Statistical Modeling

Indice

1 Errori eteroschedastici **2**

2 Errori autocorrelati **3**

3 Metodo di stima WLS, per soluzioni correlate, GLS **4**

4 Multicollinearità **6**

5 Linearità **7**

6 Non normalità **8**

7 Outlier **9**

8 Modello lineare classico multivariato **11**

9 Inferenza nella Regressione Multivariata **13**

10 Modello lineare generalizzato **15**

11 Modello SURE **16**

12 Il problema dei dati gerarchici e uso di Regressione multilevel **17**

13 Modello Multilevel: definizione e significato **20**

14 Modello Multilevel: OLS, Empty, Mixed, Total Effects **23**

15 Metodi di stima e verifica di ipotesi **27**

1 Errori eteroschedastici

Per ottenere stime efficienti per i parametri del modello lineare classico, gli errori devono essere *omoschedastici*: la varianza degli errori ε_i è costante $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$ e non dipende dal valore delle variabili indipendenti $E(\varepsilon_i) = E(\varepsilon_i|x_i)$. Questo si verifica dal momento che il valore atteso del singolo errore è nulla $E(\varepsilon_i) = 0$. Quando ciò non avviene, si è in presenza di errori *eteroschedastici*: $Var(\varepsilon_i) = \sigma_i^2$; in tal caso il vettore b dei parametri perde in efficienza[1] (non è più BLUE (*Best Linear Unbiased Estimator*). In forma matriciale, questo significa che si passa da una matrice diagonale con un valore costante a un'altra matrice diagonale i cui valori possono differire.

Inoltre, la stima campionaria s^2 di σ^2 tende a sottostimare il vero valore della distribuzione, dato che non si è più in presenza di una sola variabile casuale ma di molteplici. Ne consegue che il test T di Student ritorna valori erroneamente elevati, e quindi gli intervalli di confidenza risultano essere più stretti del reale, e i test di significatività sui parametri b_j risultano più permissivi del dovuto. Discorso analogo vale per i test basati sulla distribuzione F di Snedcor.

Si può verificare se una distribuzione è o eteroschedastica (o omoschedastica) tramite diversi metodi grafici:

- scatter plot della variabile risposta vs esplicative $y \sim x_j$ (da effettuare per ogni x_j);
- scatter plot dei valori stimati vs residui $\hat{y} \sim \varepsilon$;
- scatter plot dei residui al quadrato vs i valori predetti $\varepsilon^2 \sim \hat{y}$;
- scatter plot dei valori osservati vs predetti $y \sim \hat{y}$;
- scatter plot dei residui vs variabili esplicative $\varepsilon \sim x_j$ (da effettuare per ogni x_j).

Esistono inoltre una serie di test statistici che offrono risultati numerici e meno interpretativi per verificare l'eteroschedasticità degli errori.

Testi di White. Si basa sull'assunzione di omoschedasticità dei residui ($H_0 : Var(\varepsilon_i) = \sigma^2$ contro $H_1 : Var(\varepsilon_i) \neq \sigma^2$). Il test sfrutta la regressione OLS del quadrato dei residui con le variabili esplicative, il loro quadrato e tutte le possibili interazioni $\varepsilon^2 \sim x_j, x_i x_j \ \forall i, j < k$. Attraverso l'indice di determinazione R^2 di tale regressione, ricavato dal rapporto tra la variabilità spiegata dalla regressione e la variabilità totale $R^2 = \frac{SSE}{TSS} = \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \overline{y})^2}$, si calcola la statistica $LM = nR^2 \sim \chi^2_{k-1}$. L'ipotesi nulla verrà rigettata se LM risulterà maggiore del valore soglia della distribuzione χ^2 (ovvero con p-value basso); infatti se R^2 è maggiore di un certo valore soglia significa che le variabili esplicative sono realmente significative nello spiegare la variabilità dei residui.

$$LM = nR^2 = n\frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \overline{y})^2} \sim \chi^2_{k-1}$$

Test di Breusch-Pagan. Anche in questo test, l'ipotesi nulla è quella di omoschedasticità ($H_0 : Var(\varepsilon_i) = \sigma^2$). Il test si basa sulla regressione di ε_i^2/s^2 (dove $s^2 = \frac{1}{n}\sum \varepsilon_i^2$) con le variabili esplicative. Della regressione si calcola poi il coefficiente R^2 analogamente al test di White. Essendo ε distribuita (sotto H_0) come una normale $N(0, \sigma^2)$, $\varepsilon^2 \sim \chi^2_{k-1}$; inoltre essendo

[1] Si veda dimostrazione in appendice.

già $s^2 \sim \chi^2_{n-k-1}$ (perché anch'esso somma di normali al quadrato) e $\varepsilon \perp\!\!\!\perp X$ (sempre sotto H_0), ε^2/s^2 si distribuisce come un rapporto di χ^2 indipendenti tra di loro, ovvero come una $F_{k-1,n-k-1}$ di Snedecor. Si procede quindi effettuando un semplice test F per l'accettazione di H_0. Per risolvere tale problema si può procedere attraverso il metodo di stima WLS (Weighted Least Squares).

$$BP = nR^2 \sim F_{k-1,n-k-1}$$

2 Errori autocorrelati

Per garantire stime efficienti del modello, bisogna verificare che gli errori ε non siano tra di loro correlati, cioè non siano *autocorrelati*. Nel modello lineare classico si ipotizza infatti che:

$$Cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \qquad i \neq j$$

Tuttavia accade spesso, 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: $\varepsilon^{\#}_{i.-1}$ (errore ritardato di un tempo) e η_i (errori omoschedastici IID, ovvero indipendentemente ed identicamente distribuiti in modo normale). Una parte dell'errore è legata al suo valore ritardato, mentre l'altra è indipendente. 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 (ϱ_{-i}). Gli errori autocorrelati non incidono sulle proprietà di linearità, 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 relative inferenze non sono più corrette e affidabili (la statistica T di Student ottiene 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 esistono diversi sistemi grafici:

- scatter plot della variabile risposta vs esplicative $y \sim x_j$ (da effettuare per ogni x_j);
- scatter plot dei residui vs variabili esplicative $\varepsilon \sim x_j$ (da effettuare per ogni x_j);
- scatter plot dei residui vs ritardati $\varepsilon \sim \varepsilon_{-1}$;
- correlogramma: è un grafico imn cui sono mostrate le correlazioni a diversi gradi; analizzando l'acf (la funzione di autocorrelazione dei residui) e $pacf$ si determina il tipo di modello autoregressivo.

Il test di Durbin-Watson invece offre uno strumento analitico per verificare la presenza di autocorrelazione a diversi gradi. La sua ipotesi nulla è la mancanza di autocorrelazione: $H_0 : \varrho_{-i} = 0$; mentre l'ipotesi alternativa può essere verificata su entrambe le code della distribuzione (bidirezionale) $H_1 : \varrho_{-i} \neq 0$ o su una coda sola (unidirezionale destra o sinistra) $H_1 : \varrho_{-i} \gtrless 0$. La statistica DW per l'autocorrelazione dei residui è definita come:

$$DW = \frac{\sum (\varepsilon_i - \varepsilon_{i-1})^2}{\sum \varepsilon_i^2} \in [0, 4]$$

si dimostra inoltre che sotto ipotesi di omoschedasticità di ε, $DW = 2(1 - \varrho_{-i})$. Il valore tende a 2 in caso di mancanza di autocorrelazione, a 0 in caso di massima autocorrelazione negativa ($DW = 0 \Leftrightarrow \varrho_{-i} = -1$) e 4 positiva ($DW = 4 \Leftrightarrow \varrho_{-i} = +1$). Convenzionalmente i valori critici per definire se l'autocorrelazione è significativa o meno sono 1 e 3 (è considerata significativa se $DW < 1 \vee DW > 3$).

