Dispensa del corso di Statistical Modeling

[0]Pietro Carmine Valenti, CDLM Data Science, Università di Milano-Bicocca

Indice

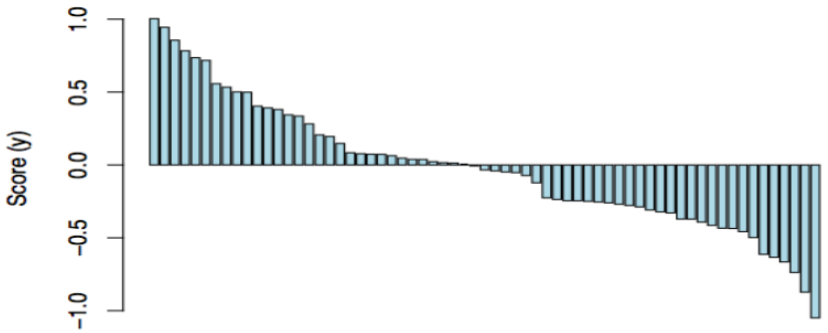
Nel procedimento di descrizione del modello lineare verrà spesso usato come esempio un set di dati composto dalle informazioni circa 4,059 studenti; le variabili rilevate sono:

Variabile	**Descrizione**	**Tipologia**
Score	Punteggio finale (Var. dipendente)	Continua
LR test Score	Punteggio intermedio	Continua
Student intake	Punteggio iniziale	Ordinale
Stud. Verbal Reasoning	Punteggio ragionamento	Ordinale
Stud. Gender	Genere degli studenti	Dicotomica
School Gender	Genere della scuola (M,F,Mix)	Discreta

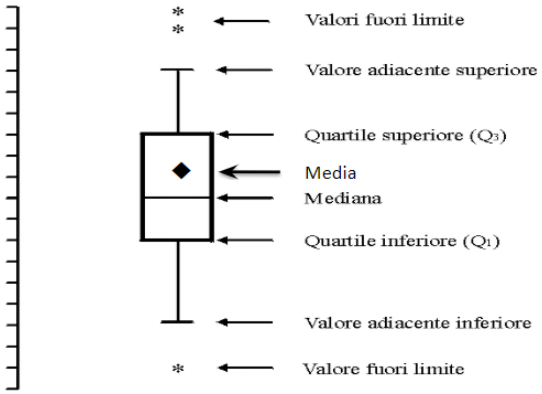
Il principale **scopo del modello lineare** è quindi determinare quanto ciascuna variabile dipendente contribuisca alla spiegazione della variabile indipendente (nell'esempio rappresentata da *Score*).

Esistono diversi passaggi da seguire prima della costruzione di un modello vero e proprio, la prima delle quali consiste nello studio delle statistiche descrittive; ne esistono molte tra le quali:

- **Studio delle distribuzioni**: nell'immagine sottostante viene mostrato il grafico della distribuzione standardizzata della variabile dipendente.



- **Range di variazione** (con minimo e massimo) e **Quartili**
- **Media** e **Mediana Frequenze assolute** e **relative** per variabili discrete e dicotomiche, e di conseguenza la **moda**
- **Box-Plot**, grafico che riassume mediana, media, *range*, quartili (IQR), *outliers*; consente di visualizzare inoltre la simmetria della distribuzione (che si ha quando media e e mediana coincidono). Solitamente si utilizza per attributi continui, ma può essere applicato ad ogni tipologia.



- **Studio delle correlazioni** tra attributi; rappresenta una delle prime analisi vere e proprie da svolgere, e può fornire un idea preliminare sull'intensità della relazione (lineare) che intercorre tra i diversi attributi. Quando si studiano le correlazioni tra la variabile dipendente e le variabili esplicative si possono identificare, in via preliminare, gli attributi che potrebbero spiegare maggiormente la variabile dipendente nel modello.

Per comprendere tuttavia cos'è un modello lineare, è necessario prima fornire la definizione generale di modello statistico: **Modello Statistico** → studia ed esplicita la relazione matematica fra diversi attributi, semplificando la realtà secondo un *trade-off* tra adattamento ai dati (precisione) e capacità di generalizzare.

In un modello statistico è di fondamentale importanza l'interpretabilità dei risultati e delle relazioni, di conseguenza richiede che il numero di variabili esplicative non sia troppo elevato (poiché ciò ridurrebbe la comprensibilità dei risultati). Al fine di selezionare solamente le variabili più interessanti ed influenti è necessario studiare il fenomeno in esame.

A differenza dei modelli matematici, i quali sono deterministici e strutturati in modo da approssimare i dati alla perfezione, i modelli statistici sono sempre caratterizzati dall'**Errore**, i quali possono derivare dalla selezione dei campioni (numerosità, metodo, porzione della popolazione), componenti sistematiche (varabili importanti) non incluse nel modello, errori di misura e altre situazioni.

La costruzione di un modello statistico segue un processo composto da diverse fasi:

- **Studio del Fenomeno**: formulazione di ipotesi sulle relazioni empiriche o di causa-effetto tra variabili e selezione delle variabili esplicative.
- **Selezione dei dati**: fonte, campionamento, trattamenti preliminari
- **Specificazione del modello**:
  - Stima dei parametri.
  - Verifica del modello.
  - Utilizzo del modello.

I modelli di regressione rappresentano una particolare tipologia di modello statistico i quali si compongono di una struttura funzionale (f) nota:

$$\boxed{y = f(x_1, ..., x_k)}$$

nel quale y rappresenta la variabile che il modello vuole spiegare (dipendente/risposta, stocastica), $x_1, ..., x_k$ rappresentano i k attributi impiegati per la costruzione del modello (variabili indipendenti/esplicative o regressori, non stocastici) e infine e rappresentano gli errori stocastici derivanti dalla non conoscenza della vera relazione tra y e i regressori. Considerando gli errori, per ogni i-esima osservazione ($i = 1, ..., n$ con n numerosità campione/popolazione) si avrà:

$$\boxed{y_i = f(x_{1i}, ...x_{ki}) \ + \ e_i \ | \ \forall i = 1, ..., n \in N^+}$$

Modello Lineare

La tipologia più comune di modelli di regressione riguarda i **Modelli di Regressione Lineare**, i quali sono caratterizzati da una forma funzionale f lineare, e possono essere:

- **Semplici**: hanno un'unica variabile dipendente e un'unica variabile esplicativa

$$\boxed{\mathbf{y} = b_0 + b_1 \cdot \mathbf{x_1} + \mathbf{e}}$$

- **Multipli**: un'unica variabile dipendente e più variabili esplicative:

$$\boxed{\mathbf{y} = b_0 + b_1 \cdot \mathbf{x_1} + ... + b_k \cdot \mathbf{x_k} + \mathbf{e} = \mathbf{Xb} + \mathbf{e}}$$

- **Multivariati**: più variabili dipendenti e più variabili esplicative:

$$\boxed{\mathbf{Y} = \mathbf{XB} + \mathbf{E}}$$

  con $\mathbf{Y} = (\mathbf{y'_1}, ..., \mathbf{y'_r})$, $\mathbf{B} = (\mathbf{b'_1}, ..., \mathbf{b'_r})$, $\mathbf{E} = (\mathbf{e'_1}, ..., \mathbf{e'_r})$.

Attraverso lo studio di $\mathbf{y}$ e $\mathbf{x}$ occorre stimare i parametri ignoti $\mathbf{b}$ e gli errori $\mathbf{e}$; in particolare la **stima dei b** risulta fondamentale, poiché essa cattura la relazione che vi è tra la variabile risposta e i regressori.

Esistono diverse tecniche per la stima di un modello; tra i più utilizzati vi è il **criterio dei minimi quadrati** (*Ordinary Least Square*, OLS), il cui concetto

fondamentale consiste nel trovare il vettore **b** che renda minima la norma del vettore degli scarti **e**:

$$\boxed{\min \sum_{i=1}^{n} (\mathbf{y_i} - \mathbf{b}\mathbf{x_i'}^2) \;=\; \min \sum_{i=1}^{n} u_i^2}$$

In caso la matrice $\mathbf{X}$ abbia rango pieno, la soluzione al problema è data da:

$$\boxed{\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

da cui possiamo ricavare le $\mathbf{y}$ stimate dal modello:

$$\boxed{\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

e gli errori $\mathbf{e}$ stimati (residui):

$$\boxed{\hat{\mathbf{e}} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

E' importante analizzare la varianza del modello, soprattutto in fase di verifica; in caso le variabili siano centrate si ha che la **devianza totale** sarà $TSS = \mathbf{y}'\mathbf{y}$. Sapendo inoltre che y=$\hat{\mathbf{y}}+\hat{\mathbf{e}}$, la **devianza spiegata** dal modello risulta $SSE = \hat{\mathbf{y}}'\hat{\mathbf{y}}$, mentre la **devianza residua** $SSR = \hat{\mathbf{e}}'\hat{\mathbf{e}}$, avremo che:

$$\boxed{TSS = SSM + SSR}$$

da cui si ricava un indicatore molto utile detto **Indice di adattamento** del modello:

$$\boxed{R^2 = \frac{SSE}{TSS}}$$

ovvero la quota di variabilità totale che il modello è riuscito a catturare; per tenere conto della *curse of dimensionality*, spesso si utilizza R^2 aggiustato (penalizzando per numero di coefficienti e di osservazioni→**Criterio della Parsimonia**):

$$\boxed{R^2_{adj} = 1 - \frac{SSE}{(n-k-1)} \Big/ \frac{SST}{(n-1)}}$$

.

Spesso per eliminare il problema dei differenti ordini di grandezza dei regressori si utlizza la standardizzazione delle variabili, ovvero si divide la matrice $\mathbf{X}$ per $\mathbf{D_x}$ (matrice diagonale delle varianze di $\mathbf{X}$), mentre si divide $\mathbf{y}$ per la varianza σ_y.

1.1 Ipotesi Classiche

Per la costruzione di un modello lineare è necessario che siano verificate alcune **Ipotesi**:

1. Linearità: sia i regressori, sia i coefficienti $\mathbf{b}$ sono lineari $\rightarrow \mathbf{y} = \mathbf{Xb} + \mathbf{e}$.
2. Non sistematicità degli errori: $E(e_i) = 0 \ \forall i = 1, ..., n$. Se non fosse rispettata $\mathbf{e}$ sarebbe parte sistematica del modello (e non casuale); inoltre valgono $E(\mathbf{e}|\mathbf{X}) = 0$ e $E(\mathbf{y}|\mathbf{X}) = \mathbf{Xb}$.
3. Sfericità degli errori: gli errori sono omoschedastici ($Var(e_i) = E(e_i^2) = \sigma^2$) ed incorrelati ($Cov(e_i \cdot e_j) = E(e_i \cdot e_j) = 0$).
4. Non stocasticità delle variabili esplicative $\mathbf{X}$, ovvero i regressori sono noti con certezza ($E(\mathbf{X}) = X$); anche in caso vi fosse una parte stocastica nella matrice $\mathbf{X}$, tale parte può essere attirbuita all'errore.
5. Non collinearità delle variabili esplicative: le variabili esplicative sono linearmente indipendenti, ovvero la matrice $\mathbf{X}$ ha rango pieno, pari al numero di variabili più la costante ($Rango(\mathbf{X}) = p + 1$). Di conseguenza la matrice $\mathbf{X'X}$ è non singolare.
6. Numerosità della popolazione: $n > p + 1$ affinché esista un unica inversa di $\mathbf{X'X}$.

Ovviamente non sempre tali ipotesi possono essere verificate, in quanto molto stringenti e spesso non coerenti con la realtà dei fenomeni (soprattutto per ciò che riguarda omoschedasticità e incorrelazione degli errori). D'ora in poi assumeremo 2,4,5 e 6 come sempre valide, mentre 1 e 3 come semplificatrici.

Esiste una settima ipotesi utilizzata nella costruzione dei test inferenziali, ovvero la **normalità degli errori**: $\epsilon_i \sim N(0, \sigma^2)$.

1.2 Stime, stimatori e test d'ipotesi

Tipicamente all'inizio della costruzione di un modello lineare bisogna valutare se e come applicare un metodo di campionamento; è infatti vero che spesso si ha la necessità di ridurre il numero di osservazioni (Parsimonia), estraendo un campione dalla popolazione, il quale sia il più rappresentativo possibile e che consenta la generalizzazione dei risultati.

Un modello campionario viene quindi costruito utilizzando una parte della popolazione di numerosità n (con $n \leq N$, N numerosità popolazione), la quale può essere estratta in maniera casuale o non; solitamente un campione casuale consente di ottenere con maggior probabilità un modello robusto. I risultati possono variare a seconda del campione utilizzato, di conseguenza bisogna generalizzare in maniera da ottenere risultati vicini a quelli della popolazione. Proprio per tale motivo si utilizzano gli stimatori dei parametri: assumendo $\mathbf{b}$ come vettore ignoto dei parametri

della popolazione e $\mathbf{B}$ stimatore dei parametri sul campione, il modello campionario avrà formulazione:

$$\boxed{\mathbf{y} = \mathbf{XB} + \epsilon}$$

con stima di $\mathbf{B}$ pari a $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Al variare del campione si genera una distribuzione stocastica del vero valore di $\mathbf{B}$; da ciò si può concludere che, mentre le stime sulla popolazione sono certe, la stime campionarie sono caratterizzate da incertezza. Gli stimatori devono possedere alcune **Proprietà** affinché siano attendibili (robusti):

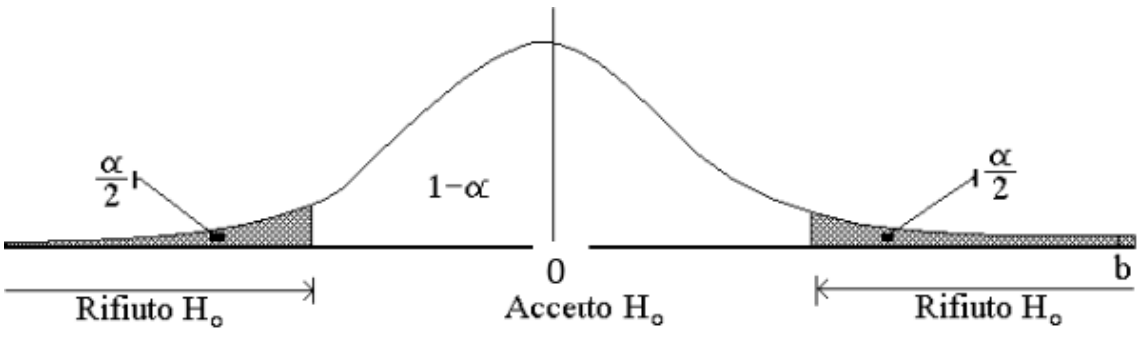
1. **Correttezza**:
$$E[\mathbf{B}] = \mathbf{b} \;\rightarrow\; E[\mathbf{B}] - \mathbf{b} = 0$$
2. **Efficiente**: il più efficiente è lo stimatore con varianza minore. Più la varianza è "piccola", più sarà probabile che lo stimatore si avvicini al valore vero di $\mathbf{b}$
3. **Consistenza**: uno stimatore si dice consistente se, al crescere dell'ampiezza campionaria, la probabilità che cada in un intervallo del valore vero tende a 1:
$$\lim_{n\to\text{inf}} P(|\mathbf{B} - \mathbf{b}| < k) = 1$$

Gli stimatori OLS sono corretti, infatti:

$$E[\hat{\mathbf{B}}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{Xb} + \mathbf{e}] = \mathbf{b}$$

Inoltre il teorema di **Gauss-Markov** dimostra che gli stimatori degli OLS hanno varianza minima nella classe degli stimatori non distorti; vengono denominati *Best Linear Unbiased Estimator* (BLUE) (dimostrato dal teorema di Gauss-Markov). Conseguentemente si può dimostrare che lo stimatore OLS è consistente.

