Modello Lineare Multiplo Non linearità

I modelli lineari sono tali nei parametri, ma le variabili esplicative possono comparire in qualsiasi forma. Nei modelli regressivi, in particolare, si ipotizza che esista una relazione tra i valori delle variabili esplicative e i valori attesi della variabile risposta, si parte normalmente da relazioni espresse mediante modelli del primo ordine, in cui le variabili esplicative compaiono sempre in termini di primo grado, ma si deve verificare che tale assunzione iniziale sia corretta. La relazione ipotizza che tra la nostra variabile dipendente y e le singole variabili esplicative x è: $y = f(x)$, con f lineare.

L'approssimazione lineare non è però sempre la migliore. Per validare la presenza di ciascun regressore all'interno dei diversi modelli dobbiamo quindi verificare la linearità di tale relazione. Dunque, la variabile risposta deve essere una combinazione lineare di variabili esplicative e di parametri lineari. Se una relazione tra y e x è non lineare, allora l'effetto su y di una variazione in x dipende puntualmente dal valore di x poiché l'effetto marginale di x non è costante.

In questo caso, una regressione lineare mal specificata: la forma è errata e lo stimatore dell'effetto su y di x non è corretto nemmeno sulla media. Può capitare, ad esempio, che l'indice R^2 sia elevato ma che non ci sia linearità perché c'è sia una componente lineare sia una non lineare. Per verificare la presenza di linearità è possibile utilizzare alcuni grafici:

1. Scatter plot della variabile risposta (y_i) in funzione di ogni esplicativa (x_j) presente nel modello
2. Scatter plot dei residui (e_i) in funzione dei valori osservati (y_i) della variabile dipendente

E' da notare che la non linearità potrebbe dipendere anche solo da una o da alcune variabili esplicative e non necessariamente da tutte. Quando è presente non linearità dei parametri, potrebbe esistere una trasformazione che li renda lineari, oppure che questi siano espressi in una forma intrinsecamente non lineare.

Si possono distinguere due tipi di non linearità:

- modelli linearizzabili:
    - $y=\beta_0+\beta_1 X+\beta_2 X^2$, il modello è lineare nei coefficienti. Si può stimare con OLS creando la variabile $W=X^2$.
    - $y=\beta_0+\exp(\beta_1)X$, il modello è lineare nelle variabili. Si può stimare con OLS stimando $y=\beta_0+\delta_1 X$ e calcolando successivamente $\beta_1=\ln(\delta_1)$.
- Modelli intrinsecamente non lineari. Per questi modelli non esiste trasformazione di uno o più parametri che li renda lineari e possono essere stimati con i minimi quadrati non lineari (NLS) ma non tramite

OLS. È un problema di minimizzazione non lineare che si risolve con gli algoritmi numerici dei software.

Volendo utilizzare funzioni di variabili indipendenti non lineari in X possiamo riformulare una vasta famiglia di funzioni di regressione lineare come regressioni multiple. La soluzione consiste nell'applicare ai dati una funzione di regressione da individuare che sia: $y_i=f(x_{1i},x_{2i},\ldots,x_{ki})+\varepsilon_i$ (i=1, …, n). La variazione in y associata ad una variazione in x_1, mantenendo x_2, …, x_k costanti è: $\Delta y=f(x_1+\Delta x_1,x_2,\ldots,x_k)-f(x_1,x_2,\ldots,x_k)+\varepsilon$. La stima e l'inferenza procedono in modo analogo al modello di regressione lineare multiplo. L'interpretazione dei coefficienti è specifica nel modello utilizzato, ma la regola generale consiste nel calcolare gli effetti confrontando i casi diversi.

Alcune delle più comuni funzioni non lineari sono:

- Polinomiali in X, Y. La variabile dipendente Y viene trasformata in una quadratica, una cubica o una polinomiale di grado più alto; la funzione di regressione nelle X viene approssimata da una quadratica, una cubica o una polinomiale di grado più alto.
- Trasformazioni logaritmiche. Le y e/o le X vengono trasformate prendendone il logaritmo, che ne dà un'approssimazione "percentuale" utile in molte applicazioni.

Nel caso di polinomiali in X (linearizzabili nelle variabili): Approssimiamo la funzione di regressione con una polinomiale: $Y_i=\beta_0+\beta_1 x_i+\beta_2 x_i^2+\ldots+\beta_r x_i^r+\varepsilon_i$, i regressori sono potenze di X.

Tra la trasformazioni logaritmiche esistono tre modelli principali:

1. Linear-log, in cui ad un incremento percentuale della variabile indipendente corrisponde un incremento nominale della variabile dipendente.
2. Log-linear, in cui ad un incremento nominale dell'esplicativa corrisponde un incremento percentuale della risposta.
3. Log-log, in cui entrambi gli incrementi sono percentuali.