# Statistical Modeling

## 1 Errori eteroschedastici

Affinché il nostro modello lineare classico ottenga delle stime efficienti abbiamo bisogno di verificare che gli errori siano omoschedastici. In virtù del fatto che ogni errore $\varepsilon_i$ è a media nulla $E(\varepsilon_i) = 0$ vale la relazione $E(\varepsilon_i^2) = \text{Var}(\varepsilon_i) = \sigma^2$, e quindi gli errori sono detti omoschedastici quando la loro varianza è costante al variare del valore dei regressori. Se ciò non accade gli errori si dicono eteroschedastici $\text{Var}(\varepsilon_i) = \sigma_i^2$; questo incide sulle proprietà degli stimatori OLS: in particolare continuano a valere correttezza e consistenza ma viene meno l'efficienza (lo stimatore non è più **BLUE**, **B**est **L**inear **U**nbiased **E**stimator). Inoltre le stime campionarie tendono a sottostimare il vero valore della varianza e non esiste più un'unica varianza, ma ce ne sono molteplici. Come conseguenza la statistica T di Student ha valori erroneamente elevati ed anche i relativi intervalli di confidenza risulteranno più stretti mentre la regione di rifiuto del test T risulterà erroneamente più ampia; verranno quindi ritenuti significativi i parametri anche quando in realtà non lo sono. Per individuare questa caratteristica, che ci porta ad un'inaffidabilità delle stime, possiamo ricorrere a diversi metodi (grafici o analitici).

Per quanto riguarda i metodi grafici:

1. Scatter plot dei valori osservati della variabile target ($y$) contro le variabili esplicative $x_j$. Ovviamente bisognerà effettuare uno scatterplot per ogni variabile $x_j$;
2. Scatter plot dei valori predetti ($\hat{y}$) contro i residui stimati ($y - \hat{y} = \varepsilon$);
3. Scatter plot dei residui al quadrato ($\varepsilon^2$) contro i valori predetti di y ($\hat{y}$);
4. Scatter plot dei valori osservati ($y$) contro quelli predetti ($\hat{y}$);
5. Scatter plot dei residui ($\varepsilon$) contro le variabili esplicative $x_j$. Ovviamente bisognerà effettuare un grafico per ogni variabile esplicativa.

Possiamo ricorrere anche ad alcuni test (metodo analitico):

- **Test di White**: questo test si basa sull'assunzione di omoschedasticità dei residui; viene perciò definita l'ipotesi nulla come $H_0 : \text{Var}(\varepsilon_i) = (\sigma^2)$ e l'ipotesi alternativa come $H_1 : \text{Var}(\varepsilon_i) = (\sigma_i^2)$. Il test sfrutta la regressione $OLS$ del quadrato dei residui $\varepsilon_i^2$ sui regressori $x_j$, i regressori al quadrato $x_j^2$ e le loro interazioni. Attraverso l'indice di determinazione $R^2$ di tale regressione, ricavato dal rapporto tra la variabilità spiegata dalla regressione $\left(SSE = \sum_i (\hat{y}_i - \bar{y})^2\right)$ e la variabilità totale $\left(TSS = \sum_i (y_i - \bar{y})^2\right)$, si calcola la statistica $LM = nR^2$ che si distribuisce come una $\chi^2$ con gradi di libertà uguale al numero di regressori $n$. L'ipotesi nulla verrà rigettata se $LM$ risulterà maggiore del valore soglia della distribuzione $\chi^2$ (ovvero con p-value

basso); infatti se $R^2$ è oltre ad un certo valore significa che le variabili esplicative sono ancora significative nello spiegare la variabilità dei residui, ovvero che i residui (al quadrato) dipendono dai valori delle variabili esplicative $x_j$, come accade tipicamente in presenza di eteroschedasticità.

- **Test di Breusch-Pagan**: anche in questo test l'ipotesi nulla è quella di omoschedasticità $\big(H_0\text{: } \mathrm{Var}(\varepsilon_i) = (\sigma^2)\big)$. Si basa su una regressione di $\varepsilon_i^2/s^2$ dove $s^2$ è uguale alla sommatoria degli errori al quadrato diviso n. La somma dei quadrati dei regressori e quella degli scarti si distribuiscono come $\chi^2$ indipendenti. Essendo *varepsilon* distribuita normalmente, la sua versione al quadrato $\varepsilon_i^2$ si distribuisce come una $\chi^2$; $s^2$ allo stesso modo in quanto somma di normali al quadrato assume una distribuzione $\chi^2$. A questo punto il loro rapporto si distribuisce quindi come una $F$ di Snedecor, in quanto rappporto tra due $\chi^2$. L'ipotesi nulla verrà rigettata quando la statistica $F$ è superiore ad un valore soglia. Per risolvere tale problema si può procedere attraverso il metodo di stima $WLS$ (Weighted Least Squares).

## 2 Errori autocorrelati

Affinché il nostro modello lineare classico ottenga delle stime efficienti abbiamo bisogno di verificare se gli errori siano correlati tra di loro. Accade spesso infatti, soprattutto in serie storiche o territoriali, che esista una correlazione tra errori in momenti successivi o territori vicini. Gli errori correlati si possono scindere in due componenti: $\rho\varepsilon_{i-1}^{\#}$ (errore ritardato di un tempo) e $\eta_i$ (errori omoschedastici IID, ovvero indipendentemente ed identicamente distribuiti in modo normale). Si nota infatti che l'errore è legato al suo valore ritardato. Possiamo classificare l'autocorrelazione in base al suo *grado*: si dice autocorrelazione di primo grado quando gli errori sono correlati con il loro valore ritardato di un tempo; allo stesso modo si dice autocorrelazione di $i$-esimo grado quando gli errori sono correlati con il loro valore ritardato di $i$ gradi. Gli errori autocorrelati non incidono sulle proprietà di correttezza e consistenza degli stimatori OLS (analogamente agli errori eteroschedastici), ma solo sull'efficienza (non sono più BLUE). Come nel caso dell'eteroschedasticità la stima della varianza dei parametri e la relativa inferenza non sono più corretti e affidabili (la statistica T di Student ottiene dei valori erroneamente più elevati; gli intervalli di confidenza tendono ad essere più stretti e l'area di rifiuto del test anomalamente più ampia). Per individuare la caratteristica di autocorrelazione si può ricorrere a rappresentazioni grafiche o a metodi analitici.

Per quanto riguarda le rappresentazioni grafiche:

1. Scatter plot dei valori osservati ($y_i$) sui valori delle variabili esplicative ($x_j$). Ovviamente si costruiranno tanti scatter plot quante sono le variabili esplicative;
2. Scatter plot dei residui ($\varepsilon_i$) sulle variabili esplicative ($x_j$), ovviamente si costruiranno tanti scatter plot quante sono le variabili esplicative;
3. Scatter plot dei residui ($\varepsilon_i$) sui residui ritardati ($\varepsilon_{i-1}$) (a seconda del grado);
4. Correlogramma: in questo grafico vengono mostrate le correlazioni a diversi gradi con relativa barra di confidenza; analizzando acf (funzione di autocorrelazione dei residui) e pacf si riesce a determinare il tipo di modello autoregressivo.

Per quanto riguarda i test invece:
**Test di Durbin-Watson**: questo test si può effettuare per verificare la presenza di autocorrelazione a diversi gradi. Prendiamo in esame il test per il primo grado. Esso si basa sull'assunzione di non correlazione degli errori: l'ipotesi nulla è quindi

$$H_0 : \rho = Corr(\varepsilon_i; \varepsilon_{i-1}) = 0$$

contro l'ipotesi alternativa di autocorrelazione degli errori che può essere:

1. Unidirezionale destra
2. Unidirezionale sinistra
3. Bidirezionale

La statistica $DW$ per l'autocorrelazione dei residui è definita come $DW = 2(1 - \rho)$. Si può notare che:

- La distribuzione di $DW$ è centrata su 2: se infatti $DW$ è uguale a 2 gli errori sono incorrelati poichè $\rho = 0$;
- $DW$ tende a 0 quando i residui sono correlati positivamente dato che $\rho = 1$;
- $DW$ tende a 4 quando i residui sono correlati negativamente con $\rho = -1$;
- I valori critici cambiano di caso in caso ma convenzionalmente se non specificati sono 1 e 3.

