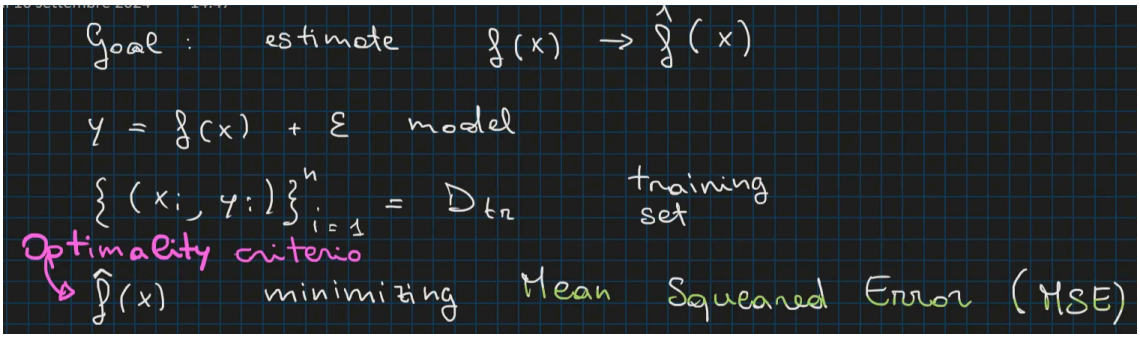
MSE, Stima del parametro deterministico A - 17/09

Problema supervisionato

Il nostro scopo è la stima di f, quindi ottenere $\hat{f}$, tale che il MSE sia minimo.



Stimare f è proprio effettuare il training, in questo caso vediamo sia x_i che y_i nel training set, quindi si tratta di un problema supervisionato.

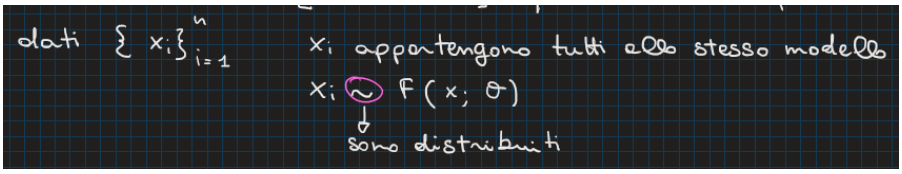
Ci sono varie metriche per la misura dell'errore, quella classica è la media quadratica visto che l'errore può essere sia positivo che negativo.

Stimare i parametri dato il modello

(qua si inizia un altro discorso da zero)

Il problema più semplice è: noto il modello stimare i parametri.

Non conoscendo μ e σ (media e varianza) il problema è bidimensionale, noi iniziamo da un caso monodimensionale dove dobbiamo trovare θ e ci interessa stimare la media della Gaussiana.

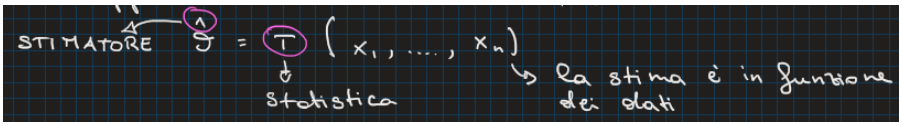


Abbiamo dei dati del modello di nostro interesse che chiamiamo x_i, questi dati sono distribuiti (tilde ~) su una certa CDF (ci potrebbe dare la PDF, o la PMF se è un caso discreto).

Mi seve uno **stimatore di** θ.

Nell'approccio classico θ è un parametro deterministico che però noi non conosciamo.

$\hat{\theta}$ lo stimiamo con il modello e i dati raccolti. L'stimatore è funzione T dei dati raccolti.