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 il più efficiente, ovvero è quello caratterizzato dalla minor 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 è costante e il valore dipende dalle variabili esplicative ($\varepsilon \not\perp\!\!\!\perp X$), 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 la varianza delle singole componenti erratiche ε_i condizionatamente al vettore dei dati x_i. Si esegue quindi una trasformazione della variabile risposta $y \to y^*$ e della matrice del disegno $X \to X^*$ per riportare la varianza degli errori ad una costante: si divide infatti ogni variabile per la radice di $h(i)$ (la varianza di ε^*, l'errore eteroschedastico). La componente erratica dunque ha valore costante e il modello assume la forma:

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

permettendo una stima b dei parametri β tramite il metodo OLS.

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 opposto (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 i-esimo ordine attraverso un modello avente come variabile risposta gli errori ε_t e come esplicative quelle già considerate più l'errore ritardato di i tempi ε_{t-i} e procedere alla stima del coefficiente di correlazione ϱ. Vale infatti:

$$\varepsilon_t = \varrho\varepsilon_{t-i} - \delta_t$$

dove δ è la componente erratica che segue le ipotesi classiche. Si moltiplica dunque ogni elemento dell'equazione ritardata per ϱ.

$$\varrho y_{t-i} = \varrho X_{t-i}\beta + \varrho\varepsilon_{t-i}$$

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

$$\begin{array}{ccccc} y_t^{\#} & = & X_t^{\#} & \beta + & \varepsilon_t^{\#} \\ (y_t - \varrho y_{t-i}) & = & (X_t - \varrho X_{t-i}) & \beta + & (\varepsilon_t - \varrho\varepsilon_{t-i}) \end{array}$$

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

In alternativa è possibile utilizzare un modello *autoregressivo* che inserisce nell'equazione iniziale un errore ritardato che tenga conto dell'autocorrelazione di i-esimo ordine:

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

con

$$\begin{aligned} AR_i + \varepsilon &= v \\ Corr(v_j, v_k) &= 0 \quad\quad j \neq k \end{aligned}$$

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 Σ_ε. Nello specifico si procede ad effettuare una *decomposizione spettrale* della matrice degli errori:

$$\Sigma_\varepsilon = \sigma^2 V V'$$

con

$$V = \sigma(\sqrt{AL})A'$$

dove A è la *matrice degli autovettori* e L è la matrice diagonale degli autovalori di Σ_ε. A questo punto premoltiplicando per V^{-1} le componenti del modello si ottiene una nuova funzione con variabili trasformate e $\Sigma_{\varepsilon^\circ}$ omoschedastica ed incorrelata:

$$\begin{aligned} V^{-1}y &= y^\circ \\ V^{-1}X &= X^\circ \\ y^\circ &= X^\circ\beta^\circ + \varepsilon \end{aligned}$$

Lo stimatore b° risulta godere delle proprietà di correttezza e consistenza; inoltre secondo il Teorema di Aitken è lo stimatore più efficiente per il modello *generalizzato* (nonostante abbia una varianza maggiore rispetto al metodo OLS $\sigma^2(X^{\circ\prime}X^\circ)^{-1} > \sigma^2(X'X)^{-1}$). Tuttavia la stima GLS necessita di assumere come nota la matrice di varianze e covarianze dei residui Σ_ε, o almeno poter calcolare una sua stima, a patto però che sia consistente al limite:

$$\lim_{n\to\infty} S_\varepsilon = \Sigma_\varepsilon$$

A questo punto si possono utilizzare gli stimatori FGLS (*Feasible Generalized Least Squares*). Spesso la soluzione di applicare i FGLS viene intrapresa anche in caso di semplice eteroschedasticità o semplice autocorrelazione, o sospette tali, poiché vige il principio di precauzione.

4 Multicollinearità

Se la matrice $(X'X)$ non è invertibile o ha determinante prossimo allo 0, le stime non possono essere calcolate (coefficienti sotto identificati, poiché non si dispone di sufficiente informazione per stimarli) o non sono affidabili (coefficienti empiricamente sotto identificati).

Tale problema si verifica quando almeno una delle variabili è correlata linearmente alle altre: si ha quindi multicollinearità. In questo caso la matrice $(X'X)$ è detta singolare e l'inversa $(X'X)^{-1}$ non è unica.

Esistono due tipi di collinearità:

- **perfetta**: sussiste quando almeno una variabile esplicativa è una combinazione lineare perfetta delle altre: questa viola le proprietà del modello lineare classico; solitamente ciò si verifica per un errore nella definizione dei regressori o per la presenza di due variabili direttamente dipendenti una dall'altra (ad esempio *titolo di studio* e *anni di studio*);
- **imperfetta**: sussiste in caso di forte correlazione tra i regressori e dunque il determinante della matrice dei coefficienti tende a 0; ciò non rende impossibilità la stima dei coefficienti ma da origine a stime fortemente distorte e caratterizzate da un'alta varianza.

L'errore di stima provoca un aumento della varianza dello stimatore dei coefficienti b, da cui deriva una sovrastima delle dimensioni degli intervalli di confidenza (che risultano più ampi del dovuto) e una maggiore zona di accettazione nei test statistici di quanto sarebbe corretto. Inoltre, l'aggiunta di una variabile fortemente correlata ad una già presente nel modello aggiunge poca informazione (sarebbe opportuno calcolare la correlazione *spuria*).

Per individuare fenomeni di multicollinearità, è buona norma, in prima istanza, generare una *matrice di correlazione* tra tutte le variabili così da identificare rapidamente possibili variabili collineari. Auspicabilmente infatti è preferibile avere forte correlazione tra y e i regressori x_j, e bassa correlazione tra questi ultimi. Esistono inoltre metodi analitici:

Indice di tolleranza. Misura il grado di interrelazione di una variabile indipendente rispetto alle altre. Si calcola come $TOL = 1-R^2 \in [0,1]$ dove R^2 è il coefficiente di correlazione della regressione di una variabile esplicativa x_j (usata come risposta) in funzione delle altre. Valori alti indicano una bassa multicollinearità tra la singola variabile x_j e le altre.

Varianza multifattoriale (o VIF). È il reciproco della tolleranza $VIF = TOL^{-1}$. Valori di tale indice variano tra 0 e ∞, ma si considera significativo già se superiore a 20 ($TOL = 0.05$) indicando uno stretto rapporto tra la variabile considerata e le altre del modello, ovvero un eccessivo grado di *multicollinearità*. Vanno considerate con attenzione anche quelle variabili con valori di VIF maggiori di 10 ($TOL = 0.1$);

L'indice di condizione. È dato dalla radice del rapporto tra l'autovalore massimo della matrice $(X'X)$ e gli altri autovalori. Quando risulta essere maggiore di 30 si considera significativa la presenza di *multicollinearità*. 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.

