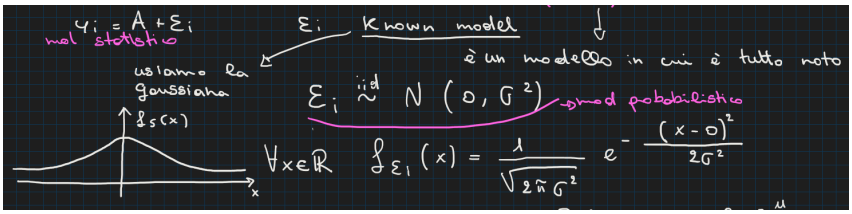
Spiegazione MLE, Approccio Bayesiano, Dimostrazione test MSE - 17/09 + 19/09

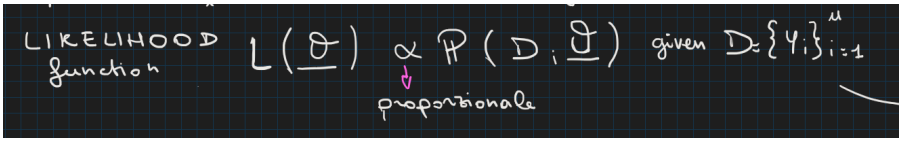
Maximum Likelihood Estimator (MLE) - 17/09

Ci serviranno strumenti per costruire un nostro stimatore a partire dai dati.
L'approccio classico richiede un po' di calcoli, si basa sul concetto di verosimiglianza e si chiama stimatore a massima
verosimiglianza. Maximum Likelihood estimator (MLE).



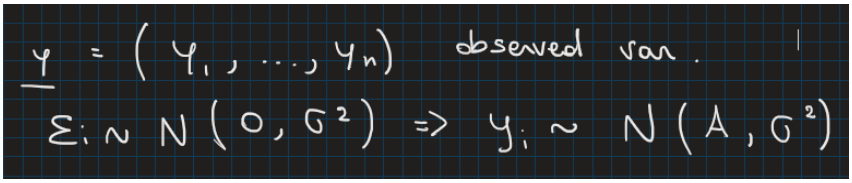
Dobbiamo conoscere il modello probabilistico di epsilon con i per usare MLE.
Questo è il classico approccio parametrico, ho tutto le forme note, il modello probabilistico eccetera.

Il nostro obiettivo è trovare A. (θ è A, ha cambiato il come ma è a stessa cosa)



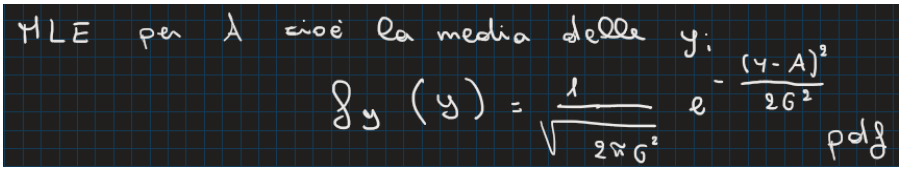
Avremo una funzione di Likelihood L che sarà funzione del parametro theta (che può essere un vettore).

Per prima cosa si osservano i dati, le mie y, possono essere un vettore ovviamente. I dati sono anche detti **variabile osservata**.



La somma di una costante non modifica la varianza, sposta solo la PDF.

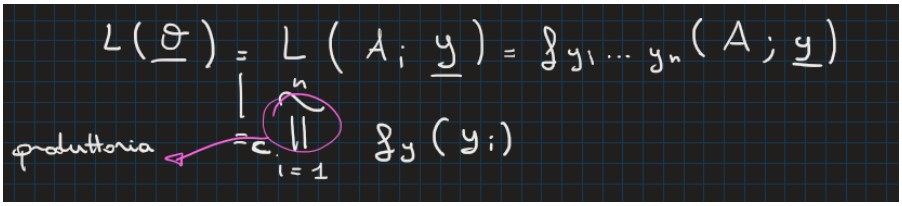
Dobbiamo costruire il MLE (Maximum Likelihood estimator) per il parametro A. Il parametro A è la media delle y_i.



Conosciamo la PDF delle y che è uguale a quella delle epsilon solo che la media non sarà 0 ma A.

> 💡 La verosimiglianza sarà funzione della A e dei dati raccolti.

