Appunti - Statistical Modeling

Alberto Filosa

30/6/2020

Indice

(PART) Modello Lineare

1 Modello Lineare Classico

I ***modelli*** esplicitano la relazione statistica e matematica tra le variabili tramite un compromesso tra l'adattamento dei dati ed il principio di parsimonia degli stessi. La differenza tra questi due modelli è che quello matematico approssima in termini esatti i dati, mentre un modello statistico è caratterizzato da errori che possono dipendere da variazioni individuali, errori di misura oppure nel caso stocastico per rapporto campione-popolazione. Il modello è specificato in riferimento alle unità statistiche di una data popolazione:

$$y_i = f(x_{1i}, \dots, x_{ki}) + e_i \quad \forall i = 1, \dots n$$

Le fasi per procedere alla stima di un modello partono da considerazioni teoriche, che stanno alla base dell'ipotesi che vogliamo verificare, passando dalla verifica pratica di quanto ipotizzato. Le fasi di stima e verifica del modello possono essere fatte fino all'identificazione del modello migliore.

Un modello di regressione è una relazione che lega la variabile dipendente a determinate variabili esplicative. All'interno di un modello sono presenti le seguenti variabili:

- Variabile *dipendente*, chiamata anche risposta, (y) la colonna di un dataset che si vuole predire;
- Variabile *esplicativa/e*, chiamata anche covariata/e, $(x_1, x_2, ..., x_k)$ le variabili utilizzate per spiegare il fenomeno che rappresentano la componente sistematica del modello con natura deterministica;
- Variabile *errore*, solitamente stocastica, (ε) che sintetizza l'errore circa la relazione che lega la variabile risposta alle esplicative.

I modelli che possono essere specificati sono molteplici e dipendono non solo dalle variabile risposta ed esplicative, ma anche dal tipo di relazione, che può essere lineare o non lineare. I modelli possono essere:

- *Semplici*, nella quale sono presenti una sola variabile dipendente ed una sola esplicativa:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- *Multipli*, nella quale sono presenti solo una variabile dipendente e più di una variabile esplicativa:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon = X\beta + \varepsilon$$

- *Multivariati*, nella quale sono presenti più di una variabile dipendente e più di una variabile esplicativa:

$$Y = XB + E$$

Nel momento in cui si specifica un modello lineare, bisogna tenere in considerazione una componente stocastica dell'errore commessa al variare del campione spiegando la variabile risposta tramite le covariate. Sia $y = X\beta + \varepsilon$ la costruzione del modello lineare, si presenta il modello in forma matriciale:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

dove:

- y rappresenta il vettore dei valori della variabile casuale dipendente;
- X rappresenta la matrice costituita dall'insieme ordinato delle n osservazioni sulle k variabili esplicative;
- β rappresenta il vettore dei parametri ignoti da stimare;
- ε rappresenta il vettore degli errori casuali non osservabili.

1.1 Criterio dei Minimi Quadrati Ordinari

Per poter stimare il vettore dei parametri β uno degli approcci utilizzati nel contesto lineare è il Metodo dei Minimi Quadrati Lineari, chiamato anche in inglese ***O***rdinary ***L***east ***S***quare (**OLS**). Si formula il problema di ricerca del vettore β che minimizza la norma del vettore degli scarti ε:

$$\min \sum_{i=1}^{k} (y_i - {x_i}^t \beta)^2 = \min \sum_{i=1}^{k} {\varepsilon_i}^2 = \min \varepsilon^t \varepsilon = \min (y - X\beta)^t (y - X$$

Lo stimatore β minimizza la somma dei quadrati degli scarti fra i valori osservati y_i sulla variabile y mediante la combinazione lineare degli elementi di β moltiplicati per i valori osservati x_i delle variabili esplicative. Procedendo al calcolo del gradiente tramite delle derivate parziali della funzione obiettivo rispetto al vettore β si ottiene:

$$\frac{\partial(\varepsilon^t \varepsilon)}{\partial \beta} = -2X^t (y - X\beta)$$

La condizione per l'esistenza di un minimo impone la nullità del gradiente $X^t(y - X\beta) = 0$ che conduce al sistema di equazioni normali $X^tX\beta = X^ty$. Se la matrice dei coefficienti sotto condizioni generali è di rango pieno, il sistema possiede un'unica soluzione che è rappresentata da $\beta^t = (X^tX)^{-1}X^ty$ e se le variabili sono centrate si ha:

$$y_1 = y^t = X\beta + \varepsilon^t = y^t + \varepsilon^t$$

dove $y^t = (X^tX)^{-1}X^ty$ e $\varepsilon^t = (I - X(X^tX)^{-1}X^ty)$

Lo stimatore rappresenta quindi lo stimatore dei minimi quadrati del vettore dei parametri; ogni parametro b_{1j} indica la variazione di y al variare unitario di x_j mentre rimangono invariati i restanti, con $i \neq j$. Sia D_x la matrice diagonale i cui elementi diagonali sono le varianze delle variabili X e le X e y sono standardizzate cioè divise per il loro scarto quadratico medio $X^* = XD_x{}^{-1/2}$, $y^* = \frac{y}{\sigma_y}$ il coefficiente standardizzato è il coefficiente di regressione fra X^* e y^* che misura la relazione tra X ed y al netto dell'ordine di grandezza:

$$B^* = (X^{*t}X^*)^{-1}X^{*t}y^*$$

1.2 Ipotesi Modello Lineare

Per specificare e stimare il modello bisogna considerare le seguenti ***assunzioni***, senza la validità delle quali il modello costruito potrebbe risultare non valido:

1. *Linearità*: un modello di regressione è lineare quando tutti i termini sono la costante o un parametro moltiplicato per una variabile indipendente. Per soddisfare questa ipotesi, il modello deve adattarsi al modello lineare;
2. *Non sistematicità degli Errori*: la casualità determina i valori del termine di errore. Affinchè il modello sia imparziale il valore medio dell'errore $E(\varepsilon|X) = 0$; altrimenti la parte del termine di errore sarebbe prevedibile e bisognerebbe aggiungere le informazioni al modello. In questo caso, si impone che la media dei residui è nulla:

   $$E(y|X) = X\beta E(y) = E(X\beta+\varepsilon) = E(X\beta)+E(\varepsilon) = X\beta$$

3. *Sfericità degli Errori*: la varianza è omoschedastica, ovvero non cambia per ogni osservazione:

   $$E(\varepsilon) = E(y-X\beta) = (X\beta-X\beta) = 0 \implies Var(\varepsilon_i) = (E(\varepsilon_i^2)) = \sigma^2 \quad Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i, \varepsilon_j) = 0$$

   all'ampiezza campionaria:

4. *Non stocasticità delle covariate*: i valori delle x_j covariate non sono soggetti a fluttuazioni stocatiche da campione a campione, perciò $E(X) = X$ e $Cov(X, \varepsilon) = 0$. La parte non fissa delle x_j finisce in ε;
5. *Non collinearità delle covariate*: le variabili della matrice X sono linearmente indipendenti. X ha rango uguale al numero delle colonne (più una costante), per cui X^tX non è singolare: in caso contrario X^tX non è invertibile ed il modello non risolvibile;
6. *Numerosità della popolazione*: il numero di osservazioni è sempre maggiore del numero di caratteri osservati: $n \geq p + 1$. Se questa proprietà non è soddisfatta, la matrice delle covariate non è invertibile e dunque non è possibile costruire alcun modello.

A queste assunzioni se ne aggiunge un'ultima, relativa agli stimatori che permette di costruire test ed intervalli di confidenza:

7. *Normalità degli errori*: il soddisfacimento dell'ipotesi che $\varepsilon_i \sim N(0, \sigma^2)$ provoca anche la normalità nella distribuzione di y e di $\hat{\beta}$. Per il teorema del limite centrale ε_i, y e $B^\wedge$ sono distribuite come una Normale:

$$y \sim N(X\beta, \sigma^2 I_n) \implies B^\wedge = (X^tX)^{-1}X^ty \sim N(\beta, \sigma^2(X^tX)^{-1})$$

1.3 Proprietà degli stimatori

Per valutare l'affidabilità di uno stimatore si considerano tre ***proprietà degli stimatori***:

1. *Correttezza*, o anche chiamata non distorsione, ovvero la differenza tra valore atteso e valore reale deve essere nulla:

   $$E[B] = \beta \implies E[B] - \beta = 0$$

2. *Efficienza*, ovvero minimizzare l'errore quadratico medio. Uno stimatore corretto è relativamente efficiente rispetto ad un altro corretto se la sua varianza è più piccola;

3. *Consistenza*, ovvero se la probabilità che cada in un intervallo del valore vero tende a una al crescere

$$n \to \infty \implies P(|B - \beta| < k) \to 1$$

Si dimostra che lo stimatore OLS è ***B***est ***L***inear ***U***nbiased ***E***stimator (**BLUE**); in particolare, non è solo corretto, ma anche consistente.

1.4 Test d'Ipotesi

Generalmente uno stimatore segue una distribuzione normale, $\beta_j \sim N(\mu, \sigma^2)$; si effettua un ***test d'ipotesi*** con probabilità dell'errore di primo grado α arbitrario per verificare la significatività di un parametro β_j con varianza σ^2 nota:

Se il valore del parametro ricavato dal campione ricade nell'area di accettazione, si accetta l'ipotesi nulla legata al valore del campione, altrimenti la si rifiuta. Solitamente nel modello lineare tale ipotesi assume che il parametro sia nullo e quindi si vuole minimizzare l'ipotesi di rifiutare H_0 quando è vera.

Esiste una relazione tra i test d'ipotesi e gli intervalli di confidenza: per i primi si rifiuta H_0 se il *p-value* ha poca probabilità che il parametro nel campione ha un valore estremo anche se H_0 è vero; per i secondi, invece, con confidenza al 95% fornisce un intervallo di valori plausibili per il coefficiente angolare in cui il 95% degli intervalli comprende il vero valore del coefficiente angolare stesso.