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 non è lineare, allora l'effetto su y (Δy) di una variazione in X (ΔX) 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 l'indice 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:

- scatter plot della variabile risposta vs esplicative $y \sim x_j$ (da effettuare per ogni x_j);
- scatter plot dei residui vs la variabile risposta $\varepsilon \sim y$ (non deve presentare andamenti sistematici);
- scatter plot dei residui vs valori previsti $\varepsilon \sim \hat{y}$ (l'andamento deve essere 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:

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

Tuttavia la trasformazione di una variabile, eccettuando casi particolari in cui il dominio lo permette (la trasformata *log-log* in un grafico quantità-prezzo indica l'*elasticità*) , rende di difficile interpretazione il modello.

6 Non normalità

Quando gli errori ε_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 β 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 non sono i più efficienti tra tutti gli stimatori corretti (non sono più VUE), ma rimangono corretti e consistenti, e i più efficienti tra tutti gli stimatori lineari (sono ancora BLUE).

Non coincidendo più le stime OLS ed ML, alcuni software statistici potrebbero fornire stime non attendibili.

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 dei residui ε: quando questi coincidono, si può supporre che la distribuzione sia normale; questa operazione può essere effettuata rapidamente utilizzando un box-plot. Si possono usare comunque altre distribuzioni grafiche, quali:

- istogramma della distribuzione dei residui ε;
- plot della distribuzione cumulata dei residui, alla ricerca di evidenti irregolarità;
- qq-plot (quantile vs quantile) della distribuzione dei residui ε vs la normale standard (i punti si dovrebbero disporre sulla bisettrice);
- pp-plot (probability vs probability) della distribuzione cumulata dei residui ε vs la cumulata della normale standard.

Esistono, infine, alcuni test non parametrici che non si basano su ipotesi sulla distribuzione ma che sono appunto detti non parametrici poichè testano la distribuzione dei parametri. Per questo motivo sono molto utili per analizzare problemi di normalità dei residui.

Test di Shapiro-Wilk. L'ipotesi nulla $H_0 : \varepsilon \sim N(0, \sigma^2)$ è accettata con valori alti dell'indice W; tuttavia essendo caratterizzato da una forte asimmetria può portare a un rifiuto dell'ipotesi nulla anche in presenza di distribuzione normale.

$$W = \frac{\sum(a_i \varepsilon_{(i)})^2}{\sum \varepsilon_i^2} \in [0, 1]$$

Test di Kolmogorov Smirnov. Anche qui $H_0 : \varepsilon \sim N(0, \sigma^2)$, ma si basa sul calcolo della statistica test D, calcolato come somma in valore assoluto della differenza tra le frequenze cumulate 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 significatività scelto comporterà il rifiuto o l'accettazione di H_0.

Skewness test. È un test di asimmetria ($H_0 : \varepsilon : P(e < 0) = P(e > 0) \forall e$) ma rigettando H_0 si rigetta l'ipotesi di normalità della distribuzione; l'accettazione dell'ipotesi nulla tuttavia non fornisce indicazioni sulla reale distribuzione.

$$S = \frac{\frac{1}{n}\sum \varepsilon_i^3}{\sigma^3} \in (-\infty, +\infty)$$

La distribuzione è considerata simmetrica con $-1 < S < +1$.

Kurtosis test. Simile a quello per l'asimmetria, fornisce solamente informazioni sulla curtosi della distribuzione:

$$K = \frac{\frac{1}{n}\sum \varepsilon_i^4}{(\frac{1}{n}\sum \varepsilon_i^2)^2} = \frac{\frac{1}{n}\sum \varepsilon_i^4}{\sigma^4} \in (-\infty, +\infty)$$

La distribuzione normale ha curtosi pari a 3, quindi il test è considerato significativo per valori prossimi.

I problemi di non normalità possono essere risolti usando una trasformata della variabile dipendente y. La trasformazione può migliorare la relazione lineare tra la variabile dipendente e le variabili indipendenti. Le trasformazioni più diffuse sono:

- $\log(y)$: quando $\sigma_\varepsilon^2 \not\perp y$ o la distribuzione dell'errore ha asimmetria *positiva*;
- y^2 quando σ_ε^2 è proporzionale a $\overline{y}$ o quando la distribuzione dell'errore ha asimmetria *negativa*;
- $\sqrt{Y}$ quando σ_ε^2 è proporzionale a $\overline{y}$;
- Y^{-1} quando σ_ε^2 cresce significativamente al crescere di y.

7 Outlier

I valori cosiddetti *outlier* possono essere distinti in valori anomali (che si discostano in modo rilevante dall'andamento generale) e punti influenti (che influenzano in misura rilevante le stime). Un'osservazione può appartenere a entrambe le categorie o a una sola in modo del tutto casuale.

Gli outliers possono essere identificati visivamente tramite rappresentazioni grafiche (box-plot e scatter-plot) su distribuzioni a due dimensioni; all'aumentare del numero di dimensioni si perde la possibilità di individuarli in modo grafico e sono necessari indici numerici.

Leverage values. Definendo la matrice di proiezione $H = X(X'X)^{-1}X'$, si considerano gli elementi $h_{i.i}$ sulla diagonale principale (chiamati *leverage*) che rappresentano l'impatto dell'osservazione i-esima sulla capacità del modello di predire tutti i casi.

Si dimostra che il valor medio del leverage è:

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

con k che rappresenta il numero di variabili esplicative e n il numero di osservazioni. Un valore *leverage* si considera significativamente alto se supera 2 o 3 volte (scelto in modo arbitrario) il suo valore medio:

$$h_{i.i} > 2\frac{k-1}{n}$$

Residui standardizzati. I residui ε possono essere calcolati come:

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

È dunque possibile scrivere la loro varianza come:

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

Si possono quindi calcolare i residui *standardizzati* grazie alla formula:

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

Si usano quindi le tavole della normale per verificare la presenza di *outliers*: il 95% della popolazione assume valori compresi tra $-2 < \varepsilon_i^* < +2$, e così via. Si considera *outlier* un valore il cui valore assoluto $|\varepsilon_i^*| > 3$ (dato che la probabilità che si verifichi casualmente è particolarmente bassa).

Residui studentizzati. Concettualmente identico ai residui standardizzati, però con varianza campionaria s^2 (formule e test sono i medesimi). La formula è la medesima ma con lo stimatore s^2 al posto del valore reale σ^2.

Residui *jack-knife*. Usati con campioni di numerosità non elevata, nei residui *jack-knife* la varianza s_i^2 è calcolata eliminando momentaneamente la i-esima variabile dal modello e stimando nuovamente i parametri.

Covrati. Misura la variazione nel determinante della matrice delle covarianze delle stime eliminando la i-esima osservazione. Eliminando infatti il valore i-esimo si provoca una variazione nel determinante che vado a quantificare.

$$\text{COVRATIO}_i = \frac{det(\Sigma_i)}{det(\Sigma)} = \frac{det(\frac{1}{n-1}\tilde{X}_i'\tilde{X}_i)}{\det(\frac{1}{n}\tilde{X}'\tilde{X})}$$

Si considera significativo se supera la soglia:

$$1 \pm 3\sqrt{\frac{k+1}{n}}$$

Dfitts. Misura l'influenza dell'i-esima osservazione sulla stima dei coefficienti di regressione e sulla loro varianza, calcolandolo sul valore stimato della variabile risposta y.