Sugli stimatori $\mathbf{B}$ si possono costruire dei **test inferenziali**, volti a verificare la significatività degli stessi; risulta molto utile per tali procedimenti l'impiego della proprietà di normalità degli errori. I test sulla significatività utilizzano come ipotesi nulla $b_j = 0$, ovvero si vuole verificare se l'ignoto parametro non è significativamente diverso da 0; in caso H_0 non venga rigettata non si può considerare il regressore x_j come utile a spiegare la variabile risposta. Esistono diversi test, il più semplice è quello a **varianza nota**: considerando $N(b_j, \frac{\sigma^2}{n\sigma_{jj}^{-1}})$ (con σ_{jj} valore j-esimo della diagonale della matrice $\mathbf{X}'\mathbf{X}$), bisogna verificare l'ipotesi $H_0 : b_j = 0$. A questo punto rifiuterò H_0 se l'intervallo centrale della statistica non comprende il valore di b_j.

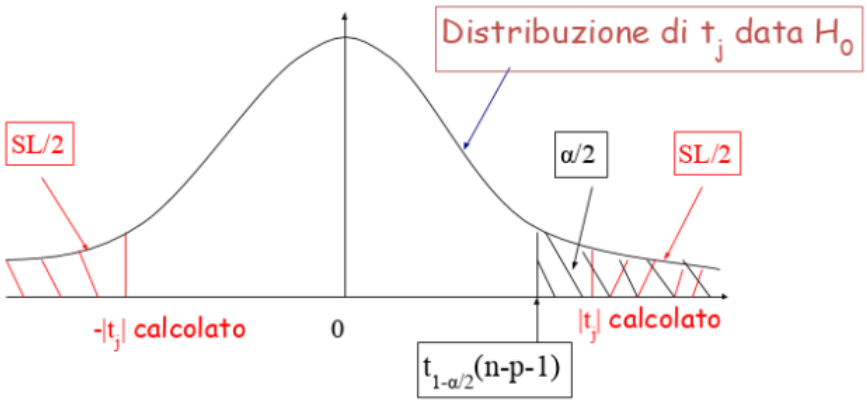


Accetterò invece H_0 se il parametro b_j cadrà nell'$IC : [-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n \cdot \sigma_{jj}^{-1}}} ; +z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n \cdot \sigma_{jj}^{-1}}}]$, il che avviene con probabilità $P(b_j \in IC) = 1 - \alpha$.

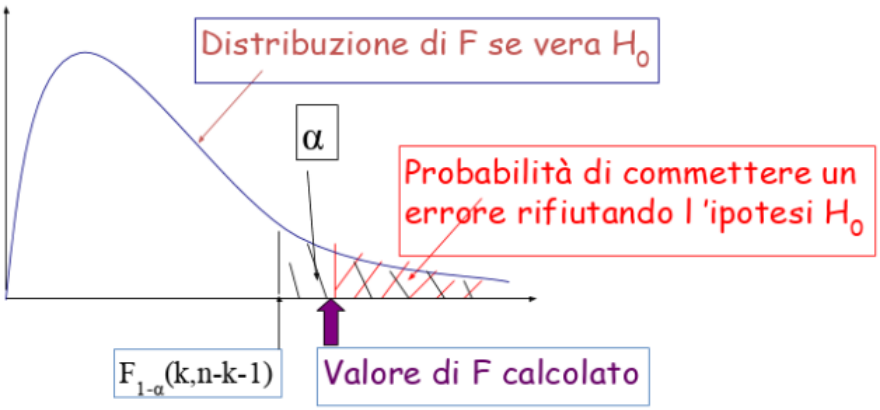
Ovviamente l'ipotesi di varianza nota non sempre risulta soddisfatta: in caso ciò non si verifichi è necessario passare a **test inferenziali a varianza ignota**. L'interpretazione del test è analoga al caso con varianza nota, con la differenza che questa volta si utilizzerà la statistica test t, la quale si serve lo scarto campionario, invece di z. La t-test viene costruita come una Normale su una Chi-Quadro ($n-k-1$ gradi di libertà):

$$\frac{b_j}{\sigma/\sqrt{n \cdot \sigma^{-1} - jj}} / (\frac{s}{\sigma}) = \frac{b_j}{s/\sqrt{n \cdot \sigma^{-1} - jj}} \sim T$$

sotto ipotesi di indipendenza tra la Normale e la Chi-Quadro. L'intervallo di confidenza prenderà forma $IC : [-t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n \cdot \sigma_{jj}^{-1}}} ; +t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n \cdot \sigma_{jj}^{-1}}}]$ con $P(b_j \in IC) = 1 - \alpha$



E' infine possibile applicare un **test inferenziale sul modello** detto test-F; il test si basa sull'indicatore di adattamento $R^2 = \frac{SSE}{SST}$, dal quale si costruisce la statistica test $F = \dfrac{SSE/k}{SSR/(n-k-1)} \sim F - Snedecor$ (rapporto tra Chi-Quadro indipendenti tra loro, divisi per i gradi di libertà). Essendo $H_0 : R^2 = 0$, accettando l'ipotesi nulla non si rifiuta l'idea che il modello non spieghi i dati. La probabilità di accettazione sarà $P(F_{oss} < F_{\alpha;(k,n-k-1)}) = 1 - \alpha$



Il test F può essere impiegato anche per testare l'ipotesi di nullità di più parametri: si ipotizzi che vi siano $q < k$ parametri nulli ($H_0 : b_1, ..., b_q = 0$), SSE_1 e SSR_1 rispettivamente devianza spiegata e residua delle ultime $k - q$ variabili, e la statistica test $F = \frac{SSE - SSE_1/(k-q)}{SSR/(n-k-1)} \sim F_{k-q,n-k-1}$ (rapporto tra chi-quadro indipendenti divise per i gradi di libertà).

Quando si fissa un livello di confidenza $1 - \alpha$ si fissa la percentuale di intervalli di confidenza dello stimatore dei parametri che conterranno il vero parametro della popolazione:

$$\boxed{P[\beta_j - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n \cdot \sigma_{jj}^{-1}}} < b_j < \beta_j + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n \cdot \sigma_{jj}^{-1}}}] = 1 - \alpha}$$

ovvero vi sarà una quota α (incertezza) di intervalli che non conterrà l'ignoto parametro della popolazione. L'intervallo di confidenza per l'ignoto parametro della popolazione b_j sarà dato da:

$$IC(b_j) : \boxed{[\beta_j - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n \cdot \sigma_{jj}^{-1}}} \ ; \ \beta_j + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n \cdot \sigma_{jj}^{-1}}}]}$$

con $P(b_j \in IC(b_j)) = 1 - \alpha$

Esistono alcuni tipi di errori quando si svolgono test sulla significatività dei parametri: l'**errore di prima specie** si riferisce al rifiuto dell'ipotesi nulla ($H_0 : b_j = 0$) quando in realtà l'ignoto parametro non è significativo. Solitamente si guarda il valore del **p-value**, ovvero della probabilità che la distribuzione Z (o altre distribuzioni) sia minore della statistica test z osservata: più sarà piccolo il p-value, maggiore sarà la mia propensione a rifiutare l'ipotesi nulla.

Nel modello lineare classico solitamente solitamente si ammette la **normalità**:

- degli errori ($\epsilon_i \sim N(0, \sigma^2)$)
- degli stimatori OLS ($\hat{B} \sim N(\mathbf{b}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$)
- delle variabili risposta ($Y \sim N(\mathbf{Xb}, \sigma^2\mathbf{I}_n)$)

Da tali ipotesi si può passare a definire il metodo di stima della **Massima Verosimiglianza**, un metodo che utilizza la distribuzione normale, il quale permette di ottenere le stesse stime dei minimi quadrati, da quali ereditano le proprietà di efficienza, consistenza e correttezza. Essi di conseguenza sono BLUE, ma hanno una caratteristica in più rispetto agli stimatori OLS: sono gli stimatori corretti con varianza minore (UMVUE, dimostrato con il teorema di Cramer-Rao).

Quando si stimano i parametri di un modello è importante lo studio dello **Standard Error** (SE), il quale fornisce una misura dell'instabilità delle stime ottenute.

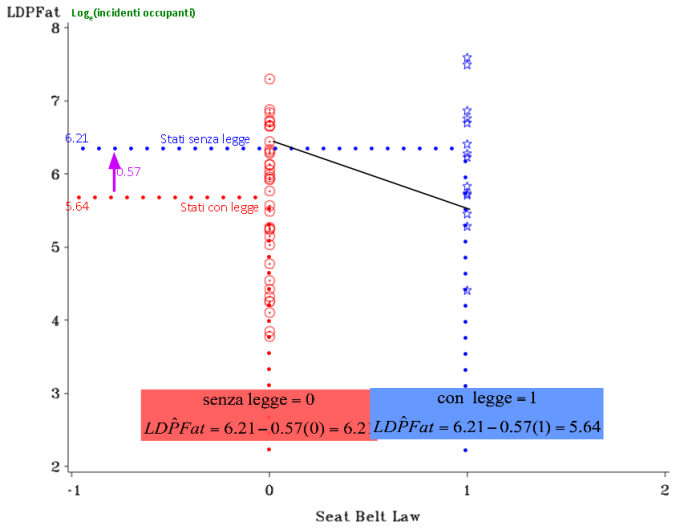
1.3 Variabili qualitative

Come anticipato prima un modello lineare è in grado di trattare anche variabili qualitative, sia che esse siano nominali (per esempio dicotomiche), sia che esse

siano ordinali (per esempio livello di istruzione). Prendiamo il caso in cui un modello abbia a disposizione una sola variabile qualitativa, per esempio una *dummy* (Sesso); dati $\mathbf{y_i}$ reddito e $\mathbf{D_i}$ sesso, avremo $\mathbf{y}_i = \beta_0 + \beta_1 \mathbf{D}_i + \epsilon_i$, con $\mathbf{D}_i = 1$ in caso il genere sia femminile, 0 in caso sia maschile. Il modello assumerà forma:

$$Y_i = \begin{cases} \beta_0 + \epsilon_i, & \text{se } D_i = 0 \text{ (Maschi)} \\ \beta_0 + \beta_1 + \epsilon_i, & \text{se } D_i = 1 \text{ (Femmine)} \end{cases}$$

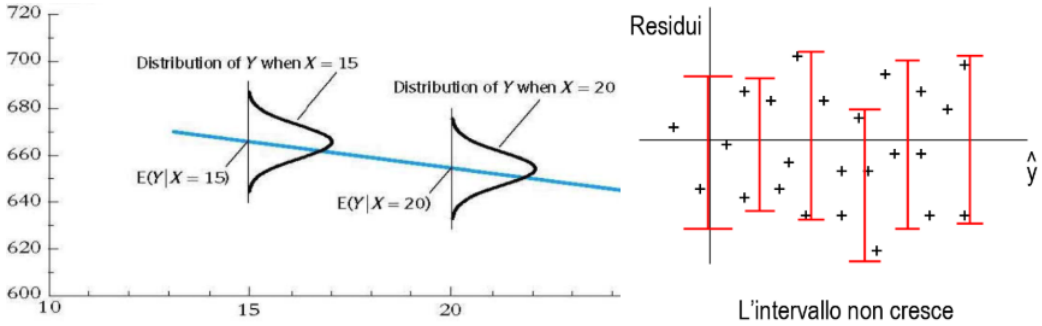
Un esempio di come appaia un modello con una sola variabile *dummy* può essere quello mostrato di seguito, in cui vengono analizzati gli incidenti stradali in paesi con o senza legge sulla cintura di sicurezza



L'interpretazione del coefficiente di una *dummy* consiste nell'aumento della variabile risposta a seguito di un aumento unitario della variabile esplicativa, quindi nel passaggio da 0 a 1 (per esempio da genere maschile a genere femminile).

1.4 Eteroschedasticità degli errori

Una delle ipotesi classiche che non sempre viene rispettata è quella dell'omoschedasticità degli errori; tale proprietà afferma che la varianza della distribuzione degli errori associati alla i-esima osservazione rimane costante al variare del campione ($\sigma^2_{i1} = ... = \sigma^2_{is} = \sigma^2$, i =osservazione, s =numero campioni considerati).



In molto casi tuttavia tale ipotesi non viene rispettata; il primo effetto in caso di **eteroschedasticità** degli errori riguarda la varianza degli stimatori:

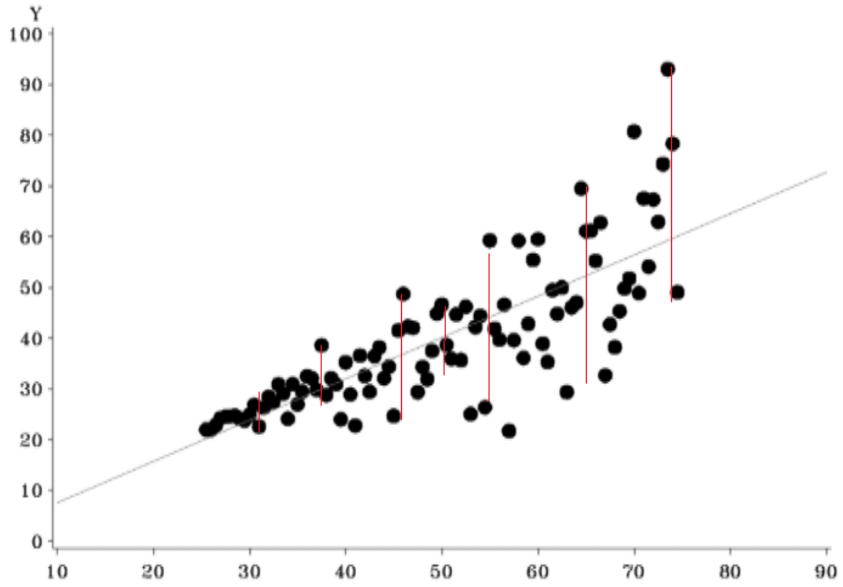
$$\boxed{Var(\mathbf{B}^*) = ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{\Sigma}_{e^*}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'}$$

In questo caso $\mathbf{B}^*$ non risulterà più BLUE.