1.4.1 Test Normale con varianza nota:

Si considera $N \sim (\beta_j, \frac{\sigma^2}{n\sigma_{jj}^{-1}})$ e si verifichi l'ipotesi $H_0 : \beta_j = 0$ contro l'ipotesi $H_1 : \beta_j \neq 0$. L'intervallo di confidenza è definito nel seguente modo:

$$P[-Z_{\frac{\alpha}{2}} < \frac{\beta_j}{\frac{\sigma}{\sqrt{n\sigma_{j.j}^{-1}}}} < +Z_{\frac{\alpha}{2}}] = P[-Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n\sigma_{j.j}^{-1}}} < \beta_j < Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n\sigma_{j.j}^{-1}}}] = 1-\alpha$$

Si rifiuta H_0 se l'intervallo centrale non comprende il valore di β_j ricavato dal campione

1.4.2 Test Normale con varianza ignota:

L'assunzione di conoscenza della varianza della popolazione è però raramente verificata nella realtà. Per verificare il test d'ipotesi si introduce una nuova variabile, chiamata t di Student, definita come il rapporto tra una variabile standardizzata ed una χ^2_{n-k-1}. Anche in questo caso si ha interesse a verificare il seguente sistema di ipotesi: $H_0 : \beta_j = 0$ contro l'ipotesi $H_1 : \beta_j \neq 0$ con varianza ignota:

$$P[-t_{\frac{\alpha}{2}} < \frac{\hat{\beta}_j - \beta_j}{\frac{s}{\sqrt{n\sigma_{j.j}^{-1}}}} < +t_{\frac{\alpha}{2}}] = P[\hat{\beta}_j - t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n\sigma_{j.j}^{-1}}} < \beta_j < \hat{\beta}_j + t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n\sigma_{j.j}^{-1}}}] = 1-\alpha$$

1.4.3 Test F per il modello

Per verificare la significatività del modello complessivamente, è necessario adottare un altro tipo di test. In questo caso si fa riferimento alla devianza complessiva e alla devianza residua. Per verificare il test d'ipotesi si introduce la distribuzione F di Snedecor:

$$\frac{Dev_{Tot}/k}{Dev_{Res}/(n-k-1)} \sim F_{k,n-k-1}$$

In questo caso il sistema di ipotesi da verificare è $H_0 : \beta_1 = ... = \beta_k = 0$ contro l'ipotesi H_1 : almeno un $\beta_j \neq 0$

Per compiere il test F su *uno o più parametri*, si ipotizza che $q < k$ siano nulli: $H_0 : \beta_1, ..., \beta_q = 0$. Si considerino $Dev_{Spieg\ 1}$ e $Dev_{Res\ 1}$ delle ultime $k - q$ variabili di X ed il seguente rapporto:

$$\frac{(Dev_{Spieg} - Dev_{Spieg\ 1})/(k-q)}{Dev_{Res}/(n-p-1)} \sim F_{k-q,n-k-1}$$

Se $q = k - 1$ il test F può essere effettuato sui singoli parametri, costruendo le seguenti ipotesi: $H_0 : \beta_j = 0$ contro $H_1 : \beta_j \neq 0$. La funzione F diventa:

$$\frac{{\beta_j}^2}{\frac{n\sigma_{j.j}^{-1}\chi^2/1}{\chi^2/(n-p-1)}} = \frac{{\beta_j}^2}{s^2/n\sigma_{jj}^{-1}} \sim F_{1,n-p-1} = t_{\alpha/2}(n-k-1)$$

2 Violazione degli Errori

Dato che *tutti i modelli sono falsi*, le ipotesi classiche non sono scorrette, tuttavia semplificano in modo eccessivo: in circostanze reali è difficile trovare fenomeni che le soddisfino tutte. Si possono però eliminare tutte le ipotesi classiche (a eccezione delle ultime due elencate precedentemente) in modo tale da ottenere un modello più veritiero.

2.1 Errori eteroschedastici

Si consideri il modello lineare classico $y_j = X_j\hat{B} + \hat{\varepsilon}_j$. Per ottenere stime efficienti per i parametri del modello lineare gli errori devono essere omoschedastici, ovvero la varianza degli errori deve essere costante e non dipendere dal valore delle variabili indipendenti:

$$Var(\varepsilon_i) = E({\varepsilon_i}^2) = \sigma^2 \implies E(\varepsilon_i) = E(\varepsilon_i|x_i)$$

Questo si verifica dal momento che il valore atteso del singolo errore è pari a zero: $E(\varepsilon_i) = 0$. Nel momento in cui ciò non avviene, si è in presenza di errori ***eteroschedastici***: $Var(\varepsilon_i) = \sigma_i^2$. In termini matriciali, la matrice di varianza-covarianza degli errori Σ_E avrà sulla diagonale valori di varianza diversi.

Le proprietà di linearità, consistenza e correttezza rimangono valide:

$$\begin{aligned}
\mathrm{E}(B^*) &= \mathrm{E}((X^tX)^{-1}X^ty) && (1)\\
&= (X^tX)^{-1}X^t\mathrm{E}(y) && (2)\\
&= (X^tX)^{-1}X^t\mathrm{E}(Xb + e^*) && (3)\\
&= (X^tX)^{-1}X^t\mathrm{E}(b) = b && (4)
\end{aligned}$$

ma lo stimatore B^* non è più BLUE e quindi non efficiente:

$$\begin{aligned}
Var(B^*) &= \mathrm{E}[(B^* - b)(B^* - b)^t] && (5)\\
&= \mathrm{E}[((X^tX)^{-1}X^ty - b)((X^tX)^{-1}X^ty - b)^t] && (6)\\
&= [(X^tX)^{-1}X^t]\mathrm{E}(e^*e^{*t})[(X^tX)^{-1}X^t]^t && (7)\\
&= [(X^tX)^{-1}X^t]\Sigma_{e^*}[(X^tX)^{-1}X^t]^t && (8)
\end{aligned}$$

Inoltre, la stima campionaria s^2 di σ^2 sottostima il vero valore della distribuzione, dato che si è in presenza di molteplici variabili casuali. Il test t-Student mostra dei valori erroneamente più elevati e di conseguenza gli intervalli di confidenza diventano più stretti e non affidabili. I test di significatività sui parametri b_j risultano più permissivi del dovuto. Analogamente si compiono gli stessi discorsi per i test basati sulla F di Snedecor.

La violazione dell'ipotesi di omoschedasticità è rilevabile attraverso due modalità: una di ispezione grafica ed una orientata verso dei test su misura. Per quanto riguarda l'ispezione grafica, preferibile come primo approccio, è necessario ricorrere a visualizzazioni quale lo Scatterplot: con quest'ultimo è possibile mettere a confronto diverse coppie di elementi sugli assi, tutte egualmente valide, tra cui le più utilizzate sono:

- La variabile risposta *vs* le covariate;
- I residui stimati *vs* i residui predetti sula risposta;
- I residui al quadrato *vs* i residui predetti;
- I valori osservati *vs* i valori predetti;
- I residui *vs* ogni regressore.

Per quanto riguarda invece i test di misura, il primo utile allo scopo è il *Test di White*, basato sull'ipotesi nulla che vi sia omoschedasticità tra i residui e che quindi $H_0 : Var(\varepsilon|X) = \sigma^2 I_n$. Il meccanismo del test prevede i seguenti passaggi:

1. Si regredisce Y rispetto alle covariate X_j e si ricava l'errore ε_i;
2. Si regredisce ε_i rispetto alle covariate x_j, al loro quadrato ${X_j}^2$ e alle loro interazioni;
3. Si determina il coefficiente R^2 della regressione e si procede alla costruzione del valore (considerando n il numero di regressori): $LM = nR^2 \sim \chi^2_n$
4. Se $LM \notin IC$ allora ε_i^2 varia al variare delle x_j e sarà quindi da confermare la presenza di eteroschedasticità (ovvero si rifiuta H_0).

Un altro test utilizzato è chiamato *Test di Breusch-Pagan*, che che invece di regredire su ε_i^2 effettua la regressione su una diversa funzione:

$$S^2 = \frac{\sum_i^n \varepsilon_i^2}{n} \sim \chi^2_k$$

Questo meccanismo punta alla normalizzazione dei residui ε_i^2, per cui se S^2 ed ε^2 divergono significa che è presente eteroschedasticità.

2.2 Errori Autocorrelati

Si consideri il modello lineare classico $y_j = X_j\hat{B}+\hat{\varepsilon}_j$. Per garantire stime efficienti, è necessario verificare che gli errori ε non siano tra di loro incorrelati, o chiamati anche ***autocorrelati***. Infatti si ipotizza che la correlazione tra gli errori è nulla:

$$Cor(\varepsilon_i, \varepsilon_j) = E(\varepsilon_j\varepsilon_i) = 0 \quad \forall i \neq j$$

Molte volte capita, soprattutto durante lo studio di serie storiche, che sia presente una correlazione tra errori in momenti successivi. Gli errori si possono dividere in errori ritardati in un tempo, $\varepsilon_i^{\#}$, ed errori omoschedastici ***i***dentici ed ***i***denticamente ***d***istribuiti (**iid**) come una Normale, η_i. L'autocorrelazione può essere ***positiva***, ovvero i residui consecutivi sono positivi, quindi con

stesso segno e simili valori, oppure ***negativa***. È possibile classificare l'autocorrelazione in base al suo grado; in particolare, un'autocorrelazione di primo grado indica errori correlati con il loro valore ritardato di un tempo, mentre un'autocorrelazione di s-esimo grado quando gli errori sono correlati con il loro valore ritardato di s gradi:

$$\varepsilon_i^{\#} = \rho\hat{\varepsilon}_{i-1}^{\#} + \eta_i : \varepsilon_i^{\#} = \rho\hat{\varepsilon}_{i-s}^{\#} + \eta_i$$

Essi non incidono sulle proprietà di linearità, correttezza e consistenza degli stimatori OLS:

$$\begin{aligned} \mathrm{E}(B^*) &= \mathrm{E}(X^tX)^{-1}X^ty && (9)\\ &= (X^tX)^{-1}X^t\mathrm{E}(y) && (10)\\ &= (X^tX)^{-1}X^t\mathrm{E}(Xb + e^{\#}) && (11)\\ &= (X^tX)^{-1}X^t\mathrm{E}(b) = b && (12) \end{aligned}$$

Tuttavia, lo stimatore non è più BLUE in quanto lo stimatore non è efficiente:

$$\begin{aligned} Var(B^{\#}) &= \mathrm{E}[(B^* - b)(B^* - b)^t] && (13)\\ &= \mathrm{E}[((X^tX)^{-1}X^ty - b)((X^tX)^{-1}X^ty - b)^t] && (14)\\ &= [(X^tX)^{-1}X^t]\mathrm{E}(e^{\#}e^{\#t})[(X^tX)^{-1}X^t]^t && (15)\\ &= [(X^tX)^{-1}X^t]\Sigma_{e^{\#}}[(X^tX)^{-1}X^t]^t && (16) \end{aligned}$$

Inoltre, le stime della varianza dei parametri non sono più corrette ed affidabili; di conseguenza, la t-Student ottiene valori erroneamente più elevati ed i relativi intervalli di confidenza diventano più stretti con un'area di rifiuto più ampia.