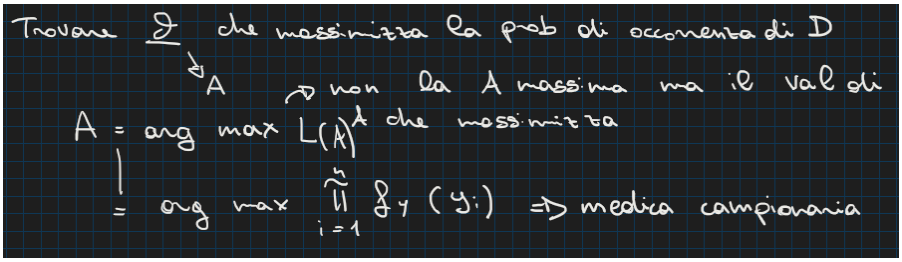
Per scrivere la verosimiglianza bisogna ricordare come si scrive un modello probabilistico, la verosimiglianza è la PDF congiunta delle nostre y.



I dati sono iid, questo ci importa perché se le x sono indipendenti le y sono indipendenti. La congiunta di variabili aleatorie indipendenti è il prodotto delle marginali.

Visto che le marginali sono tutte uguali possiamo scrivere una produttoria.

La costante moltiplicativa sposta in basso o in alto il tutto ma non è importante, a noi interessa trovare il massimo della funzione ma questo massimo non è influenzato dalla costante c (c indica una costante generica, penso sia riferita al contributo di A che sposta la Gaussiana ma non ne cambia il massimo quindi è irrilevante).



Dobbiamo trovare il valore A per cui la funzione likelihood viene massimizzata.

(theta è in generale, nel nostro caso è A)

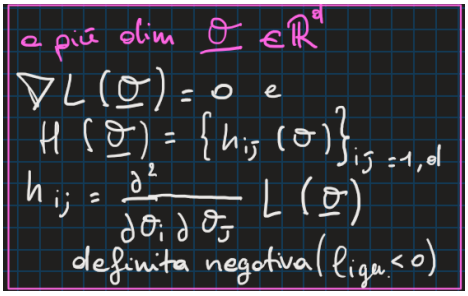
Arg max sarebbe il valore di A che massimizza la funzione.

Piccolo spoiler: il valore che ne uscirà sarà la media campionaria.

19/09

Riprendendo il discorso sul massimizzare $L(A)$, dobbiamo trovare il massimo di una funzione, il che solleva la domanda di come si trova il massimo di una funzione, ovviamente basta porre la derivata uguale a 0.
A questo punto calcolo la derivata seconda e ci interessa che sia negativa. Questo nel caso monodimensionale.

Esempio multidimensionale:



Per theta vettore a più dimensioni il gradiente (equivalente a più dimensioni della derivata prima, sono tipo tutte le derivate) di L di theta deve essere uguale al vettore nullo e la matrice Hessiana (matrice delle derivate rispetto a theta i e theta j, è l'equivalente a più dimensioni della derivata seconda) deve essere definita negativa (visto che è quadrata pari simmetrica basta che gli autovalori siano tutti dello stesso segno e negativi).

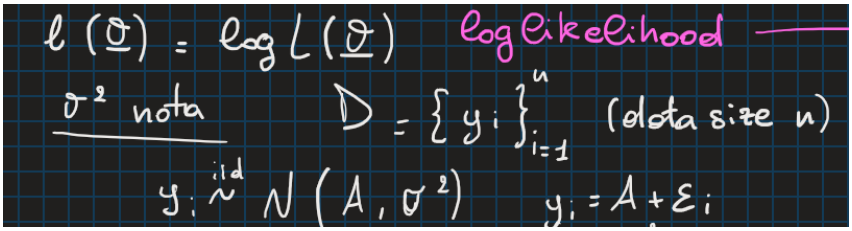
Invece della funzione verosimiglianza usiamo il log della funzione verosimiglianza, si può fare perché il log non cambia il massimo.

▼ Perché si usa la log likelihood?

- **Semplificazione dei calcoli**:
  - La likelihood è spesso il prodotto di molte probabilità (o densità), che possono diventare numeri molto piccoli, portando a problemi di precisione numerica. La log-likelihood trasforma questo prodotto in una somma, che è più facile da gestire matematicamente e computazionalmente.
  - Le derivate della log-likelihood (necessarie per ottimizzare o stimare parametri) sono più semplici da calcolare rispetto a quelle della likelihood.
- **Equivalenza nell'ottimizzazione**:
  - La log-likelihood conserva l'ordine di grandezza della likelihood. Massimizzare la log-likelihood è equivalente a massimizzare la likelihood, poiché il logaritmo è una funzione monotona crescente. Quindi, entrambe portano agli stessi parametri stimati.
- **Stabilità numerica**:

- La likelihood può diventare estremamente piccola (vicina a zero), soprattutto con dataset grandi, il che può causare underflow numerico. Prendere il logaritmo di valori piccoli evita questi problemi.