Un'altra componente che cambierà riguarda le statistiche utilizzate per i test sui parametri: per esempio nel test t ($t = \frac{\beta_j}{s_{i*}/\sqrt{n \cdot \sigma - 1_{jj}}}$) l'intervallo di confidenza varierà in base al campione i. La varianza campionaria s tende a sottostimare il valore di σ^2, di conseguenza avrò delle regioni di rifiuto più ampie, con un aumento del rischio di incorrere in errori di prima specie.

1.4.1 Analisi grafica

Una prima tecnica per notare se vi siano situazioni di eteroschedasticità è quello di analizzare i grafici di dispersione dei regressori (presi singolarmente) e la variabile risposta; una dispersione non costante e disomogenea è un primo segnale di possibile presenza di eteroschedasticità (ovviamente ipotesi che andrà verificata).



Di analoga interpretazione sono i grafici di dispersione dei valori previsti $\hat{\mathbf{y}}$ (o dei regressori) rispetto ai residui (normali o al quadrato) e dei valori osservati $\mathbf{y}$ rispetto ai predetti $\hat{\mathbf{y}}$. A seconda della grandezza del campione la dispersione può essere maggiore (nel caso di campioni piccoli, solitamente concentrata nell'intorno della media) o minore (campioni grandi).

Un alternativa ai grafici di dispersione si ottiene dall'analisi dei boxplot per le diverse regioni dei dati (divido in regioni e guardo dispersione Box-Plot).

1.4.2 Test sull'eteroschedasticità

Ovviamente l'analisi grafica è solo indicativa; per confermare o rigettare l'ipotesi di eteroschedasticità bisogna svolgere dei test d'ipotesi. La presenza di eteroschedasticità e sempre legata a una sistematicità tra il set di regressori e gli errori. Uno tra i test più utilizzati è il **Test di White** (H_0= Errori omoschedastici), il quale segue i seguenti passaggi:

1. Costruisco la regressione campionaria $y_i = b_0 + b_1 \cdot x_{i1} ... + b_p \cdot x_{ip} + \epsilon_i^2$
2. Regredisco ϵ_i^2, mediante OLS, su x_j, x_k^2 con le interazioni.
3. Costruisco la statistica test: $LM = n \cdot R^2 \sim \chi_n^2$
4. Utilizzo un test Chi-Quadro: se LM cade nella regione di rifiuto ($p - value < (0.05, 0.01, 0.001)$) allora gli ϵ_i variano al variare degli x_i (del campione). Ciò è indice di eteroschedasticità.

Uno dei metodi per risolvere il problema di eteroschedasticità è l'utilizzo dei minimi quadrati pesati (**Weighted Least Squares**, WLS), i quali consistono nel dividere i regressori per lo *standard error*.

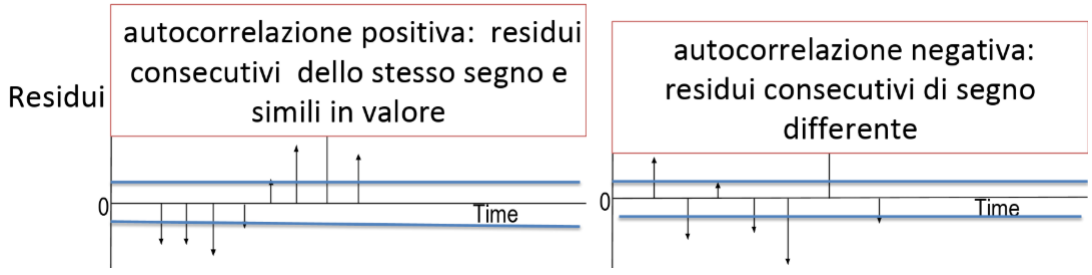
1.5 Autocorrelazione degli errori

La seconda violazione dell'ipotesi di sfericità degli errori riguarda la presenza di autocorrelazione tra errori ($cov(\epsilon_i, \epsilon_j) = \rho_{ij} \neq 0$). Con autocorrelazione degli errori si intende la situazione nel quale errori del passato ($t-1$ o $i-1$) influenzano gli errori successivi (sono quindi determinati da essi). Come nel caso dell'eteroschedasticità gli stimatori OLS non saranno più BLUE.

1.5.1 Analisi grafica

Un primo metodo per identificare una possibile presenza di autocorrelazione tra errori è l'analisi grafica dei correlogrammi (o dei grafici di dispersione degli errori). Per esempio nel grafico di dispersione gli errori, affinchè vi sia assenza di autocorrelazione, dovrebbero disporsi casualmente intorno allo 0, in maniera da non poter identificare una sistematicità nella disposizione.

Un esempio invece di come appare un grafico dei residui in presenza di autocorrelazione è fornito dalla seguente immagine:



1.5.2 Test sull'autocorrelazione

Tra i test maggiormente utilizzati per la verifica di autocorrelazione degli erorri è il **test di Durbin-Watston**, il quale utilizza la statistica-test DW al fine di verificare l'ipotesi nulla di assenza di autocorrelazione seriale del primo ordine ($H0 : \rho = 0$). La statistica test ha forma: $DW = \frac{\sum_i (\epsilon_i \epsilon_{i-1})^2}{\sum_i \epsilon_i^2}$ che tuttavia può essere scomposta e riportata alla forma intuitiva $DW = 2 \cdot (1 - \rho)$; poichè $-1 \leq \rho \leq 1$, la statistica apparterrà alll'intervallo $DW \in [0, 4]$, con:

- $DW > 3 \rightarrow$ Autocorrelazione negativa ($DW = 4$,ρ=-1)
- $DW < 1 \rightarrow$ Autocorrelazione positiva ($DW = 0$,ρ=1)
- $1 \leq DW \leq 3 \rightarrow$ Incorrelazione ($DW = 2$,$\rho = 0$)

In caso di autocorrelazione degli errori si utilizzano i *Weighted Least Squares* (WLS) ovvero vengono pesate la variabile risposta, i regressori e gli errori, dividendoli per $\sqrt{s_i}$, ottenendo così un modello trasformato.

1.6 Minimi quadrati generalizzati (GLS)

Uno dei metodi per risolvere il problema della non sfericità degli errori consiste nell'impiego dei **Minimi Quadrati Generalizzati** (*Generalized Least Squares*, GLS): gli stimatori GLS hanno interpretazione analoga ai minimi quadrati ordinari, tuttavia si basano su trasformate del modello lineare ($\mathbf{y} = \mathbf{X}\hat{\mathbf{B}}^* + \hat{\mathbf{E}}^*$, $\hat{\mathbf{E}}^*$ errori correlati ed eteroschedastici). La matrice di varianze/covarianze campionaria degli errori diventerà quindi $\mathbf{S}_{\hat{E}^*} = \frac{1}{n}(\hat{\mathbf{E}}^* \hat{\mathbf{E}}^{*\prime})$; ipotizzando una matrice non singolare $\mathbf{V}$ tale che $\mathbf{S}_{\hat{E}^*} = \sigma^2 \mathbf{V}\mathbf{V}'$ (ovvero una decomposizone spettrale di $\mathbf{S}$), possiamo definire una trasformata degli errori $\hat{\mathbf{E}} = \mathbf{V}^{-1}\hat{\mathbf{E}}^*$, la quale consentirà di ottenere errori incorrelati e omoschedastici:

$$\boxed{(\mathbf{V}^{-1})\mathbf{S}_{\hat{E}^*}(\mathbf{V}^{-1\prime}) = (\mathbf{V}^{-1})\sigma^2\mathbf{V}\mathbf{V}'(\mathbf{V}^{-1})' = \sigma^2\mathbf{I}_n'}$$

ovvero la matrice di varianze/covarianze di errori sferici. Ovviamente tale trasformata inciderà sulla formulazione degli stimatori:

$$\boxed{\hat{\mathbf{B}}^* = (\mathbf{X}'\mathbf{S}_{\hat{E}^*}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}_{\hat{E}^*}^{-1}\mathbf{y}}$$

con $\hat{\mathbf{B}}^*$ stimatore GLS. Ovviamente tutto ciò si basa sull'assunzione che esista la matrice $\mathbf{V}$, la quale deve essere di conseguenza definita; partendo dalla matrice di varianze e covarianze di $\mathbf{y}$ ($\mathbf{\Sigma_y} = \mathbf{B}^{*\prime}\mathbf{\Sigma_{X^*}}\mathbf{B}^* + \mathbf{S_{\hat{E}}} = \mathbf{B}^{*\prime}\mathbf{\Sigma_{X^*}}\mathbf{B}^* + \sigma^2\mathbf{I}_n$) si può ricavare la matrice $\mathbf{V}$ utilizzando le proprietà degli autovettori e autovalori, ovvero attraverso la decomposizione spettrale:

$$\boxed{\mathbf{V} = \sigma\mathbf{A}\mathbf{L}^{\frac{1}{2}}\mathbf{A}'}$$

con $\mathbf{A}$ matrice degli autovettori e $\mathbf{L}$ matrice diagonale degli autovalori, entrambe riferite a $\mathbf{S}_{\hat{E}^*}$.

Possiamo passare ora a definire le **Proprietà degli stimatori dei minimi quadrati generalizzati**, le quali sono fortemente legate al **Teorema di Aitken**:

- $\mathbf{B}^*$ è uno stimatore corretto ($E(\mathbf{B}^*) = \mathbf{b}^*$);
- è consistente;
- è efficiente ($E((\mathbf{B}^* - \mathbf{b}^*)(\mathbf{B}^* - \mathbf{b}^*)') = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}) = \sigma^2(\mathbf{X}'\mathbf{\Sigma}_{\mathbf{E}^*}\mathbf{X})^{-1})$) per le variabili trasformate, come dimostra il teorema di Aitken (analogo a Gauss-Markov). Non sarà il più efficiente in generale (né BLUE, né UMVUE) ma solo nel campo delle trasformate.

Dal momento che la matrice di varianze e covarianze degli errori ($\mathbf{\Sigma}_{E^*}$) non è sempre nota, spesso bisogna impiegare la matrice stimata $\mathbf{S}_{\hat{E}^*}$; in tal caso il metodo di stima verrà denominato **Feasible Generalized Least Squares** (FGLS). Lo stimatore $\mathbf{S}_{\hat{E}^*}$ godrà della proprietà di convergenza in probabilità alla matrice $\mathbf{\Sigma}_{\mathbf{E}^*}$:

$$\boxed{p - \lim_{n\to\infty} \mathbf{S}_{\hat{E}^*} = \mathbf{\Sigma}_{\mathbf{E}^*}}$$

con n numerosità del campione, $\mathbf{E}^*$ errori correlati e eteroschedastici.

1.7 Multicollinearità

Con collinearità si intende la situazione in cui due variabili sono fortemente correlate tra loro (per esempio coefficiente di correlazione maggiore di 0.9); quando una variabile è fortemente correlata a diverse altre variabili esplicative, ci si trova in presenza di **multicollinearità**. Tale condizione può generare forti distorsioni nei risultati di un modello lineare, di conseguenza è sempre bene individuarla e applicare tecniche volte a mitigarne gli effetti. Una delle principali criticità riguardati la multicollinearità riguarda l'impossibilità di identificare una variabile collineare a più variabili solamente dalla matrice di correlazione, la quale mostra solo la collinearità semplice tra due variabili.

In presenza di multicollinearità nella matrice dei regressori $\mathbf{X}$, la matrice $\mathbf{X}'\mathbf{X}$ risulta singolare, e di conseguenza non invertibile, il che può generare:

- Problemi di stime inesistenti;
- soluzioni non uniche;
- forte aumento della varianza dei coefficienti e di conseguenza stime molto instabili e difficoltà nell'impiego del t-test (si allarga di molto IC).

Un'altra forma di collinearità difficile da identificare si ha quando una o più variabili sono combinazione lineare di altre variabili; esistono diversi indici per identificare tali situazioni:

- **Indici di Tolleranza** ($\in [0,1]$): sono indici che misurano il grado di interrelazione di una variabile con le altre. Si costruiscono prendendo una variabile e regredendola rispetto a tutte le altre:

$$\boxed{Tol(x_j) = 1 - R^2(x_j|x_1, ..., x_{j-1}, x_{j+1}, ..., x_n)}$$

  il quale assume valore 1 in caso tutte le variabili siano incorrelate, 0 se x_j è collineare.

- **Varianza Multifattoriale**:

$$\boxed{Vif = Tol^{-1} = \frac{1}{1 - R^2}}$$

  quando è uguale a 0 si hanno variabili linearmente indipendenti, quando è superiore a 10 si ha multicollinearità.

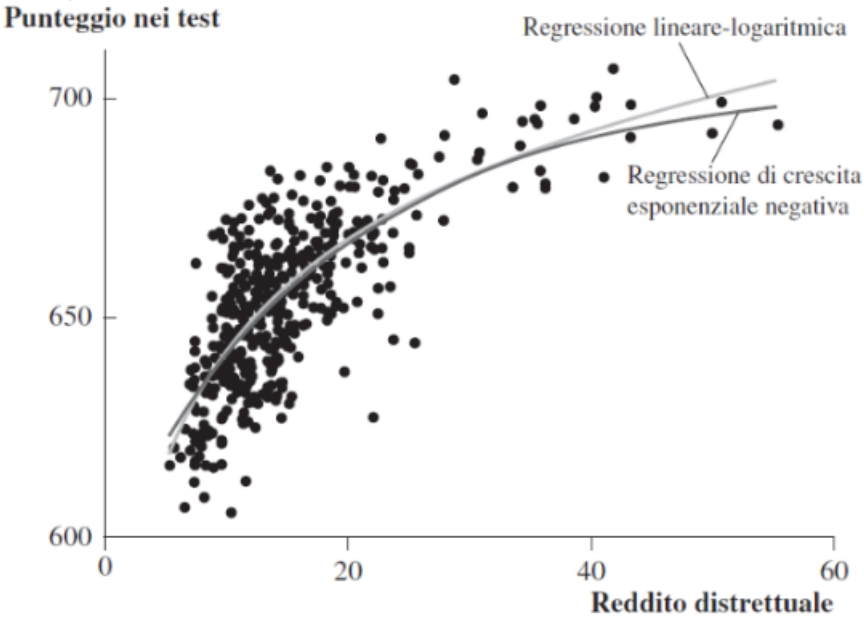
- **Indice di collinearità** (*Condition index*): è dato dal rapporto tra l'autovalore massimo di $\mathbf{X'X}$ e ogni altro autovalore. Quando l'indice è maggiore di 10 indica possibile presenza di collinearità; per confermare bisogna inoltre bisogna verificare che la quota di varianza associata ai *condition index* elevati sia anch'essa eleavata (solitamente > 0.9).

Diagnostiche di collinearità						
Numero	**Autovalore**	**Indice condizione**	**Proporzione di variazione**			
			Interc	**RunTime**	**MaxPulse**	**RunPulse**
1	3.98642	1.00000	0.00015612	0.00088747	0.00002206	0.00002750
2	0.01136	18.73287	0.01794	0.93969	0.00301	0.00270
3	**0.00202**	**44.40634**	0.80384	0.00004079	0.01670	0.05420
4	**0.00019910**	**141.49923**	0.17806	0.05938	**0.98026**	**0.94307**

Il *condition index* risulta il più sofisticato, utile per verificare la presenza di multicollinearità in caso si abbia una forte necessità di ottenere campioni robusti (per esempio per studi in campo medico).