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

Si considera significativo se supera il valore soglia:

$$\pm 2\sqrt{\frac{k+1}{n}}$$

Dfbetas. Misura l'influenza dell'i-esima osservazione sulle stime dei coefficienti di regressione (considerandoli singolarmente). Valori elevati indicano che l'osservazione influisce molto sulla stima dei parametri.

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

Si considera significativo se è superato il valore 2 (o $2\sqrt{n}$) per almeno un parametro β: risulta essere un indice più stringente dei precedenti.

Distanza di Cook. Misura l'influenza dell'i-esima osservazione sulla stima dei coefficienti di regressione nel loro complesso.

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

Si considerano significative le distanze che superano il valore soglia 1 o $^4/_n$.

8 Modello lineare classico multivariato

Si intende con modello lineare classico *multivariato* l'estensione del modello classico multiplo a m variabili risposta y_j; ogni variabile risposta è legata alle stesse variabili esplicative.

Per l'i-esimo individuo si ha:

$$\begin{aligned} y_i &= [y_{i.1} \dots y_{i.j} \dots y_{i.m}] \\ z_i &= [1, z_{i.1} \dots z_{i.k} \dots z_{i.r}] \\ \varepsilon_i &= [\varepsilon_{i.1} \dots \varepsilon_{i.j} \dots \varepsilon_{i.m}] \end{aligned}$$

Mentre la matrice dei parametri β, di dimensioni $m \times r + 1$, diventa:

$$\beta = \begin{bmatrix} \beta_{1.0} & \cdots & \beta_{1.k} & \cdots & \beta_{1.r} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \beta_{j.0} & \cdots & \beta_{j.k} & \cdots & \beta_{j.r} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ \beta_{m.0} & \cdots & \beta_{m.k} & \cdots & \beta_{m.r} \end{bmatrix}$$

In sintesi ogni *riga* della matrice dei parametri β si riferisce ad una variabile risposta $y_{1\ldots m}$ mentre ogni colonna si riferisce ad una variabile esplicativa $z_{1\ldots r}$.
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:

$$y_j = \beta_j Z + \varepsilon_j \qquad \forall j \in [1,m]$$

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

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

- I parametri sono lineari.
- I valori attesi degli errori casuali sono nulli: $E(\varepsilon_{i.j}) = 0$.
- 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:

  $$\Sigma_E = \begin{bmatrix} \sigma^2 I_n & \cdots & 0_{n\times n} & \cdots & 0_{n\times n} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ 0_{n\times n} & \cdots & \sigma^2 I_n & \cdots & 0_{n\times n} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ 0_{n\times n} & \cdots & 0_{n\times n} & \cdots & \sigma^2 I_n \end{bmatrix}$$

  Con dimensione $nm \times nm$ poiché ognuna delle m matrici Σ_{ε_j} relativa ad ogni singola equazione j è di dimensione $n\times n$; mentre le matrici sulla diagonale principale rappresentano gli errori relativi alla medesima equazione, le matrici 0 racchiudono le correlazioni fra gli errori relativi ad equazioni diverse. Infatti le variabili risposta si presume siano indipendenti le une dalle altre sia per lo stesso individuo ($y_{i.j} \perp\!\!\!\perp y_{i.k}\ \forall j,k : i \neq k$, sulla diagonale della matrice) che per individui diversi; eventuali correlazioni sono da considerarsi *spurie* data la correlazione alle stesse variabili esplicative.
- Le variabili esplicative sono non stocastiche: per ogni osservazione i, i valori $z_{i.k}$ sono costanti, mentre il corrispondente valore di ogni $y_{i.j}$ è una variabile casuale influenzata dagli errori casuali.
- La matrice Z ha rango pieno $rk(Z) = r+1$, ciò vuol dire che nessuna variabile esplicativa è una combinazione lineare delle altre; contrariamente la matrice $Z'Z$ non sarebbe invertibile e non si potrebbe calcolare lo stimatore dei minimi quadrati.
- La numerosità della popolazione n è maggiore del numero dei parametri stimati più l'intercetta $n > r+1$, perciò per ogni equazione le stime dei minimi quadrati di $\hat{\beta}$ sono trovate in modo analogo al caso univariato:

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

  Di conseguenza, nel modello multivariato classico calcolare le soluzioni per ogni variabile dipendente y singolarmente o in gruppo è identico.

- Gli errori E si distribuiscono come una normale multivariata:

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

Permane la condizione di ortogonalità poiché i residui sono incorrelati sia con le variabili esplicative Z che con i valori predetti della variabile dipendente $\hat{Y}$.

Inoltre poiché $Y = \hat{Y} + \hat{E}$, si ha che:

$$\begin{aligned} YY' &= \hat{B}ZZ'\hat{B}' + \hat{E}\hat{E}' \\ \Sigma_Y &= \quad \hat{H} \quad + \quad \hat{\Sigma_E} \end{aligned}$$

Con Σ_Y matrice di varianze e covariante di Y, $\hat{H}$ matrice di varianze e covariante spiegate e Σ_E matrice di varianze e covariante residue.

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

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

$$\begin{aligned} YY' &= \hat{B}ZZ'\hat{B}' + \sigma^2 I_{nm} \\ \Sigma_Y &= \quad \hat{H} \quad + \sigma^2 I_{nm} \end{aligned}$$

L'R^2 complessivo del modello è una media degli R^2 delle singole equazioni pesata per la numerosità dei casi (che può mutare a causa dei valori nulli).

9 Inferenza nella Regressione Multivariata

Gli stimatori OLS sono corretti ed efficienti, poiché 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 b dei minimi quadrati è quello a *varianza minore*. Inoltre, per il modello di regressione multivariata con rango pieno ed errori E normalmente distribuiti anche le m variabili dipendenti Y sono distribuite secondo una normale multivariata come anche i parametri stimati b

$$\begin{aligned} Y &\sim N(BZ, \Sigma_Y) \\ b &\sim N(\beta, \hat{H}) \end{aligned}$$

con $\hat{H}$ matrice di varianze e covarianze spiegate della popolazione, positiva definita ed efficiente, e che si dimostra essere distribuita in modo indipendente da E matrice degli errori. È però il caso di notare che Σ_Y e $\hat{H}$ sono entrambe 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 $det(\hat{H})$; 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.

Considerando che $\hat{H}$ si distribuisce come $\hat{H} \sim W_r$ (variabile casuale di Wishart, una generalizzazione multivariata della Γ) e $\Sigma_E \sim W_{r-n}$ (e inoltre $\hat{H} \perp\!\!\!\perp \Sigma_E$), si può quindi definire il loro rapporto come test di verosimiglianza Lambda di Wilks:

$$\Lambda = \frac{det(\Sigma_E)}{det(\Sigma_E + \hat{H})}$$

Che si distribuisce *asintoticamente* come una χ^2_{mr}. Sempre da Λ, inoltre, si ricava una distribuzione asintotica di F

$$F = \frac{1 - \Lambda}{\Lambda}$$

che nel caso sia rispettata l'ipotesi 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:

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

per cui sotto H_0, $\Lambda \to 1$ dato che 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.

