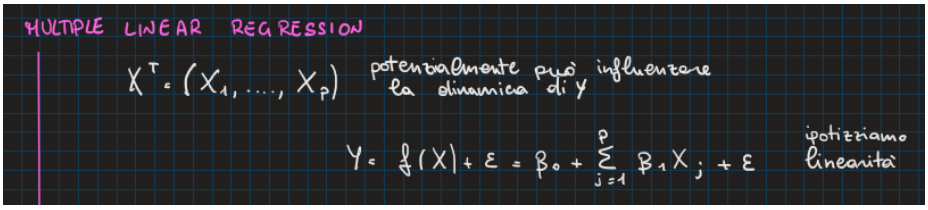
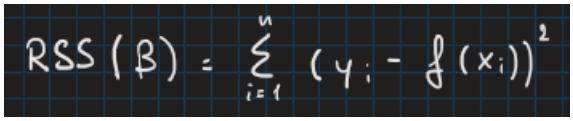
Dimostrazione formula del calcolo di Beta cappello

Prima parte, gradiente

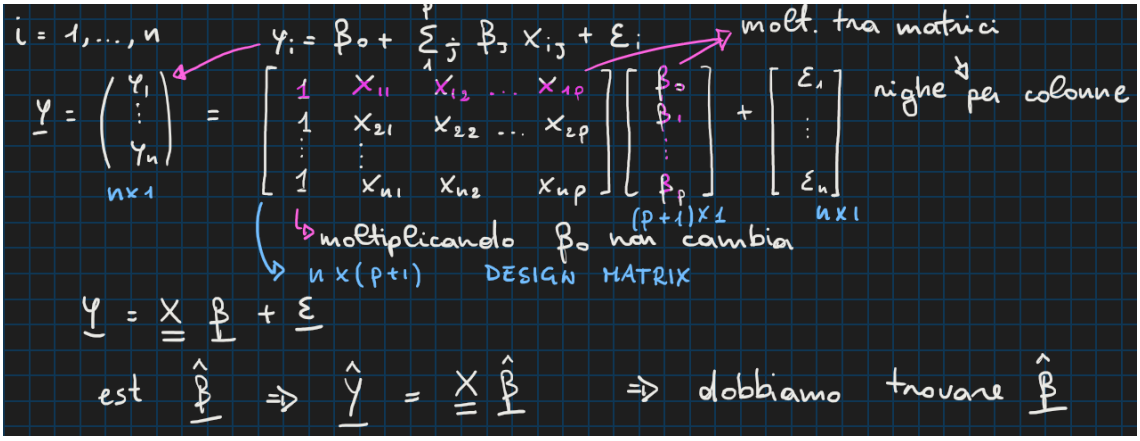
Nel contesto:



Usando Least Squares, quindi puntando alla minimizzazione di:



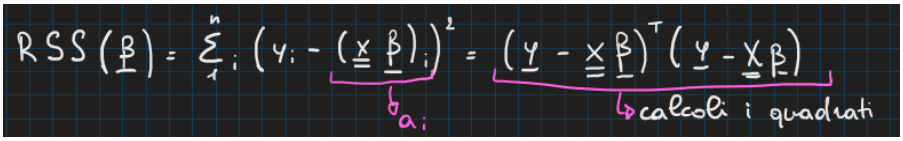
Valgono le formule:



Essenzialmente abbiamo preso l'espressione e l'abbiamo scritta come matrici e vettori.

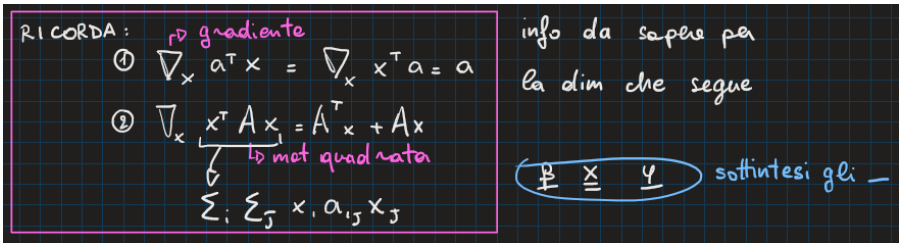
Nella matrice delle x, o matrice di design, le x hanno due pedici, il primo distingue i diversi campionamenti, il secondo indica il numero del regressore.