1.8 Relazioni non lineari

Quando il modello no non adatta bene ai dati bisogna verificare, oltre alle eventuali violazioni delle ipotesi classiche enunciate in precedenza, la possibilità di una forma non lineare nel modello, per esempio per ciò che concerne i parametri e/o i regressori. Quando le relazioni non lineari risultino estremamente evidenti è possibile verificarne la presenza già da alcune analisi grafiche, come per esempio dal grafico di dispersione della variabile dipendente rispetto alle esplicative. Uno dei metodi per trattare la presenza di relazioni non lineari è l'impiego dei **Modelli Lineari Generalizzati** (*Generalized Linear Models*, GLM).



Possiamo avere casi in cui le funzioni siano linearizzabili (allora applico una linearizzazione e mi riconduco al modello classico e alle stime OLS), o funzioni che risultano intrinsecamente non lineari. Nel secondo caso, non esistendo trasformazioni che rendano i parametri lineari, si applicano le stime dei **Minimi Quadrati non Lineari** (*Non-linear Least Squares*, NLS), basati su approssimazioni progressive/iterative e algoritmi numerici.

Quando la funzione non è nota a priori si valutano diverse possibili trasformazioni, confrontando gli errori di ciascuna con l'obiettivo di minimizzarli. Solitamente si utilizzano trasformazioni polinomiali, esponenziali e logaritmiche; le diverse tipologie di trasformate logaritmiche per il modello lineare sono fornite nella seguente tabella

I.	lineare-log	$y_i = \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i$	(26a)
II.	log-lineare	$\ln(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i$	(26b)
III.	log-log	$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i$	(26c)

1.9 Violazione della Normalità

Estraendo casualmente da una determinata popolazione una serie di k campioni si generano diverse collezioni di osservazioni; definendo ϵ_{ij} (con $i = 1, ..., n$ osservazioni, $j = 1, ..., k$ campioni) come l'elemento i-esimo del j-esimo campione, si generano n manifestazioni di variabili casuali $(E_1, ..., E_n)$ identicamente distribuite, ciascuna delle quali rappresenta la distribuzione degli errori per l'osservazione i-esima.

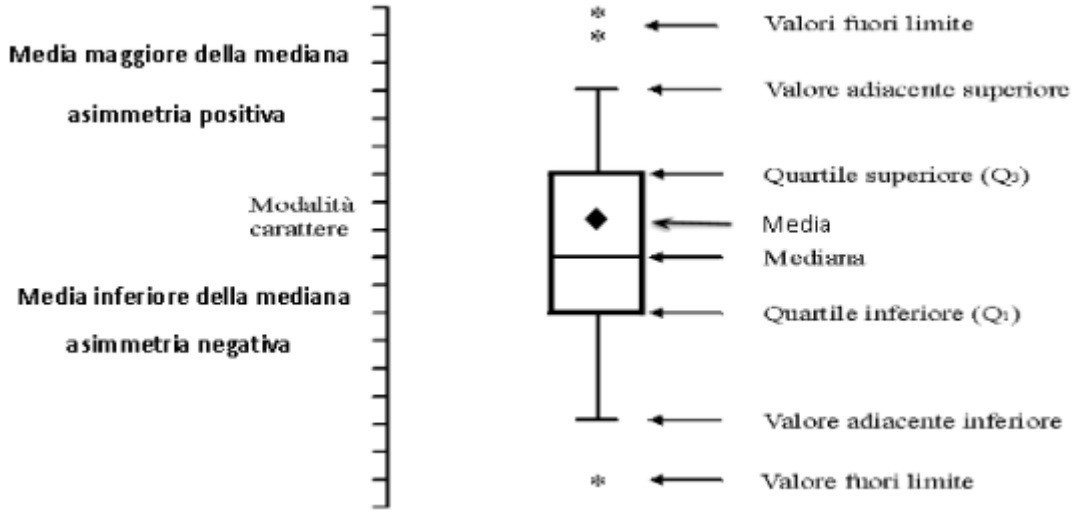
Nel modello lineare classico solitamente si assume che la distribuzione di $E_1, ..., E_n$ sia data da una normale standard con varianza σ^2 ($\mathbf{E} \sim N(0, \sigma^2)$); ciò si verifica (perlomeno asintoticamente) con maggiore facilità in campioni molto ampi, come conseguenza del Teorema del limite centrale (Theil, 1978); per campioni di dimensioni ridotte è invece difficile assumere tale proprietà, e ciò genera alcuni problemi durante la fase di test e costruzione degli intervalli di confidenza per i parametri. I principali **effetti del non rispetto della normalità** degli errori sono:

- Non è possibile assumere la normalità dei parametri **b**.
- Non è possibile ricavare i test basati sulla Normale standardizzata, sulla T di Student e sulla F di Snedecor (significatività parametri, bontà/adattamento del modello).
- Non è possibile ricavare intervalli di confidenza basati sulla Normale standardizzata e sulla T di Student.
- Le stime OLS non coincideranno con le stime di massima verosimiglianza (ML); non varrà più il Teorema di Cramer-Rao e di conseguenza gli OLS non saranno più UMVUE (poichè diversi da ML), rimanendo comunque BLUE ed efficienti.
- A livello computazionale si possono avere problemi con i pacchetti statistici che forniscono stime basate sulla ML.

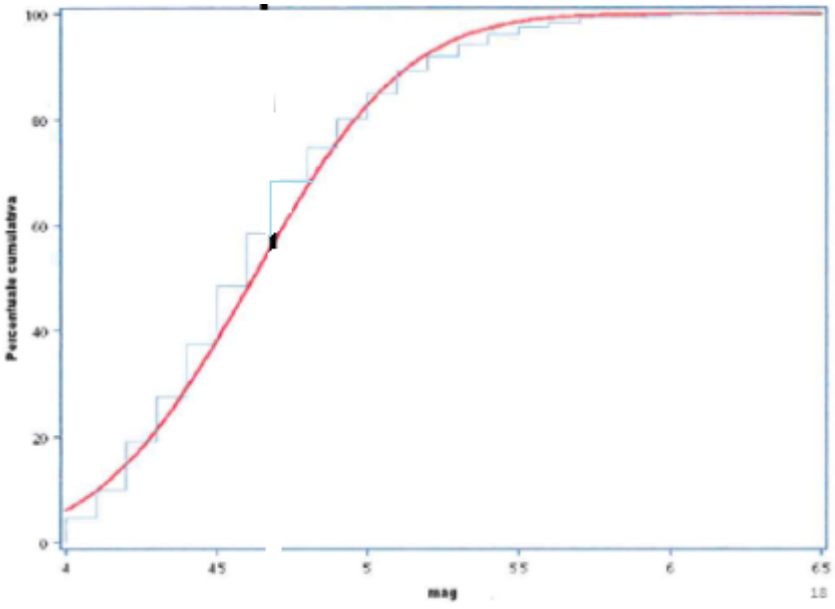
Esistono alcuni indici descrittivi che possono far sorgere dubbi sulla normalità di una distribuzione, per esempio l'**asimmetria** (*skweness*) (ovvero media e mediana assumono valori diversi), la quale più si discosta dallo 0, più fa sorgere dubbi sulla normalità; un secondo indicatore è invece la **curtosi**, la quale assume valore 0 in caso di distribuzione normale. Ovviamente bisogna interpretare gli indici con un certo grado di flessibilità, e svolgere successivamente test per valutare l'ipotesi di normalità della distribuzione.

1.9.1 Analisi Grafica

Graficamente è possibile valutare la normalità utilizzando il Box-Plot, il quale evidenzia la presenza di *outliers* e lo scostamento della media dalla mediana, o mediante istogramma.



Un altro grafico utilizzabile per la verifica della normalità è quello che confronta la curva empirica dei residui cumulati, con la curva teorica di una normale standard; più la prima approssimerà correttamente la seconda, maggiore sarà la probabilità che la distribuzione degli errori sia normale.



Un ultimo grafico che viene spesso impiegato confronta le probabilità cumulate di una distribuzione normale con quelle dei residui (**PP-Plot**); in una situazione ideale, ovvero di normalità perfetta, tutti i punti sarebbero allineati sulla bisettrice del grafico (ovvero le probabilità della normale e della distribuzione dei residui sono sempre coincidenti). Di analoga interpretazione è il grafico che confronta i quantili (**QQ-Plot**).

1.9.2 Test sulla normalità

Successivamente allo studio degli indicatori descrittivi e dei grafici vi è la necessità di passare allo svolgimento di test che possano confermare o rigettare l'ipotesi di normalità degli errori:

- Tra i più utilizzati vi è il **test di Shapiro-Wilk**. Si definisce la statistica test:

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

  la quale tenderà ad 1 in caso i residui provengano da una normale, altrimenti a 0. L'ipotesi nulla del test è H_0 : normalità.

- Il secondo test è il **test di Kolmogorov-Smirnov**, il quale si costruisce nel seguente modo:

  1. L'intervallo di variazione viene divisi in classi di uguale ampiezza.
  2. A ciascuna classe viene assegnata la frequenza cumulata della distribuzione empirica da testare e si calcola la differenza tra essa e la probabilità della normale (per ogni classe).
  3. Si costruisce la statistica D come somma di tutte le differenze per classe.
  4. Si confronta con la distribuzione della D e si valuta accettazione (*p-value*).

- Il terzo è il **test skewness**, il quale utilizza come statistica test l'indicatore di asimmetria:

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

  la quale ha valore atteso $E(S) = 0$ sotto l'ipotesi nulla di normalità dei residui. Il test necessità un campione estremamente ampio.

- L'ultimo è il **test kurtosis**, il quale utilizza come statistica test l'indicatore di curtosi:

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

  il quale, sotto ipotesi di normalità, ha valore atteso pari a $E(K - 3) = 0$

Test di normalità			
Equazione	Statistica di test	Valore	Prob
mag	Shapiro-Wilk W	0.81	0.0143
Sistema	Skewness Mardia	8.31	0.0039
	Curtosi Mardia	2.08	0.0376
	Henze-Zirkler T	0.43	0.5449

Test di normalità			
Equazione	Statistica di test	Valore	Prob
Life_expectancy	Shapiro-Wilk W	0.96	0.1644
Sistema	Skewness Mardia	0.00	0.9554
	Curtosi Mardia	0.16	0.8735
	Henze-Zirkler T	0.49	0.1053

Per risolvere i problemi legati all'assenza di normalità si può utilizzare il teorema centrale del limite (per campioni grandi) o applicare delle trasformate; viene di seguito fornito un esempio delle principali trasformate per ovviare al problema della non normalità.

Partendo dal modello lineare semplice $\mathbf{y} = b\mathbf{x}$ (per semplicità senza intercetta)

Una breve lista delle principali trasformazioni è:

1) $\mathbf{z}=\log \mathbf{y}$ (con y>0)(7) $\longrightarrow$ $(\log \mathbf{y}=\log \mathbf{x}^b)=(\log \mathbf{y}=b \log \mathbf{x})=(\mathbf{z_1}=b\,\mathbf{k_1})$ ove $\mathbf{z_1}=\log \mathbf{y}$, $\mathbf{k_1}=\log \mathbf{x}$

Si usa quando s_ε cresce con $\mathbf{y}$, o se la distribuzione dell'errore ha una asimmetria positiva

2) $\mathbf{w} = \mathbf{y}^2$ (8) $\longrightarrow$ $(\mathbf{y}^2 = b^2\mathbf{x}^2) = (\mathbf{z_2}= c\mathbf{k_2})$ ove $\mathbf{z_2}= \mathbf{y}^2$, $\mathbf{k_2} = \mathbf{x}^2$, $c = b^2$

Si usa se s_ε è proporzionale a $E(\mathbf{y})$ nei diversi campioni o se la distribuzione dell'errore ha asimmetria negativa

3) $\mathbf{v}= \mathbf{y}^{1/2}$ (con $\mathbf{y} > 0$). (9) $\longrightarrow$ $(\mathbf{y}^{1/2} =b^{1/2}\mathbf{x}^{1/2})= (\mathbf{z_3}= d\mathbf{k_3})$ ove $\mathbf{z_3}= \mathbf{y}^{1/2}$, $\mathbf{k_3}= \mathbf{x}^{1/2}$, $d = b^{1/2}$

Si usa quando s_ε è proporzionale a $E(\mathbf{y})$ nei diversi campioni

4) $\mathbf{v} =\mathbf{1/y}$ (10) $\longrightarrow$ $(\mathbf{1/y} = 1/(\mathbf{bx}))= (\mathbf{z_4}= f\,\mathbf{k_4})$ ove $\mathbf{z_4}= 1/\mathbf{y}$, $\mathbf{k_4}= 1/\mathbf{x}$, $f=1/b$

Si usa quando s_ε cresce significativamente al crescere di y nei diversi campioni

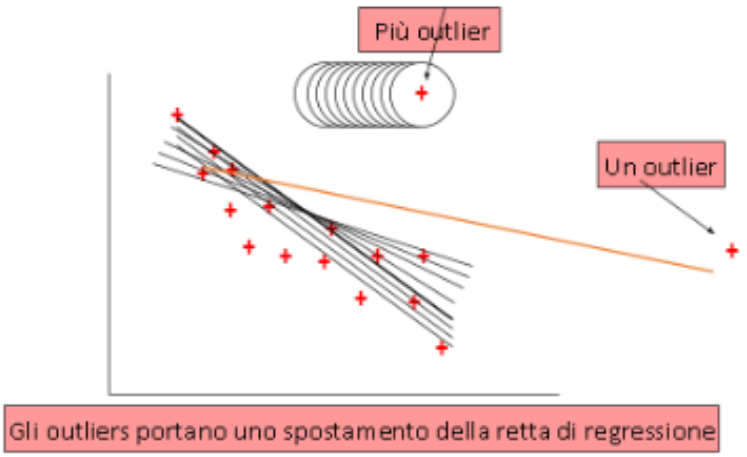
---

1.10 Valori anomali, *outiliers* e punti influenti

Si può partire fornendo alcune definizioni:

- Con **valori anomali** si intendono tutte quelle osservazioni che si discostano significativamente dall'andamento generale.
- Con **punti influenti** si intendono tutte quelle osservazioni che influenzano significativamente le stime.

Entrambe le tipologie di osservazioni sono fonte di distorsione per quanto concerne la stima e la robustezza di un modello statistico. Risulta infatti chiaro come i diversi metodi di stima (OLS, WLS, GLS, FGLS, ...) sono basati su valori medi; un certo numero di valori che si discostano in maniera rilevante dalla media possono influenzare pesantemente i risultati.



Solitamente gli *otuliers* possono essere identificati graficamente sia dal Box-Plot (osservazioni che sono esterne ai baffi), sia dai grafici di dispersione (Scatter-Matrix). Possiamo tuttavia definire dei grafici più sofisticati, i quali ci consentono di identificare gli eventuali *outliers* con maggior certezza; uno tra tutti è il grafico dei residui studentizzati rispendo all'indice i dell'osservazione associata.