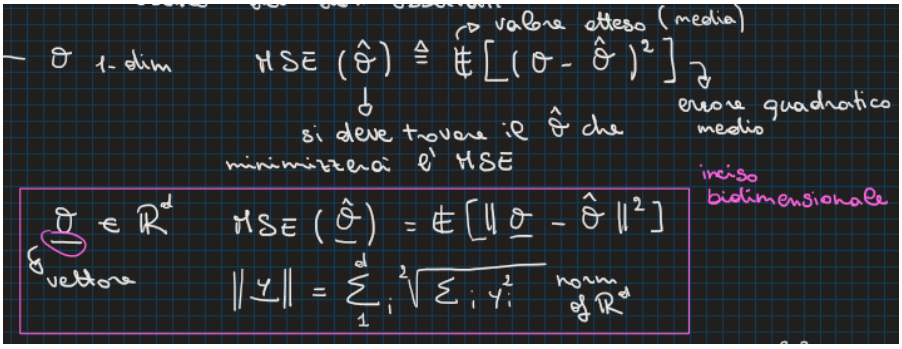
Lo stimatore se considero le maiuscole (variabili aleatorie) abbiamo una funzione di variabili aleatorie e quindi $\hat{\theta}$ sarà una variabile aleatoria.
Se invece considero proprio i valori numerici avremo una funzione che ci restituisce una variabile deterministica.

> Lo **stimatore** è la funzione che, essenzialmente, se ci mettiamo i dati otteniamo dei risultati cioè la stima.

> La **stima** è un numero prodotto dallo stimatore quando si usano i dati osservati

Errore quadrativo medio - MSE (mean squared error)

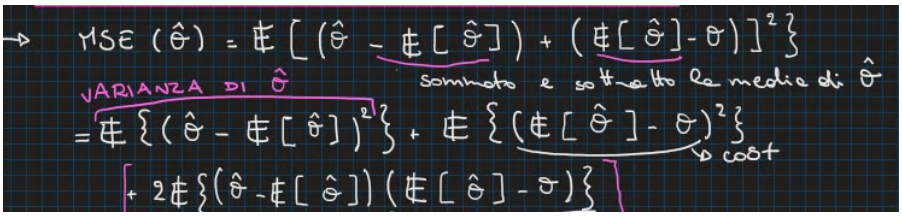
> Lo **stimatore ottimo** è quello che minimizza l' MSE per il parametro theta. Theta è monodimensionale cioè un numero.



definizione di MSE($\hat{\theta}$) in caso monodimensionale e multidimensionale

MSE in funzione di varianza e bias

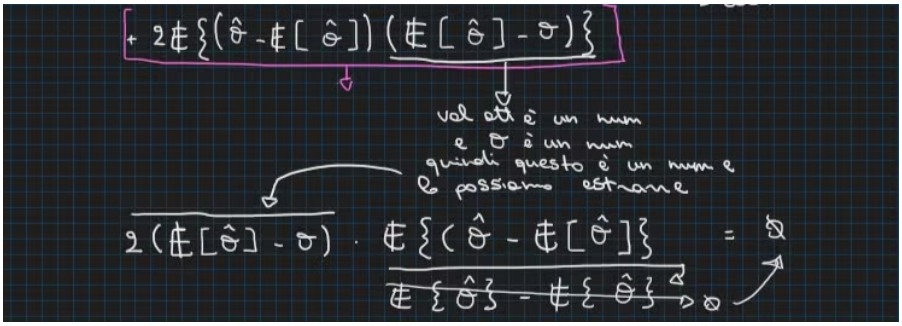
Segue la dimostrazione



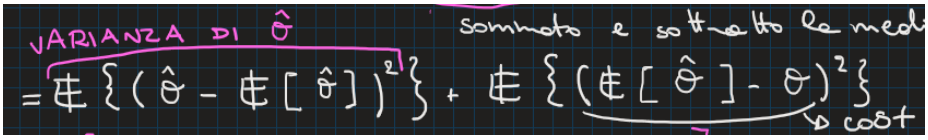
Sommo e sottraggo lo stesso valore (E[theta cappello]), poi associo la somma e sottrazione per comodità.

Essendo la media quadratica lineare dividiamo le diverse E.

Osserviamo il doppio prodotto.