Per individuare nelle osservazioni la caratteristica di autocorrelazione si costruiscono diversi grafici:

- Scatterplot della risposta con le covariate: $y \sim x_j \ \ \forall j$;
- Scatterplot dei residui con le covariate: $\varepsilon \sim x_j \ \ \forall j$;
- Scatterplot dei residui con i residui ritardati: $\varepsilon \sim \varepsilon_{-1}$;
- *Correlogramma*, nella quale si visualizzano le correlazioni a diversi gradi. Analizzando la funzione di autocorrelazione dei residui (**ACF**) si determina il tipo di modello autoregressivo.

Il *Test di Durbin-Watson* offre uno strumento analitico per verificare la presenza di autocorrelazione a diversi gradi. Il test considera come ipotesi nulla la mancanza di autocorrelazione: $H_0 : \rho = Corr(\varepsilon_i{}^{\#}, \varepsilon_{i-1}{}^{\#}) = 0$, mentre l'ipotesi alternativa può essere verificata su entrambe le code della distribuzione od in modo unidirezionale: $H_1 : \rho \neq, >, < 0$.

$$\begin{aligned} DW &= \frac{\sum_{i=1}^{n}(\varepsilon_i{}^{\#} - \varepsilon_{i-1}{}^{\#})^2}{\sum_{i=1}^{n}\varepsilon_i{}^{\#2}} = && (17)\\ &= \frac{E(\varepsilon_i{}^{\#})^2 + E(\varepsilon_{i-1}{}^{\#})^2 - 2E(\varepsilon_i{}^{\#}, \varepsilon_{i-1}{}^{\#})}{E(\varepsilon_i{}^{\#})^2} = && (18)\\ &= \frac{2\sigma^2 - 2\rho\sigma^2}{\sigma^2} = && (19)\\ &= 2(1-\rho) \in [0,4] && (20) \end{aligned}$$

Il valore tende a 2 in caso di assenza di autocorrelazione, a 0 in caso di autocorrelazione positiva e 4 in caso di autocorrelazione negativa. Convenzionalmente se $DW \in [0,3]$ non è presente autocorrelazione.

2.3 Metodo di stima WLS

Nel momento in cui si presentano errori eteroschedastici ed incorrelati, ovvero la varianza dell'errore del modello non è costante, si violano delle ipotesi del modello lineare classico. Di conseguenza non è possibile utilizzare lo stimatore OLS, in quanto non più efficiente, ma lo stimatore ***W***eighted ***L***east ***S***quares (**WLS**). Esso permette di stimare la varianza delle singole componenti d'errore ε_i dati x_i. Da un modello $y_i = x_iB^* + \varepsilon_i^*$, si passa ad un modello per riportare la varianza degli errori ad una costante:

$$Y^+ = X^+B + \varepsilon^+$$

con variabili:

$$y_i \to y_i^+ = \frac{y_i}{\sqrt{s_i}} \quad x_{ij} \to x_{ij}^+ = \frac{x_{ij}}{\sqrt{s_i}} \quad \varepsilon_i \to \varepsilon_i^+ = \frac{\varepsilon_i^*}{\sqrt{s_i}}$$

In caso di eteroschedasticità ed incorrelazione, la matrice inversa degli errori è quindi più semplice, $S_\varepsilon^{-1} = diag(1/s_1^2, \dots, 1/s_k^2)$. Tuttavia, esistono dei problemi di ordine computazionale nella quale le varianze diventano negative per alcune osservazione. Per correggere l'eteroschedasticità si utilizza la funzione esponenziale e successivamente lo si riporta alla normalità con il logaritmo.

2.4 Modello di stima GLS

Nel caso in cui gli errori siano autocorrelati ed eteroschedastici, si ipotizza che esista una correlazione fra errori in momenti successivi. Si parla di autocorrelazione se al variare di X si osserva una fluttuazione dei valori di Y con lo stesso segno (autocorrelazione positiva) o di segno opposto (autocorrelazione negativa), oltre un certo intervallo di confidenza. È possibile ricavare stime per errori correlati tramite una stima dei parametri in una equazione che tenga conto della struttura di autocorrelazione seriale, ovvero il metodo proposto da Durbin.

Il metodo di stima ***G***eneralized ***L***east ***S***quare (**GLS**) consiste nello stimare il coefficiente di autocorrelazione di i-esimo ordine attraverso un modello lineare semplice con errori correlati di tempo t, $y_t = \beta_0 + \beta_1 x_t + \hat{\varepsilon}_t^{\#}$ per stimare il coefficiente di correlazione ρ del t-esimo ordine:

$$\hat{\varepsilon}_i^{\#} = a_0 + a_1^{\#} x_1 + \cdots + a_k^{\#} x_k + \rho \hat{\varepsilon}_{i-1}^{\#}$$

Si moltiplica ogni elemento dell'equazione ritardata per ρ, ottenendo:

$$\rho y_{t-i} = \rho X_{t-i} \beta + \rho \hat{\varepsilon}_{t-i}$$

Infine si sottrae l'equazione precedente all'equazione in forma normale $Y - t - \rho y_{t-i}$, ottenendo un modello OLS per i parametri trasformati e con errori incorrelati:

$$y_t^{\#} = X_t^{\#} \beta + \varepsilon_t^{\#} \tag{21}$$
$$(y_t - \rho y_{t-i}) = (X_t - \rho X_{t-i})\beta + (\varepsilon_t - \rho \varepsilon_{t-i}) \tag{22}$$

Il modello rispetta tutte le proprietà classiche di correttezza, consistenza ed efficienza. Dato che $\mathrm{E}(w_i) = \mathrm{E}(\hat{\varepsilon}_i^{\#} - \rho \hat{\varepsilon}_{i-1}^{\#}) = 0$, si ha:

$$\mathrm{cov}(w_i, w_{i-1}) = \mathrm{cov}(\hat{\varepsilon}_i^{\#} - \rho \hat{\varepsilon}_{i-1}^{\#}, \hat{\varepsilon}_{i-1}^{\#} - \rho \hat{\varepsilon}_{i-2}^{\#}) \tag{23}$$
$$= \mathrm{cov}(\hat{\varepsilon}_i^{\#}, \hat{\varepsilon}_{i-1}^{\#}) - \rho \mathrm{cov}(\hat{\varepsilon}_i^{\#}, \hat{\varepsilon}_{i-2}^{\#}) - \rho \mathrm{cov}(\hat{\varepsilon}_{i-1}^{\#}, \hat{\varepsilon}_{i-1}^{\#}) + \rho^2 \mathrm{cov}(\hat{\varepsilon}_{i-1}^{\#}, \hat{\varepsilon}_{i-2}^{\#}) \tag{24}$$
$$= \rho - \rho^3 - \rho + \rho^3 = 0 \tag{25}$$

In alternativa, è possibile utilizzare un modello autoregressivo per la quale si inserisce nell'equazione iniziale un errore ritardato che tiene conto dell'autocorrelazione dell'i-esimo ordine:

$$y = X\beta + AR_i + \varepsilon$$

con $AR_i + \varepsilon$ e $\rho_{v_j, v_k} = 0 \quad \forall j \neq k$

2.5 Stimatore GLS

Nel caso in cui gli errori non siano sferici in quanto eteroschedastici e correlati con correlazione diversa tra diversi errori si utilizza lo stimatore GLS, interpretabile come stimatore OLS basato su variabili trasformate per mezzo delle proprietà degli autovettori ed autovalori ricavati dalla matrice di var-cov dei residui Σ_ε. Si procede con la decomposizione spettrale della matrice degli errori:

$$\Sigma_\varepsilon = \begin{bmatrix} s_1^2 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{12} & s_1^2 & \cdots & \rho_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1k} & \rho_{12} & \cdots & s_k^2 \end{bmatrix} = \sigma^2 V V^t$$

con $V = \sigma(\sqrt{AL})A^t$, A la matrice degli autovettori e L la matrice degli autovettori di Σ_ε. Premotiplicando per V^{-1} le componenti del modello si ottiene una nuova funzione con variabili trasformate e $\Sigma_{\varepsilon^\circ}$ omoschedastica ed incorrelata:

$$V^{-1} y = y^\circ \;\text{ e }\; V^{-1} X = X^\circ \implies y^\circ = X^\circ \beta^\circ + \varepsilon$$

Lo stimatore β° risulta essere corretto e consistente:

$$E(B^\circ) = \mathrm{E}[(X^{\circ t} X^\circ)^{-1} X^{\circ t} y^\circ] \tag{26}$$
$$= \mathrm{E}[b^\circ + (X^{\circ t} X^\circ)^{-1} X^{\circ t} E^\circ] \tag{27}$$
$$= b^\circ + (X^{\circ t} X^\circ)^{-1} X^{\circ t} \mathrm{E}(E^\circ) = b^\circ \tag{28}$$

Inoltre, secondo il *Teorema di Aitken* è lo stimatore più efficiente per il modello generalizzato nonostante abbia una varianza maggiore rispetto al metodo OLS:

$$\mathrm{E}[(B^\circ - b^\circ)(B^\circ - b^\circ)^t] = \mathrm{E}[((X^{\circ t} X^\circ)^{-1} X^{\circ t} y^\circ - b^\circ)((X^{\circ t} X^\circ)^{-1} X^\circ \tag{29}$$
$$= [(X^{\circ t} X^\circ)^{-1} X^{\circ t}][\mathrm{E}(EE^t)][X^\circ (X^{\circ t} X^\circ)^{-1} \tag{30}$$
$$= \sigma^2 (X^t \Sigma_{E^\circ} X)^{-1} \tag{31}$$

Tuttavia, la stima GLS necessita di assumere come nota la matrice di varianza-covarianza dei residui Σ_ε, oppure si può calcolare una sua stima S_{E° a patto che sia consistente a $\lim_{n\to\infty} S_{\varepsilon^\circ} = \Sigma_{\varepsilon^\circ}$

Di conseguenza è possibile utilizzare lo stimatore ***F***easibile ***G***eneralized ***L***east ***S***quare (**FGLS**). Questa soluzione solitamente viene utilizzata per il controllo di eteroschedasticità o di semplice autocorrelazione.