Data la matrice di proiezione $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ (capacità del modello di predire la variabile dipendente $\mathbf{y}$), possiamo definire gli errori $\epsilon = (\mathbf{I} - \mathbf{H})\mathbf{y}$, i quali hanno varianza $Var(\epsilon) = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y})$; dalla diagonale della matrice $\mathbf{H}$ possiamo definire gli elementi h^*_{ii} i quali identificano il **leverage**, ovvero l'impatto di ciascuna osservazione sulla capacità previsiva del modello. A questo punto i può ridefinire la varianza degli errori in funzione del *leverage*, la quale assumerà forma $Var(\epsilon_i) = (1 - h^*_{ii})\sigma^2$, da cui ricaviamo:

- I **residui standardizzati**:

$$\boxed{\epsilon^*_i = \frac{\epsilon_i}{\sigma\sqrt{(1 - h^*_{ii})}}}$$

  ovvero i residui di ogni osservazione divisi per lo scarto quadratico medio. Poiché il 99% dei residui standardizzati risulta compreso tra -2.5 e 2.5, possiamo affermare con che un osservazione cui viene associato un $\epsilon^*_i > 2$ (2.5 o 3 a seconda della confidenza) sarà probabilmente un *outlier*.

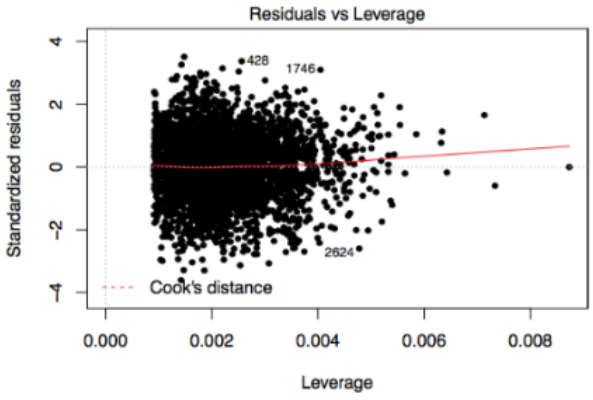
- Dal momento che lo scarto quadratico medio σ spesso non è noto, solitamente si possono utilizzare i **residui studentizzati**, i quali hanno forma:

$$\boxed{\epsilon^s_i = \frac{\epsilon_i}{s_\epsilon\sqrt{(1 - h^*_{ii})}}}$$

- Un'ultima tipologia di residui, i quali sono sempre studentizzati, sono i **residui studentizzati JackKnife**, i quali hanno forma:

$$\boxed{\epsilon^{jk}_i = \frac{\epsilon_i}{s_{\epsilon(i)}\sqrt{(1 - h^*_{ii})}}}$$

  con $s_{\epsilon(i)}$ varianza campionaria dei residui, ottenuta eliminando l'i-esima osservazione. Si possono analizzare graficamente attraverso n confronto con i *leverage*; in tale grafico, in presenza di *outliers*, la retta nera e la curva rossa sono molto distanti.

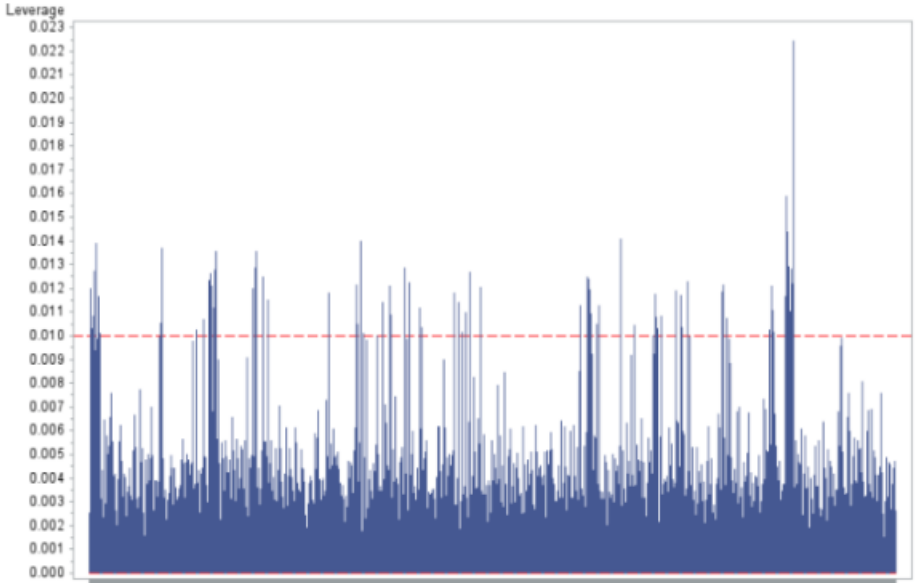


L'impiego dei residui standardizzati o studentizzati, al posto dei residui ordinari, consente di eliminare distorsioni dovute all'ordine di grandezza dell'errore stesso.

I **leverage** possono essere a loro volta impiegati come metodo di identificazione degli *otuliers*; l'interpretazione è semplice, ovvero valori di h^*_{ii} vicini a 1 fanno pensare a un effetto predominante dell'osservazione i sulla capacità di predire il modello. Dato il valore medio dei *leverage* $\frac{k+1}{n}$ (k numero variabili esplicative, n numero osservazioni), possiamo definire un valore soglia dato da $2 \cdot \frac{k+1}{n}$: cataloghereremo quindi come *otuliers* un osservazione i che abbia *leverage*

$$\boxed{h^*_{ii} > 2 \cdot \frac{k+1}{n}}$$

ovvero i ha un influenza predominante sulla stima del modello. Nel grafico sottostante viene mostrato un esempio di grafico dei *leverage* con soglia pari a 3 volte la media dei *leverage* stessi.



Esistono inoltre degli **Indici complessi**, i quali si basano sull'eliminazione dell'i-esima osservazione; ovviamente in caso di campioni molto grandi risultano computazionalmente onerosi, e di conseguenza non rispettano il principio di parsimonia.

- Un primo esempio è fornito dal **CovRatio**, ovvero le variazione nel determinante delle matrici di varianze/covarianze delle stime, eliminando l'i-esima osservazione; se lo scostamento a seguito dell'eliminazione di i risulta elevato, l'osservazione i viene considerata anomala. L'indice ha forma:

$$\boxed{\text{COVRATIO} = \frac{det(\sigma^2_{(-i)}(\mathbf{X}'_i\mathbf{X}_i)^{-1})}{det(\sigma^2(\mathbf{X}'_i\mathbf{X}_i)^{-1})}}$$

  e in genere utilizza il valore soglia $1 \pm 3(\frac{k+1}{n})^{\frac{1}{2}}$

- Un altro indice è il **DFfits**, il quale misura l'influenza dell'i-esima osservazione sulla stima dei coefficienti di regressione e sulla loro varianza, eliminandola dai dati. Se lo scostamento è elevato, i sarà un'osservazione anomala

$$\boxed{\text{DFFITS} = \frac{\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)}}{s_{\epsilon(i)}\sqrt{h_i}}}$$

- Il terzo indice è il **DFbetas**, il quale misura l'influenza dell'i-esima osservazione sulle stime di ogni coefficiente di regressione (separatamente), eliminado i dai dati. Uno scostamento elevato indica che i è potenzialmente un'osservazione anomala

$$\boxed{\text{DFBETAS} = \boldsymbol{\beta} - \boldsymbol{\beta}_{(-i)} = \mathbf{X}_{(-i)}(\mathbf{X}'\mathbf{X})^{-1}\frac{\epsilon_i}{1-h_i}}$$

  utilizzando 2 o $2 \cdot (n)^{\frac{1}{2}}$ come valori di soglia.

- L'ultimo indice è la **Distanza di Cook**, la quale misura l'influenza della i-esima osservazione sulla stima dei coefficienti di regressione, calcolata i termini di capacità del modello di predire i casi eliminando i dai dati. Un valore elevato indica che i è probabilmente un'osservazione influente. L'indice ha forma:

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

  un valore di D_i maggiore di 1 indica presenza di *otuliers*.

Modello lineare multivariato

Il modello lineare multivariato rappresenta un'estensione dei modelli classici; al posto una una singola variabile risposta ($\mathbf{y}$ vettore) avremo più variabili, ovvero una matrice $\mathbf{Y}$. Il modello assumerà quindi forma :

$$\boxed{\mathbf{Y} = \mathbf{BZ} + \mathbf{E}}$$

Un esempio di modello lineare multivariato potrebbe essere la relazione tra un cocktail di r medicine $\mathbf{Z}$ su un certo numero m di indicatori di salute $\mathbf{Y}$.

Partendo dal presupposto che gli errori, al variare del campione, rappresentano manifestazioni di variabili casuali identicamente distribuite, dove alla variabile risposta j viene associato il vettore degli errori $\mathbf{E}_j = (E_{1j}, ..., E_{ij}, ..., E_{nj})$, è possibile identificare una matrice $\hat{\mathbf{E}}^*$ $(m, n \cdot m)$ che ha sulla diagonale i vettori $E_1, ...E_j, ...E_m$ appena descritti. Ciascuno degli m vettori si riferirà ad una differente variabile risposta $\mathbf{y}_j$.

2.1 Ipotesi Classiche

Come nel modello classico, anche il multivariato si basa su una serie di **assunzioni o ipotesi**:

1. **Linearità** nei parametri e nelle variabili.
2. **Non sistematicità** degli errori (errori stocastici)$\rightarrow E(\hat{\mathbf{E}}_{\mathbf{ij}} = 0)$; $E(\hat{\mathbf{E}}) = \mathbf{0}$; $E(\hat{\mathbf{E}}^*) = \mathbf{0}$ da cui possiamo dedurre $E(\mathbf{Y}) = \mathbf{Z}\hat{\mathbf{B}}$
3. **Sfericità degli Errori**: la matrice di varianze e covarianze è data da

$$\Sigma_Y = \hat{\mathbf{B}}\mathbf{Z}\mathbf{Z}'\hat{\mathbf{B}}' + \mathrm{VAR}(\hat{\mathbf{E}}^*) = \hat{\mathbf{B}}\mathbf{Z}\mathbf{Z}'\hat{\mathbf{B}}' + E(\hat{\mathbf{E}}^*\hat{\mathbf{E}}^{*\prime})$$

La matrice di varianze e covarianze degli errori $\Sigma_{\hat{E}}$avrà forma:

$$\begin{bmatrix} \boldsymbol{\Sigma}_{E\hat{}1} \ldots\ldots \boldsymbol{\Sigma}_{E\hat{}1j} \ldots\ldots\ldots \boldsymbol{\Sigma}_{E\hat{}1m} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \boldsymbol{\Sigma}_{E\hat{}1j} \ldots\ldots \boldsymbol{\Sigma}_{E\hat{}j} \ldots\ldots\ldots \boldsymbol{\Sigma}_{E\hat{}jm} \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \boldsymbol{\Sigma}_{E\hat{}mj} \ldots\ldots \boldsymbol{\Sigma}_{E\hat{}mj} \ldots\ldots\ldots \boldsymbol{\Sigma}_{E\hat{}m} \end{bmatrix}$$

Nel caso più generico, ovvero con autocorrelazione e eteroschedasticità, ciascuna matrice $\Sigma_{\hat{\mathbf{E}}_j}$ avrà forma:

$$\begin{bmatrix} \sigma^2_1 & \rho_{12} & \rho_{1n} \\ \rho_{12} & \sigma^2_2 & \rho_{2n} \\ \\ \rho_{1n} & \rho_{2n} & \sigma^2_n \end{bmatrix}$$

Dove sulla diagonale troviamo la parte di varianza che il modello non riesce a spiegare per l'i-esima osservazione (σ_i^2). Le matrici $\Sigma_{\hat{\mathbf{E}}_{jl}}$(ovvero quelle non sulla diagonale di $\Sigma_{\hat{\mathbf{E}}}$) avranno forma:

$$\begin{bmatrix} \sigma^2_{1(a)} & \rho_{1(a,b)} & \rho_{2(a,n)} \\ \rho_{1(a,b)} & \sigma^2_{2(b)} & \rho_{3(b,n)} \\ \\ \rho_{2(a,n)} & \rho_{3(b,n)} & \sigma^2_{p(n)} \end{bmatrix}$$

le quali mostrano le correlazioni tra lo stesso individuo (a,b,...,n) rispetto a variabili dipendenti diverse , e tra la stessa variabile dipendente e individui diversi. Tale matrice mostrerà a sua volta eteroschedasticità e correlazione.

Detto questo l'ipotesi di sfericità nel modello multivariato assumerà il seguente significato:

- Gli errori sono **omoschedastici all'interno della stessa equazione**, ovvero per ogni individuo la parte spiegata all'interno della stessa variabile risposta è uguale (restrittiva).
- Gli errori sono **omoschedastici tra equazioni differenti** (molto restrittiva).
- Gli errori sono **incorrelati all'interno della stessa equazione**, ovvero il comportamento di ogni individuo, all'interno della stessa variabile dipendente, non è legato agli altri (molto restrittiva).
- Gli errori sono **incorrelati tra equazioni differenti** (nuova rispetto al modello classico), ovvero il comportamento di ogni individuo rispetto a diverse variabili non è legato al proprio e a quello degli altri.

La sfericità degli errori è tuttavia spesso troppo restrittiva, e di conseguenza non verificata.

4. Le variabili esplicative $\mathbf{Z}$ sono **non stocastiche**.
5. Le variabili esplicative $\mathbf{Z}$ sono **non collineari** (altrimenti $\mathbf{ZZ}'$ sarebbe non invertibile). Può avvenire tuttavia che vi sia quasi multicollinearità (matrice invertibile ma stime non robuste).
6. Spesso viene aggiunta un'ultima proprietà, utile per la costruzione dei test inferenziale, che è quella della **normalità degli errori**:

$$\mathbf{E} \sim N_m(0, \sigma^2 \mathbf{I}_{n \cdot m})$$

Tutte le ipotesi classiche vanno verificate affinché il modello possa essere applicato, oppure affinché si possa mettere in atto una strategia di correzione per ipotesi non verificate.

2.2 Stime, stimatori e test d'ipotesi

A partire dalla formulazione teorica del modello lineare multivariato si può passare alla definizione del modello campionario:

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

con $\hat{\mathbf{B}}$ e $\hat{\mathbf{E}}$ rispettivamente le stime OLS dei parametri e degli errori. La j-esima componente della matrice dei parametri stimati $\hat{\mathbf{B}}$ sarà data da:

$$\boxed{\beta_j = \mathbf{y}_j \mathbf{Z}'(\mathbf{ZZ}')^{-1}}$$

Gli stimatori OLS $\hat{\mathbf{B}}$ saranno:

- **Corretti** $\rightarrow E(\hat{\mathbf{B}} = \mathbf{B})$, o espresso in termini probabilistici: $P_{n\rightarrow\infty}(|\hat{\mathbf{B}} - \mathbf{B}| < \lambda) = 1$
- **A varianza minima** (efficienti): $\rightarrow E((\hat{\mathbf{B}} - \mathbf{B})'(\hat{\mathbf{B}} - \mathbf{B}) = \sigma^2(\mathbf{ZZ}')^{-1}$, per il Teorema di Gauss-Markov.
- **Consistenti**.

