74.07	2.93
12.6	75	84.42	-9.42
14.7	99	93.88	5.12
18.3	104	110.09	-6.09



Come abbiamo visto la quantità ε^2 indica quanto il modello non è in grado di spiegare la variabilità delle Y complessivamente e in questo caso:

$$\varepsilon^2 = \sum_{k=1}^{m} \varepsilon_k^2 = \sum_{k=1}^{m} \left(y_k - y(x_k)\right)^2 = 541.7$$

e dunque la varianza non spiegata dal modello è:

$$\sigma_0^{\ 2} = \frac{\sum_{k=1}^{m} \varepsilon_k^2}{m-n} = \frac{\sum_{k=1}^{m} \left(y_k - y(x_k)\right)^2}{m-n} = \frac{541.7}{8} = 67.7$$

Come abbiamo visto la quantità ε^2 indica quanto il modello non è in grado di spiegare la variabilità delle Y complessivamente e in questo caso:

$$\varepsilon^2 = \sum_{k=1}^{m} \varepsilon_k^2 = \sum_{k=1}^{m} \left(y_k - y(x_k)\right)^2 = 541.7$$

e dunque la varianza non spiegata dal modello è:

$$\sigma_0^{\ 2} = \frac{\sum_{k=1}^{m} \varepsilon_k^2}{m-n} = \frac{\sum_{k=1}^{m} \left(y_k - y(x_k)\right)^2}{m-n} = \frac{541.7}{8} = 67.7$$

I parametri stimati α_1 e α_2 che sono stati valutati sono:

$$\alpha_1' = \frac{\sum_{k=1}^{m} y(x_k)}{m} = \frac{652}{10} = 65.2 \quad \alpha_2' = \frac{\sum_{k=1}^{m} y(x_k)\cdot(x_k - \bar{x})}{\sum_{k=1}^{m}(x_k - \bar{x})^2} = \frac{1282.44}{284.841} = 4.50$$

e la varianza dei due parametri sono:

$$\sigma_{\alpha_1'}{}^2 = \frac{\sigma_0^2}{\sum_{k=1}^{m} \varphi_1'(x_k)^2} = \frac{67.7}{10} = 6.77 \quad \sigma_{\alpha_2'}{}^2 = \frac{\sigma_0^2}{\sum_{k=1}^{m} \varphi_2'(x_k)^2} = \frac{67.7}{284.841} = 0.2377$$

Retta ai minimi quadrati:
esempio numerico – varianza del modello

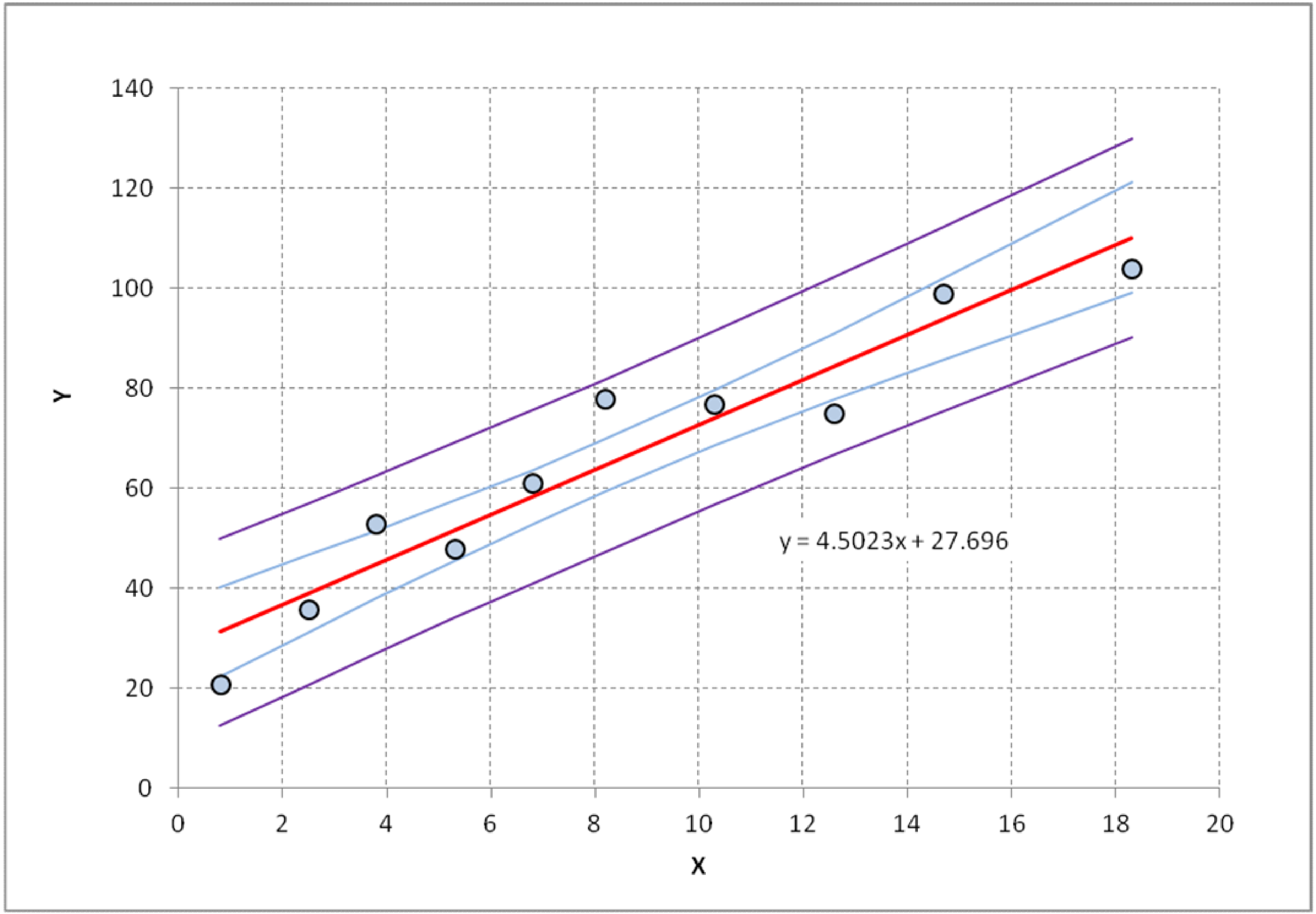
Come è stato indicato precedentemente la varianza dei parametri può essere dunque propagata sul modello y(x). Analogamente per la varianza della previsione effettuata a partire dal modello.