3 Violazioni del modello lineare

3.1 Multicollinearità

La ***multicollinearità*** sorge nel momento in cui è presente una elevata correlazione tra due o più variabili esplicative del modello o linearmente dipendente da altre. In questo caso, le varianze delle stime dei parametri diventano molto alte. Di conseguenza, i parametri stimati non sono più attendibili e l'intervallo di confidenza costruito con la t Student (o la F di Snedecor) tende ad avere una ampiezza maggiore, diminuendo la probabilità di stima del vero valore del parametro. Esistono diversi modi per studiare questo fenomeno: il primo è quello di costruire la *matrice di correlazione*, in modo da osservare la presenza di valori molto alti dell'indice di correlazione ρ tra coppie di variabili. Un secondo approccio è il calcolo dell'*indice di tolleranza* (***TOL***), che misura il grado di dipendenza lineare di una variabile esplicativa rispetto alle altre:

$$TOL_{x_j} = 1 - R^2_{x_j \mid x_1,\dots,x_{j-1},x_{j+1},\dots,x_p} \in [0,1]$$

Se una variabile è multicollineare con le altre $TOL_{x_j} = 0$, con il coefficiente di determinazione corretto della regressione della variabile in funzione delle altre variabili presenti pari ad 1; se è incorrelata $TOL_{x_j} = 1$.

Un terzo approccio è il calcolo della *varianza multifattoriale* (***VIF***), calcolato come il reciproco del TOL:

$$VIF_{x_j} = \frac{1}{TOL_{x_j}} \in [1, +\infty)$$

In questo caso se una variabile è incorrelata con le restanti covariate $VIF_{x_j} = 1$. Una variabile è collineare nel caso in cui il valore $VIF_{x_j} > 10$.

La principale differenza tra questi due indici è che il TOL è un indice normalizzato, mentre il VIF si estende su quasi tutti i numeri reali positivi con soglia 10.

Un ultimo approccio è il calcolo dell'*indice di collinearità*, calcolato come la radice del rapporto tra l'autovalore massimo della matrice X^tX e ogni autovalore. Un valore superiore a 10 della quota di varianza di ogni variabile associata indica la presenza di multicollinearità.

3.2 Linearità

Per linearità si intende una relazione di tipo polinomiale tra la variabile risposta e le covariate del modello, secondo una combinazione lineare. Prima di affermare che il modello utilizzato non è adatto anche introducendo nuove variabili, bisogna verificare la linearità del modello tramite strumenti grafici, tra i quali:

- Scatterplot variabile risposta vs covariata $y \sim x_j \ \ \forall j$;
- Scatterplot residui vs valori osservati $\varepsilon \sim x_i \ \ \forall i$;
- Scatterplot residui vs valori previsti $\varepsilon \sim \hat{x}_i \ \ \forall i$.

Un altro modo per verificare la linearità del modello è l'utilizzo di indici, come l'indice di adattabilità R^2, e di test F e t, osservando la significatività delle variabili esplicative. È da notare che la non linearità potrebbe dipendere anche da poche variabili esplicative e non da tutte. Nel momento in cui il modello non è lineare, si possono utilizzare delle trasformazioni dei parametri o delle variabili tali che li renda lineari e stimare i parametri applicando la trasformazione inversa per ottenere il parametro originale. Nel caso in cui le componenti siano intrinsecamente non lineari, si utilizza lo stimatore ***N***on ***L***inear ***S***quare (***NLS***), che sfrutta algoritmi numerici per risolvere il problema di minimizzazione non lineare.

Volendo utilizzare funzioni di variabili risposte non lineari in X si possono riformulare una vasta famiglia di funzioni di regressione lineare come regressioni multiple. Quelle più utilizzate sono funzioni polinomiali e trasformazioni logaritmiche, nella quale esistono tre modelli principali:

- *Log-lineare*, in cui ad un incremento dell'1% di x corrisponde un incremento di β_1% di log_y;
- *Log-log*, in cui ad un incremento dell'1% di log_x corrisponde un incremento di β_1% di log_y;
- *Linear-log*, in cui ad un incremento dell'1% di log_x corrisponde ad un incremento di β_1 di y.

3.3 Non Normalità

Nel caso in cui gli errori ε_i sono i.i.d. $\sim N(0, \sigma^2)$ è possibile ricavare la distribuzione degli stimatori ed i test statistici con i relativi intervalli di confidenza. Nel caso in cui gli errori ***non*** fossero ***normali***, in campioni sufficientemente elevati con $n > 25$, per il Teorema del Limite Centrale la distribuzione degli errori tende asintoticamente alla normalità. Per campioni più piccoli non è possibile applicare test ed intervalli di confidenza perché sono basati sull'ipotesi di normalità degli errori. Se vengono violati i principi di normalità:

1. I parametri β, espressi come combinazione lineare degli errori, non sono distribuiti come una Normale, come le rispettive stime $\hat{\beta}$;

2. Non è possibile ricavare test basati sulla Normale standardizzata, t Student e F Snedecor per i parametri;
3. Non è possibile ricavare intervalli di confidenza per i parametri basati sulla Normale standardizzata e t Student;
4. Le stime OLS non coincidono con le stime di massima verosimiglianza (ML); di conseguenza, gli stimatori non sono più corretti a minima varianza tra tutti quelli corretti (VUE), ma sono ancora BLUE;

Tuttavia, gli stimatori OLS sono corretti:

$$\begin{aligned} \mathrm{E}(\hat{B}) &= \mathrm{E}(X^tX)^{-1}X^ty && (32)\\ &= (X^tX)^{-1}X^t\mathrm{E}(y) && (33)\\ &= (X^tX)^{-1}X^t\mathrm{E}(X\beta + e) && (34)\\ &= (X^tX)^{-1}X^t\mathrm{E}(\beta) = b && (35) \end{aligned}$$

Dato che le stime OLS non coincidono con quelle ML, alcuni software statistici potrebbero fornire stime non attendibili.

Per individuare i casi di non normalità, in prima istanza, è opportuno calcolare indici descrittivi: se moda, media e mediana coincidono, allora la distribuzione è Normale; per osservare questo fenomeno si costruisce un boxplot. Gli indici più utilizzati sono l'indice di simmetria, calcolato come il rapporto tra il momento terzo intorno alla media ed il cubo della varianza, e l'indice di curtosi, calcolato come il rapporto tra il momento quarto intorno alla media ed il cubo della varianza. Il secondo modo è attraverso le rappresentazioni grafiche, quali:

- Istogramma della distribuzione dei residui;
- Distribuzione cumulata dei residui, in modo da osservare evidenti irregolarità;
- qq-plot della distribuzione dei residui contro la Normale Standard, con i punti che si devono distribuire sulla retta;
- pp-plot della distribuzione cumulata dei residui contro la cumulata della Normale Standard in termini probabilistici.

Un'ultima modalità è con la costruzione di test statistici: essi sono detti non parametrici poiché testano la distribuzione dei parametri, perciò sono molto utili per analizzare problemi di normalità dei residui.

Il *Test di Shapiro-Wilk* prende in considerazione l'indice W calcolato come la combinazione lineare degli errori in rango con i parametri del modello costanti fratto le varianze campionarie relative all'errore, con ipotesi nulla $H_0 : \varepsilon \sim N(0, \sigma^2)$ ed alternativa $H_1 : \varepsilon \nsim N(0, \sigma^2)$. Gli estremi dell'indice W corrispondono rispettivamente alla regione di rifiuto e alla regione di accettazione:

$$W = \frac{\sum_{i=1}^{n}(\beta_i\varepsilon_i)^2}{\sum_{i=1}^{n}{\varepsilon_i}^2} \in [0, 1]$$

W è fortemente asimmetrico ed i suoi valori elevati possono portare ugualmente al rifiuto dell'ipotesi di normalità.

Il *Test di Kolmogorov-Smirnov* si basa sul calcolo della statistica test D, definita come la somma in valore assoluto della differenza tra le frequenze cumulate della distribuzione empirica da testare e quelle della Normale. Infine viene confrontata con le tavole numeriche: in caso di superamento del valore critico in base al livello di significatività scelto si rifiuta o accetta l'ipotesi nulla.

Il *Test di Asimmetria* è basato sulla ipotesi che la distribuzione sia Normale. In questo caso il valore atteso dell'indice di simmetria è $E(S) = 0$. Bisogna assicurarsi che il campione sia sufficientemente grande perché il test non riconosce che la distribuzione del modello è Normale quando il campione è piccolo. Per questo a partire dai dati il valore dell'indice è spesso ottenuto da una simulazione con 1000 iterazioni. Se S supera una certa soglia a cui è associato il livello di significatività α prescelto e quindi il suo p-value è inferiore a α si rifiuta l'ipotesi H_0 di normalità.

Il *Test di Curtosi* verifica se la distribuzione empirica ha la stessa curtosi della Normale, perciò in questo caso $K - 3$ ha valore atteso $E(K - 3) = 0$. Se K supera una certa soglia cui è associato il livello di significatività α prescelto e quindi il suo p-value è inferiore a α si rifiuta l'ipotesi H_0 di normalità.