Gli stimatori ML e OLS (con normalità degli errori) sono UMVUE anche nel caso multivariato. Gli stimatori dei minimi quadrati, sotto ipotesi di distribuzione normale degli errori, avranno forma:

$$\hat{\mathbf{B}} \sim N(\mathbf{B}, \sigma^2(\mathbf{ZZ}')^{-1})$$

mentre le variabili risposta:

$$\mathbf{Y} \sim N(\mathbf{BZ}, \mathbf{\Sigma}_Y)$$

Poiché le soluzioni possono essere ricavate equazione per equazione, possiamo impiegare i test della Normale e della T per verificare la significatività dei singoli parametri, F per gruppi di parametri e significatività del modello.

Per la costruzione dei test, ove la varianza della popolazione non sia nota, è necessario ricavare la varianza del campione:

$$\boxed{\mathbf{\Sigma}_Y = (\frac{1}{n}\hat{\mathbf{B}}\mathbf{ZZ}'\hat{\mathbf{B}}') + \mathrm{VAR}(\hat{\mathbf{E}})}$$

Sotto le ipotesi classiche del modello, al variare del campione, $\hat{\mathbf{E}}$ si distribuirà come una Normale multivariata $\mathrm{N}(0, \frac{1}{n-r}\mathrm{VAR}(\hat{\mathbf{E}}))$. La *hat-matrix* ($\mathbf{H} = \hat{\mathbf{B}}\mathbf{ZZ}'\hat{\mathbf{B}}'$, la quale rappresenta la matrice di varianze covarianze spiegate campionaria), si distribuirà come una v.c Wishart con r gradi di libertà, e sarà inoltre indipendente da $\mathrm{VAR}(\hat{\mathbf{E}})$ (anch'essa distribuita come una W con $n - r$ gradi di libertà).

2.2.1 Varianza generalizzata e test di significatività di Wilks

Definiamo ora la **varianza generalizzata di Wilks** di $\hat{\mathbf{H}}$ e di $\mathrm{VAR}(\hat{\mathbf{E}})$ i rispettivi determinanti matriciali; tale indicatore viene impiegato nei modelli multivariati al fine di catturare sinteticamente la variabilità e la correlazione tra le variabili. Avrà valore 0 quando il rango della matrice di covarianze è minore di m; ciò avviene quando almeno una variabile è costante, oppure quando una variabile è correlata perfettamente (o combinazione lineare) ad un'altra. Il valore massime della varianza di Wilks di $\mathrm{VAR}(\hat{\mathbf{E}})$ sarà $\prod_j \sigma_j$ (il prodotto delle varianze, ovvero vi è incorrelazione).

Possiamo ora passare a definire il **test di Wilks**, il quale si serve della statistica test:

$$\boxed{\Lambda = \frac{|\mathrm{Var}(\hat{\mathbf{E}})|}{|\mathrm{Var}(\hat{\mathbf{E}}) + \hat{\mathbf{H}}|} = \prod_i (\frac{1}{(1+\lambda_i)^{-1}})}$$

ovvero il rapporto tra il determinante della matrice di varianze/covarianze dell'errore e il determinante di quella generale. Gli elementi $\lambda_1 > \lambda_2 > ... > \lambda_p$ rappresentano gli autovalori non nulli di $|\text{VAR}(\hat{\mathbf{E}})|^{-1}\hat{\mathbf{H}}$; Λ avrà distribuzione Lambda di Wilks con (n,m,r) gradi di libertà. Viene tuttavia spesso più comodo utilizzare l'**Approssimazione di Bartlett** data da:

$$\boxed{W = -v \log_e \Lambda} \quad v = \frac{n-1}{2 \cdot (m+r+1)}$$

la quale ha distribuzione asintotica $\chi^2_{m \cdot r}$.

Si può partire dal test di base, ovvero in cui nessuna variabile indipendente è significativa per spiegare nessuna variabile risposta ($H_0 : \hat{\mathbf{B}} = \mathbf{0}$), contro l'ipotesi alternativa ($H_1 : \beta_j \neq 0$) che almeno una variabile esplicativa sia significativa per almeno una variabile dipendente. Sotto ipotesi nulla ($\hat{\mathbf{B}} \to \mathbf{0}$, e quind $\hat{\mathbf{H}} \to 0$) Λ tenderà ad 1; poichè il rapporto $\frac{(1-\Lambda)}{\Lambda}$ si distribuisce asintoticamente come una F, quando Λ tende a 1 la F tenderà a 0. e quindi all'accettazione dell'ipotesi nulla di non significatività di tutti i parametri. D'altro canto, quanto più $\hat{\mathbf{H}}$ sarà grande, tanto più Λ tenderà a 0, e di conseguenza la F divergerà a $+\infty$, cadendo di conseguenza nella regione di rifiuto dell'ipotesi nulla di non significatività dei parametri. Per riassumere:

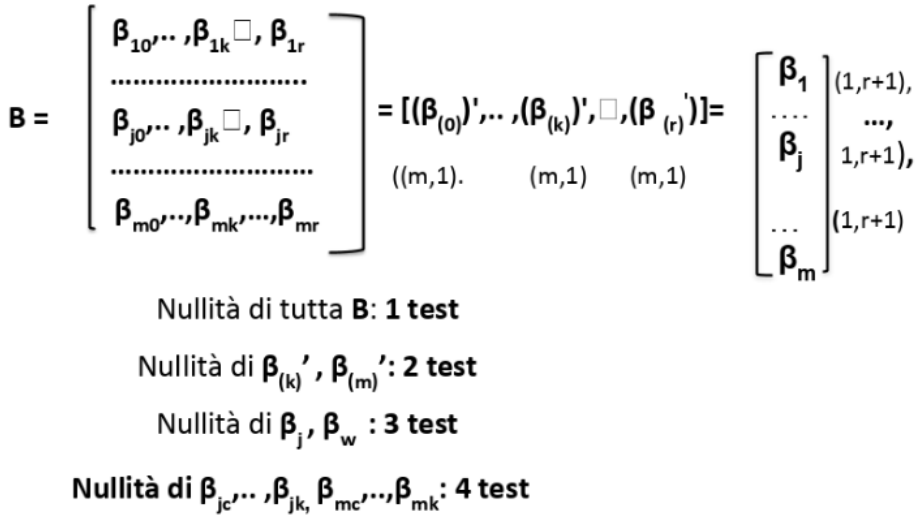
- Quando $\Lambda \to 1$, la distribuzione asintotica $F = \frac{1-\Lambda}{\Lambda} \to 0$ cade nella regione di accettazione e quindi sarà probabile l'accettazione di H_0. Ciò deriva dal fatto che, sotto H_0, i parametri $\hat{\mathbf{B}} \to \mathbf{0}$ e di conseguenza anche la matrice di varianza/covarianza spiegata $\hat{\mathbf{H}} \to 0$.

- Quando $\Lambda \to 0$, la distribuzione asintotica $F = \frac{1-\Lambda}{\Lambda} \to \infty$ cade nella regione di rifiuto e quindi sarà probabile l'accettazione di H_1. Ciò deriva dal fatto che, sotto H_0, i parametri $\hat{\mathbf{B}} \neq \mathbf{0}$ e di conseguenza la matrice di varianza/covarianza spiegata $\hat{\mathbf{H}} \neq \mathbf{0}$.

2.2.2 Test sulla significatività delle variabili esplicative Z

1. Il primo test è quello descritto nella precedente sezione (tutti i regressori rispetto a tutte le equazioni).
2. Definendo un certo gruppo di regressori come $\mathbf{Z}_1$, è possibile costruire un test di significatività di tale gruppo rispetto a tutte le variabili dipendenti $\mathbf{Y}$. L'ipotesi nulla sarà quindi $H_0 : \mathbf{B}_{(1)} = \mathbf{0}$, con $\mathbf{B}_{(1)}$ parametri dei regressori appartenenti al gruppo testato; l'ipotesi alternativa sarà che almeno un regressore è significativo per spiegare almeno una variabile risposta.
3. Si può inoltre capovolgere il test, ovvere prendendo tutte le variabili esplicative $\mathbf{Z}$ e testarne la significatività rispetto ad un determinato gruppo di variabili risposta $\mathbf{Y}_2$; l'ipotesi nulla sarà ora $H_0 : \mathbf{B}_{(2)} = \mathbf{0}$, con $\mathbf{B}_{(2)}$ gruppo di parametri dei regressori, riferiti alle variabili risposta del gruppo Y_2.
4. Il caso più generale è infine il test che riguarda la significatività di alcuni regressori $\mathbf{Z}_3$ nello spiegare un certo gruppo di variabili risposta $\mathbf{Y}_3$; l'ipotesi nulla assumerà forma $H_0 : \mathbf{B}_{(3)} = \mathbf{0}$,

Per riassumere i **quattro test sulla Nullità dei parametri** (considerando anche il test Λ, ovvero sulla significatività di tutti i regressori rispetto a tutte le risposte), viene fornito il seguente schema:



Per la costruzione di tutti i test è possibile rifarsi alla forma del test di Wilks precedentemente descritta; definito un certo gruppo g di regressori e variabili risposta, basterà sostituire:

- $\hat{\mathbf{H}}$ con $\hat{\mathbf{H}}_g$, ovvero matrice di varianze/covarianze spiegate riferita al gruppo g.
- VAR($\hat{\mathbf{E}}$) con VAR($\hat{\mathbf{E}}_g$), ovvero matrice di varianze/covarianze degli errori riferita al gruppo g.
- Λ con Λ_g (e di conseguenza anche la asintotica F con $F_g = \frac{1-\Lambda_g}{\Lambda_g}$)
- L'ipotesi nulla $H_0 : \ \hat{\mathbf{B}} = \mathbf{0}$ diventerà $H_0 : \ \hat{\mathbf{B}}_g = \mathbf{0}$

Lo sviluppo e l'interpretazione del test rimane analoga a quella del test sulla nullità di tutti i parametri rispetto a tutte le variabili risposta. Vi sono inoltre alcuni **test sull'Uguaglianza dei parametri**, che hanno analoga interpretazione

2.3 GLS per modelli lineari multivariati

Per la violazione di tutte le ipotesi si applicano soluzioni analoghe al modello lineare classico; l'unico caso che prevede tecniche differenti è la **violazione della sfericità degli errori**. Il concetto più stringente dell'ipotesi di sfericità degli errori è che considerando regressori uguali, rispetto alle diverse variabili risposta, la parte di varianza non spiegata dal modello (ovvero la parte dell'errore) risulta la medesima (ovvero VAR($\hat{E}$) $= \boldsymbol{\sigma}^2\mathbf{I}_n$, e quindi $\Sigma_Y = \mathbf{BZZ'B} + \boldsymbol{\sigma}^2\mathbf{I}_n$); come diretta conseguenza di tale ipotesi anche la parte spiegata dalle diverse equazioni del modello (ovvero rispetto alle diverse variabili risposta) è uguale.

Essendo molto irrealistica e quasi mai verificabile, vi è la necessità, come nel caso classico, di definire un metodo basato sui **Minimi Quadrati Generalizzati** per ovviare a tale problematica. Partendo dalla formulazione generale, il modello GLS avrà forma:

$$\textbf{GLS} \rightarrow \boxed{\mathbf{Y}^{\circ} = \mathbf{B}^{\circ}\mathbf{Z}^{\circ} + \mathbf{E}^{\circ}}$$

La matrice di varianze/covarianze degli errori ($\mathbf{\Sigma_E} = E(\mathbf{E}^{\circ *\prime}\mathbf{E}^{\circ *})$) avrà ora forma (come già mostrato nella sezione delle ipotesi classiche 2.1):

$$\underset{(nm,\,m)(m,\,mn)}{E(\,\mathbf{E}^{\circ *\prime}\mathbf{E}^{\circ *}\,)} = \underset{(nm,\,nm)}{\mathbf{\Sigma}_{\mathbf{E}^{*}*}} = \begin{bmatrix} \underset{(n,1)}{\mathbf{E}_1^{*\prime}} & & \\ \cdots & & \\ & \mathbf{E}_j^{*\prime} & \\ & & \mathbf{E}_m^{*\prime} \end{bmatrix} \begin{bmatrix} \underset{(1,n)}{\mathbf{E}_1^{*}} & & \\ \cdots & & \\ \cdots & \mathbf{E}_j^{*} & \\ & & \mathbf{E}_m^{*} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{\Sigma}_{E^*1} \cdots \mathbf{\Sigma}_{E*1j} \cdots & \mathbf{\Sigma}_{E^*1m} \\ \cdots & \\ \mathbf{\Sigma}_{E*1j} \cdots \mathbf{\Sigma}_{E*j} \cdots & \mathbf{\Sigma}_{E^*jm} \\ \cdots & \\ \mathbf{\Sigma}_{E*mj} \cdots \mathbf{\Sigma}_{E*mj} \cdots & \mathbf{\Sigma}_{E*m} \end{bmatrix}$$

Come soluzione al problema si applica una trasformazione agli errori mediante l'utilizzo dell'ipervettore $\mathbf{W}$, in maniera da ottenere errori sferici $\mathbf{E}^{\circ} = \mathbf{E}^{\circ *}(\mathbf{W})^{-1}$. A questo punto, essendo $\mathbf{W}$ matrice non stocastica, la matrice di varianze e covarianze degli errori diventerà:

$$\boxed{\Sigma_E = (\mathbf{W})^{-1\prime} E(\mathbf{E}^{\circ *\prime}\mathbf{E}^{\circ *})(\mathbf{W})^{-1} = \sigma^2 \mathbf{I}_{nm}}$$

ottenendo di conseguenza errori sferici (incorrelati ed eteroschedastici).

Conseguentemente alla trasformazione degli errori, anche le altre componenti del modello subiranno delle modifiche, e il modello assumerà forma:

$$\boxed{\mathbf{Y}^{\circ}(\mathbf{W})^{-1} = \mathbf{Y}^{\circ *} = \mathbf{B}^{\circ *}\mathbf{Z}^{\circ}(\mathbf{W})^{-1} + \mathbf{E}^{\circ}(\mathbf{W})^{-1} = \mathbf{B}^{\circ *}\mathbf{Z}^{\circ *} + \mathbf{E}}$$

con ipervettore dei parametri:

$$\boxed{\mathbf{B}^{\circ *} = \mathbf{Y}^{\circ *}\mathbf{Z}^{\circ *\prime}(\mathbf{Z}^{\circ *}\mathbf{Z}^{\circ *\prime})^{-1} = \mathbf{Y}^{\circ}\mathbf{\Sigma}_E^{-1}\mathbf{Z}^{\circ\prime}(\mathbf{Z}^{\circ}\mathbf{\Sigma}_E^{-1}\mathbf{Z}^{\circ\prime})^{-1}}$$

Tale soluzione consente il rispetto delle ipotesi, come nel modello classico, con la differenza che i GLS per modelli multivariati ammettono il diverso comportamento degli individui rispetto a variabili dipendenti diverse, pur avendo le stesse caratteristiche; il modello trasformato sarà di conseguenza meno rigido di un modello non trasformato quando quest'ultimo rispetti l'ipotesi di sfericità (poiché, a parità di caratteristiche, non ammette il diverso comportamento degli individui rispetto a variabili dipendenti differenti).