- **Additività**:
  - Il logaritmo trasforma il prodotto delle probabilità in una somma. Questo rende i modelli più facili da analizzare e permette di applicare tecniche di ottimizzazione standard (come il gradiente o il metodo di Newton).

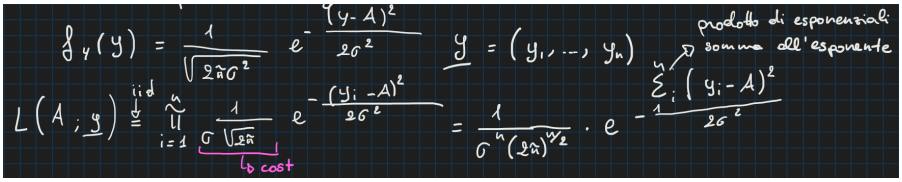


Vogliamo stimare A.

Supponiamo che sigma quadro è noto, il caso incognito è quando sia A che sigma quadro sono da individuare, qui invece supponiamo di averlo per semplificare ma è poco realistico averlo.

Vettore y raccoglie tutti i dati da uno ad n, è il data set.

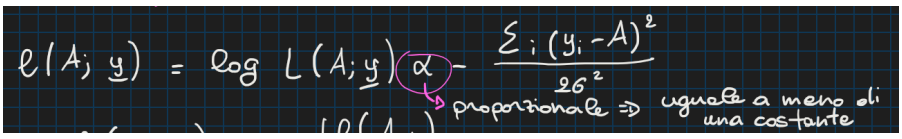
Dimostrazione che la stima MLE è uguale alla media campionaria per il caso del calcolo della media di una gaussiana



Possiamo cacciare fuori dalla produttoria la parte costante, ora abbiamo la produttoria di esponenziali che sarebbe l'esponenziale con esponente la somma degli esponenti.

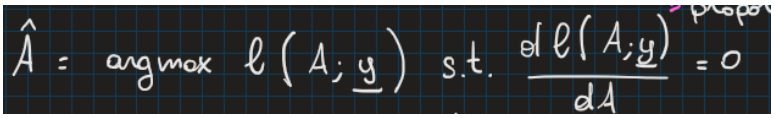
Questa espressione ci fa capire perché passiamo ai logaritmi.

Inoltre la costante visto che dobbiamo derivare per A si perderebbe non ci serve. Il fatto che ignoriamo la costante si dice che scriviamo un termine di proporzionalità ad A.

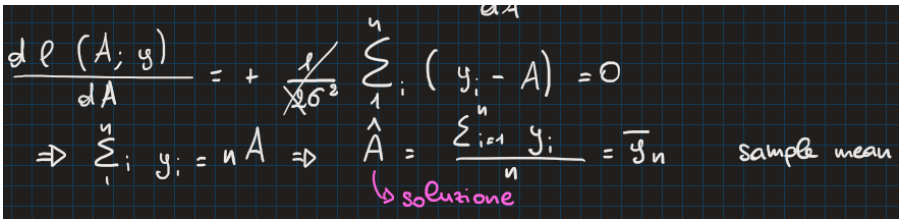


Ora calcoliamo il massimo facendo la derivata della log likelihood.
Cerchiamo A cappello che non sarà altro che l'arg max della log likelihood cioè il punto tale che la derivata di L rispetto ad A sia uguale a 0.

Cerchiamo A tale che questa espressione è uguale a 0, questa sarà la A che massimizza la funzione (dopo aver verificato le condizioni sulla derivata seconda).



s.t. = such that



Abbiamo portato la parte di sommatoria relativa ad A dall'altro lato (nA va a destra).

La stima della media (sappiamo per dimostrazioni precedenti che A è la media) è la media campionaria (sample mean) intuitivamente, tutti questi calcoli lo confermano.

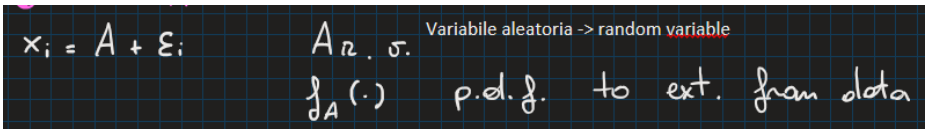
La A che risolve l'equazione è la A cappello.