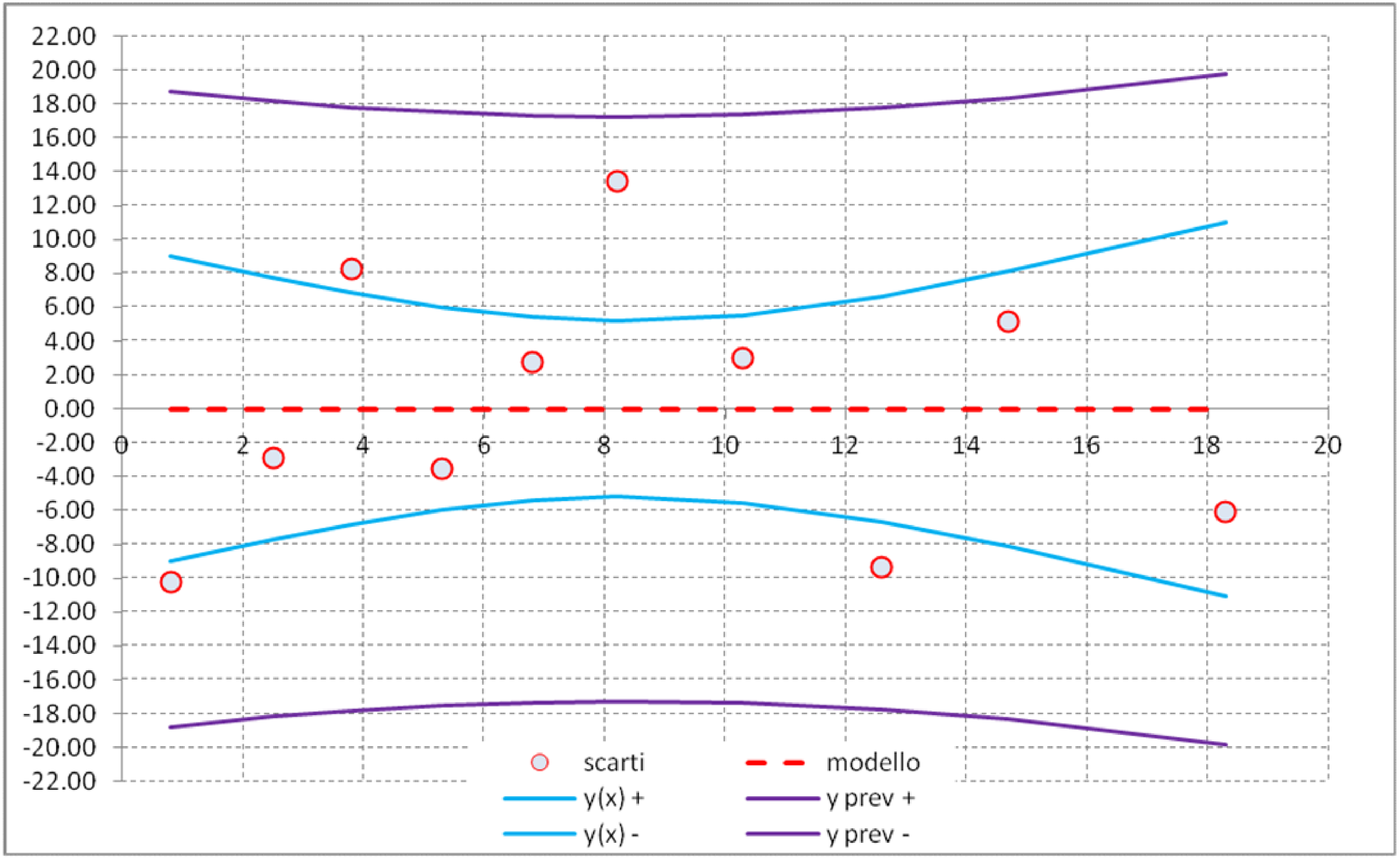
Per ogni valore di x si ha:

$$\sigma_{y(x)}{}^2 = \sum_{i=1}^{n} \varphi_i'(x)^2 \cdot \sigma_{\alpha_i'}^2 \qquad \sigma_{y_{prev}}{}^2 = \sigma_{y(x)}{}^2 + \sigma_0^2$$

attorno al modello possono essere individuate due fasce di ampiezza $k \cdot \sigma$ che individuano rispettivamente:

- inviluppo dei modelli probabili per la relazione y=y(x)
- luogo dei punti ipotizzabili come previsioni di y a partire dal modello y(x)


140
120
100
80
60
40
20
0
Y
y = 4.5023x + 27.696
0
2
4
6
8
10
12
14
16
18
20
X


22.00
20.00
18.00
16.00
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
-2.00
-4.00
-6.00
-8.00
-10.00
-12.00
-14.00
-16.00
-18.00
-20.00
-22.00
0
2
4
6
8
10
12
14
16
18
20
scarti
modello
y(x) +
y prev +
y(x) -
y prev -