I problemi di non normalità possono essere risolti mediante trasformazione della variabile risposta y, migliorando l'adattabilità del modello ai dati. Le trasformazioni più diffuse sono:

- $log(y)$ quando la varianza dei residui cresce con il modello o la distribuzione dell'errore ha una asimmetria positiva;
- y^2 quando la varianza dei residui è proporzionale a $E(y)$ o la distribuzione dell'errore ha una asimmetria positiva;
- $\sqrt{y}$ quando la varianza dei residui è proporzionale a $E(y)$;
- y^{-1} quando la varianza dei residui cresce significativamente al crescere di y nei diversi campioni.

3.4 Outlier

I valori *outlier* sono delle osservazioni all'interno di un dataset che alterano le stime dei parametri nella costruzione di un modello lineare. In particolare, bastano uno o pochi valori molto lontani dagli altri perché cambi completamente la somma delle differenze al quadrato e quindi le stime dei parametri. Essi possono essere distinti in punti *anomali*, che si discostano in modo rilevante dall'andamento generale, e punti *influenti*, che influenzano in misura rilevante le stime.

Gli outlier possono essere identificati graficamente tramite la costruzione di boxplot, se sono presenti punti superiori al baffo, e scatterplot univariati, nella quale i punti si discostano maggiormente dalla linea di tendenza delle osservazioni. All'aumentare del numero di dimensioni, tuttavia, si perde la possibilità di individuare gli outlier in termini grafici e sono necessari indici numerici.

Definendo la matrice di proiezione $H = X(X^tX)^{-1}X^t$, si chiamano *valori di leva* gli elementi h_{ii} sulla diagonale principale che rappresentano l'impatto della i-esima osservazione sulla capacità del modello di predire tutti i casi. Il valore medio del leverage è:

$$\overline{h} = \frac{k-1}{n}$$

con k il numero di covariate ed n il numero di osservazioni nel modello. Un valore di leva si considera significativamente alto se supera 2 o 3 volte il suo valore medio:

$$h_{11} > \frac{2(k-1)}{n}$$

I residui possono essere calcolati nella seguente formula matriciale:

$$\varepsilon = (I-H)y \implies Var(\varepsilon_i) = (1-h_{i.i})\sigma^2$$

Di conseguenza è possibile calcolare anche i *residui standardizzati*:

$$\varepsilon_i^* = \frac{\varepsilon_i}{\sigma\sqrt{1-h_{ii}}} \sim N(0,1)$$

Per verificare la presenza di outlier si utilizzano le tavole della distribuzione Normale. Il 95% della popolazione assume valori compresi tra $-2 < \varepsilon_i^* < +2$. Si considera outlier un valore il cui valore assoluto è $|\varepsilon_i^*| > 3$

I *valori studentizzati* sono utilizzati per verificare la presenza di osservazioni anomale in campioni di non elevata numerosità. La differenza con i residui standardizzati è che sono divisi per lo s.q.m di ogni osservazione invece che per quella generale:

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

I residui studentizzati *Jacknife* sono residui divisi per una stima della deviazione standard ottenuta eliminando dal dataset la i-esima osservazione dal modello e stimando nuovamente i parametri:

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

Si presentano ora indici complessi basati sulla eliminazione della i-esima osservazione dal dataset.

Il *covratio* misura la variazione del determinante della matrice Σ delle stime eliminando la i-esima osservazione. Il valore è significativo se supera la soglia $1 \pm 3\sqrt{\frac{k+1}{n}}$:

$$\text{COVRATIO}_i = \frac{det(\Sigma_i)}{det(\Sigma)} = \frac{det(\frac{1}{n-1}\tilde{X}_i^t\tilde{X}_i)}{\det(\frac{1}{n}\tilde{X}'\tilde{X})}$$

Il *Dfitts* misura l'influenza dell'i-esima osservazione sulla stima dei coefficienti di regressione e sulla loro varianza eliminandola dal dataset. Se lo scostamento super la soglia$\pm 2\sqrt{\frac{k+1}{n}}$ l'osservazione è da considerarsi outlier:

$$\text{DFITTS}_i = \frac{\hat{y} - \hat{y}_i}{s_i\sqrt{h_{i.i}}}$$

Il *Dfbetas* misura l'influenza della i-esima osservazione sulle stime di ogni singolo coefficiente di regressione eliminandola dal dataset. Se lo scostamento supera la soglia 2 o $2\sqrt{n}$ per almeno un parametro β, l'osservazione è da considerarsi outlier:

$$\text{DFBETAS}_i = \beta - \beta_i = X_i(X'X)^{-1}\frac{\varepsilon_i}{1-h_{ii}}$$

La *Distanza di Cook* misura l'influenza della i-esima osservazione sulla stima dei parametri del modello nel loro complesso. Se lo scostamento dei coefficienti di regressione supera la soglia 1, il valore è da considerarsi outlier:

$$D_i = \frac{\sum(\hat{y}_j - \hat{y}_{j.i})^2}{ks^2} \sim F_{(k,n-k)}$$

(PART) Modello Multivariato

4 Modello Lineare Multivariato

Per modello lineare multivariato si intende l'estensione multivariata della regressione lineare multipla a m variabili risposta y_j, con essa legata alle stesse variabili esplicative. Per l'i-esima osservazione si ha:

$$y_i = [y_{i1}, \dots, y_{ij}, \dots, y_{im}] \tag{36}$$
$$z_i = [1, z_{i1}, \dots, z_{ik}, \dots, z_{ir}] \tag{37}$$
$$\varepsilon_i = [\varepsilon_{i1}, \dots, \varepsilon_{ij}, \dots, \varepsilon_{i.m}] \tag{38}$$

Nella matrice dei parametri B di dimensioni $m \times r+1$ ogni riga si riferisce ad una variabile risposta e per ogni colonna ad una variabile esplicativa diventa:

$$B = \begin{bmatrix} \beta_{10} & \cdots & \beta_{1k} & \cdots & \beta_{1r} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ \beta_{j0} & \cdots & \beta_{jk} & \cdots & \beta_{jr} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{m0} & \cdots & \beta_{mk} & \cdots & \beta_{mr} \end{bmatrix}$$

Nel suo complesso perciò il modello multivariato appare come:

$$Y_{m\times n} = B_{m\times r+1} Z_{r+1\times n} + E_{m\times n}$$

o altrimenti scritto come:

$$y_j = \beta_j Z + \varepsilon_j \quad j = 1 \dots m$$

Ogni colonna della matrice Y rappresenta un carattere, mentre ogni riga rappresenta i valori dei caratteri per un singolo individuo.

La matrice dei valori predetti è calcolata nel seguente modo:

$$\hat{Y} = \hat{B}Z = YZ^t(ZZ^t)^{-1}Z$$

mentre la matrice degli errori è $\hat{E} = Y - \hat{B}Z = Y(I - Z^t(ZZ^t)^{-1}Z)$. Le condizioni di ortogonalità rimangono le medesime rispetto al caso univariato in quanto i residui sono incorrelate con le variabili esplicative e con i valori predetti:

$$\hat{E}Z^t = Y[I - Z^t(ZZ^t)^{-1}Z]Z^t = YZ^t - YZ^t = 0 \qquad \hat{Y}\hat{E}^t = YZ^t(ZZ^t)^{-1}ZZ^t - YZ^t(ZZ^t)^{-1}ZZ^t = YZ^t - YZ^t = 0$$

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

1. I parametri e le variabili sono *lineari*;
2. *Non sistematicità degli errori*, ovvero il valore atteso degli errori casuali sono nulli (nella seguente ipotesi): $\mathrm{E}(\hat{E}_{ij}) = 0$
3. Sfericità degli errori: gli errori casuali in ogni e tra diverse equazioni sono omoschedastici ed incorrelati. La matrice di var-cov $\Sigma_E = \mathrm{diag}(\sigma^2 I_n, \dots, \sigma^2 I_n)$. Gli elementi diagonali della matrice rappresentano gli errori relativi della stessa equazione, gli altri elementi rappresentano la correlazione tra errori relativi ad equazioni diverse;
4. Le variabili esplicative Z *non* sono *stocastiche*, in quanto per ogni osservazione il valore della covariata Z è una costante mentre il corrispondente valore di ogni y_j è una v.c influenzata dagli errori casuali $\hat{\varepsilon}_j$;
5. Le variabili esplicative sono *non collineari* con rango $r+1$, altrimenti la matrice ZZ^t non sarebbe invertibile e non potremmo calcolare lo stimatore OLS;
6. La numerosità della popolazione n è maggiore del numero dei r parametri stimati più l'intercetta ($n > r+1$) per la stessa ragione, per cui per ogni equazione le stime OLS dei parametri si calcolano analogamente al caso univariato;
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 var-cov della v.c multivariata E.

4.1 Inferenza

Dato il modello lineare multivariata della popolazione è possibile definire il modello campionario e stimare con il metodo OLS la matrice dei parametri ($\hat{B}$) e gli errori campionari ($\hat{E}$):

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

Se il modello è omoschedastico e gli errori sono incorrelati, la matrice Σ_ε è diagonale ed analogamente al caso multiplo la matrice dei parametri è pari a $\hat{B} = YZ^t(ZZ^t)^{-1}$.

Lo stimatore OLS è corretto, consistente ed efficiente.

$$\mathrm{E}[(\hat{B}-B)^t((\hat{B}-B))] = \mathrm{E}[(YZ^t(ZZ^t)^{-1}-B)^t(((YZ^t(ZZ^t)^{-1}-B))] \quad (39)$$

$$= (ZZ^t)^{-1}Z\mathrm{E}(\hat{E}^t\hat{E})Z^t(ZZ^t)^{-1}Z \quad (40)$$

$$= \sigma^2(ZZ^t)^{-1}ZZ^t(ZZ^t)^{-1} \quad (41)$$

$$= \sigma^2(ZZ^t)^{-1} \quad (42)$$

, di conseguenza il teorema di Gauss-Markov è valido anche nel contesto multivariato, lo stimatore dei parametri $\hat{B}$ è BLUE e lo stimatore ML di massima verosimiglianza, uguale a quello OLS imponendo anche l'ipotesi di normalità, è UNVUE.