Nel caso di autocorrelazione, il teorema di Aitken stabilisce che nella classe degli stimatori lineari per il modello di regressione *generalizzato* lo stimatore GLS è efficiente in quanto caratterizzato dalla minima varianza.

# 3 Metodo di stima WLS; per soluzioni correlate, GLS

*Errori eteroschedastici e incorrelati: modello WLS*

Per errori eteroschedastici si intende quando la varianza dell'errore non rimane costante al variare del valore delle variabili esplicative, violando quindi una delle ipotesi della regressione lineare classica. Per tale motivo gli stimatori OLS non possono essere usati (in quanto non più efficienti); al contrario si possono utilizzare gli stimatori Weighted Least Squares (WLS) che permettono di stimare un modello per la varianza degli errori condizionata ai regressori. Si tratta di definire le seguenti nuove variabili che danno luogo al modello trasformato dividendo ogni variabile contenuta nel modello di partenza per la radice di $h(i)$ (corrisponde alla varianza di $\varepsilon^*$, ovvero l'errore eteroschedastico). Infatti si tratta di stimare i parametri del modello trasformato con il metodo OLS regredendo $y^*$ su $X^*B$ riportando così la varianza degli errori ad una costante ottenendo la forma $y^* = X^*B + \varepsilon$.

*Errori omoschedastici e correlati: modello GLS*

Nel caso invece ci si trovi davanti ad errori autocorrelati come accade in serie storiche e territoriali è ragionevole ipotizzare che esista correlazione fra errori in momenti successivi o territori vicini. Si parla di autocorrelazione se al variare di $X$ c'è fluttuazione dei valori di $Y$ con lo stesso segno (autocorrelazione *positiva*), o segno alternato (autocorrelazione *negativa*) oltre un certo intervallo di confidenza. Si possono ricavare stime per

errori correlati in modo più semplice tramite una stima dei parametri in una equazione che tenga conto della struttura di autocorrelazione seriale (metodo proposto da Durbin). Bisogna innanzitutto stimare il coefficiente di autocorrelazione di primo ordine attraverso un modello avente come variabile risposta gli errori $\varepsilon_t^{\#}$ e come esplicative quelle già considerate più l'errore ritardato di un tempo $\varepsilon_{t-1}^{\#}$ e procedere alla stima del coefficiente di correlazione $\rho$. Una volta ottenuta la stima di $\rho$ si procede a moltiplicare ogni elemento dell'equazione ritardata per $\rho$ stesso:

$$\rho y_{t-1} = \rho\beta_0 + \rho\beta_1 x_{t-1} + \rho\varepsilon_{t-1}^{\#} \ .$$

Infine si procede a sottrarre l'equazione ritardata moltiplicata per $\rho$ all'equazione nella forma normale $y_t - \rho y_{t-1}$ ottenendo un modello OLS per i parametri trasformati

$$y_t^{\#} = \beta_0^{\#} + \beta_1 x_t^{\#} + w_i$$

che rispetta tutte le classiche proprietà di correttezza, consistenza ed efficienza.

In alternativa è possibile utilizzare un modello *autoregressivo* (proprio del software SAS) per inserire nell'equazione iniziale un errore ritardato che tenga conto dell'autocorrelazione di ordine 1 (o anche ordini superiori):

$$y_i = b_0 + b_1 x_i + AR1_i + \varepsilon_i.$$

con $AR1 + \varepsilon_i = v_i$ e $Corr(v_i; v_j) = 0$.
*Errori eteroschedastici e correlati: stimatore GLS*

Nel caso in cui gli errori non siano sferici in quanto eteroschedastici e correlati si utilizzano gli stimatori dei minimi quadrati generalizzati ($GLS$) interpretabili in modo analogo al modello classico in quanto stimatori $OLS$ basati su variabili trasformate per mezzo delle proprietà degli autovettori e autovalori ricavati dalla matrice dei residui $\Sigma_\varepsilon$. Nello specifico si procede ad effettuare una *decomposizione spettrale della matrice degli errori* $\Sigma_\varepsilon = \sigma^2 VV'$. A questo punto moltiplicando per $V^{-1}$ il modello si ottiene un nuovo modello nelle variabili trasformate ottenendo $\Sigma_\varepsilon$ omoschedastica ed incorrelata:

$$V^{-1}y = y^\circ$$

$$V^{-1}X\beta + V^{-1}\varepsilon^\circ = X^\circ\beta^\circ + \varepsilon$$

$$y^\circ = \beta^\circ X^\circ + \varepsilon$$

Lo stimatore risulta godere delle tre proprietà:

1. Correttezza;
2. Consistenza;
3. Efficienza in quanto il teorema di Aitken stabilisce che nella classe degli stimatori lineari per il modello di regressione *generalizzato* lo stimatore $GLS$ è caratterizzato dalla minima varianza, che risulta comunque maggiore di quella ottenuta attraverso il modello $OLS$ per i modelli lineari ma, condizionatamente ai modelli lineari generalizzati è il migliore. Infatti $\sigma^2(X^{\circ'}X^\circ)^{-1} > \sigma^2(X'X)^{-1}$, e la differenza tra le due risulta essere una matrice semidefinita *positiva*;

4. Lo stimatore assegna un peso maggiore alle osservazioni caratterizzate da una minore varianza da considerarsi più "affidabili".

Tutto questo è possibile assumendo come nota la matrice di varianze e covarianze dei residui $\Sigma_\varepsilon$. Nel caso in cui questa non fosse conosciuta allora è possibile ricorrere alla matrice campionaria $S_\varepsilon$ in modo che rispetti la condizione $\lim_{n\to\infty} S_\varepsilon = \Sigma_\varepsilon$. A questo punto si possono utilizzare gli stimatori FGLS (Feasible Generalized Least Squares). Spesso la soluzione di applicare i FGLS viene applicata anche in caso di semplice eteroschedasticità o semplice autocorrelazione, o sospette tali, poichè vige il principio di precauzione.

# 4 Multicollinearità

Se la matrice $(X'X)^{-1}$ non è invertibile oppure ha determinante prossimo allo 0 le stime non esistono (coefficienti sotto identificati, poiché non si dispone di sufficiente informazione per stimarli) o non sono stabili (coefficienti empiricamente sotto identificati).

Tale problema si verifica quando almeno una delle variabili è correlata linearmente alle altre e quindi si ha multicollinearità. In questo caso la matrice $(X'X)$ è singolare e le soluzioni non sono uniche.

Esistono due tipi di collinearità:

1. **Perfetta**: sussiste quando almeno una variabile esplicativa è una combinazione lineare perfetta delle altre. Essa viola le proprietà del modello lineare classico solitamente per un errore nella definizione dei regressori o per una stranezza nei dati o ancora per la presenza di due variabili che sono direttamente dipendenti una dall'altra (ad esempio *titolo di studio* e *anni di studio*).
2. **Imperfetta**: sussiste quando 2 o più regressori sono fortemente correlati e il determinante della matrice dei coefficienti tende a 0. Questa condizione non provoca l'impossibilità della stima dei coefficienti come per la collinearità perfetta ma dà origine a coefficienti fortemente distorti.

Le principali conseguenze sono:

1. Un aumento della varianza delle stime dei coefficienti;
2. Gli intervalli di confidenza al cui interno sta il valore vero del parametro con *confidenza* $1-\alpha$ risultano essere più grandi di quanto non siano in realtà mentre la regione di accettazione del test si amplia notevolmente, ciò implica che i parametri vengano ritenuti non significativi anche quando in realtà lo sarebbero;
3. Con due variabili fortemente correlate se aggiungo la seconda, l'incremento di $R^2$ è inferiore all'incremento che avrei aggiungendo una seconda variabile debolmente correlata con la prima. Quindi, siccome le due variabili hanno molta varianza in comune, non posso dire quale delle due è più influente rispetto all'outcome.