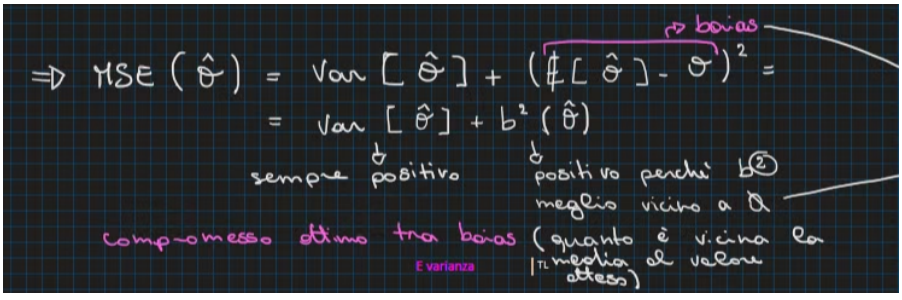
Dopo aver tolto la costante di cui sopra usando la linearità dividiamo di nuovo il contenuto della stima, il risultato sarà 0 in quanto ne esce differenza di E[theta cappello] e E[E[theta cappello]], E[theta cappello] è una costante e la stima di una costante è uguale alla costante stessa quindi i due termini sono uguali.



(rigo dell'immagine di sopra)

L'MSE di theta cappello è pari al primo termine più il secondo termine, il primo termine è proprio la varianza di theta cappello, del secondo termine eliminiamo la E perché al suo interno c'è una costante (E[theta cappello] e theta sono entrambe costanti), per la quale il valore atteso è irrilevante.

Il secondo termine prende il nome di bias al quadrato dello stimatore theta cappello.



Notiamo che non compare la varianza di ϵ, perché il problema non ha errore, se ci fosse stato quello sarebbe stato il termine irriducibile, perché se con la magia si azzeravano varianza e bias quadro di $\hat{\theta}$ quello restava.

> 💡 Il **bias** è la differenza tra la media dello stimatore e il valore vero che voglio stimare.

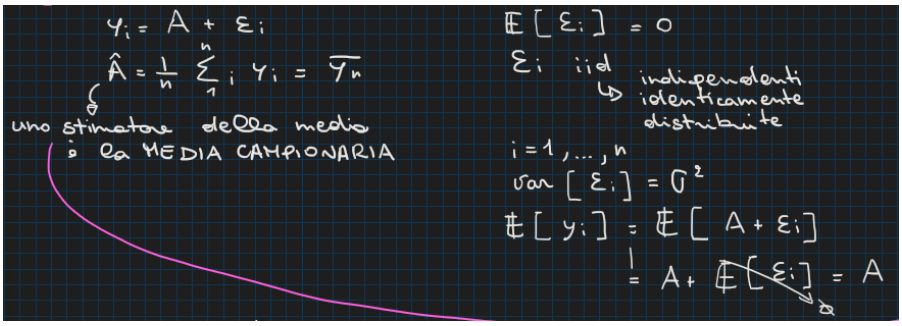
La maggior parte della teoria si studia con bias 0, in teoria possiamo parlare di stimatori unbiased.

L'MSE lo vogliamo 0, la varianza deve essere positiva, il bias è per forza positivo, quindi vogliamo ridurre il più possibile questi termini.

Mettere bias 0 non è per forza il caso migliore, magari con bias diverso da 0 otteniamo una varianza molto più piccola.

Si cerca il compromesso ottimo tra bias e varianza per ridurre l'MSE.

Stima del parametro deterministico A



A destra dimostriamo che la media di y_i è uguale ad A. Vogliamo quindi stimare A per stimare la media di y_i.

Uno stimatore della media è la media campionaria.

A cappello è lo stimatore di A.

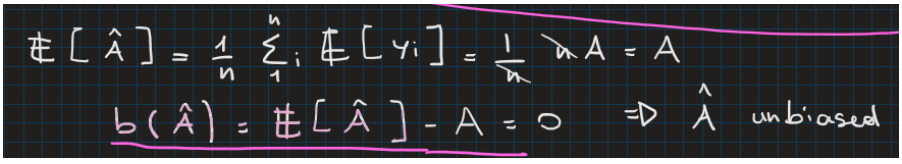
Come pedice i va da 1 alla nostra cardinalità n.

Assumiamo che la varianza per le nostre variabili aleatorie sarà sempre la stessa. (Omoschedasticità)

Non ci ha detto CDF o PDF o cose simili ma in questo caso ci ha fornito media e varianza.