Esistono inoltre altri test che possiedono la stessa distribuzione, implicazioni e ipotesi H_0 della Λ di Wilks:

- **Traccia di Lawney-Hotelling**

$$LH = \frac{det(\hat{H})}{det(\Sigma_E)}$$

- **Traccia di Pillai**

$$P = \frac{det(\hat{H})}{det(\hat{H} + \Sigma_E)}$$

- **Massimo autovalore di Roy**

$$\max(\lambda_i) \quad \text{con } \lambda \text{ autovettore di} \frac{\hat{H}}{\hat{H} + \Sigma_E}$$

In modo analogo, si possono costruire altri test con altre ipotesi nulle H_0:

- 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

Il modello lineare multivariato generalizzato supera le ipotesi, molto stringenti, del modello lineare multivariato classico. 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* infatti abbiamo che:

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.

Infatti ipotizzando 3 variabili dipendenti:

- Spesa viaggi
- Spesa partite
- Spesa concerti

La parte di variabilità non spiegata dalle variabili esplicative (Σ_E) è identica per tutti gli individui, per ognuna delle 3 variabili dipendenti (omoschedasticità). Inoltre la stessa Σ_E per l'individuo i-esimo non è influenzata dall'individuo k-esimo per ogni variabile dipendente. La parte non spiegata dalle variabili esplicative dell'individuo i-esimo rispetto alla spesa viaggi (ad esempio) non è influenzata dalla parte non spiegata dalle variabili esplicative per lo stesso individuo i rispetto alla spesa partite (incorrelazione).

Nell'ipotesi *intermedia*, invece:

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 **eteroschedastici** tra equazioni diverse: per ogni individuo rispetto alle diverse variabili dipendenti la parte spiegata è diversa;
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 **correlati** fra equazioni diverse: il comportamento di ogni individuo rispetto a diverse variabili dipendenti è legato al proprio e a quello degli altri individui.

Infatti facendo riferimento all'esempio precedente potremmo dire che:

- La parte di variabilità non spiegata dalle variabili esplicative è identica per tutti gli individui per spesa viaggi ad esempio (omoschedasticità nella stessa equazione);
- La parte di variabilità non spiegata dalle variabili esplicative per tutti gli individui è diversa per spesa viaggi, partite, concerti (eteroschedasticità fra diverse equazioni);
- La parte non spiegata dalle variabili esplicative per l'individuo i non è influenzata da quella dell'individuo k per ogni variabile dipendente e per le varie voci di spesa (incorrelazione fra individui diversi);
- La parte non spiegata dalle variabili esplicative dell'individuo i rispetto alla spesa è influenzata dalla parte non spiegata dalle variabili esplicative per lo stesso individuo i rispetto alla spesa partite ad esempio (correlazione per medesimo individuo).

Nell'ipotesi *estrema*:

1. Gli errori sono **eteroschedastici** all'interno delle stesse equazioni: per ogni individuo rispetto alla medesima variabile dipendente la parte spiegata è diversa;
2. Gli errori sono **eteroschedastici** tra equazioni diverse: per ogni individuo rispetto alle diverse variabili dipendenti la parte spiegata è diversa;
3. 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;
4. 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.

Continuando l'esempio precedente avremo quindi tutto l'opposto dell'ipotesi classica:

- La parte di variabilità non spiegata dalle variabili esplicative è diversa per tutti gli individui per spesa viaggi ad esempio (eteroschedasticità nella stessa equazione);
- La parte di variabilità non spiegata dalle variabili esplicative per tutti gli individui è diversa per spesa viaggi, partite, concerti (eteroschedasticità fra diverse equazioni);
- La parte non spiegata dalle variabili esplicative per l'individuo i è influenzata dall'individuo k per ogni variabile dipendente come ad esempio fra spesa viaggi e spesa partite (correlazione fra individui diversi);
- La parte non spiegata dalle variabili esplicative dell'individuo i rispetto alla spesa è influenzata dalla parte non spiegata dalle variabili esplicative per lo stesso individuo i rispetto ad esempio alla spesa partite (correlazione per medesimo individuo)

Quindi occorre usare non le singoli sottomatrici di correlazione degli errori $\Sigma_{E(i)}$, ma la matrice Σ_E relativa all'intero modello.

11 Modello SURE

Nel modello **S**eemengly **U**ncorrelated **R**egression **E**quation, anche detto **SURE**, si segue un approccio più realistico: degli r regressori si usano solo i regressori effettivamente legati alle diverse variabili dipendenti, che potrebbero anche essere tutti, come nel caso classico del modello lineare multivariato, ma che in caso di non significatività potrebbero portare ad una differenziazione delle diverse equazioni. Inoltre questo modello permette di risolvere anche il problema di una numerosità diversa delle osservazioni tra le diverse equazioni.

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.
Quindi la somma di tutti i regressori nelle diverse equazioni è uguale a

$$\sum_{j=1}^{n} r_j$$

Data n_j come la numerosità delle osservazioni per l'equazione j-esima allora il complesso delle numerosità è dato da

$$\sum_{j=1}^{m} n_j$$

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 Z_A, Z_B, Z_C 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.

Considerando B^* come stimatore dei coefficienti del modello OLS e $\hat{B}^*$ come lo stimatore del modello SURE, possiamo affermare che i due risultano identici ma per la costruzione del modello SURE $\hat{B}^*$ possiede dei parametri uguali a zero proprio per permettere di ottenere equazioni differenziate nel numero delle variabili esplicative. La possibilità di differenziare il numero di covariate risiede infatti nel porre uguale a zero il valore della variabile non presente nell'equazione x così che questa, moltiplicata per il rispettivo coefficiente B generi un influenza nulla sulla stima di Y.

NB: Quando si omettono delle variabili, essendo i coefficienti di regressione parziali (influenzati, cioè, dalle altre variabili esplicative poichè stimati sulla totalità di esse) cambiano tutti i coefficienti di regressione in base all'equazione di riferimento. Infatti ipotizzando di avere una variabile X_j presente in 2 equazioni distinte il coefficiente β corrispondente sarà diverso proprio in virtù del fatto che le stime sono influenzate le une dalle altre.

12 Il problema dei dati gerarchici e uso di Regressione multilevel

I modelli statistici, di solito, si basano sull'assunzione di indipendenza delle osservazioni, ottenuta per mezzo di un *campionamento casuale semplice* da popolazione infinita o finita con reinserimento. In questo caso si dice che le osservazioni si distribuiscono come una variabile casuale e che sono **IID**, ovvero **I**denticamente ed **I**ndipendentemente **D**istribuite.

In molti casi, però i dati risultano essere raggruppati in cluster ovvero presentano una struttura gerarchica come ad esempio:

- Ospedale - Pazienti;
- Classi - Studenti;
- Imprese - Impiegati.

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 diversi (livelli gerarchici della struttura dei dati). Questo tipo di campionamento implica infatti **dipendenza** tra le osservazioni appartenenti allo stesso gruppo. Ad esempio gli studenti appartenenti alla stessa scuola condividono stesso ambiente, stessi insegnanti, stesso quartiere di provenienza oltre a scambi e comunicazioni tra essi.