Esistono tre metodi analitici per verificare la presenza di multicollinearità:

1. **Indice di tolleranza** che misura il grado di interrelazione di una variabile indipendente rispetto alle altre. Nella pratica, $TOL = 1 - R_j^2$ dove $R_j^2$ è calcolato dalla regressione della variabile esplicativa $X_j$ (usata come risposta) in funzione di tutte

le altre esplicative. Per questo può assumere valori compresi tra 0 (che indica la massima *collinearità*) e 1 (che indica la massima *indipendenza* tra le variabili);
2. **Varianza multifattoriale** o **VIF**, ovvero il reciproco della tolleranza. Valori di tale indice variano tra 0 e $\infty$ perciò se superiori a 20 indicano uno stretto rapporto tra la variabile considerata e le altre ovvero un eccessivo grado di *multicollinearità*. Vanno considerate con attenzione anche quelle variabili con valori di VIF maggiori di 10;
3. **L'indice di condizione** è dato dalla radice del rapporto tra l'autovalore massimo della matrice $(X'X)$ e ogni autovalore. Quando risulta essere maggiore di 30 si ritiene esistere *collinearità*. Tale convinzione viene rafforzata se un autovalore con condition index maggiore di 30 contribuisce a spiegare elevate quote di varianza di due o più variabili.

Oltre all'utilizzo di queste misure analitiche è buona norma, in prima istanza, generare una *matrice di correlazione* tra tutte le variabili così da identificare rapidamente possibili variabili collineari. Auspicabilmente infatti vorremmo forte correlazione tra $y$ e le singole $x_j$ con bassa correlazione tra la singole $x_j$.

# 5 Linearità

La relazione ipotizzata tra la nostra variabile dipendente $y$ e le singole variabili esplicative $x$ è di tipo: $y = f(x)$, con $f$ lineare.

L'approssimazione lineare non è 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 $y$ $(\Delta y)$ di una variazione in $X$ $(\Delta X)$ dipende puntualmente dal valore di $X$ poiché l'effetto marginale di $X$ non è costante.

In questo caso, una regressione lineare è mal specificata: la forma funzionale è errata e lo stimatore dell'effetto su $y$ di $X$ non è corretto nemmeno sulla media. Può capitare ad esempio che $R^2$ sia elevato ma che non ci sia linearità perchè c'è sia una componente lineare sia una non lineare.

Per verificare la presenza (o meno) di linearità è possibile ricorrere ad 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 $(\varepsilon_i)$ in funzione dei valori osservati $(y_i)$ della variabile dipendente; non deve essere un andamento sistematico;
3. Scatter plot dei residui $(\varepsilon_i)$ in funzione dei valori previsti $(\hat{y}_i)$; deve esserci un andamento regolare.

È 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 (caso linearizzabile) oppure che questi siano espressi in una forma intrinsecamente non lineare.

Nel primo caso si procede innanzitutto alla linearizzazione del parametro (o della variabile) *non lineare* con una trasformazione che lo renda *lineare*, poi si procede alla stima OLS ed infine si applica la trasformazione inversa ricavando la stima del parametro originale. Nel caso invece di componenti intrinsecamente non lineari si procede allora alla stima attraverso gli stimatori NLS (minimi quadrati non lineari) che sfruttano algoritmi numerici nei software per affrontare il problema di minimizzazione non lineare.

Volendo utilizzare funzioni di variabili indipendenti non lineari in X possiamo riformulare una vasta famiglia di funzioni di regressione lineare come regressioni multiple.

Tra le funzioni non lineari le più utilizzate sono le polinomiali e le trasformazioni logaritmiche.

Tra la trasformazioni logaritmiche esistono tre modelli principali:

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

# 6 Non normalità

Quando gli errori $\varepsilon_i$ sono indipendenti e identicamente distribuiti come $N(0, \sigma^2)$ si possono ricavare la distribuzione degli stimatori, i test statistici, gli intervalli di confidenza e le proprietà ottimali (inoltre stima di massima verosimiglianza $ML$ coincide con stima dei minimi quadrati $OLS$). Nel caso in cui gli errori non siano normali, se tuttavia i campioni sono sufficientemente larghi per il **teorema del limite centrale** la distribuzione degli errori tende *asintoticamente* alla normalità. Se ciò non accade non è possibile applicare test e intervalli di confidenza perchè essi sono basati tutti sull'ipotesi di normalità degli errori.

Conseguenze della violazione della normalità:

1. I parametri $\beta$ possono essere espressi come combinazione lineare degli errori, per cui se gli errori non sono normali anch'essi non sono più normali;
2. Non è più possibile ricavare test basati sulla normale standardizzata;
3. Non è più possibile ricavare intervalli di confidenza per i parametri basati sulla normale standardizzata;
4. Le stime OLS non coincidono con le stime ML ottenute attraverso il metodo della massima verosimiglianza, quindi gli stimatori OLS non sono più gli stimatori corretti a minima varianza *fra tutti gli stimatori corretti* cioè non sono più **VUE**. Il fatto che le stime ML ed OLS non coincidano più rende meno affidabili le stime attraverso software statistici, che comunemente effettuano la stima attraverso il metodo della massima verosimiglianza. Nonostante gli stimatori OLS non siano più **VUE**, conservano le proprietà di correttezza, consistenza ed efficienza condizionatamente ai dati. Essendo i dati affetti da un bias sulla distribuzione dei residui $\varepsilon$ anche le stime OLS erediteranno tale bias ma, proprio in virtù di ciò, possono essere ancora

considerati gli stimatori a minima varianza *tra tutti gli stimatori lineari* e perciò sono considerati **BLUE**.

Per individuare casi di non normalità è opportuno:

- Osservare indici descrittivi;
- Effettuare rappresentazioni grafiche;
- Effettuare test non parametrici (ovvero realizzati con lo scopo di testare la distribuzione del parametro sotto osservazione).

Tra gli indici descrittivi possiamo, in prima istanza, osservare indicatori quali **moda**, **media** e **mediana**. Banalmente quando queste corrispondono possiamo affermare che la distribuzione dei residui $\varepsilon_i$ è normale. Questi indicatori sono anche visualizzabili in maniera diretta utilizzando un box-plot.

Tra la rappresentazioni grafiche utili rientrano:

- Plot della distribuzione dei residui, per cui se la media risulta maggiore della mediana allora sarà possibile visualizzare una distribuzione caratterizzata da asimmetria *positiva*(a destra), mentre in caso di media inferiore alla mediana sarà possibile visualizzare una distribuzione affetta da asimmetria *negativa*(a sinistra).
- Plot della distribuzione cumulata dei residui, che è possibile ispezionare alla ricerca di evidenti irregolarità.
- P-P plot che mette a confronto la distribuzione cumulata dei residui (sulle ascisse) con la distribuzione cumulata della normale (sulle ordinate). Il risultato di ciò è che in caso di distribuzione normale allora i punti si distribuiranno in modo ordinato lungo la *bisettrice*.
- Q-Q plot, molto simile al precedente, mette a confronto i quantili della distribuzione normale (sulle ascisse) con i residui $\varepsilon$ (sulle ordinate). Anche in questo caso la distribuzione dei punti lungo la *bisettrice* indica il soddisfacimento dell'assunzione di normalità dei residui. La diversa forma assunta dai punti sulla bisettrice può inoltre indicare una distribuzione leptocurtica, platicurtica oppure asimmetrica (a destra o a sinistra a seconda della forma assunta).

Esistono infine alcuni test non parametrici che non si basano su ipotesi sulla distribuzione. Per questo motivo sono molto utili per analizzare problemi di normalità dei residui.