Caratteristiche dello stimatore $\hat{A}$

BIAS



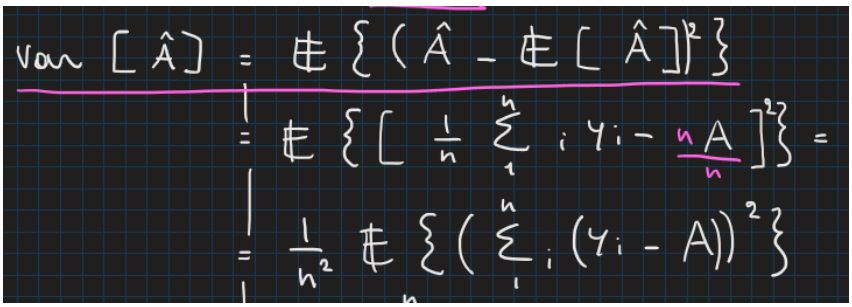
Proviamo a calcolare la media dello stimatore, se ci metto i numeri diventa un valore ma qui non mettiamo i numeri.

Lo stimatore A cappello ha come media proprio A, quindi il bias è 0, si dice che lo stimatore è non polarizzato o unbiased.

La linearità del valore atteso la usiamo in tutti i passaggi essenzialmente.

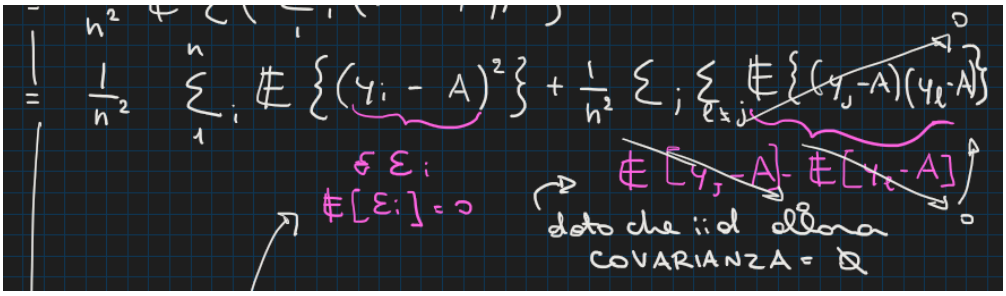
VARIANZA $\hat{A}$ che è lo simatore media campionaria

Calcoliamo anche la varianza dello stimatore usando la formula normale della varianza.



Moltiplico e divido per n la A che sta da sola al fine di poter fare i passaggi successivi, per prima cosa per prendere in evidenza 1/n che diventa fuori dal quadrato $1/n^2$.

Eseguiamo il quadrato:

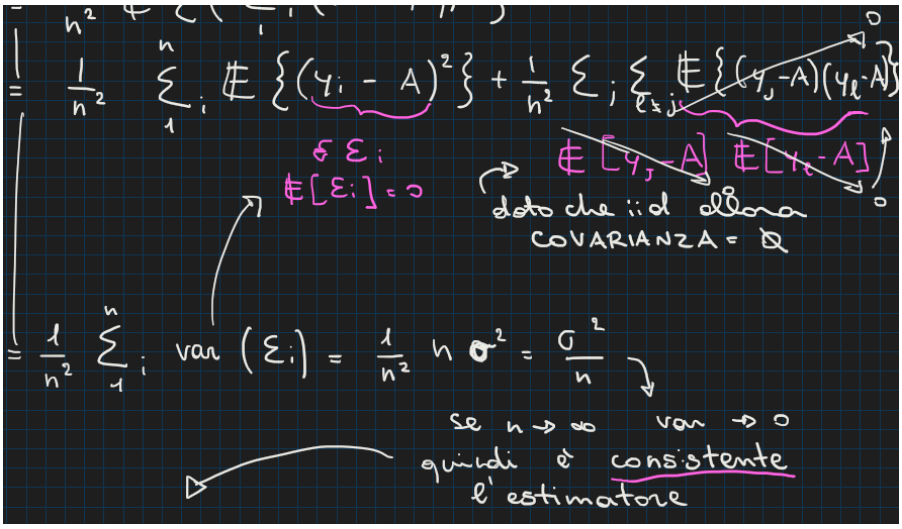


Abbiamo, non a caso, trovato la covarianza.