Definito il comportamento del modello quando viene applicata una trasformazione, è necessario definire le caratteristiche della matrice $\mathbf{W}$ che consentano di ottenere errori sferici. Risulta possibile definirla in termini di decomposizione spettrale della matrice $\mathbf{\Sigma}_{E^{\circ *}}$:

$$\boxed{\mathbf{W} = \sigma \mathbf{K} \mathbf{L}^{\frac{1}{2}} \mathbf{K}' = \frac{1}{\sigma} \mathbf{K}^{-1} \mathbf{L}^{-\frac{1}{2}} \mathbf{K}^{-1}}$$

con $\mathbf{K}$ matrice degli autovettori (ortogonale$\rightarrow \mathbf{K}\mathbf{K}' = \mathbf{I}$) e L matrice diagonale degli autovalori di $\mathbf{\Sigma}_{E^*}$. Varrà infatti la relazione $\sigma \mathbf{K} \mathbf{L}^{\frac{1}{2}} \mathbf{K}' \sigma \mathbf{K} \mathbf{L}^{\frac{1}{2}} \mathbf{K}' = \mathbf{W}\mathbf{W}' = \mathbf{\Sigma}_{E^*}$.

Come nel modello lineare classico bisogna considerare il caso, estremamente realistico, della non conoscenza della matrice di varianze e covarianze $\mathbf{\Sigma}_{E^*}$; in tal caso si impiega la matrice campionaria $\mathbf{S}_{E^*}$ (la quale tende alla matrice della popolazione all'allargarsi del campione), utilizzando in tal modo il metodo **Feasible Generalized Least Squares** (FGLS). L'ipervettore dei parametri sarà ora:

$$\boxed{\mathbf{B}^{\circ *} = \mathbf{Y}^{\circ *} \mathbf{Z}^{\circ *\prime} (\mathbf{Z}^{\circ *} \mathbf{Z}^{\circ *\prime})^{-1} = \mathbf{Y}^{\circ} \boldsymbol{S}_E^{-1} \mathbf{Z}^{\circ \prime} (\mathbf{Z}^{\circ} \boldsymbol{S}_E^{-1} \mathbf{Z}^{\circ \prime})^{-1}}$$

con ciascuna equazione j pari a:

$$\boxed{\hat{\beta}_j^* = \mathbf{y}_j^{\circ} \boldsymbol{S}_E^{-1} \mathbf{Z}^{\circ \prime} (\mathbf{Z}^{\circ} \boldsymbol{S}_E^{-1} \mathbf{Z}^{\circ \prime})^{-1}}$$

2.4 Modelli *Seemingly Uncorrelated Regression Equations* (SURE)

Quando si hanno a disposizione diverse variabili risposta, non è sempre utile impiegare gli stessi regressori per costruire le diverse equazioni; se per esempio la propensione al mangiare può essere un ottimo regressore per spiegare la spesa per il cibo, potrebbe invece risultare inutile nello spiegare la spesa per la casa o per la macchina. E' possibile tuttavia risolvere tale inconsistenza attraverso l'impiego di diverse variabili esplicative per le diverse variabili risposta, come mostrato dall'esempio sottostante:

$$\mathbf{y_1} = \beta_{10} + \beta_{11}\mathbf{z_1} + \beta_{1m}\mathbf{z_m} + \boldsymbol{\varepsilon}_1$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\mathbf{y_j} = \beta_{j0} + \beta_{j2}\mathbf{z_2} + \ldots + \beta_{jr}\mathbf{z}_r + \boldsymbol{\varepsilon_j}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\mathbf{y_m} = \beta_{m0} + \beta_{mm}\mathbf{z_m} + \ldots + \beta_{mr}\mathbf{z_r} + \boldsymbol{\varepsilon_m}$$

Tra le principali differenze dal caso classico troviamo che le diverse equazioni avranno gruppo di regressori di numerosità variabile. I modelli **Seemingly Uncorrelated Regression Equations** (SURE) sono un esempio di tale situazione.

Nonostante l'impiego di regressori differenti possa essere applicato a situazioni con matrici di varianze e covarianze differenti (eteroschedastici correlati, omoschedastici correlati, ...), il modello SURE vero e proprio ha matrice di varianze e covarianze con:

- **Errori omoschedastici all'interno della stessa equazione**, ovvero tutti gli individui hanno la stessa parte non spiegata dal modello, riguardo la stessa variabile esplicativa $(\sigma_1^2, ..., \sigma_n^2)$.
- Ogni individuo è **correlato con se stesso anche rispetto ad equazioni differenti** (per esempio la spesa per i viaggi e la spesa per gli alimentari di uno stesso individuo sono legate).
- Ogni individuo è **incorrelato dagli altri individui**, sia all'interno della stessa equazione, sia tra equazioni differenti.
- La forma della matrice di varianze e covarianze $\mathbf{\Sigma}_E^\circ$ (matrice $nm \cdot nm$) avrà la seguente struttura (quasi-incorrelata):

$$\begin{bmatrix} \sigma^2_1 \mathbf{I}_n & \sigma^2_{12} \mathbf{I}_n & \cdots & \cdots & \sigma^2_{1m} \mathbf{I}_n \\ \sigma^2_{12} \mathbf{I}_n & & \cdots & \cdots & \sigma^2_{2m} \mathbf{I}_n \\ \cdots & \cdots & \cdot & \cdot & \cdots \\ \cdots & \cdots & \cdots & \cdot & \cdots \\ \sigma^2_{1m} \mathbf{I}_n & \sigma^2_{2m} \mathbf{I}_n & \cdots & \cdots & \sigma^2_m \mathbf{I}_n \end{bmatrix}$$

La soluzione per la stima dei parametri, poiché non abbiamo errori sferici, avverrà nuovamente attraverso i GLS:

$$\boxed{\mathbf{B}^\circ = \mathbf{Y}^\circ \mathbf{Z}^{\circ\prime} (\mathbf{Z}^\circ \mathbf{Z}^{\circ *\prime})^{-1} = \mathbf{Y}^\circ \mathbf{\Sigma}_{E^\circ}^{-1} \mathbf{Z}^{\circ\prime} (\mathbf{Z}^\circ \mathbf{\Sigma}_{E^\circ}^{-1} \mathbf{Z}^{\circ\prime})^{-1}}$$

con $\mathbf{B}^{\circ *}$ di dimensione $(1, \sum_j r)$, $\mathbf{Y}^{\circ *}$ $(1, nm)$, $\mathbf{\Sigma}_{E^\circ}$ (nm, nm) e $\mathbf{Z}$ $(nm, \sum_j r)$

2.5 Scelta del modello

Per scegliere il miglior modello in termini di adattamento ai dati bisogna seguire tali criteri:

1. Valutare la sfericità degli errori attraverso la matrice di varianze/covarianze e gli appositi test statistici.
2. Valutare la significatività singola delle variabili esplicative, in particolar modo per il modello SURE e GLS.
3. Valutare l'adattamento ai dati e la parsimonia $(R^2, R^2_{\text{adj}}, F, ...)$.
4. Impiego di intuizione, creatività, esperienza, conoscenza di dominio...

Modello di regressione multilivello

La **Regressione Multilevel** si principalmente basa su due concetti:

- **Analisi contestuale** (*contextual analysis*), ovvero lo sviluppo delle scienze sociali inerenti agli effetti del contesto sociale sul comportamento individuale.
- **Modelli ad effetti misti** (*mixed effects models*), ovvero modelli costruiti su parametri fissi e parametri casuali.

Nei modelli lineari classici infatti si ignorano le **strutture gerarchiche dei dati**, dal momento che si estraggono campione casuali da una popolazione, assumendo l'indipendenza e l'identica distribuzione di tutte le osservazioni. I modelli multilivello ammettono invece l'idea che alcune caratteristiche possano non essere indipendenti ma appartenenti ad un gruppo (classe-studenti, famiglie-figli, ...), avendo quindi a disposizione dati nidificati/gerarchici. Quando vengono ignorate delle strutture gerarchiche all'interno delle variabili esplicativi si rischia di generare forti distorsioni.

Un esempio esplicativo riguarda una serie di test di un certo farmaco, il quale viene somministrato a due gruppi di persone divisi per età (giovani-anziani); molto probabilmente l'analisi giungerà alla conclusione che il farmaco ha maggiore effetto sulle persone giovani, risultato distorto poiché non viene considerata la maggiore resistenza di tale gruppo rispetto agli anziani (ovvero non è stata applicata un'analisi contestuale).

Uno dei principali errori a cui può portare all'errata considerazione del fenomeno in esame viene causata dall'**aggregazione**, ovvero quando si mettono insieme grandi quantità di dati e si considerano le caratteristiche macroscopiche, dimenticandosi di tenere conto delle peculiarità delle singole (o di gruppi più piccoli) osservazioni. Quando non si presta attenzione a tale condizione si rischia di incorrere nella **fallacia ecologica**, ovvero l'errore di utilizzare le correlazioni di livello macro per fare asserzioni a livello micro. E' possibile inoltre commettere l'errore opposto, ovvero applicare una **disaggregazione** e utilizzare i risultati ottenuti a livello micro per trarre conclusioni a livello macro, commettendo quindi una **fallacia atomistica**.

Avendo quindi analizzato l'esistenza di gruppi con alcune peculiarità all'interno della popolazione, un campionamento casuale risulta alquanto distorcente; tale metodo infatti si serve dell'assunzione di indipendenza e identica distribuzione (affinché tutte le osservazioni abbiano la stessa probabilità di estrazione), condizioni che non si verificano in caso di strutture gerarchiche/nidificate. Il **campionamento** avviene quindi a **diversi stadi**, tenendo in tale modo conto della diversa distribuzione tra osservazioni di gruppi diversi (i quali sono indipendenti) e della dipendenza tra

osservazioni all'interno dello stesso gruppo (le quali sono identicamente distribuite); il risultato finale dev'essere un metodo che estragga con la stessa probabilità ogni osservazione, mitigando di conseguenza l'effetto gerarchico.

Il modello multilevel avrà quindi un doppio indice, uno riferito all'osservazione e uno al gruppo:

$$\boxed{y_{ij} = b_0 + b_1 \cdot x_{ij} + e_{ij}}$$

con $(i = 1, ..., n_j)$, $(j = 1, ..., p)$ e. $\sum_{j=1}^{p} n_j = n$.

E' possibile definire diverse **regressioni multilevel**, distinte in base alla relazione che considerano:

- Regressioni basate su **relazioni disaggregate**, nel quale i raggruppamenti delle unità vengono ignorati.

- Regressioni basate **su relazioni aggregate tra gruppi**, ovvero regressioni basate sulle medie, le quali ignorano le singole unità all'interno dei gruppi; la struttura sarà quindi data da:

  $$\boxed{\mu(y_j) = \beta_0^* + \beta_1^* \cdot \mu(x_j) + \epsilon_j^*}$$

  con parametri pari a: $\hat{\beta}_1^* = \frac{\text{COV}(\mu(x_j), \mu(y_j))}{\sigma^2 \mu(x_j)}$; $\hat{\beta}_0^* = \mu(\mathbf{y}) - \hat{\beta}_1^* \cdot \mu(\mathbf{x})$

- Regressioni basate su **relazioni entro ciascun gruppo**, le quali utilizzano le differenze tra le caratteristiche delle osservazioni e le medie dei gruppi. Il modello assumerà quindi la seguente forma

  $$\boxed{y_{ij} - \mu(y_j) = \beta_1^\circ (x_{ij} - \mu(x_j)) + \epsilon_j^\circ)}$$

  con parametri pari a: $\hat{\beta}_1^\circ = \frac{\text{COV}(x_{ij} - \mu(x_j), y_{ij} - \mu(y_j))}{\sigma^2(\mathbf{x})}$; $\hat{\beta}_0^\circ = \mu(y_{ij} - \mu(y_j)) - -\beta_1^\circ \cdot \mu(x_{ij} - \mu(x_j)) = 0$

- Regressioni basate su **relazioni multilevel** (modello di Cronbach), nel quale vengono catturate sia le relazioni entro i gruppi, sia le relazione fra gruppi diversi; avrà forma:

  $$\boxed{y_{ij} = \beta_0^+ + \beta_1^+ (x_{ij} - \mu(x_j)) + \beta_2^+ \mu(x_j) + \epsilon_j^+}$$

  con β_1^+ parametro relativo alla variabilità interna al gruppo (*within parameter*) e β_2^+ parametro relativo ai risultati medi dei gruppi (*between parameter*); il primo fornisce informazioni circa le differenze individuali all'interno dei gruppi, il secondo invece sull'effetto delle medie dei gruppi sulla variabile dipendente. La differenza tra il parametro *between* e il parametro *within* viene definito *contextual effect* ($\delta = \beta_{\text{wt}} - \beta_{\text{bw}}$), il quale rappresenta l'effetto della media dei gruppi sulla variabile dipendente, al netto delle differenze all'interno dei gruppi (ovvero al netto delle caratteristiche individuali).

Anche se fin'ora tutti gli esempi sono stati informa univariata, è possibile estendere i concetti ai casi multiplo e multivariato. Nel caso multiplo per esempio avremo un modello con la seguente formulazione:

$$\boxed{y_{ij} = \beta_0 + \sum_k (\beta_{1,k}(x_{ij} - \mu(x_{jk}))) + \beta_{2,k}\mu(x_{jk}) + \epsilon_{ij}}$$

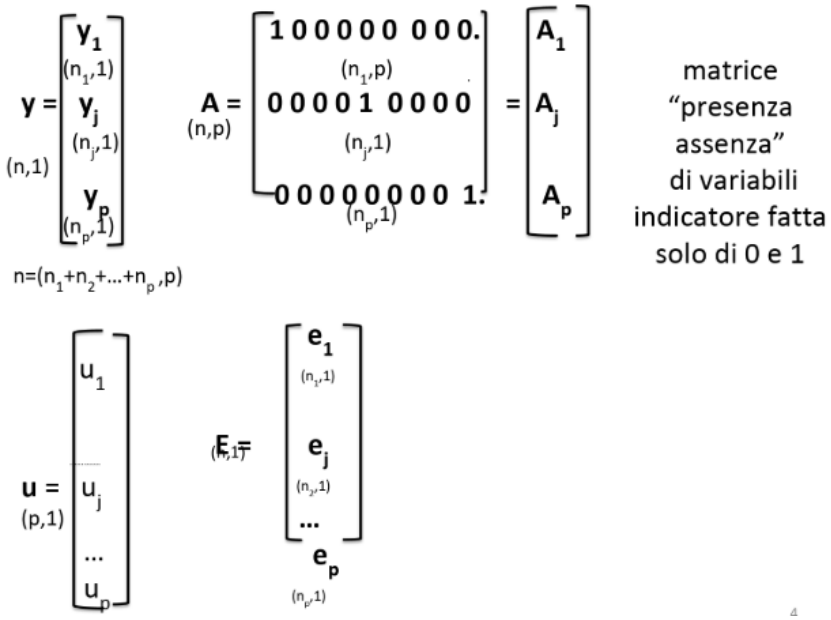
con $\beta_{1,k}$ e $\beta_{2,k}$ rispettivamente i coefficienti *within* e *between* per il k-esimo attributo. E' possibile riscrivere tale equazione in forma matriciale $\mathbf{y} = \mathbf{XB} + \mathbf{E}$, con $\mathbf{X}$ di dimensioni $n, 2 \cdot r + 1$ (1 colonna di valori unitaria e 2 colonne per ogni attributo, sia per i valori medi che per le differenze), $\mathbf{B}$ di dimensioni $2 \cdot r + 1, 1$ (una colonna per il termine noto, due colonne per β_1 e β_2 per ogni attributo). Ovviamente la soluzione OLS per modelli lineari multilevel è analoga ai precedenti casi ($\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$), e sia i parametri sia il modello possono essere soggetti ai test inferenziali e di verifica delle ipotesi.