1. **Test di Shapiro-Wilk**, che assume valori compresi tra 0 e 1 e gli estremi corrispondono rispettivamente al rifiuto e all'accettazione dell'ipotesi di normalità. Il test parte dell'ipotesi $H_0 : \varepsilon \sim N(0, \sigma^2)$. In ogni caso, il test $W$ essendo caratterizzato da una forte asimmetria potrebbe comunque portare ad un rifiuto dell'ipotesi di normalità.
$$W = \frac{(\sum_i \beta_i \varepsilon_i)^2}{\sum_i \varepsilon_i^2}$$
2. **Test di Kolmogorov Smirnov**, in cui $H_0 : \varepsilon \sim N(0, \sigma^2)$ e si basa sul calcolo della statistica test $D$ come la somma in valore assoluto della differenza tra le frequenze cumulative della distribuzione empirica da testare e quelle della normale, una volta definite delle classi di eguale ampiezza. $D$ viene poi messa a confronto con le apposite

tavole (essendo una statistica tabulata) ed in caso di superamento del valore critico in base al livello di significativà scelto comporterà il rifiuto di $H_0$;

3. **Skewness test** (test di asimmetria), ovvero un test direzionale basato sul fatto che la distribuzione della normale è simmetrica; si basa perciò su un indice di simmetria; rigettando $H_0$ si rigetta la normalità, non rigettandola si dice solo che la distribuzione è simmetrica, ma non per forza normale.

$$S = \frac{(E[X-\mu]^3)^2}{(E[X-\mu]^2)^3}$$

   Sinteticamente si tratta di mettere a rapporto il quadrato del momento terzo intorno alla media di $X$ con il cubo della varianza.
   Quando l'ipotesi di normalità è rispettata allora $E(S) = 0$;
4. **Test della Kurtosis**, simile nella forma a quello per l'asimmetria

$$K = \frac{E(X-\mu)^4}{(E[X-\mu]^2)^2}$$

   Anche in questo caso sinteticamente si tratta di mettere a rapporto il momento quarto intorno alla media di $X$ con il quadrato della varianza.
   Sotto l'ipotesi di normalità $E(K-3) = 0$, poichè la curtosi della normale è appunto uguale a 3.

I problemi di non normalità possono essere risolti usando una *trasformazione* della variabile dipendente $Y$. La trasformazione può migliorare la relazione lineare tra la variabile dipendente e le variabili indipendenti.

Tra le trasformazioni disponibili vi sono:

- *log(Y)* quando $S_\varepsilon$ cresce con $y$ o quando la distribuzione dell'errore ha asimmetria *positiva*;
- $Y^2$ quando $S_\varepsilon$ è proporzionale a $E(y)$ o quando la distribuzione dell'errore ha asimmetria *negativa*;
- $\sqrt{Y}$ quando $S_\varepsilon$ è proporzionale a $E(y)$;
- $Y^{-1}$ quando $S_\varepsilon$ cresce significativamente al crescere di $y$.

# 7 Outlier

I valori cosiddetti **outlier** possono essere distinti in:

1. Valori anomali: valori che si discostano in modo rilevante dall'andamento generale.
2. Punti influenti: punti che influenzano in misura rilevante le stime.

Non sempre un valore anomalo è anche influente; per contro esistono punti non anomali che influiscono in misura rilevante sul risultato.

Come identificare gli outlier:

- Rappresentazioni grafiche per mezzo di box-plot e scatter-plot.

- Indicatori

Tra questi **indicatori** è possibile distinguere tra:

1. **Leverage values**: Definita $H = X(X'X)^{-1}X'$, nota come matrice di proiezione, gli elementi $h_{ii}$ sulla diagonale, chiamati leverage, possono essere usati per verificare l'impatto dell'osservazione $i$-esima sulla capacità del modello di predire tutti i casi.

   Si dimostra che il valor medio del leverage è:

   $$\frac{(k-1)}{n} \ ,$$

   con $k = n^\circ$ variabili esplicative ed $n = n^\circ$ osservazioni. Può dunque essere considerato

   $$h_{ii} > \frac{2(k-1)}{n}$$

   come valore soglia per individuare osservazioni potenzialmente anomale con un'eccessiva influenza sulla stima complessiva di tutte le osservazioni

2. **Residui standardizzati**: Assumendo per i residui $\varepsilon$

   $$\varepsilon = (I - H)y$$

   allora è possibile scrivere la varianza esplicitata come

   $$Var(\varepsilon_i) = (1 - h_{ii})\sigma^2 \ .$$

   Come conseguenza di ciò i residui *standardizzati* sono definiti come

   $$\varepsilon_i^* = \frac{\varepsilon_i}{\sigma\sqrt{(1-h_{ii})}}$$

   In un campione distribuito normalmente il 95% dei valori dei residui standradizzati $\varepsilon_i^*$ dovrebbe assumere valori compresi tra $-2$ e $+2$ mentre il 99% dovrebbe assumere valori compresi tra $-2.5$ e $+2.5$; nel caso in cui il valore del residuo standardizzato sia maggiore di 3 probabilmente l'osservazione è un outlier.

3. **Residui studentizzati**: è la versione dei residui standardizzati ma relativamente al campione. Di conseguenza le forme analitiche saranno le medesime facendo però riferimento non alla varianza $\sigma^2$ ma alla varianza campionaria $s^2$.

   $$\varepsilon_i^* = \frac{\varepsilon_i}{s_{\varepsilon i}\sqrt{(1-h_{ii})}}$$

   Sono utilizzati per verificare la presenza di osservazioni anomale in campioni di non elevata numerosità. La versione dei residui studentizzati cosiddetta *jacknife* è ricavata calcolando il rapporto dei residui sulla deviazione standard dei residui ottenuta eliminando dal dataset l'$i$-esima osservazione, così per ogni residuo studentizzato.

4. **Covrati**: indica la variazione nel determinante della matrice delle covarianze delle stime eliminando la $i$-esima osservazione. Eliminando infatti il valore $i$-esimo provoco una variazione nel determinante che vado a quantificare.

$$\text{COVRATIO} = \frac{(det(\sigma_i X_i' X_i^{-1}))}{det(\sigma^2 (X_i' X_i)^{-1}))}$$

   Il valore di soglia è determinato da

$$1 \pm 3 \sqrt{\left(\frac{(k+1)}{n}\right)}$$

5. **Dfitts**: misura l'influenza dell'$i$-esima osservazione sulla stima dei coefficienti di regressione e sulla loro varianza, eliminandola dal dataset. Osservazioni con valori elevati di Dfitts sono associati a punti influenti. Con $\hat{y} - \hat{y_{(i)}}$ verifico infatti l'impatto che la rimozione dell'$i$-esima osservazione ha sul valore finale dell'output del modello.

$$\text{DFITTS} = \frac{\hat{y} - \hat{y}_{(i)}}{S_{e(i)} \sqrt{h_{ii}}}$$

   con valore soglia

$$\pm 2 \sqrt{\left(\frac{(k+1)}{n}\right)}$$

6. **Dfbetas**: misura l'influenza dell'$i$-esima osservazione sulle stime di ogni coefficiente di regressione separatamente, eliminandola dal dataset. Ancora una volta valori elevati indicano che l'osservazione influisce molto sulla stima dei parametri. Per questo indice infatti un valore è ritenuto anomalo non se cambia il valore previsto ma se cambia anche solo uno dei coefficienti, per sua natura è quindi un indice molto più stringente rispetto al precedente.

$$\text{DFBETAS} = \beta - \beta_{(i)} = X_{(i)} (X'X)^{-1} \frac{\varepsilon_i}{1 - h_{ii}}$$

   con valore soglia 2 oppure $2\sqrt{n}$ in caso si voglia tenere conto della numerosità delle osservazioni.

7. **Distanza di Cook**: misura l'influenza dell'$i$-esima osservazione sulla stima dei coefficienti di regressione *nel loro complesso*, in termini di capacità del modello di predire tutti i casi quando la singola osservazione viene rimossa dal dataset, per questo motivo è molto simile al Dfitts. Valori superiori a 1 ( o eventualmente a $4/n$, essendo $n$ il numero di osservazioni) indicano che il punto è influente.

$$D_i = \frac{(\beta - \beta_{(i)})(X'X)(\beta - \beta_{(i)})}{k \sigma^2_{(i)}}$$

# 8 Modello lineare classico multivariato

Consideriamo l'estensione multivariata (con più di una variabile dipendente) della regressione lineare multipla (con più di un regressore) che modella la relazione fra un insieme di $r$ variabili esplicative $z_1$,..,$z_r$, e $m$ variabili dipendenti $y_1$,..,$y_m$. Ognuna delle $m$ variabili dipendenti è legata a una particolare regressione multipla.