In questo caso A, sigma quadro e i modelli sono noti, io devo arrivare alla regressione, in futuri casi avremo tutto incognito e aleatorio e in quei casi la verosimiglianza non vale.

Ci servirà l'approccio supervisionato che tramite le y e le x ci permette di ottenere la f(x) che per noi ora è aleatoria.

Ci serve un approccio diverso che è l'approccio Bayesiano.

Bayesian approach



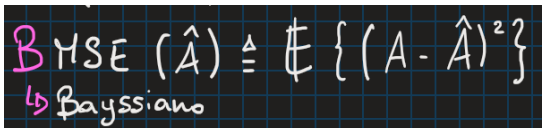
Nel mio modello semplice semplice, $x_1 = A + \epsilon_i$, ora anche A è una variabile aleatoria.

Così posso inglobare il fatto che può esserci aleatorietà anche in quello che devo stimare. La A avrà quindi una sua PDF che volendo va stimata a partire dai dati.

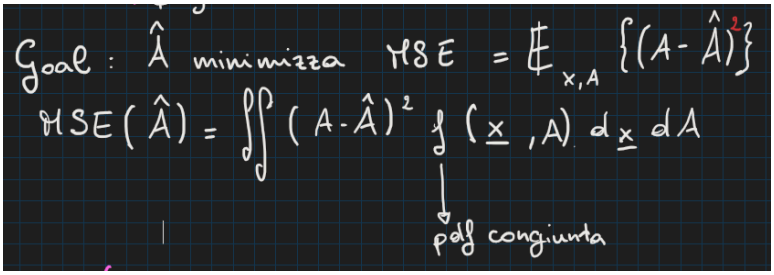
Cerchiamo sempre lo stimatore, lo stimatore Bayesiano in questo caso.
Lo stimatore sarà sempre funzione dei dati (che possiamo scrivere come x o come y, è solo una lettera).

$\hat{A}$ visto che va ricavata dai dati sarà quindi funzione del vettore dei dati x.

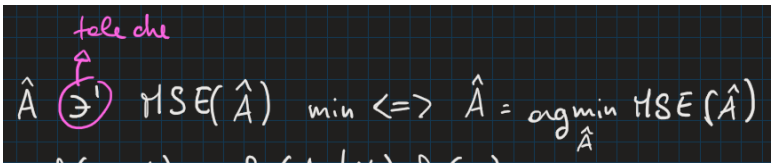
Anche ora cerchiamo lo stimatore che minimizza l'errore. Continuiamo a usare l'errore quadratico ma ce ne sarebbero altri.



Talvolta si segna BMSE per indicare che è da trattare con l'approccio Bayesiano.
BMSE resta la media statistica di A meno A cappello al quadrato.
Obiettivo: trovare A cappello che minimizza l'MSE.

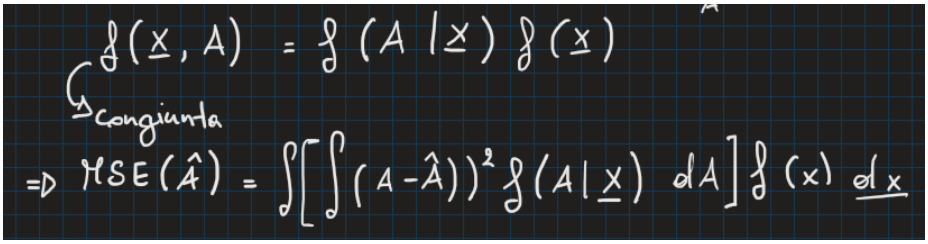


Prima il valore atteso in questione agiva solo sui dati x ma ora che anche A è random variable questo agisce anche su di A e quindi c'è un doppio integrale.



Arg min e arg max sono il minimizzatore e il massimizzatore.