Detto ciò, anche la matrice della variabile risposta Y e delle variabili esplicative Z si distribuiscono come una Normale Multivariata:

$$Y \sim N(BZ, \Sigma_Y) \quad (43)$$

$$\hat{B} \sim N(B, \sigma^2(ZZ^t)^{-1}) \quad (44)$$

Essendo il caso multivariato analogo a quello univariato, si possono ricavare le soluzioni equazione per equazione ed utilizzare i test N, t-Student e F di Snedecor per verificare la nullità dei singoli parametri o a gruppi per la non significatività del modello. Inoltre, si possono utilizzare le tre distribuzioni per costruire intervalli di confidenza verificandole per ogni equazione. È comunque possibile compiere test su una dimensione multivariata dato il carattere del modello.

Dato $Y = \hat{B}Z + \hat{E}$ è possibile definire la matrice di var-cov campionarie:

$$\Sigma_Y = \frac{1}{n}\hat{B}ZZ^t\hat{B}^t + \sigma_{\hat{E}} = \hat{H} + \sigma_{\hat{E}}$$

La matrice di var-cov spigata $\hat{H}$ è positiva ed efficiente. Inoltre si distribuisce come una v.c. di Wishart, W_r in modo indipendente da $\sigma_{\hat{E}}$. Quest'ultima è definita come la matrice di var-cov residua distribuita come una v.c. W_{n-r}.

Si definisce invece varianza generalizzata di Wilks $\hat{H}$ il suo determinante $\det(\hat{H})$; per le distribuzioni multivariate si usa questo indice di variabilità in quanto ha il vantaggio di essere univariato e permette quindi di costruire facilmente test t ed F. La varianza generalizzata inoltre considera la correlazione delle variabili, e si dimostra infatti uguale a 0 in caso di presenza di:

- Variabile costante nelle unità statistiche;
- Variabile perfettamente correlata con un'altra (rango non pieno);
- Variabile combinazione lineare di altre variabili.

Analogamente si definisce varianza generalizzata di Σ_E il determinante della matrice di varianza-covarianza residua.

4.1.1 Test di Wilks

Il test di Wilks è definito come il test del rapporto di verosimiglianza:

$$\Lambda = \frac{\det(\Sigma_E)}{\det(\Sigma_E + \hat{H})} = \prod_{i=1}^{n}\frac{1}{1+\lambda_i}$$

con $\lambda_1 \geq \cdots \geq \lambda_p$ gli autovalori non nulli di $(\Sigma_E)^{-1}\hat{H} \sim \Lambda_{n,m,r}$. Da Λ si ricava una distribuzione asintotica $F = \frac{1-\Lambda}{\Lambda}$.

Quando $\varepsilon \sim N$ si possono costruire test del rapporto di verosimiglianza Λ, analogo al test F nel modello univariato. Il test ha come ipotesi nulla che tutte le Z non sono significative rispetto a tutte le variabili risposte, mentre l'alternativa è che almeno una è significativo rispetto ad una variabile dipendente:

$$H_0 : \hat{B} = 0 \quad H_1 : \hat{\beta}_j \neq 0$$

Sotto l'ipotesi nulla $\Lambda \to 1$, dato che il numeratore e denominatore tenderebbero a coincidere, mentre $F \to 0$. Analogamente si possono costruire intervalli di confidenza per i parametri e per i valori predetti delle Y.

Si possono costruire anche diversi test su diverse ipotesi e possono essere verificati mediante il rapporto di varianze generalizzate di Wilk:

1. Test sulla non significatività di un gruppo di variabili Z_1 rispetto a tutte le variabili dipendenti Y, con $H_0 : B_{(1)} = 0$;
2. Test sulla non significatività di tutte le variabili Z rispetto a ad un gruppo di variabili dipendenti Y_2, con $H_0 : B_{(2)} = 0$;
3. Test sulla non significatività di alcune variabili Z_3 rispetto ad alcune variabili dipendenti Y_3, con $H_0 : B_{(3)} = 0$;
4. Test sull'uguaglianza dei parametri relativi a 2 gruppi di variabili Z_k e Z_g rispetto a un gruppo di variabili dipendenti Y_j, con $H_0 : B_{kj} = B_{gj}$;
5. Test sull'uguaglianza dei parametri relativi alle stesse variabili Z_c rispetto a un insieme di variabili dipendenti Y_A, con $H_0 : B_{cA} = B_{vA}$.

5 Modello Lineare Generalizzato

Il modello lineare multivariato generalizzato supera le ipotesi, molto stringenti, del modello lineare multivariato classico, rimuovendo le ipotesi sugli errori. Delle 7 ipotesi del modello lineare classico generalizzato quella inerente la sfericità degli errori è l'unica che deve essere riproposta. Quando si modificano le ipotesi sugli errori si ha il modello lineare generalizzato in cui la matrice di var-cov degli errori $\Sigma_{\hat{E}}$ non è più diagonale. Il modello è formulato nel seguente modo:

$$Y = BZ + E$$

in cui la matrice di var-cov degli errori $\Sigma_{\hat{E}}$ non è più necessariamente diagonale e con gli errori che possono essere eteroschedastici. Nel caso di errori omoschedastici ed incorrelati $\sigma^2_E = \sigma^2 I_n$, con la matrice di varianze del modello pari a $\Sigma_Y = BZZ^tB^t + \sigma^2 I_n$. La sfericità degli errori nel caso multivariato è un'ipotesi molto più restrittiva rispetto al caso univariato in quanto fa si che la parte non spiegata del modello sia identica per ogni variabile risposta e che necessiti di modelli generalizzati per affrontare la maggior parte dei casi reali.

5.1 Soluzione dei minimi quadrati generalizzati

Come nel caso del modello lineare classico multiplo non è possibile utilizzare lo stimatore OLS dato che la matrice degli errori non è diagonale La soluzione viene, ma è possibile ricavare le stime con lo stimatore GLS.

Data la matrice di var-cov della popolazione Σ_E, si ipotizza l'esistenza di una matrice $W_{(nm,nm)}$ non singolare, ottenuta con la decomposizione spettrale di $\Sigma_{E^{\circ *}}$, tale per cui:

$$\Sigma_{E^{\circ *}} = \mathrm{E}(E^{\circ *t} E^{\circ *}) = \sigma^2 W^t W$$

Si definiscono gli errori trasformati E tali che:

$$E^{\circ} = E^{\circ *} W^{-1} \implies \Sigma_{E^{\circ}} = W^{-1}\sigma^2 W^t W W^{-1} = \sigma^2 I_{nm}$$

Gli errori E° trasformati sono quindi sferici, omoschedastici e incorrelati rispetto alla risposta e covariate del modello, con il modello che diventa

$$Y^{\circ} = B^{\circ} Z^{\circ} + E^{\circ} \tag{45}$$
$$Y^{\circ} W^{-1} = B^{\circ *} Z^{\circ} W^{-1} + E^{\circ} W^{-1} \tag{46}$$
$$Y^{\circ *} = B^{\circ *} Z^{\circ *} + E \tag{47}$$

Perciò la matrice dei parametri diventa:

$$B^{\circ} * = Y^{\circ *} Z^{\circ * t} \times (Z^{\circ *} Z^{\circ * t})^{-1} \tag{48}$$
$$= Y^{\circ} \Sigma_{E^*} Z^{\circ t} \times (Z^{\circ} {\Sigma_{E^*}}^{-1} Z^{\circ t})^{-1} \tag{49}$$

La differenza principale con il modello multivariato classico è che per calcolare le soluzione dei parametri è necessario considerare tutte le equazioni attraverso la matrice di var-cov degli errori.

5.2 Forme della matrice di var-cov degli errori

Esistono diversi modelli a seconda della struttura della matrice Σ_E, presentando le più importanti.

Modello lineare multivariato **classico**. Gli errori sono *omoschedastici* all'interno delle stesse equazioni e anche tra equazioni diverse, cioè per ogni individuo rispetto alla stessa/diversa variabile risposta la parte spiegata è identica. Inoltre, gli errori sono *incorrelati* all'interno delle stesse equazioni e tra equazioni diverse, ovvero il comportamento di ogni individuo rispetto alla stessa/diversa variabile risposta non è legato ad altri individui.

Modello lineare multivariato **intermedio**. Gli errori sono *omoschedastici* ed *incorrelati* all'interno delle stesse equazioni, come nel modello classico, ma sono *eteroschedastici* e *correlati* tra equazioni diverse: la varianza spiegata è unica per ogni individuo e il valore di una variabile dipendente può influenzare gli altri dello stesso individuo.

Modello lineare **generalizzato**. Gli errori sono *eteroschedastici* e *correlati* sia all'interno delle stesse equazioni, ovvero il comportamento di un individuo può influenzare gli altri e ognuno ha una varianza unica, sia tra equazioni diverse, come nel modello intermedio

5.3 Soluzioni FGLS

Come nel caso multiplo raramente si conosce la matrice Σ_E della popolazione. Di conseguenza, è possibile utilizzare la matrice di var-cov degli errori campionaria S_{E^*}, con essa che tende a Σ^*_E se $n \to \infty$:

$$B^{\circ *} = Y^{\circ *} Z^{\circ * t} \times (Z^{\circ *} Z^{\circ * t})^{-1} \tag{50}$$
$$= Y^{\circ} {S_{E*}}^{-1} Z^{\circ t} \times (Z^{\circ} {S_{E*}}^{-1} Z^{\circ t})^{-1} \tag{51}$$

5.4 Modello SURE

Il modello Seemengly Uncorrelated Regression Equation (SURE) è un tipico modello lineare generalizzato più realistico. Il modello costruito è il seguente:

$$Y^\circ = B^\circ Z^\circ + E^\circ$$

Il modello ha regressori diversi e può essere presentato con diverse versioni della matrice di var-cov degli errori Σ_{E° omoschedastici ed incorrelati, ma è possibile costruire modelli diversi in ogni equazione utilizzando una diversa matrice Σ_{E°. Il modello si risolve con il metodo dei minimi quadrati generalizzati:

$$B^\circ = Y^\circ Z^{\circ t} \times (Z^\circ Z^{\circ t})^{-1} \quad (52)$$
$$= Y^\circ \Sigma_{E^\circ} Z^{\circ t} \times (Z^\circ {\Sigma_{E^\circ}}^{-1} Z^{\circ t})^{-1} \quad (53)$$

Il modello è caratterizzato dalla presenza delle variabili esplicative diverse da equazione a equazione, modificando i coefficienti di regressione. Gli errori sono omoschedastici ed incorrelati nella stessa equazione, ma eteroschedastici, correlati per lo stesso individuo ed incorrelati tra individui diversi tra diverse equazioni. Se nelle diverse equazioni gli errori non fossero correlati, si avrebbe un sistema di equazioni di regressione multipla individuali stimate separatamente come nel modello classico multivariato.