Dobbiamo trovare $\hat{\beta}$ che minimizza l'RSS:



Se ho un vettore e voglio calcolarne i quadrati facciamo riga per colonna, la sommatoria la riscriviamo.

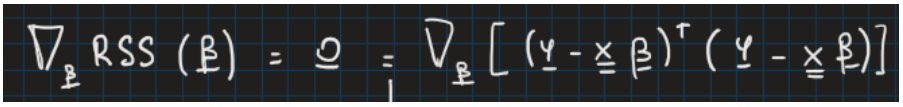
Ricordando che:



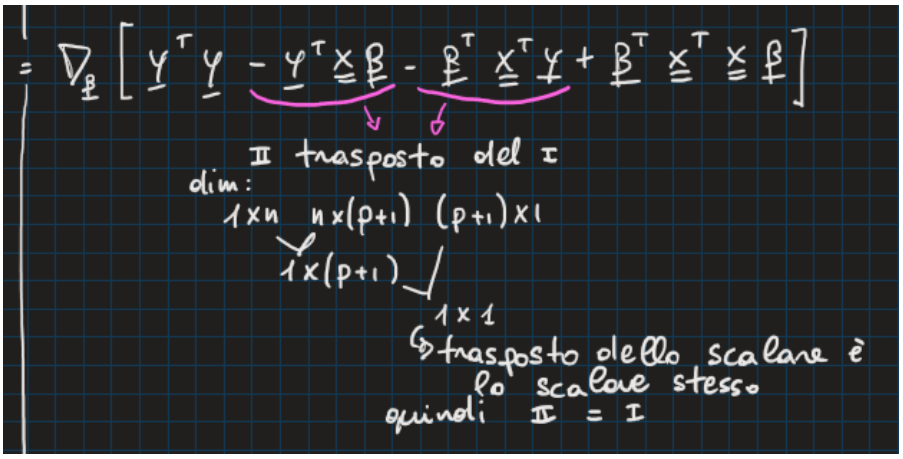
Dimostriamo la formula:

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

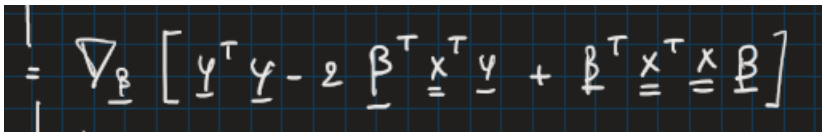
Per trovare il minimo RSS iniziamo ponendo il gradiente a 0:



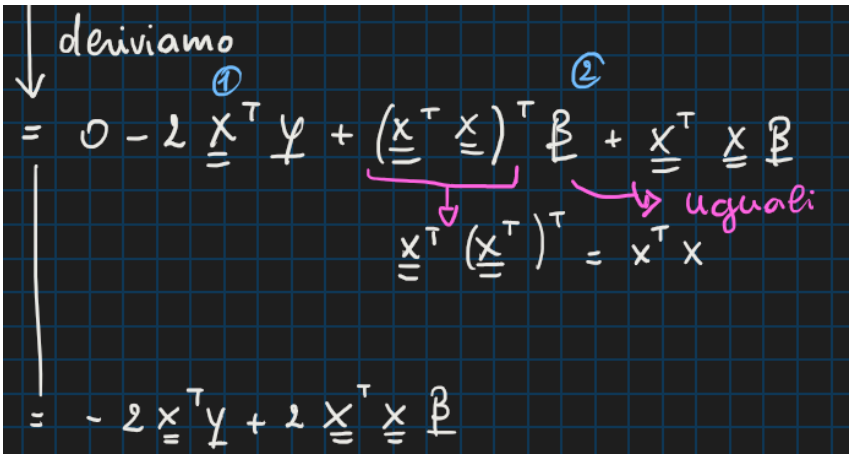
Eseguiamo il prodotto:



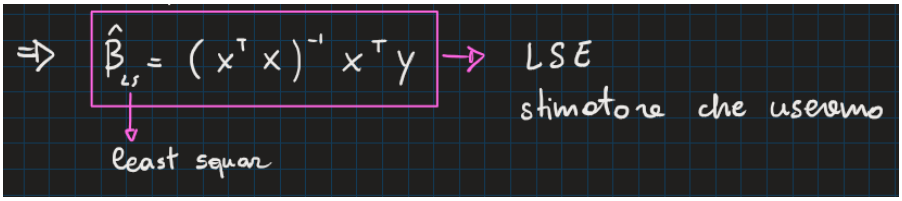
Sommiamo i due termini rosa:



Svolgiamo il gradiente:



Ricordando che quest'ultima espressione va posta a 0 ricaviamo il valore $\hat{\beta}$ che minimizza l'RSS:



Invertibilità

Nella formula che abbiamo ottenuto c'è l'inversione di una matrice ma non tutte le matrici sono invertibili (se il rango è 0 non si può).

Risulta che la matrice è invertibile se la matrice di design è a rango pieno (full rank).

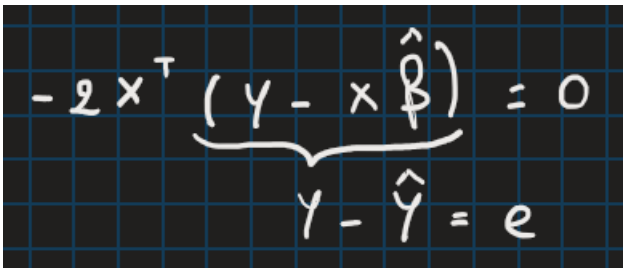
> 💡 La matrice di design è full rank se $n > p$, cioè il numero di punti del dataset deve essere maggiore del numero di regressori.

Questa affermazione può essere interpretata come il fatto che ci serve avere abbastanza punti per permettere la stima dei parametri che vogliamo stimare.

Ortogonalità di errore e span generato dalle x

Ripartiamo dall'espressione:

Mettiamo in evidenza -2 e specializziamo l'equazione per $\hat{\beta}$.



> 💡 Il fatto che il risultato dell'equazione sia 0 ci dice che l'errore è ortogonale allo SPAN generato dai vettori x (visto che moltiplicato per la trasposta di x fa 0).

Il sottospazio generato dai vettori x si chiama SPAN. L'errore è ortogonale a quello spazio, dove? nel punto che minimizza l'RSS. Questa è una proprietà interessante.

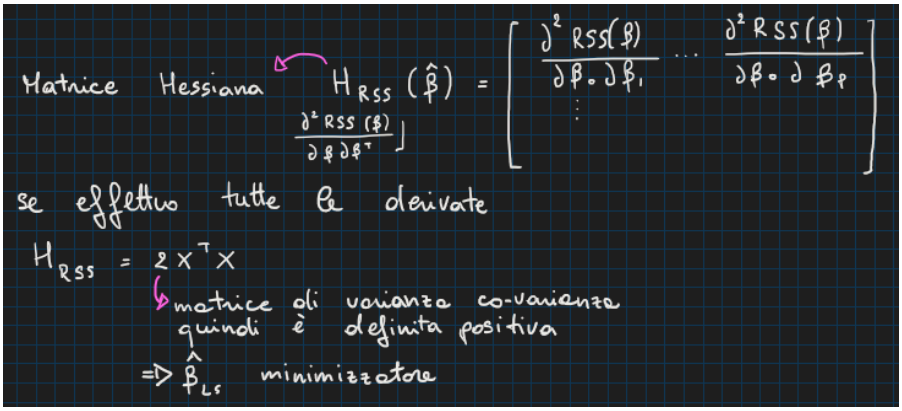
Seconda parte, matrice Hessiana

Ripartiamo dall'espressione:

Abbiamo detto che a noi serve lo stimatore di β che minimizza l'RSS, per ottenere questo non basta porre il gradiente pari a 0 ma bisogna anche fare considerazioni sulla Matrice Hessiana.

Valutiamo la Matrice Hessiana nel punto $\hat{\beta}$.

Nello specifico la Matrice Hessiana deve essere definita positiva, quindi visto che è simmetrica basta che gli autovalori siano tutti positivi.

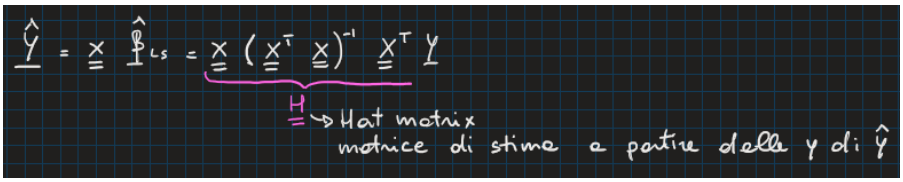


La matrice $H_{RSS} = 2X^TX$ è una somma di quadrati, è quindi proporzionale alla matrice di varianza-covarianza dei dati x. Si può quindi dimostrare che questa matrice è una stima della matrice di var-covar e sapendo che quest'ultima ha per definizione tutti i valori positivi questo varrà anche per la matrice che stiamo cercando.

Possiamo quindi confermare che il nostro $\hat{\beta}_{LS}$ è un punto di minimo.

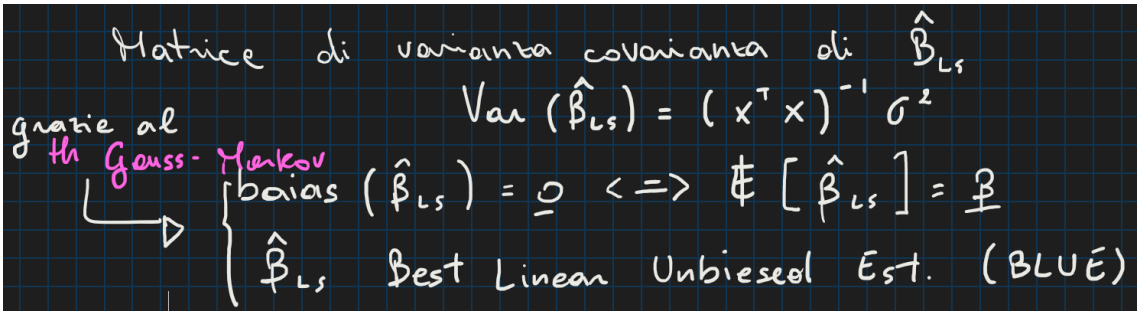
Ultime considerazioni

Avendo confermato la validità di $\hat{\beta}_{LS}$ possiamo sostituire l'espressione dello stimatore nell'espressione.



La matrice H, Hat Matrix, è una matrice che prende i dati y e li trasforma in $\hat{y}$, il modo in cui effettua questa azione è proiettando la y nello SPAN generato dalle x.

Constatiamo che la varianza del nostro stimatore dipende da σ^2, questo ha senso perché la varianza dell'errore implica la dispersione più o meno grande delle mie stime.



In virtù del teorema di Gauss-Markov ha media che coincide con il valore vero, quindi bias nullo. Questo è il miglior stimatore unbiased tra gli stimatori lineari, BLUE.

La stima di σ^2 è un problema affrontato nel continuo di 2 - Linear Regression B.