Quando i dati possiedono una struttura gerarchica significa che possono essere scomposti in dati *dell'unità* e dati *del gruppo*. 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. Potrei **aggregare** i dati micro (le unità) a livello macro (le sovrastrutture), ad esempio le condizioni di lavoro degli impiegati di un'azienda non possono essere attribuite ai singoli impiegati, così facendo si andrebbe incontro a quella che è definita **fallacia ecologica**: se vi è correlazione tra variabili a livello *macro* non può essere usata per fare asserzioni a livello *micro*.
2. Potrei **disaggregare**, ovvero utilizzare i dati *macro* a livello *micro* ignorando la variabilità tra i gruppi (ad esempio, le caratteristiche degli studenti non possono dire nulla sulla scuola se non è messa in luce esplicitamente la loro appartenenza all'una o all'altra scuola), andando incontro a quella che è definita **fallacia atomistica**: se vi è correlazione tra variabili a livello *micro* non può essere usata per fare asserzioni a livello *macro*.

Ipotizzando una **regressione Multilevel**, in cui la regressione di Y su X è la funzione lineare di x che meglio spiega y ed ipotizzando che i dati abbiano struttura ad un livello, nulla cambia rispetto al Modello lineare classico univariato, singolare o multiplo, infatti assume forma:

$$y = \beta_0 + \beta_1 x + r$$

Ipotizzando invece una struttura a più livelli con:

- j identificativo del gruppo;
- i identificativo dell'unità entro il gruppo;
- x_{ij} ed y_{ij} osservazioni di X e Y sulle unità micro del gruppo j;
- $\bar{x}_{.j}$ e $\bar{y}_{.j}$ medie di gruppo j per X e Y.

ad esempio il soggetto 1 del primo gruppo sarà diverso dal soggetto 1 del secondo gruppo.
In questo contesto la variabile dipendente Y, quindi, ha sia un aspetto individuale sia uno 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

$$\bar{y}_{.j} = \beta_0 + \beta_1 \bar{x}_{.j} + r_{.j}$$

e sono quindi diverse dalle regressioni a livello *micro* tra X ed Y

$$y_{ij} + \bar{y}_{.j} = \alpha_1 (x_{ij} - \bar{x}_{.j}) + r$$

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 un altro modo: facendo variare i coefficienti della regressione da gruppo a gruppo

$$Y_{ij} = \beta_{0j} + \beta_{ij} x_{ij} + r_{ij}$$

con diverse intercette β_{0j} e coefficienti di regressione β_{ij} per ciascun gruppo:

- Se i coefficienti β_{0j} e β_{ij} sono entrambi costanti allora la struttura gerarchica non ha effetto, ovvero non vi è differenza con una regressione OLS;
- Se i due coefficienti dipendono entrambe da j, ovvero dal gruppo di cui sono espressione, allora la regressione OLS non può essere utilizzata.

Nello specifico riguardo quest'ultimo punto:

1. Se varia solo β_{0j} con j allora si ha un modello **Random Intercept**;
2. Se anche β_{ij} varia con j allora il modello è detto **Random Coefficient**.

Ci sono diversi modi di approcciare lo studio della relazione tra Y ed X:

- **Relazione diseggregata**: è una relazione in cui il raggruppamento delle unità (ovvero la sovrastruttura) viene ignorato

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + r_{ij}$$

- **Relazione aggregata fra i gruppi**: si può infatti essere interessati alla relazione aggregata, ovvero a livello macro, fra i gruppi, quindi alla relazione tra $\bar{x}_{.j}$ e $\bar{y}_{.j}$

$$\bar{y}_{.j} = \beta_0 + \beta_1 \bar{x}_{.j} + r_j$$

- **Relazione entro ciascun gruppo**: poichè si può essere interessati alla relazione tra x_{ij} ed y_{ij} entro ciascun gruppo j, ovvero a livello micro. Questo avviene assumendo α_1 costante entro ciascun gruppo

$$y_{ij} - \bar{y}_{.j} = \alpha_1 (x_{ij} - \bar{x}_{.j}) + r$$

- **Relazione Multilevel**: data la struttura dei dati, si può pensare di porre assieme la *regressione tra i gruppi* e la *regressione entro i gruppi* per cui y_{ij} sarebbe funzione sia delle relazioni **entro** i gruppi, sia di quelle **tra** i gruppi.

$$y_{ij} = \beta_0 + \beta_1 \bar{x}_{.j} + \alpha_1 (x_{ij} - \bar{x}_{.j}) + r$$

13 Modello Multilevel: definizione e significato

Per effettuare l'analisi della **covarianza** (**ANCOVA**) occorre prima di tutto partire da quello che è definito modello **ANOVA**, ovvero il modello di analisi della varianza:

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

con:

- u_j uguale all'effetto dell'unità macro j, ovvero la differenza tra la media generale e la media del gruppo j;
- r_{ij} residuo relativo all'unità micro i appartenente al gruppo j;
- $\gamma_{00} + u_j$ media relativa al gruppo j;
- γ_{00} media di y relativa all'intera popolazione.

Il modello di analisi della varianza ANOVA cerca di spiegare in che misura la variabilità della variabile dipendente y è dovuta a differenze delle medie fra i gruppi. E' infatti il modello utilizzato in caso di stima di y con sole covariate x_j di tipo qualitativo.

Nel complesso dato:

$$\begin{gathered} y_{(1,n)} = [y_{1(1,n)}, \ldots, y_{p(1,n_p)}] \\ r_{(1,n)} = [r_1, \ldots, r_p] \\ \text{con} \ \ (n_1 + \cdots + n_p) = n \end{gathered}$$

e costruendo A matrice *presenza-assenza*, composta solo di 0 e 1

$$A_{p,n} = \begin{bmatrix} 1_{(n_1,1)} & 0 & 0 \\ 0 & 1_{(n_2,1)} & 0 \\ 0 & 0 & 1_{(n_p,1)} \end{bmatrix}$$

possiamo procedere a calcolare

$$\begin{aligned} & u_1 = u_1 A_1 \\ & u_2 = u_2 A_2 \\ & \ldots \ \ldots \ \ldots \\ & u_p = u_p A_p \end{aligned}$$

ottenendo cosi per il gruppo j-esimo

$$y_{j_{(1,n)}} - \gamma_{00_{(1,p)}} = u_{j_{(p,n)}} + r_{j_{(1,n)}}$$

$$(y - \gamma_{00})(y - \gamma_{00})' = [uA - e][uA - e]' = uAAu' + rr'$$

Per avere un'idea della variabilità nei gruppi, su tutti i gruppi j si considerano le devianze intra gruppo e, nell'ipotesi di omoschedasticità, essendo la varianza di ogni errore pari a σ^2 otteniamo che la devianza residua totale $SSR = n\sigma^2$.

Allo stesso modo la devianza tra i gruppi corrisponde alla devianza delle medie di gruppo perciò $SSE = \sum_{j=1}^{p}(u_j A_j)^2$.

Da cui devianza totale $SST = SSE + SSR$.

Definendo a questo punto

$$\tau^2 = \frac{uAA'u'}{n}$$

$$\sigma^2 = \frac{rr'}{n}$$

Otteniamo il **coefficiente di correlazione intraclasse** ρ

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

ovvero il rapporto tra **devianze tra i gruppi** e **devianza complessiva**.