Tra le svariate possibilità, bisogna individuare il modello ottimale. I passaggi da seguire sono:

1. Verificare omoschedasticità e correlazione all'interno delle equazioni. Utilizzando le stime OLS, se la matrice di varianza non è diagonale, con varianze diverse (molto diverse da 0) e correlate cade l'ipotesi di sfericità;
2. Riguardo alla scelta del modello lineare generalizzato o SURE è necessario osservare la significatività univariata delle variabile esplicative, eliminando le variabili esplicative;
3. Verificare l'adattamento ai dati con l'indice R^2 insieme alla parsimonia equazione per equazione;
4. La scelta di ipotesi e la specificazione dei modelli non è meccanico, ma bisogna avere intuizione, creatività e conoscere bene il fenomeno analizzato.

(PART) Modello Multilevel

6 Modello Lineare Multilevel

I modelli statistici si basano su un *c.c.s.* da una popolazione finita od infinita CR, nella quale vige l'ipotesi di *i.i.d.* di tutte osservazioni. In certi casi i dati presentano una struttura gerarchica: la tipologia di campionamento non è più efficiente, ma è preferibile effettuare un campionamento a stadi, scegliendo i cluster e le unità condizionate dalla prima selezione, in modo da analizzare le relazioni tra le variabili misurate a livelli i raggruppamento diversi. Questo campionamento implica una dipendenza tra le osservazioni appartenenti alla stesso gruppo.

Ignorando la struttura dei dati, a livello descrittivo i valori osservati saranno distorti e perderanno informazioni in due modi:

- *Fallacia ecologica*, che consiste nell'interpretare i dati aggregati come se fossero dati individuali. In tal modo si utilizza la correlazione tra variabili a livello di gruppo per fare affermazioni su relazioni di livello micro;
- *Fallacia atomistica*, nella quale si disaggregano i dati ignorando la variabilità tra i gruppi. Se è presente correlazione tra variabili micro, essa non può essere usata a livello macro.

A livello stocastico, le osservazioni non sono più *i.i.d.* e non hanno la stessa probabilità di essere estratte:

- Aggregando i dati micro a livello macro, le unità sono *i.i.d.* con uguale probabilità di estrazione, quando in realtà non lo sono per la presenza dei gruppi;
- Disaggregando i dati macro a livello micro, i gruppi sono *i.i.d.* con stessa probabilità di estrazione, quando in realtà non lo sono per la eterogeneità di distribuzione interna.

Esistono due modelli che tengono conto dei dati gerarchici e che superano i problemi precedenti: la regressione multilevel ed il modello multilevel.

6.1 Regressione Multilevel

Considerando una ***Regressione Multilevel*** e dati con struttura a 1 livello, la costruzione del modello e le stime sono le medesime del modello lineare univariato e multivariato, con p gruppi e n_p osservazioni:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \forall i = 1, \dots, n_j \quad \forall i = 1, \dots, p \quad n_1 + \cdots + n_p = n$$

Le medie della variabile risposta e l'esplicativa, pur misurate a livello individuale, possono essere diverse da quelle di un altro gruppo, in quanto di diversa composizione. Si possono avere diversi tipi di regressioni:

- *Regressione disaggregata* a livello micro con stimatore OLS: in esse il raggruppamento delle unità viene ignorato;
- *Relazione aggregata* fra gruppi a livello macro: sono diverse dalle regressioni micro in quanto ignorano le unità all'interno dei gruppi e sono basate sulle medie di x ed y;
- *Relazione entro ciascun gruppo*, considerando le deviazioni dalle medie;
- *Relazione multilevel*, nella quale si costruiscono regressioni tra i gruppi ed entro i gruppi congiuntamente.

Si definisce il modello di Cronbach, dovuto all'effetto della media dei gruppi di y al netto delle differenze all'interno dei gruppi:

$$y_{ij} = \alpha + \beta_{within}(x_{ij} - \overline{x}_{.j}) + \beta_{between}\overline{x}_{.j} \quad \begin{cases} \overline{y}_{.j} = \alpha + \beta_{between}\overline{x}_{.j} + \overline{\varepsilon}_{.j} \\ y_{ij} - \overline{y}_{.j} = \beta_{within}(x_{ij} - \overline{x}_{.j}) + (\varepsilon_{ij} - \overline{\varepsilon}_{.j}) \end{cases}$$

In termini di regressione multipla si ha:

$$y_{ij} = \beta_0 + \sum_{k=1}^{p} \beta_{kj} x_{ijk} + \varepsilon_{ij} = \tag{54}$$

$$= \beta_0 + \sum_{k=1}^{p} \beta_{k\ (within)}(x_{ijk} - \mu(x_{jk})) + \beta_{k\ (between)}\mu(x_{jk}) + \varepsilon_{ij} \tag{55}$$

ed in termini matriciali $y = X^{@}B^{@} + E$, con soluzione OLS pari a $B^{@} = (X^{@t}X^{@})^{-1}X^{@t}y$

6.2 Modello Multilevel: definizione e significato

Innanzitutto, si definisce ***modello ANOVA*** il modello di analisi della varianza, che spiega in che misura la variabilità della y è dovuta a differenze delle media tra gruppi con covariate x_j qualitative. Il modello è definito nel seguente modo:

$$y_{ij} = \gamma_{00} + u_j + e_{ij}$$

- γ_{00} media di y di tutta la popolazione;
- $\gamma_{00} + u_j$ media relativa al gruppo *j-esimo*;
- u_j effetto dell'unità macro j su y_{ij};
- e_{ij} residuo dell'unita micro $i \in j$.

A livello matriciale, il modello ANOVA si costruisce nel seguente modo:

$$y - \gamma_{00} = Au + e \implies \begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_p \end{bmatrix} - \gamma_{00} = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ \vdots \\ u_j \\ \vdots \\ u_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_j \\ \vdots \\ e_p \end{bmatrix}$$

L'ANOVA è perciò un caso particolare del modello lineare con covariate qualitative in cui la matrice A si sostituisce con la matrice X ed u con il vettore dei parametri β.

È possibile calcolare la devianza totale (SST) del modello, che è possibile scomporre in devianza spiegata (SSF) e devianza residua (SSE):

$$SST = (y - \gamma_{00})^t(y - \gamma_{00}) = \tag{56}$$

$$= (Au + e)^t(Au + e) = \tag{57}$$

$$= u^t A^t A + e^t e = \tag{58}$$

$$= SSF + SSE \tag{59}$$

In caso di omoschedaticità, essendo la varianza di ogni errore pari a σ^2, $SSF = \tau^2$ e $SSE = \sigma^2$ ed è possibile calcolare il *coefficiente di correlazione intraclasse*:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

Esso indica la quota di varianza totale spiegata dovuta ai gruppi: quanto più è elevato il valore, tanto più l'appartenenza delle osservazioni ai gruppi spiega la varianza totale.

Se le caratteristiche x delle osservazioni appartenenti ai diversi gruppi sono differenti tra i gruppi l'analisi della varianza viene distorta. Il ***modello ANCOVA*** elimina l'effetto delle caratteristiche della popolazione X sulla variabile risposta ed effettua successivamente l'analisi della varianza. Il modello Questo processo avviene in 4 fasi:

1. Calcolare i parametri del modello lineare $\hat{B}$ ed i residui $\hat{w}$;
2. Ricavare la soluzione OLS v°;

3. Calcolare le varianze tra gruppi SSV e nei gruppi *SSE$;
4. Calcolare il coefficiente intraclasse.

Successivamente, l'ANOVA cattura la relazione aggregata tra gruppi descrivendone la varianza tra i gruppi. Questo processo avviene in 3 fasi:

1. Effettuare l'ANOVA su SSR_{yc} (devianza residua corretta di y)
2. Calcolare la devianza spiegata del fattore sperimentale corretta per l'effetto della covariata X, ottenuta come $SSE_{yc} = SST_{yc} - SSR_{yc}$;
3. Calcolare la devianza totale $SST_y = SSE_x + SST_{yc} = SSE_x + SSE_{yc} + SSR_{yc}$.

In definitiva , il modello ANCOVA elimina gli aspetti casuali, analizza gli effetti di gruppo ed attribuisce la varianza al gruppo di appartenenza.

Il modello ANCOVA ad effetti variabili è definito modello Multilevel: la regressione lineare cattura la relazione disaggregata tra i dati, eliminando l'effetto distorsivo della varianza tra i gruppi e ricavando quella nei gruppi, mentre l'ANOVA cattura la relazione aggregata tra i gruppi descrivendone la varianza tra i gruppi.

In una prima tipologia di modelli, chiamati *Mixed Models*, la relazione disaggregata tra i dati e la varianza nei gruppi sono descritte mediante parametri fissi, mentre la relazione aggregata tra gruppi e la varianza tra gruppi sono descritte come *v.c.*

In un secondo tipo di modelli, chiamati *Random Models*, anche la relazione disaggregata tra i dati e la varianza nei gruppi sono descritte come *v.c.*

I modelli finora studiati possono essere visti come sottocasi del modello multilevel:

- se $u_j = 0$ e non è presente alcuna gerarchia nei dati, si ottiene un modello lineare;
- se $u_j = 0$ si ottiene una regressione multilevel;
- se u_j è fisso e l'effetto della media nei gruppi è nulla, si ottiene un modello ANOVA;
- se u_j è casuale e l'effetto della media nei gruppi è nulla, si ottiene un modello ANOVA ad effetti casuali.