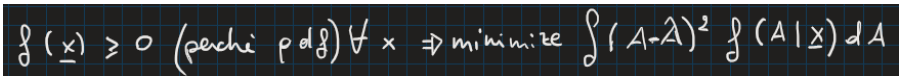
Consideriamo la pdf congiunta visto che abbiamo detto che il valore atteso è su due variabili.
(Congiunta = joint distribution)



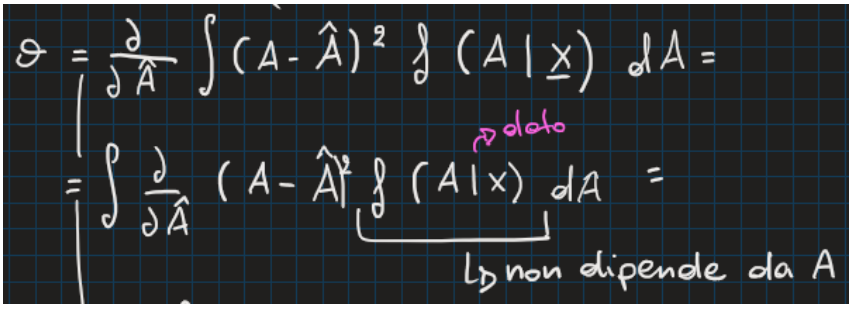
La congiunta si può sempre scrivere come probabilità di eventi condizionati.

Con integrali doppi può essere comodo dividerli facendo prima all'interno un'espressione e poi il risultato nell'integrale esterno.

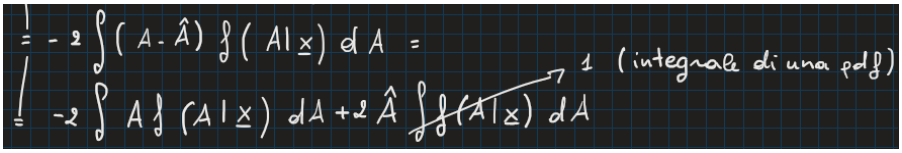
Una proprietà di una pdf è che non può essere negativa. Questo vuol dire che se voglio minimizzare l'MSE mi basta minimizzare la funzione interna.



Per trovare il minimo di nuovo devo fare la derivata e porla uguale a 0.



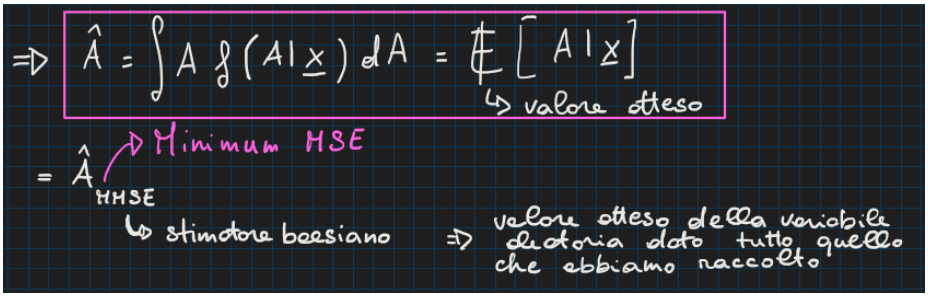
La PDF condizionata è indipendente da A cappello quindi la derivata non ha effetto su di lei.



il -2 viene dalla derivazione

L'integrale della pdf su tutto il suo dominio fa 1.
Quindi rimane solo A cappello.

A questo punto ponendo tutto questo uguale a 0 trovo subito A cappello.



Quello che abbiamo trovato è il valore atteso di A dato vettore x.

> 💡 Lo stimatore Bayesiano è chiamato anche in ingegneria stimatore a minimo MSE.

f(A|x) o la conosco o in qualche modo devo stimarla.

Supponiamo di doverla stimare.

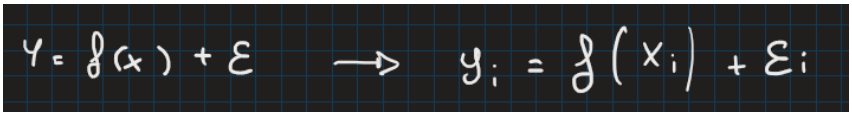
Questa è la pdf di A dopo aver raccolto i dati, quella prima si chiamava conoscenza a priori (f(A)), ora è conoscenza a posteriori.

In genere priori e posteriori dovrebbero somigliarsi.

Teorema di Bayes legava, del resto, conoscenze a priori e conoscenze a posteriori, il tramite è la verosimiglianza.