Questa struttura è fondamentale per ricondurre il modello ANOVA alla struttura del modello lineare con

$$u = \beta$$

$$A = X$$

$$y - \gamma_{00} = y$$

$$r = \varepsilon$$

Il **modello ANCOVA** è un modello di analisi della **covarianza**: se infatti le caratteristiche x delle osservazioni appartenenti ai diversi gruppi sono differenti tra gruppo e gruppo l'analisi della varianza viene distorta e si attribuiscono alla varianza fra i gruppi effetti che dipedono da tali caratteristiche. Occorre quindi prima di tutto eliminare l'effetto di queste caratteristiche sulla variabile dipendente attraverso regressione OLS e, successivamente, effettuare l'analisi della varianza depurata.

Ciò avviene attraverso 4 fasi:

1. Prima si procede al calcolo della devianza totale di Y;

2. Si stimano i coefficienti di regressione;

3. Si calcola la devianza spiegata di X (SSE);

4. Infine si stima la devianza residua corretta di y uguale a $SSR_{yc} = SST_y - SSE_x$.

Successivamente, **l'analisi della varianza** (ANOVA) cattura la relazione aggregata fra i gruppi e quindi descrive la **varianza fra gruppi**.

Ciò avviene in 3 fasi:

1. Si effettua l'analisi della varianza su SSR_{yc};

2. Si calcola la devianza spiegata del **fattore sperimentale** corretta per l'effetto della covariata X, ottenuta come $SSE_{yc} = SST_{yc} - SSR_{yc}$;

3. Nel complesso quindi si ottiene

$$SST_y = SSE_x + SST_{yc} = SSE_x + SSE_{yc} + SSR_{yc}$$

Questo modello quindi spezza in due l'analisi mettendo insieme la covarianza:

1. Elimina gli aspetti individuali;

2. Analizza gli effetti di gruppo;

3. Attribuisce la varianza al gruppo di appartenenza.

Il modello ANOVA è specificabile anche in una versione ad *effetti casuali*, per cui abbiamo la stessa struttura citata in precedenza ma con U_j ipotizzata variabile casuale con distribuzione $N(0, \tau^2)$ di cui u_j è una particolare manifestazione e E_{ij} un'altra variabile casuale con distribuzione $N(0, \sigma^2)$. In questo caso il modello è definito ad *effetti misti*. Mentre quindi non cambia la forma, cambia la sostanza. Sotto il profilo interpretativo significa infatti che le medie parziali u_j sono determinazioni di una variabile casuale U_j. In questo caso il Test F serve a verificare H_0 che le medie parziali ottenute dal campione possano essere ritenute nel complesso equivalenti, ma per confrontare tra loro le strutture di secondo livello, non si utilizzano più i valori delle medie campionarie, ma i loro intervalli di confidenza; ciò significa probabilizzare la gerarchia.

Nel caso si consideri l'intera popolazione, oppure U_j ed E_j abbiano distribuzione non normale è meglio utilizzare l'analisi della varianza ad effetti fissi.

Per concludere, dopo le lunghe premesse, il modello ANCOVA ad effetti variabili è appunto definito **modello Multilevel**: restando valide tutte le considerazioni riguardo il modello ANOVA ad effetti casuali in prima istanza la **regressione lineare** (OLS) cattura la relazione disaggregata tra i dati, così da eliminare l'effetto distorsivo sulla varianza tra i gruppi e ricavare la varianza nei gruppi, mentre l'analisi della varianza cattura la relazione aggregata fra i gruppi e descrive quindi la varianza fra gruppi.

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, abbiamo un **modello Lineare**:

$$y_i = \gamma_{00} + \sum_k \beta_k (x_{ik} - \bar{x}_{.k}) + \varepsilon_i \ ;$$

2. Per $u_j = 0$ otteniamo una **regressione Multilevel**:

$$y_{ij} = \gamma_{00} + \sum_k \beta_k (x_{ijk} - \bar{x}_{.k}) + \varepsilon_{ij};$$

3. Per $\sum_k \beta_k(x_{ik} - \bar{x}_{.k}) = 0$ e u_j fisso abbiamo un'**ANOVA**:

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

4. Per $\sum_k \beta_k(x_{ik} - \bar{x}_{.k}) = 0$ e u_j stocastico otteniamo un'**ANOVA ad effetti casuali**:

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

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;
2. Si propone poi l'Empty model (anche detto **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.

1. Stima modello lineare con metodo di stima 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

$$y_{ij} = \beta_0 + \sum_{jk} \beta_k x_{ijk} + \varepsilon_{ij}$$

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.

Si possono proporre anche regressioni Multilevel introducendo variabili esplicative Z misurate sui gruppi, e quindi rappresentanti il livello 2, invece che sugli individui. Questo aspetto inoltre può essere esteso anche all'interazione *cross-level*, ciò significa che nel modello si possono introdurre variabili prodotto originate dall'interazione tra variabili misurate sull'individuo e misurate sui gruppi cui gli individui appartengono $z_w x_k$. Per risolvere la regressione multilevel si può scomporre il coefficiente di regressione in parte *between* e parte *within*.

Il **modello di Cronbach** fa esattamente questo

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

dove il *contextual effect* δ è l'effetto della media dei gruppi che non è contemplato dal valore individuato

$$\delta = \beta_{between} - \beta_{within}$$

dove $\beta_{between} - \beta_{within}$ sono gli effetti dovuti alla sovrastruttura al netto degli effetti individuali.

2. 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, V_j, distribuiti in modo normale $N(\gamma_{00}, \tau^2)$
- a livello individuale, dai residui R_{ij}, distribuiti in modo normale $N(0, \sigma^2)$

con V_j ed R_{ij} **indipendenti** e **mutualmente incorrelati**.
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 γ_{00} e la misura della sua deviazione attorno alla media tra i gruppi di tipo casuale u_j

$$v_j = \gamma_{00} + u_j$$

Possiamo a questo punto riscrivere il modello nel seguente modo:

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

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

$$Var(y) = Var(U_j) + Var(R_{ij}) = \tau^2 + \sigma^2$$

Si può quindi definire il coefficiente di correlazione intraclasse, come visto in precedenza

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

Il coefficiente di correlazione intraclasse ρ misura quindi la quota di varianza di y spiegata dall'appartenenza ai gruppi dei singoli individui. Se $\rho = 0$, ovvero tutti gli u_j sono nulli, allora 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 come in ogni analisi della varianza 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 il Test F 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 (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 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 probabilizzare 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.

NB A differenza dell'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 u_j di una struttura J si considera superiore a un'altra media parziale u_g di una struttura G se e solo se l'estremo inferiore del suo intervallo di confidenza è più grande dell'estremo superiore dell'altra in quanto, solo in questo caso, con un elevato grado di probabilità, il valore vero di J sarà più grande di G.