Per l'$i$-esimo individuo abbiamo:

$$\begin{aligned} y_i &= [y_{i1}, ..., y_{ij}, ..., y_{im}] \\ z_i &= [1, z_{i1}, ..., z_{ik}, ..., z_{ir}] \\ \varepsilon_i &= [\varepsilon_{i1}, ..., \varepsilon_{ij}, ..., \varepsilon_{im}] \end{aligned}$$

Mentre la matrice dei parametri $\beta$ (m, r+1) per le m equazioni è:

$$\beta = \begin{bmatrix} \beta_{10} & \beta_{1k} & \beta_{1r} \\ \dots & \dots & \dots \\ \beta_{j0} & \beta_{jk} & \beta_{jr} \\ \dots & \dots & \dots \\ \beta_{m0} & \beta_{mk} & \beta_{mr} \end{bmatrix}$$

In sintesi ogni *riga* della matrice dei parametri $\beta$ si riferisce ad una variabile risposta $y_{1,\dots,m}$ mentre ogni colonna si riferisce ad una variabile esplicativa $z_{1,\dots,r}$.
Nel suo complesso perciò il modello multivariato appare come

$$Y_{(m,n)} = B_{(m,r+1)} Z_{(r+1,m)} + E_{(m,n)}$$

Con il contenuto della matrice delle variabili dipendenti interpretabile come:

- Ogni colonna rappresenta un individuo con i valori assunti dalle $y_m$ variabili dipendenti per quell'individuo.
- Ogni riga rappresenta il valore assunto dalla singola variabile dipendente $y_i$ su tutti gli individui.

Le ipotesi del modello sono analoghe a quelle formulate per il modello univariato ma, essendo applicate su più variabili dipendenti risultano molto più stringenti:

1. Parametri lineari;
2. Valori attesi degli errori casuali sono nulli $E(\varepsilon_{ij}) = 0$;
3. Gli errori casuali all'interno gli ogni equazione e *anche tra diverse equazioni* sono omoschedastici e incorrelati La matrice di varianze e covarianze dei residui assume infatti la forma

$$\Sigma_E = \begin{bmatrix} \sigma^2 I_n & 0 & \dots & \dots & 0 \\ 0 & \sigma^2 I_n & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & \sigma^2 I_n \end{bmatrix}$$

   Con dimensione $(nm, nm)$ poichè ogni matrice $\Sigma_\varepsilon$ relativa ad ogni singola equazione è di dimensione $(n, n)$ ed essendo $m$ il numero di variabili dipendenti $y$ otteniamo

appunto una matrice di varianze e covarianze di questa dimensionalità. Mentre gli elementi diagonali di questa matrice rappresentano gli errori relativi alla medesima equazione, le matrici 0 che non si trovano sulla diagonale, racchiudono le correlazioni fra gli errori relativi ad equazioni diverse. Per le matrici 0 infatti abbiamo sulla diagonale la correlazione di *ogni individuo* con sè stesso relativamente alle diverse variabili dipendenti $y$ (ovvero le scelte dell'$i$-esimo individuo riguardo una determinata $y_i$ non influenzerebbe le scelte dello stesso individuo riguardo un altra $y_j$, ipotesi molto forte) mentre per gli elementi non diagonali abbiamo la correlazione di ogni individuo *con un altro* (questa ipotesi molto meno forte rispetto alla precedente) ;

4. Le variabili esplicative $Z$ sono non stocastiche: per ogni osservazione, il valore delle $Z$ è una costante mentre il corrispondente valore di ogni $y$ è una variabile casuale influenzata dagli errori casuali;
5. Le $Z$ variabili esplicative sono non collineari con rango $(z = r + 1)$, contrariamente la matrice $Z'Z$ non sarebbe invertibile e non sarebbe calcolabile lo stimatore dei minimi quadrati;
6. La numerosità della popolazione $n$ è maggiore del numero degli $r$ parametri stimati più l'intercetta $(n > r + 1)$ per la stessa ragione, perciò per ogni equazione le stime dei minimi quadrati di $\hat{\beta}$ sono trovate in modo analogo al caso univariato:

$$\hat{\beta} = y_j Z'(Z'Z)^{-1}$$

   Di conseguenza **nel modello multivariato classico calcolare le soluzioni per ogni variabile dipendente $y$ singolarmente oppure tutte insieme, dal punto di vista descrittivo, è identico**.
7. Gli errori $E$ si distribuiscono come una normale multivariata:

$$E \sim N(0, s^2 I_{nm})$$

   con 0 vettore delle medie e $s^2 I_{nm}$ matrice di varianze e covarianze della variabile casuale multivariata $E$. Permane la condizione di ortogonalità poichè i residui sono incorrelati sia con le variabili esplicative $Z$ che con i valori predetti della variabile dipendente $\hat{Y}$.

Inoltre poichè

$$Y = \hat{Y} + \hat{E}$$

abbiamo che

$$(\Sigma_Y = YY') = (\hat{H} = \hat{B}ZZ'\hat{B}') + (\hat{\Sigma}_E = \hat{E}\hat{E}')$$

Con $\Sigma_Y$ matrice di varianze e covariante di $Y$, $\hat{H}$ matrice di varianze e covariante *spiegate* e $\hat{\Sigma_E}$ matrice di varianze e covariante *residue*.

La grossa differenza tra soluzione univariata e multivariata, però, sta nelle *covarianze*, poichè a differenza della soluzione univariata varianze spiegate e residue non sono scalari ma, appunto matrici; occorre quindi tenere conto delle correlazioni tra le soluzioni.

Nel caso multivariato classico infatti le parti diagonali di $\hat{\Sigma}_E$ e $\hat{H}$ sono *identiche*.

$$(\Sigma_Y = YY') = (\hat{H} = \hat{B}ZZ'\hat{B}') + \sigma^2 I_{nm}$$

L'$R^2$ in quest'ottica è una media pesata degli $R^2$ delle singole equazioni (sempre tenendo conto della numerosità dei casi che per tipo di rilevazione e missing value può non essere uguale nelle diverse equazioni).

**Osservazione.**

- *La dipendenza della variabile dipendente $y_j$ da $Z$* ***non influenza*** *la dipendenza delle altre variabili $y_m$.*
- *Abbiamo le* ***stesse*** *variabili esplicative in tutte le equazioni del sistema.*
- *La correlazione simultanea tra i disturbi è* ***costante*** *nel tempo.*

# 9 Inferenza nella Regressione Multivariata

Gli stimatori OLS sono corretti ed efficienti, poichè il teorema di Gauss-Markov vale anche per il caso multivariato. Infatti nell'ambito degli stimatori lineari e corretti del vettore dei parametri, lo stimatore $\beta$ dei minimi quadrati è quello a *varianza minore*. Inoltre per il modello di regressione multivariata con rango pieno con errori $E$ normalmente distribuiti anche le $m$ variabili dipendenti $Y$ sono distribuite secondo una normale multivariata

$$Y \sim N(BZ, \Sigma_Y)$$

mentre i parametri stimati $\beta$

$$\beta \sim N(B, \hat{H})$$

con $\hat{H}$ matrice di varianze e covarianze spiegata della popolazione , positiva definita ed efficiente che si dimostra essere distribuita in modo indipendente da $E$ matrice degli errori. É però il caso di notare che sia $\Sigma_Y$ che $\hat{H}$ sono entrambi matrici di varianza e covarianza *non diagonali* e, di conseguenza, sono influenzate dalle correlazioni.

Si definisce invece *varianza generalizzata* di $\hat{H}$ il suo determinante. Decidiamo di utilizzare la varianza generalizzata di $\hat{H}$ perchè ci è impossibile usare sia $t$ che $F$ in quanto misure univariate. A proposito di quanto detto riguardo la non diagonalità di $\hat{H}$, la sua varianza generalizzata è proprio una misura di variabilità che considera la correlazione tra le variabili.

La varianza generalizzata si dimostra infatti uguale a 0 in caso di presenza di:

- variabile costante nelle unità statistiche;
- variabile perfettamente correlata con un'altra;
- variabile combinazione lineare di altre variabili.

Analogamente si definisce *varianza generalizzata* di $\Sigma_E$ il determinante della matrice di varianza-covarianza residua.