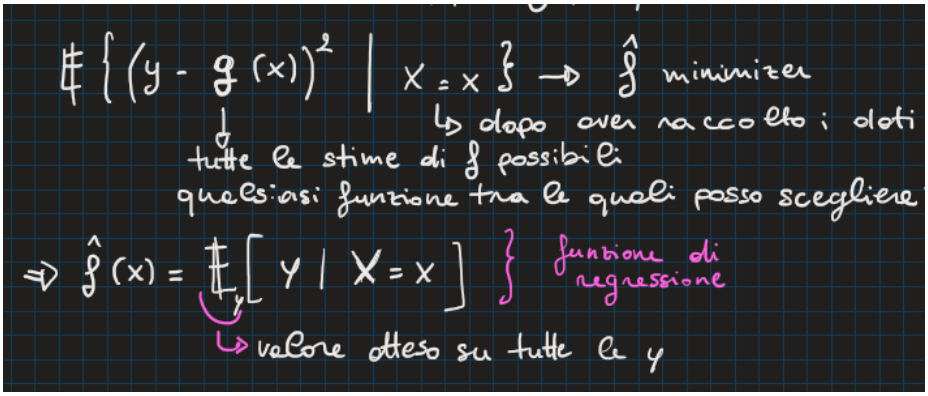
Approccio classico (verosomiglianza) vs Approccio Bayesiano (che ha anche le informazioni a priori).

Errore di generalizzazione - dimostrazione



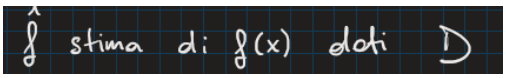
Modello della popolazione. L'espressione campionaria. Ora al posto di A c'è f(x) che è una funzione incognita.

Per approssimare bene f faccio il valore atteso della differenza tra la realtà e l'approssimazione.

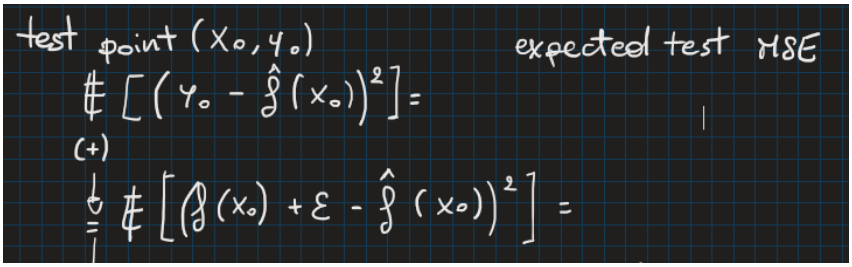


F(x) resta una variabile aleatoria, per la variabile aleatoria abbiamo detto che lo stimatore migliore è quello che minimizza il BMSE.

F è teorica e ha generato i dati, non la conoscerò mai al 100 per cento, conoscerò f cappello, g di x invece sono tutte le funzioni dello spazio delle funzioni di f.

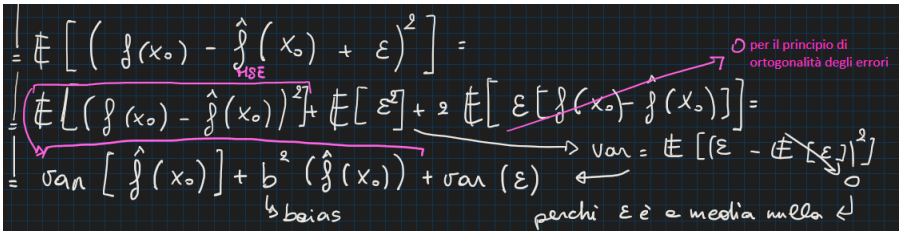


Consideriamo un test set costituito da un solo punto un test point. (vogliamo trovare il test mse).



sostituiamo y_0 con la sua definizione

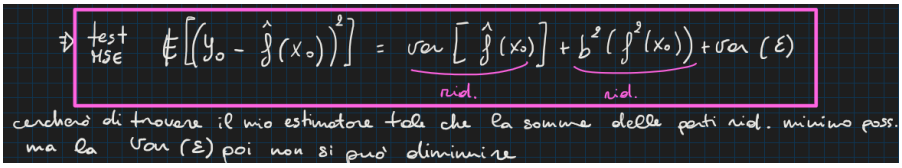
Svolgiamo i quadrati.



Si può dimostrare che il termine epsilon per la differenza tra f ed f cappello di x_0 è uguale a 0, si chiama principio di ortogonalità degli errori per epsilon, la distanza tra f ed f cappello.

F cappello di x_0 meno $f(x_0)$ è un termine che riguarda il valore atteso di questa quantità.

Ricordiamo che Epsilon è a media nulla.



Oltre al classico termine di compromesso bias varianza c'è anche la varianza di epsilon, un errore irriducibile.

In statistica si sostituiscono alle medie le medie campionarie.