3. Random intercept model

Se si inserisce nel modello Empty una variabile esplicativa x_k il modello diventa il vero e proprio random intercept model (mixed model)

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

dove u_j è la determinazione della *variabile casuale* U_j distribuita normalmente $N(\gamma_{00}, \tau^2)$ a rappresentazione dei residui di secondo livello. Ess sono indipendenti e quindi incorrelati con i residui di primo livello ε_{ij} determinazioni della variabile casuale normalmente distribuita $E_{ij} \sim N(0, \sigma^2)$

In questo caso la variabile dipendente y dipende:

- dalle variabili X e dai relativi parametri fissi β;
- dall'effetto casuale a livello di gruppo u_j, che si distribuisce in modo normale $N(\gamma_{00}, \tau^2)$;
- dall'effetto casuale a livello individuale ε_{ij}, che si distribuisce in modo normale $N(0, \sigma^2)$.

La correlazione intraclasse ρ misura la quota di varianza di y spiegata dall'appartenenza ai gruppi dei singoli 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 rispetto al caso del modello Empty.

Il modello comprende 4 parametri da stimare:

- i coefficienti di regressione γ_{00} e β;
- le componenti della varianza σ^2 e τ^2.

Il coefficiente di regressione β 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 dai residui ε_{ij}.

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

- La v.c. U_j può essere vista come le variabile casuale che descrive i residui a livello di gruppo, ovvero gli effetti di gruppo non spiegati da X;
- La v.c. E_{ij} può essere vista come variabile casuale che descrive i residui a livello di individuo, 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

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

Con

- parte fissa del modello $y_{ij} = \gamma_{00} + \beta_1 x_{ij}$
- parte casuale (random part) del modello $U_{0j} + R_{ij}$

e come varianze e covarianze al primo livello σ^2, ed al secondo livello τ^2.

NB Variabili esplicative relative al secondo livello possono essere misurate direttamente sulle unità di secondo livello o derivate dalle misurazioni effettuate sulle unità di primo livello. Inoltre spesso la numerosità dei vari gruppi è diversa da gruppo a gruppo, questo però non rappresenta un problema grazie all'**effetto shrinkage**, che tiene conto della diversa numerosità dei gruppi facendo pesare di più per la stima complessiva, i gruppi più numerosi.

4. Random slopes and intercepts model

La relazione tra variabile dipendente Y e variabili esplicative X_j può variare tra i gruppi in modi diversi: si può infatti 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.

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

A seconda del comportamento di β_{0j} e β_{1j} possiamo ottenere:

- Con diversi β_{0j} in base ai gruppi otteniamo un **modello Random Intercept**;
- Con anche diversi β_{1j} in base ai gruppi otteniamo un **modello Random Coefficient**;
- Se i coefficienti β_{0j} e β_{1j} sono entrambi costanti la struttura gerarchica non ha effetto ed otteniamo un **modello OLS**.

Considerando

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

con β_{0j} e β_{1j} variabili, ognuno può essere scomposto in 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;
- $U_{0j} + U_{10} x_{ij} + R_{ij}$ parte random del modello generale.

15 Metodi di stima e verifica di ipotesi

Specificazione del modello e stima dei parametri

La specificazione del modello comporta la scelta del modello più soddisfacente. Nel caso di modelli lineari gerarchici tutto ciò implica:

- Scelta delle variabili esplicative x_j e delle interazioni della parte fissa;
- Scelta dei coefficienti casuali con le strutture di covarianza per la parte random del modello.

I parametri da stimare nel modello random intercept sono:

- Coefficienti di regressione γ_{00} e β;
- Componenti di varianza, σ^2 e τ^2;
- Gli effetti casuali U_{0j} non sono parametri ma variabili casuali latenti, ovvero non direttamente osservabili.

I metodi comunemente utilizzati per la stima dei parametri sotto l'assunzione che i residui U_{0j} e R_{ij} siano distribuiti normalmente sono il metodo del *maximum likelihood* (ML) ed il *restricted maximum likelihood* (REML).

Il metodo REML massimizza la verosimiglianza (likelihood) dei residui osservati ottenendo le stime degli effetti fissi usando metodi «non likelihood-like» 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.

Verifica di ipotesi

Test sui parametri fissi del modello

Per testare i parametri fissi del modello si utilizza la seguente ipotesi nulla (ipotesi di significatività) su ciascun parametro

$$H_0 : \gamma_h = 0$$

Questa ipotesi viene verificata con un Test t

$$\mathrm{T}(\gamma_h) = \frac{\hat{\gamma}_h}{\mathrm{s.e.}(\hat{\gamma}_\mathrm{h})}$$

noto come WALD TEST.

Sotto l'ipotesi nulla il test ha approssimativamente una distribuzione t con g.d.l. basati sulla struttura multilevel dell'analisi.

Test su più parametri della parte fissa del modello e parte random

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:

$$\mathrm{DEVIANCE} = -2 \cdot ln(\mathrm{Likelihood})$$

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. l'uno con le variabili esplicative di interesse e l'altro senza alcuna variabile (**empty-model**);
2. l'uno 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 D_0, D_{mod}, D_{sat} le devianze calcolate per il **modello vuoto** (empty-model), il **modello considerato** e il **modello saturo**, valori di D_{mod} più prossimi a 0 che non a D_0 faranno propendere per ritenere "buono" il modello considerato.

Ognuna delle devianze ha distribuzione asintotica χ^2 (con j gradi di libertà pari al numero delle variabili esplicative). Le loro differenze avranno distribuzione χ^2 (con k gradi di libertà pari alla differenza del numero di variabili esplicative).

Se l'obiettivo è quello di sottoporre a verifica l'ipotesi che riguarda la nullità congiunta di tutti i coefficienti (esclusa l'intercetta),

$$H_0 : \beta_1 = \beta_2 = \cdots = 0$$

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.

Dimostrazioni

Non Efficienza con Errori Eteroschedastici

Siano ε^* gli errori eteroschedastici, $E[b^*] = E[((X'X)^{-1}X'y - b)((X'X^{-1})X'y - b)'] =$ $E[((X'X)^{-1}X'(X\cancel{b}+\varepsilon^*)-\cancel{b})((X'X^{-1})X'(X\cancel{b}+\varepsilon^*)-\cancel{b})'] = (X'X)^{-1}X'E[\varepsilon^*(\varepsilon^*)']((X'X)^{-1}X')' =$

$(X'X)^{-1}X'\Sigma_{\varepsilon^*}((X'X)^{-1}X')'$ e abbiamo che Σ_{ε^*} varia al variare dell'indice. $\square$

Non collinarietà dopo la trasformazione ritardata

Il modello trasformato è

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

quindi abbiamo che

$$cov(w_t, w_{t-1}) = cov(\varepsilon_t^{\#} - \rho\varepsilon_{t-1}^{\#}, \varepsilon_{t-1}^{\#} - \rho\varepsilon_{t-2}^{\#}) =$$

$$= cov(\varepsilon_t^{\#}, \varepsilon_{t-1}^{\#}) - \rho cov(\varepsilon_{t-1}^{\#}, \varepsilon_{t-1}^{\#}) - \rho cov(\varepsilon_t^{\#}, \varepsilon_{t-2}^{\#}) + \rho^2 cov(\varepsilon_{t-1}^{\#}, \varepsilon_{t-2}^{\#}) =$$

$$= \rho - \rho - \rho^3 + \rho^3 = 0$$

il risulta poteva anche essere intuitivo poiché ho eliminato la parte ritardata che era presente nella variabile e portava alla multicollinarietà. $\square$