Considerando che $\hat{H}$ si distribuisce come una variabile casuale di **Wishart** con $r$ *gradi di libertà* e $\Sigma_E$ sempre come una Wishart con $r - n$ *gradi di libertà* e considerando la

Wishart una generalizzazione multivariata di $F$ possiamo quindi definire come **test del rapporto di verosimiglianza Lambda di Wilks**:

$$\wedge = \frac{|\Sigma_E|}{|\Sigma_E + \hat{H}|}$$

Che si distribuisce *asintoticamente* come una $\chi^2$ con *mr gradi di libertà*.
Sempre da $\wedge$, inoltre, si ricava una distribuzione asintotica di $F$

$$F = \frac{(1 - \wedge)}{\wedge}$$

che nel caso sia rispettata l'ipotesì di normalità dei residui

$$E \sim N(0, \sigma^2 I_{nm})$$

permette di costruire *test multivariati* per i parametri del modello analoghi a quelli costruiti utilizzando $F$ nel caso univariato.
Il test del rapporto di verosimiglianza Lambda di Wilks assume come ipotesi nulla $H_0$:

$$H_0 : \hat{B} = 0$$

per cui nel caso $H_0$ si rivelasse vera allora $\wedge$ tenderà ad 1 per la struttura stessa di $F$.
Se infatti $H_0$ è vera il numeratore e il denominatore di $\wedge$ tenderanno a coincidere poichè $\hat{H}$ tenderà a 0. Perciò la regione di accettazione di $H_0$ (nullità dei parametri $\hat{B}$) è per valori di $\wedge$ vicini all'1. Tenendo quindi conto di queste circostanze e per la struttura di $F$, per $\hat{H}$ che tende a 0 anche $F$ tenderà a 0, cadendo così nella *regione di accettazione* del test.

La regione di rifiuto di $H_0$ è per valori di $\wedge$ più piccoli di 1, in cui il numeratore è più piccolo del denominatore per la presenza di $\hat{H}$. Data la struttura della $F$, al crescere di $\wedge$ decresce il numeratore e cresce il denominatore. In sintesi, quindi, se $H_0$ falsa allora $\hat{H}$ diventa più grande ed $F$ tende ad infinito cadendo nella *regione di rifiuto* del test.

Perciò per il test basato su $F$ asintotica: regione di accettazione di $H_0$ è per $p - value$ inferiori ad $\alpha$; regione di rifiuto di $H_0$ per $p - value$ superiori ad $\alpha$; analogamente si costruiscono intervalli di confidenza per i parametri e per i valori predetti delle $Y$.

Esistono inoltre altri test che possiedono la stessa distribuzione, impalcature ed $H_0$ della Lambda di Wilks:

- **Traccia di Lawney-Hotelling**

$$LH = \frac{|\hat{H}|}{|\Sigma_E|}$$

- **Traccia di Pillai**

$$P = \frac{|\hat{H}|}{|\hat{H} + \Sigma_E|}$$

- **Massimo autovalore di Roy**

$$\text{Max autovalore di } \frac{|\hat{H}|}{|\hat{H} + \Sigma_E|}$$

In modo analogo ad $F$, si possono costruire altri test con $H_0$ particolari:

- Test sulla non significatività di un gruppo di variabili esplicative rispetto a tutte le variabili dipendenti.
$$H_0 : \hat{B} = 0$$
- Test sull'uguaglianza dei parametri relativi a diversi gruppi di variabili esplicative nelle singole equazioni.
$$H_0 : B_{kj} = B_{gj}$$
- Test sull'uguaglianza dei parametri relativi alle stesse variabili in coppie di diverse equazioni.
$$H_0 : B_{cA} = B_{vA}$$

# 10 Modello lineare generalizzato

Guardare Modello lineare classico multivariato.
Quando cambiano le ipotesi sugli errori si ha il modello lineare generalizzato: $Y = BZ + E$ in cui la matrice di covarianza degli errori non è più necessariamente diagonale e gli errori potrebbero essere eteroschedastici.

Nell'ipotesi classica:

1. Gli errori sono omoschedastici all'interno delle stesse equazioni: per ogni individuo rispetto alla medesima variabile dipendente la parte spiegata è uguale
2. Gli errori sono omoschedastici tra equazioni diverse: per ogni individuo rispetto alle diverse variabili dipendenti la parte spiegata è uguale
3. Gli errori sono incorrelati all'interno delle stesse equazioni: il comportamento di ogni individuo rispetto alla medesima variabile dipendente non è legato a quello degli altri individui
4. Gli errori sono incorrelati fra equazioni diverse: il comportamento di ogni individuo rispetto a diverse variabili dipendenti non è legato al proprio e a quello degli altri individui.

Nell'ipotesi intermedia:

5. Gli errori sono omoschedastici all'interno delle stesse equazioni: per ogni individuo rispetto alla medesima variabile dipendente la parte spiegata è uguale
6. Gli errori sono eteroschedastici tra equazioni diverse: per ogni individuo rispetto alle diverse variabili dipendenti la parte spiegata è diversa
7. Gli errori sono incorrelati all'interno delle stesse equazioni: il comportamento di ogni individuo rispetto alla medesima variabile dipendente non è legato a quello degli altri individui

8. Gli errori sono correlati fra equazioni diverse: il comportamento di ogni individuo rispetto a diverse variabili dipendenti è legato al proprio e a quello degli altri individui.

Nell'ipotesi estrema:

9. Gli errori sono eteroschedastici all'interno delle stesse equazioni: per ogni individuo rispetto alla medesima variabile dipendente la parte spiegata è diversa
10. Gli errori sono eteroschedastici tra equazioni diverse: per ogni individuo rispetto alle diverse variabili dipendenti la parte spiegata è diversa
11. Gli errori sono correlati all'interno delle stesse equazioni: il comportamento di ogni individuo rispetto alla medesima variabile dipendente è legato a quello degli altri individui
12. Gli errori sono correlati fra equazioni diverse: il comportamento di ogni individuo rispetto a diverse variabili dipendenti è legato al proprio e a quello degli altri individui.

Quindi occorre usare non le singoli sottomatrici di correlazione degli errori $\Sigma\varepsilon(i)$, ma la matrice $\Sigma\varepsilon$ relativa all'intero modello.

# 11 Modello SURE

Secondo un approccio più realistico, degli r regressori si usano solo i regressori effettivamente legati alle diverse variabili dipendenti: r1 nella prima, r2 nella seconda, ..., rm nell'ultima.
In altre parole nel Modello SURE abbiamo regressori diversi per ogni equazione all'interno dell'insieme complessivo dei regressori per l'insieme delle equazioni del modello.
$\Sigma$jrj è quindi la somma di tutti i regressori nelle diverse equazioni.
Ciò permette di risolvere anche il problema di una numerosità diversa delle osservazioni delle diverse equazioni. Data nj è la numerosità delle osservazioni per l'equazione j-esima il complesso delle numerosità è dato da $\Sigma$jnj.
La soluzione dei minimi quadrati per la stima dei coefficienti sembra simile a quella dei minimi quadrati generalizzati ma solo in apparenza:

- Il modello è caratterizzato dalla presenza delle variabili esplicative ZA, ZB, ZC diverse da equazione ed equazione.
- Gli errori sono:
  - omoschedastici e incorrelati nella stessa equazione
  - eteroschedastici fra diverse equazioni
  - correlati per lo stesso individuo e incorrelati tra individui diversi fra diverse equazioni

# 12 Il problema dei dati gerarchici e uso di Regressione multilevel

I modelli statistici si basano su campionamento casuale semplice da popolazione infinita o finita con reinserimento. In tal caso vige l'ipotesi di indipendenza tra le singole osservazioni.
In molti casi, però i dati risultano essere raggruppati in cluster ovvero presentano una struttura gerarchica (ad esempio Ospedale - Pazienti). In tali casi il campionamento casuale semplice non risulta efficiente, ma appare preferibile effettuare un campionamento a più stadi perché si desidera analizzare le relazioni tra le variabili che possono essere misurate a livelli di raggruppamento dei dati diversi (livelli gerarchici della struttura dei dati).
Il campionamento a stadi implica la dipendenza tra le osservazioni appartenenti allo stesso gruppo (esse hanno medesima probabilità di essere estratte).
Ad esempio gli studenti appartenenti alla stessa scuole condividono stesso ambiente, stessi insegnanti, stesso quartiere di provenienza oltre a scambi e comunicazioni tra essi.
La dipendenza tra le unità di primo livello (micro) appartenenti alla stessa unità di secondo livello (macro) è cruciale per l'analisi.
Cosa succede se si ignora la struttura gerarchica dei dati?