Se le variabili sono incorrelate la covarianza è 0.

La media del prodotto è uguale al prodotto delle medie (scritte in rosa) + la covarianza, visto che tutti questi termini valgono 0 la doppia sommatoria vale 0.

y_i—A è uguale ad ϵ_i ma ϵ_i ha media nulla quindi abbiamo come primo termine la varianza di epsilon con i (secondo la formula normale per la varianza, solo che la media come detto è 0).



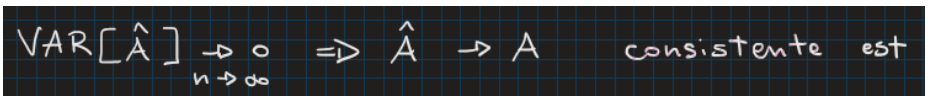
$$E[(\epsilon_i - E[\epsilon_i])^2] = Var(\epsilon_i)$$

visto che la media di ϵ_i è 0:

$$E[(\epsilon_i)^2] = Var(\epsilon_i)$$

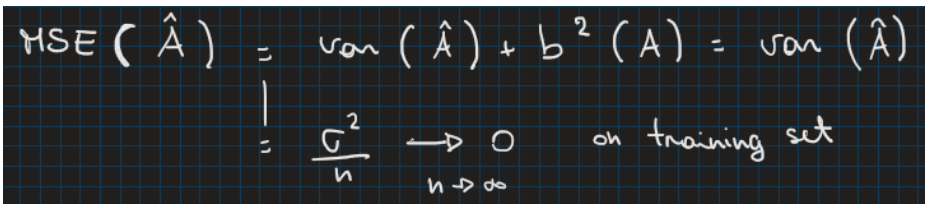
💡 Se la varianza tende a 0 lo stimatore si dice consistente.

💡 La **consistenza** consiste nella capacità dello stimatore di produrre risultati sempre più affidabili (varianza sempre più piccola) al crescere della quantità di dati.



Quindi già sappiamo che era unbiased, poi più è grande il campione e più la stima è vicina al valore deterministico, cioè è consistente.

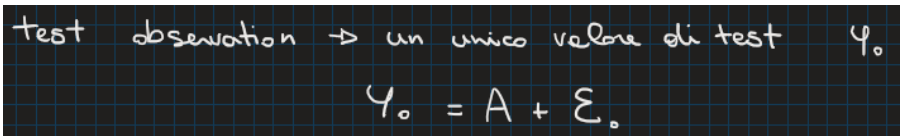
Da cui consegue che l'MSE tende a 0 se n tende all'infinito.



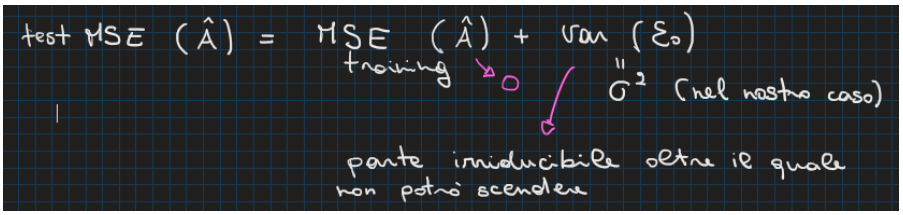
Quindi in queste condizioni è facile portare l'MSE vicino a 0 nel training set, ma a noi interessa minimizzare l'**errore di generalizzazione**, che ha a che fare con il **test set**.

Errore di generalizzazione - formula

Supponiamo di avere un test set con un unico campione y_0.



Epsilon con 0 ovviamente è diversa da prima ma è sempre in accordo con la famiglia delle epsilon di prima (quindi varianza e media come prima).



> 💡 Ci sarà una componente dell'MSE di test, sigma quadro, che non possiamo azzerare. Si chiama **parte irriducibile del MSE di test**.