6.3 Modello Multilevel: OLS, Empty , Mixed, Total Effects

La stima del modello multilevel si compone di 4 passaggi:

1. Modello Lineare OLS Si consideri un modello Multilevel in cui appare solo la parte del modello lineare stimata con lo stimatore dei parametri OLS con una sola variabile ed ipotizzando che X ed Y siano centrate:

$$y_{ij} = \beta_0 + \beta_1 x_{1jk} + \cdots + \beta_k x_{njk} + \varepsilon_{ij}$$

Così facendo, si osserva l'effetto delle covariate sulla risposta nel caso in cui i dati non fossero centrati, con $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Si possono proporre anche regressioni Multilevel introducendo variabili esplicative Z misurate sui gruppi, rappresentanti il livello 2, invece che sugli individui. Questo aspetto può essere esteso anche all'interazione *cross-level*, cioè nel modello si possono introdurre variabili originate dall'interazione tra variabili misurate sull'individuo e sui gruppi appartenenti. Per risolvere la regressione Multilevel si può scomporre il coefficiente di regressione tra gruppi e nei gruppi utilizzando il modello di Cronbach:

$$y_{ij} = \alpha + \beta_{whithin}(x_{ij} - \overline{x}_{.j}) + \beta_{between}\overline{x}_{.j} + \varepsilon$$

2. Empty Model Nel modello ANOVA ad effetti casuali, chiamato anche Empty model, si ha il seguente modello:

$$y_{ij} = v_j + r_{ij}$$

In questo caso la variabile risposta dipende dagli effetti casuali a livello di gruppo, $V_j \sim N(\gamma_{00}, \tau^2)$, mentre a livello individuale dai residui $R_{ij} \sim N(0, \sigma^2)$, con V_j e R_{ij} indipendenti e mutualmente incorrelati.

La variabilità all'interno di ogni gruppo è dovuta solamente alla distribuzione casuale della risposta. L'intercetta casuale a livello di gruppo può essere scomposta in due parti: l'intercetta fissa media tra tutti i gruppi γ_{00} e la misura della sua deviazione attorno alla media tra i gruppi di tipo casuale u_j. Il modello si può scrivere nel seguente modo:

$$y_{ij} = \gamma_{00} + u_+ r_{ij}$$

In questo modello la variabilità totale della risposta può essere scomposta nella somma delle varianze ai due livelli, con varianza nei gruppi e varianza tra i gruppi:

$$\sigma_y^2 = \sigma_{U_j}^2 + \sigma_{R_{ij}}^2 = \tau^2 + \sigma^2$$

Si può quindi definire il coefficiente di correlazione intraclasse:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$$

Il coefficiente di correlazione intraclasse ρ misura la quota di varianza di y spiegata dall'appartenenza ai gruppi dei singoli individui. Se $\rho = 0$, ovvero tutti gli u_j sono nulli, il raggruppamento è irrilevante ed è inutile utilizzare altri modelli oltre il modello lineare semplice. Nel caso invece ρ fosse positivo, è necessario considerare un modello di tipo gerarchico.

Il Test F può essere utilizzato per verificare in termini inferenziali l'ipotesi che le intercette casuali u_j siano nel complesso tra loro equivalenti (nel caso non ci fosse differenza fra gruppi). In questo caso, serve per capire se nel complesso vale l'ipotesi nulla H_0 che le medie parziali ottenute nel campione possano essere ritenute nel complesso equivalenti. Per confrontare tra loro le strutture di secondo livello come nell'analisi della varianza casuale non si utilizzano i valori delle medie campionarie, non informative del vero valore di U_j ma i loro *intervalli di confidenza* che comprendono con una probabilità del 90%, 95%, 99% i valori veri ignoti di U_j. Ciò significa potabilizzare la gerarchia fra medie parziali in quanto, tanto più sono piccoli gli intervalli di confidenza, maggiore è la loro capacità di fornire informazioni sui valori veri ignoti di U_j.

3. Random Intercept Model Se si inserisce nel modello Empty una variabile esplicativa x_k il modello diventa un Random Intercept Model, anche chiamato Mixed Model:

$$y_{ij} = \gamma_{00} + \beta_1 x_{ij} + u_j + \varepsilon_{ij}$$

dove u_{ij} è la della determinazione della v.c. $U_j \sim N(\gamma_{00}, \tau^2)$ a rappresentazione dei residui di secondo livello. Essi sono indipendenti e quindi incorrelati con i residui del primo livello ε_{ij} della v.c. $E_{ij} \sim N(0, \sigma^2)$.

In questo caso la variabile risposta dipende dalle variabili X e dai relativi parametri fissi β, dall'effetto casuale a livello di gruppo $u_j \sim N(\gamma_{00}, \tau^2)$ e dall'effetto casuale a livello individuale $\varepsilon_{ij} \sim N(0, \sigma^2)$.

La correlazione intraclasse ρ misura la quota di varianza spiegata dall'appartenenza ai gruppi dei singoli individui al netto della quota di varianza spiegata da x, che a differenza del modello nullo X spiega una parte della variabilità non dovuta all'appartenenza ad un gruppo. Il suo valore può diminuire di molto rispetto al modello precedente.

Il modello RIM può essere costruito in:

- Modello micro: $y_{ij} = \beta_1 x_{ij} + R_{ij}$;
- Modello macro: $\beta_{0j} = \gamma_{00} + U_{0j}$.

Come unica equazione, il modello RIM si costruisce nel seguente modo:

$$y_{ij} = \beta_1 x_{ij} + \gamma_{00} + R_{ij} + U_{0j}$$

con la parte fissa del modello $\beta_1 x_{ij} + \gamma_{00}$ e la parte casuale $R_{ij} + U_{0j}$ e come varianze e covarianze al primo livello σ^2 ed al secondo livello τ^2.

4. Random Slope and Intercept Model La relazione tra Y e le covariate X_j può variare tra i gruppi in diversi modi: è possibile avere una eterogeneità delle regressioni tra i diversi gruppi (interazione gruppo-covariate). La struttura dei dati ed il loro raggruppamento può essere spiegato anche facendo variare i coefficienti della regressione da gruppo a gruppo:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + R_{ij}$$

A seconda del comportamento dei parametri è possibile ottenere:

- Con diversi β_{0j} in base ai gruppi un modello RIM;
- Con diversi β_{1j} in base ai gruppi un modello RC;
- Se i coefficienti β_{0j} e β_{1j} sono entrambi costanti la struttura gerarchica non ha effetto e si ottiene un modello lineare con stime OLS.

Considerando il modello precedente, è possibile scomporre i parametri in una parte costante e deviazione dalla media a livello di gruppo:

$$\beta_{0j} = \gamma_{00} + U_{0j} \beta_{1j} = \gamma_{10} + U_{1j}$$

ottenendo quindi l'equazione completa

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + U_{0j} + U_{1j} x_{ij} + R_{ij}$$

con effetti di gruppo dati da:

- U_{0j} intercetta random;
- $U_{1j} x_{ij}$ interazione random tra i gruppi e la variabile esplicativa x_{ij};
- $\gamma_{00} + \gamma_{10} x_{ij}$ parte fissa del modello generale;
- $U_{0j} + U_{1j} x_{ij} + R_{ij}$ parte random del modello generale.

6.4 Metodi di Stima e Verifica di Ipotesi

La specificazione del modello comporta la scelta dello stesso più soddisfacente. Nel caso di modelli lineari gerarchici implica la scelta delle covariate x_{ij} e delle interazioni della parte fissa e la scelta dei coefficienti casuali con le strutture di covarianza per la parte random del modello.

I parametri da stimare nel modello RIM sono i coefficienti di regressione γ_{00} e β, i componenti di varianza σ^2 e τ^2 e gli effetti casuali U_{0j} non direttamente osservabili.

I metodi per la stima dei parametri assumendo che i residui U_{0j} e R_{ij} siano distribuiti come una Normale sono il metodo di massima verosimiglianza (ML) e la massima verosimiglianza ristretta (REML). Quest'ultimo massimizza la verosimiglianza dei residui osservati ottenendo le stime degli effetti fissi usando metodi come OLS o GLS e dopodiché le utilizza per massimizzare la verosimiglianza dei residui, sottraendo gli effetti misti, per ottenere le stime dei parametri della varianza.

Test sui Parametri Fissi del Modello Per testare i parametri fissi del modello si utilizza l'ipotesi di significatività su ciascun parametro, $H_0 : \gamma_h = 0$. Questa ipotesi viene verificata con un test T noto come Test di Wald:

$$T(\gamma_h) = \frac{\hat{\gamma}_h}{\text{se}_{\hat{\gamma}_h}}$$

Sotto l'ipotesi nulla il test ha approssimativamente una t con i gradi di libertà basati sulla struttura Multilevel.

Test sui Parametri Fissi e Random del Modello Per testare più parametri, in questo caso fissi e random, del modello si utilizza il Test di Devianza. Dalla stima del modello lineare con il metodo ML si ottiene la verosimiglianza del modello:

$$\text{Deviance} = -2ln(ML)$$

Solitamente viene interpretata in termini differenziali, calcolando la differenza tra le Deviance dei modelli alternativi. Si tratta di confrontare i valori osservati della y con i valori teorici di due modelli:

1. X_j di interesse e l'altro senza variabile (modello nullo);
2. X_j di interesse e l'altro chiamato modello saturo.

Il confronto si basa sulla funzione di log-verosimiglianza: indicate con D_0 la devianza del modello vuoto, D_{mod} del modello scelto e D_sat del modello saturo, valori più vicini a 0 che non a D_0 indicano un modello considerato buono.

Tutte le devianze $\sim \chi^2$, con j gradi di libertà pari al numero delle x_j, di conseguenza anche le loro differenze saranno distribuite allo stesso modo.

Se l'obiettivo è verificare la nullità congiunta di tutti i coefficienti, l'ipotesi nulla sarà $H_0 : \beta_1 = \cdots = \beta_k = 0$ confrontando il modello nullo ed il modello ipotizzato (applicabile sia alla parte fissa che a quella casuale del modello).