1. FALLACIA ECOLOGICA: quando ad una variabile riferita al livello macro si vuole dare validità micro.
2. FALLACIA ATOMISTICA: quando ad una variabile riferita al livello micro si vuole dare validità macro.

la variabile dipendente y ha sia un aspetto individuale sia di gruppo; la variabile x pur essendo misurata a livello individuale contiene anche una quota di variabilità imputabile al gruppo, infatti la media di x in un gruppo può essere diversa dalla media di x in un altro gruppo poiché la composizione della x nei gruppi può essere diversa.
Le regressioni a livello macro considerano i dati aggregati dati dalla media di x ed y e sono quindi diverse dalle regressioni a livello micro tra x ed y.
L'analisi delle relazioni entro i gruppi può portare a risultati molto diversi da quelli ottenuti considerando le relazioni tra i gruppi. In altri termini la struttura dei dati ed il loro raggruppamento può avere effetto anche in altro modo: facendo variare i coefficienti della regressione da gruppo a gruppo.
Yij = $\beta$0j + $\beta$ijxij + rij
Diverse intercette $\beta$0j e coefficienti di regressione $\beta$1j per ciascun gruppo:

- se i coefficienti $\beta$0j e $\beta$1j sono entrambe costanti allora la struttura gerarchica non ha effetto = regressione OLS;
- se i due coefficienti dipendono entrambe da j allora la regressione OLS non può essere utilizzata:
  - Se varia solo $\beta$0j con j allora si ha un modello random intercept
  - Se anche $\beta$1j varia con j allora il modello è detto random coefficient

Pertanto data la struttura dei dati, si può pensare di porre assieme la regressione tra i gruppi e la regressione entro i gruppi.

## 13 Modello multilevel: definizione e significato

Il Modello Multilevel è un Modello di analisi della covarianza a effetti casuali, la regressione cattura la relazione disaggregata tra i dati e quindi descrive la varianza nei gruppi, mentre l'analisi della varianza cattura la relazione aggregata fra i gruppi e quindi descrive la varianza fra gruppi. Spezza in due l'analisi mettendo insieme la covarianza: elimina gli aspetti individuali, analizza gli effetti di gruppo e riesce ad attribuire la varianza al gruppo di appartenenza.
In un primo tipo di modelli (mixed models) la relazione disaggregata tra i dati e la varianza nei gruppi sono descritte mediante parametri fissi mentre la relazione aggregata fra i gruppi e la varianza fra gruppi sono descritte come variabili casuali.
In un secondo tipo di modelli (random models) anche la relazione disaggregata tra i dati e la varianza nei gruppi sono descritte come variabili casuali.
I modelli finora studiati possono essere visti come sottocasi del modello Multilevel:

1. Per $u_j = 0$ e nessuna gerarchia dei dati, Modello Lineare:

$$y_i = \gamma_{00} + \Sigma_k \beta_k (x_{ik} - \bar{x}_k) + \varepsilon_i \ ;$$

2. Per $u_j = 0$ Regressione Multilevel:

$$y_{ij} = \gamma_{00} + \Sigma_k \beta_k (x_{ijk} - \bar{x}_k) + \varepsilon_{ij};$$

3. Per $\Sigma_k \beta_k (xik - \bar{x}_k) = 0$ e $u_j$ fisso Analisi Varianza:

$$y_{ij} = \gamma_{00} + u_j + \varepsilon_{ij};$$

4. Per $\Sigma_k \beta_k (x_{ik} - \bar{x}_k) = 0$ e $u_j$ stocastico Analisi Varianza Casuale:

$$y_{ij} = \gamma_{00} + u_j + \varepsilon_{ij}.$$

## 14 Modello Multilevel: OLS, Empty, Mixed, Total Effects

La stima del modello multilevel si compone di 4 step:

1. Si stima innanzitutto il modello lineare solitamente con il metodo di stima OLS (Unstructured ordinary least squares model (OLS)).
2. Poi si propone l'empty model (Unconditional means model UMM) vale a dire l'analisi della varianza a effetti casuali.
3. Random intercepts model (RIM) cioè l'analisi della covarianza a effetti casuali per l'analisi della varianza.
4. Random slopes and intercepts model (UGM) analisi della covarianza a effetti casuali sia per il modello lineare che per l'analisi della varianza.

OLS: Si consideri un modello Multilevel in cui appare solo la parte del Modello Lineare che viene stimata mediante metodo OLS con una sola variabile e ipotizzando che le variabili X e Y siano centrate ($y_{ij} = \beta_0 + \Sigma_{jk}\beta_k x_{ijk} + \varepsilon_{ij}$).
In questo modo si vede quale sia l'effetto delle variabili esplicative sulla variabile dipendente se i dati non fossero centrati. Naturalmente gli errori si distribuiscono come una normale.
EMPTY MODEL: Nel modello ANOVA ad effetti casuali detto anche empty model si ha che: $y_{ij} = v_j + r_{ij}$. In questo caso la variabile dipendente y dipende dagli effetti casuali:

- a livello di gruppo, Vj, distribuiti in modo normale $N(\gamma_0, \tau_2)$
- a livello individuale, dai residui $R_{ij}$, distribuiti in modo normale $N(0, \sigma_2)$

La variabilità all'interno di ogni gruppo è quindi dovuta solamente alla distribuzione casuale della variabile dipendente.
L'intercetta casuale a livello di gruppo può essere scomposta in due parti: l'intercetta fissa media tra tutti i gruppi e la misura della sua deviazione attorno alla media tra i gruppi di tipo casuale: $v_j = \gamma_0 + u_j$.
Possiamo riscrivere il modello nel seguente modo: $y_{ij} = \gamma_0 + u_{jj} + r_{ij}$.
In questo modello quindi la variabilità totale di y può essere scomposta nella somma delle varianze ai due livelli, varianza fra i gruppi e varianza nei gruppi: $var(y) = var(U_j) + var(R_{ij}) = \sigma_2 + \tau_2$.
Si può quindi definire il coefficiente $r^2/(r^2 + \sigma^2)$ di correlazione intraclasse:
Il coefficiente di correlazione intraclasse misura quindi la quota di varianza di y spiegata dall'appartenenza ai gruppi degli Individui. Se il coefficiente di correlazione intraclasse è nullo, ovvero tutti gli uj sono nulli, allora il raggruppamento è irrilevante ed è inutile utilizzare altri modelli rispetto alla regressione semplice. Se invece il coefficiente di correlazione intraclasse è positivo è necessario considerare un modello gerarchico.
Il test F come in ogni analisi della varianza può essere utilizzato per verificare in termini inferenziali l'ipotesi che le intercette casuali uj siano nel complesso tra loro equivalenti (se non c'è differenza fra gruppi). In questo caso il test F serve per capire se nel complesso vale l'ipotesi nulla che le medie parziali ottenute nel campione possano essere ritenute nel complesso equivalenti. Per confrontare tra loro le strutture di secondo livello (ad esempio scuole, ospedali, università) come nell'analisi della varianza casuale non si utilizzano i valori delle medie campionarie, non informative del vero valore di Uj ma i loro intervalli di confidenza che comprendono con una probabilità del 90%, 95%, 99% i valori veri ignoti di Uj.
Ciò significa probabilizzare la gerarchia fra medie parziali in quanto, quanto più sono piccoli gli intervalli di confidenza è maggiore la loro capacità di fornire informazioni sui valori veri ignoti di Uj. A differenza che nell'analisi della varianza casuale nel Modello Multilevel che è come ricordato un'analisi della covarianza casuale, tali intervalli di confidenza sono al netto dell'influenza delle variabili X del modello lineare. Questa probabilizzazione della gerarchia influenza e rende più robusto il confronto fra strutture di secondo livello in quanto una media parziale uj di una struttura J si considera superiore a un'altra media parziale ug di una struttura G se e solo se l'estremo inferiore del suo intervallo di confidenza inf

(uj) è più grande dell'estremo superiore dell'altra sup(ug) in quanto solo in questo caso con un elevato grado di probabilità il valore vero di J sarà più grande di G.

MIXED MODEL:

Se si inserisce nel modello una variabile esplicativa $x_k$ il modello diventa il vero e proprio random intercept model (mixed model) $y_{ij} = \gamma_0 + \beta_1 x_{ij} + u_j + \delta_{ij}$ dove $u_j$ è la determinazione della variabile casuale $U_j$ $(j = 1, ..., p)$ variabili casuali indipendenti e identicamente distribuite normalmente dalla formula $N(\gamma_0, \tau_2)$ e rappresentano i residui di 2 livello. Esse sono indipendenti e quindi incorrelate con i residui di primo livello $\delta_{ij}$ determinazioni delle variabili casuali normalmente distribuite $\Delta_{ij} \sim N(0, \sigma_2)$

In questo caso la variabile dipendente y dipende:

- dalle variabili x e dai relativi parametri fissi $\beta_1$.
- dall'effetto casuale a livello di gruppo, $u_j$, che si distribuisce in modo normale $N(\gamma_0, \tau_2)$.
- dall'effetto casuale a livello individuale $\delta_{ij}$, che si distribuisce in modo normale $N(0, \sigma_2)$.

La correlazione intraclasse misura la quota di varianza di y spiegata dall'appartenenza ai gruppi degli Individui al netto della quota di varianza spiegata da x (a differenza del modello empty in questa circostanza ho x che spiega una parte della variabilità non dovuta all'appartenenza a un gruppo di un individuo). Per questa ragione il suo valore può decrescere anche molto dal caso empty. Il modello comprende 4 parametri da stimare:

- i coefficienti di regressione $\gamma_0$ e $\beta_1$ e le componenti della varianza $\sigma^2$ e $\tau_0^2$.
- Il coefficiente di regressione $\beta_1$ può essere interpretato come variazione di Y corrispondente ad una variazione unitaria di x.
- In un modello di regressione semplice la variabilità di Y non spiegata dalla regressione è semplicemente data dal residui $\delta_{ij}$.

La variabilità in un modello multilevel fa riferimento a più popolazioni:

- Le v.c. $U_j$ possono essere viste come le variabili casuali che descrivono i residui a livello di gruppo, ovvero gli effetti di gruppo non spiegati da x.
- La v.c. $\Delta$ può essere vista come variabile casuale che descrive i residui a livello individuale, ovvero gli effetti individuali non spiegati da x.

Rappresentazioni di un modello random intercept:

- Micro model: $y_{ij} = \beta_1 x_{ij} + R_{ij}$
- Macro model: $\beta_{0j} = \gamma_{00} + U_{0j}$

Come unica equazione multilevel: $y_{ij} = \gamma_{00} + \beta_1 x_{ij} + U_{0j} + R_{ij}$

- parte fissa del modello $y_{ij} = \gamma_{00} + \beta_1 x_{ij}$

- parte casuale (random part) del modello $U_{0j} + R_{ij}$

Come varianze e covarianze:

- livello 1 $\sigma^2$
- livello 2 $\tau_0^2$

TOTAL MODEL:
La relazione tra variabile dipendente ed esplicative può variare tra i gruppi in modi diversi: si può avere un'eterogeneità delle regressioni tra i diversi gruppi (si parla anche di interazione gruppo – covariate). Ad esempio nel caso dell'analisi delle performance degli studenti appartenenti alle scuole, si può assumere che l'effetto dello stato socio economico o dell'intelligenza individuale sulle performance possa essere diverso nelle singole scuole.
La struttura dei dati ed il loro raggruppamento può essere spiegato quindi anche facendo variare i coefficienti della regressione da gruppo a gruppo.
$y_{ij} = \beta_{0j} + \beta_{1j}x_{ij} + R_{ij}$

- diversi $\beta_{0j}$ (INTERCETTE)
- diversi $\beta_{1j}$ (COEFFICIENTI DI REGRESSIONE: l'effetto di X su Y può essere diverso nei singoli gruppi)
- se i coefficienti $\beta_{0j}$ e $\beta_{1j}$ sono entrambi costanti e la struttura gerarchica non ha effetto: regressione OL
- se solo il coefficiente dell'intercetta $\beta_{0j}$ varia con j allora si ha un modello random intercept
- mentre se anche il coefficiente di regressione $\beta_{1j}$ varia con j allora il modello è detto random coefficient

# 15 Metodi di stima e Verifica di ipotesi

I parametri da stimare nel modello random intercept sono: coefficienti di regressione $\gamma$ e componenti di varianza, $\sigma_2$ e $\tau_2$. Gli effetti casuali $U_{0j}$ non sono parametri ma variabili casuali latenti, ovvero non direttamente osservabili. La letteratura riporta metodi per la stima dei parametri sotto l'assunzione che i residui $U_{0j}$ e $R_{ij}$ siano distribuiti normalmente quali: ML e REML. Esistono due metodi di stima (sotto l'assunzione di normalità dei residui) per la stima dei parametri, il full maximum likelihood (ML) e il restricted (anche chiamato residual) maximum likelihoo (REML).
REML è alternativo a ML. Questo metodo massimizza la verosimiglianza (likelihood) dei residui osservati ottenendo le stime degli effetti fissi usando metodi ≪non likelihoodlike≫ come ordinary least squares (OLS) o generalized least squares (GLS)) e successivamente usa queste per massimizzare la verosimiglianza dei residui (sottraendo gli effetti misti) per ottenere le stime dei parametri della varianza. Diversi algoritmi sono disponibili per ottenere queste stime: EM (expectation-maximisation), Fischer Scoring, IGLS e

RIGLS. Questi sono algoritmi iterativi che convergono dopo alcune iterazioni alle stime ML o REML. Per testare i parametri fissi del modello si utilizza la seguente ipotesi nulla (ipotesi di significatività) su ciascun parametro $\rightarrow H_0 : \gamma_h = 0$.
Questa ipotesi viene verificata con un test t; questo test è noto come WALD TEST. Sotto l'ipotesi nulla il test ha approssimativamente una distribuzione t con d.f. basati sulla struttura multilevel dell'analisi. Per testare più parametri (fissi e random) del modello invece viene utilizzato il deviance test. Dalla stima del modello lineare con il metodo ML si ottiene la verosimiglianza del modello, da cui: DEVIANCE = -2 lnL (misura della bontà di adattamento ai dati del modello).
Solitamente la deviance viene interpretata in termini differenziali, ovvero si calcola la differenza tra le deviance di modelli alternativi. Si tratta di confrontare i valori osservati della variabile dipendente con i valori teorici di due modelli:

1. con le variabili esplicative di interesse e l'altro senza alcuna variabile (empty-model);
2. con le variabili esplicative di interesse e l'altro che contiene "tanti parametri quante sono le osservazioni" (saturated model).

Il confronto si basa sulla funzione di log-verosimiglianza. Perciò indicate rispettivamente con D0, Dmod, Dsat le devianze calcolate per il modello vuoto (empty-model), il modello considerato e il modello saturo, valori di Dmod più prossimi a 0 che non a D0 faranno propendere per ritenere "buono" il modello considerato. Ognuna delle devianze ha distribuzione asintotica Chi-quadrato (con j gradi di libertà pari al numero delle esplicative). Le loro differenze avrà distribuzione Chi-quadrato (con k gradi di libertà pari alla differenza del numero di esplicative).
Se l'obiettivo è quello di sottoporre a verifica l'ipotesi che riguarda la nullità congiunta di tutti i coefficienti (esclusa l'intercetta), si può pensare a ragion veduta di operare un confronto fra due modelli, l'empty-model e il modello ipotizzato. Questo test può essere applicato sia alla parte fissa sia a quella random del modello.