3.1 Analisi della varianza (ANOVA)

Per comprendere il modello multilevel è necessario definire cosa si intende per analisi della varianza, iniziando a fornirla nella sua forma descrittiva; è possibile definire

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

dove γ_{00} rappresenta la media di $\mathbf{y}$ su tutta la popolazione, $\gamma_{00} + u_j$ la media relativa al gruppo j-esimo (u_j differenza tra la media della popolazione e la media del gruppo) e infine e_{ij} il residuo dell'osservazione i appartenente al gruppo j. In tale ottica l'analisi della varianza cerca di misurare la variabilità della $\mathbf{y}$ (variabile risposta), dovuta alle differenze nelle medie dei gruppi (*between*) e alla fluttuazione individuale (*within*). Possiamo ridefinire le componenti della precedente relazione in forma matriciale:



Definite tali matrici risulta ora possibile ridefinire la forma descrittiva della variabilità, precedentemente mostrata per le singole unità, in forma vettoriale:

$$\boxed{\mathbf{y}_j - \boldsymbol{\gamma}_{00} = \mathbf{A}_j\mathbf{u}_j + \mathbf{e}_j}$$

con $\mathbf{y}_j - \boldsymbol{\gamma}_{00}$ gli scarti tra le medie del gruppo j e quella della popolazione, $\mathbf{A}_j\mathbf{u}_j$ uguale a $\mathbf{u}_j$, a causa della forma di $\mathbf{A}$, e infine $\mathbf{e}_j$ vettore dei residui del gruppo j. L'intero modello andrà infine espresso in forma matriciale:

$$\boxed{\mathbf{y} - \boldsymbol{\gamma}_{00} = \mathbf{A}\mathbf{u} + \mathbf{e}}$$

il quale rappresenta un modello lineare in cui $\mathbf{u}$ rappresenta i parametri e $\mathbf{A}$ rappresenta le variabili esplicative. Dalla forma matriciale ricaviamo la devianza totale:

$$\boxed{\text{SST} = (\mathbf{y} - \boldsymbol{\gamma}_{00})'(\mathbf{y} - \boldsymbol{\gamma}_{00}) = (\mathbf{A}\mathbf{u} - \mathbf{e})'(\mathbf{A}\mathbf{u} - \mathbf{e}) = \mathbf{u}'\mathbf{A}'\mathbf{A}\mathbf{u} + \mathbf{e}'\mathbf{e}}$$

con $\mathbf{u}'\mathbf{A}'\mathbf{A}\mathbf{u}$ =SSF somma della devianza fra gruppi e $\mathbf{e}'\mathbf{e}$ =SSE devianza nei gruppi; dividendo per la numerosità n si ottengono rispettivamente le varianze τ^2 e (sotto ipotesi di omoschedasticità) σ^2. Il coefficiente di **correlazione intraclasse** sarà infine:

$$\boxed{\rho = \frac{\tau^2}{\tau^2 + \sigma^2} = \frac{\text{varianza tra i gruppi}}{\text{varianza complessiva}}}$$

il quale rappresenta la quota di varianza dovuta ai gruppi, ovvero l'intensità della variabilità dovuta all'appartenenza ai diversi gruppi (indice di separazione dei gruppi).

3.2 Analisi della covarianza (ANCOVA)

Quando si effettua un'analisi della varianza bisogna individuare alcune caratteristiche della popolazione $\mathbf{X}$, e depurare dai possibili effetti che potrebbero avere sulla variabile dipendente. A tale scopo viene impiegato l'**analisi della covarianza**, la quale si basa sul modello:

$$\boxed{y_{ij} = (\beta_0 + \sum_k \beta_{kj} \cdot x_{ijk}) + \gamma_{00} + v_j + e_{ij}}$$

da cui possiamo ricavare il modello di analisi della varianza al netto del modello lineare (che cattura le caratteristiche della popolazione):

$$\boxed{y_{ij} - (\beta_0 + \sum_k \beta_{kj} \cdot x_{ijk}) = \gamma_{00} + v_j + e_{ij}}$$

oppure in termini di modello multiplo nella forma matriciale:

$$\boxed{(\mathbf{y} - \boldsymbol{\gamma}_{00}) - \mathbf{X}\mathbf{b} = \mathbf{A}\mathbf{v} + \mathbf{e} = \mathbf{y}_e}$$

con $\mathbf{y}_e$ uguale alla parte non spiegata (residua) dal modello lineare, ovvero non spiegata dalle caratteristiche della popolazione. E' possibile, come nel modello lineare, definire gli stimatori OLS:

$$\boxed{\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \boldsymbol{\gamma}_{00})}$$

con matrice di varianze/covarianze (ovvero la parte di variabilità spiegata dal modello) pari a $\boldsymbol{\Sigma}_{\hat{b}} = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}$.

La matrice $\mathbf{A}$, poiché ha colonne costruite sulla base delle altre, e quindi non indipendenti, non avrà rango pieno e non sarà invertibile Nell'ipotesi migliore, in cui Rango($\mathbf{A}$) $= p-1$ (minore di un unità rispetto al rango pieno), è possibile porre un vincolo annullando una componente v_j ($\sum_j v_j = 0$) e eliminare una componente da $\mathbf{A}$ in maniera da avere una matrice $(n, p-1)$; è possibile porre vincoli e ridurre la dimensionalità fino al raggiungimento del rango pieno. Definendo $\mathbf{A}^\circ$ e $\mathbf{v}^\circ$ le nuove componenti del modello ($\mathbf{y}_e^\circ = \mathbf{A}^\circ\mathbf{v}^\circ + \mathbf{e}$) ridotte, grazie all'invertibilità raggiunta è possibile applicare la soluzione OLS a $\mathbf{v}$:

$$\boxed{\hat{\mathbf{v}}^\circ = (\mathbf{A}^{\circ\prime}\mathbf{A}^\circ)^{-1}\mathbf{A}^{\circ\prime}\hat{\mathbf{w}}}$$

con $\hat{\mathbf{v}}^\circ$ stima OLS ottenuta ponendo i vincoli ai parametri. Risulta ora possibile definire il **modello di varianze-covarianze**:

$$\boxed{\text{SST} = (\mathbf{y} - \boldsymbol{\gamma}_{00})'(\mathbf{y} - \boldsymbol{\gamma}_{00}) = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}'\mathbf{X} + \hat{\mathbf{v}}^{\circ\prime}\mathbf{A}^{\circ\prime}\mathbf{A}^\circ\hat{\mathbf{v}}^\circ + \mathbf{e}'\mathbf{e}}$$

con SSV $= \hat{\mathbf{v}}^{\circ\prime}\mathbf{A}^{\circ\prime}\mathbf{A}^\circ\hat{\mathbf{v}}^\circ$ devianza spiegata (fra i gruppi/*between*) dell'analisi della covarianza, SSE $= \mathbf{e}'\mathbf{e}$ la devianza residua (o nei gruppi/*within*) dell'analisi della covarianza e SSX $= \hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}'\mathbf{X}$ devianza della regressione; varrà allora:

$$\boxed{\text{SST} - \text{SSX} = \text{SSV} + \text{SSE} = SSY_e}$$

ovvero la devianza residua di regressione (non spiegata da SSX). Sotto l'ipotesi di omoschedasticità dei residui avrò che $\frac{\text{SSV}}{n} = \tau^2$ e $\frac{\text{SSE}}{n} = \sigma^2$ è possibile definire nuovamente un **coefficiente di correlazione interclasse**, il quale misura la quota della varianza totale (SST) dovuta ai gruppi, al netto delle caratteristiche della popolazione (SST-SSX=SSY$_e$):

$$\boxed{\rho^2 = \frac{\tau^2}{\tau^2 \cdot \sigma^2}}$$

il quale sarà tanto più elevato, quanto più sarà forte l'appartenenza ai gruppi al netto delle caratteristiche $\mathbf{X}$.

3.2.1 Analisi della covarianza campionaria

Dal momento che spesso risulta difficile definire modelli utilizzando come osservazione l'intera popolazione, tutti i procedimenti fin'ora definiti vanno riconsiderati in termini di modello campionario $(\mathbf{y} - \gamma_{00}) - \mathbf{X}\hat{\mathbf{B}} = \mathbf{A}^\circ\hat{\mathbf{v}}^\circ + \hat{\boldsymbol{\epsilon}} = \hat{\mathbf{y}}_e$ con $\hat{\boldsymbol{\epsilon}}_j$ che al

variare del campione definisce la variabile casuale E_j con valore atteso nullo e (sotto ipotesi di omoschedasticità) varianza uguale (E$\sim N(0, \sigma^2)$). Come diretta conseguenza anche $\mathbf{Y} = \mathbf{Y}_1, ..., \mathbf{Y}_n$ avranno distribuzione $\mathbf{Y}_i \sim N(\boldsymbol{\gamma}_{00} + \mathbf{u}_j, \sigma_y^2)$, tutte mutualmente indipendenti. Sotto tali ipotesi di normalità, rapportando le devianze SST, SSX, SSV E SSE a σ^2, otterremo delle distirbuzioni χ^2:

- SST$\sim \chi^2_{(n-1)}$
- SSV$\sim \chi^2_{(p-1)}$
- $\dfrac{\text{SSE}}{\sigma^2} \sim \chi^2_{(n-p-r)}$
- $\dfrac{\text{SSX}}{\sigma^2} \sim \chi^2_{(r)}$
- $\dfrac{\text{SSY}_e}{\sigma^2} \sim \chi^2_{(n-r-1)}$

Dal momento che $\mathbf{X}\hat{\mathbf{B}}$, $\mathbf{A}^\circ\hat{\mathbf{v}}^\circ$ e $\hat{\boldsymbol{\epsilon}}$ (componenti di SSX, SSV e SSE) sono incorrelate, al variare del campione, le devianze SSX, SSV e SSE saranno indipendenti tra loro; è possibile perciò definire, attraverso i rapporti di χ^2 indipendenti, delle F di Snedecor:

- $\dfrac{SSX}{r} / \dfrac{SSY_e}{n-r-1} \sim F_{(r\,;\,n-1-r)}$
- $\dfrac{SSV}{p-1} / \dfrac{SSE}{n-p-r} \sim F_{(p-1\,;\,n-p-r)}$

le quali possono essere impiegate per la costruzione di **test statistici F**. Prendendo in esame il primo test, sotto H_0 che i regressori $\mathbf{X}$ non spieghino la variabilità della risposta $\mathbf{Y}$, al crescere di F si cadrà nella regione di rifiuto. Analogamente ma con interpretazione opposta si definisce il secondo test il quale, sotto H_0 che i gruppi non spiegano la variabilità di $\mathbf{Y}$, al crescere della F cade nella regione di rifiuto.

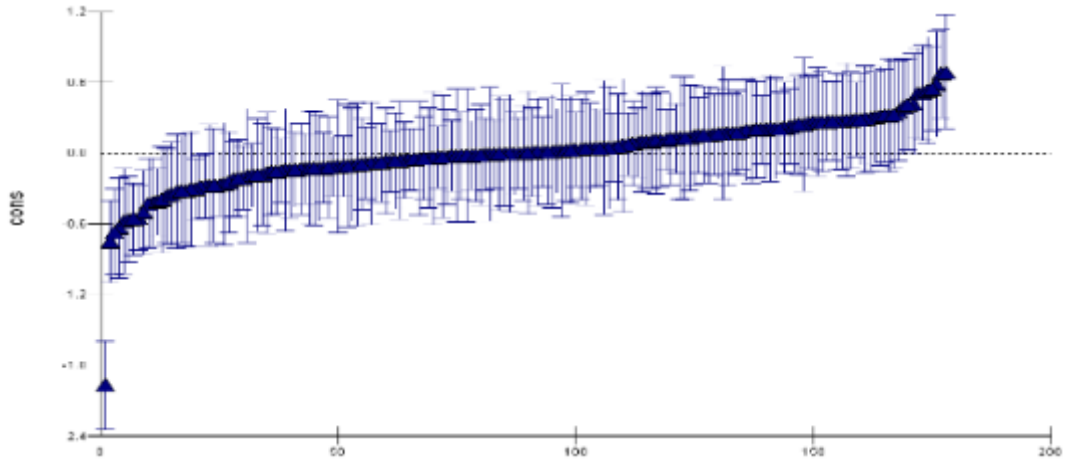
3.2.2 Analisi della covarianza casuale (Ancova ad effetti casuali)

Fino ad ora abbiamo considerato $\mathbf{v}$, ovvero la differenza tra la media del gruppo e la media della popolazione, come una componente fissa/determinata; tale assunzione non è tuttavia realistica poiché in ogni gruppo $j = 1, ...p$, al variare del campione, cambierà la componente $\mathbf{v}_j$ e sarà determinazione di una variabile casuale $\mathbf{V}_j$, la quale avrà media nulla (poiché la fluttuazione è 0), e come varianza (sotto ipotesi di omoschedaticità) τ^2. Il motivo della nuova struttura casuale del problema dipende dal secondo stadio del campionamento (estrazione dai gruppi), e in particolare dalla dipendenza dei parametri $\mathbf{v}_j$ (e quindi della variabile casuale di cui sono manifestazione) dal gruppo j cui fanno riferimento. Riprendendo la forma proposta nella sezione precedente, con l'aggiunta della casualità di $\mathbf{v}$, il modello avrà struttura identica:

$$\boxed{y_{ij} - (\beta_0 + \sum_k \beta_{kj} \cdot x_{ijk}) = \gamma_{00} + v_j + \epsilon_{ij} = \mathbf{y}_{eij}}$$

con ϵ_{ij} residuo casuale relativo all'unità i dal gruppo j, manifestazione anch'essa di una variabile casuale $\mathbf{E}_{ij}$ ($i = 1, ..., n$; $j = 1, ..., p$, $\mathbf{E}_{ij}$ identicamente distribuite a $\mathbf{E}_j$ al variare di i) con media 0 e varianze σ^2.

Poiché a questo punto $\mathbf{v}_j$ è un risultato campionario, non valido per l'intera popolazione, vi è la necessità di costruire intervalli di confidenza che, ad un certo livello $1-\alpha$ (0.9, 0.95, 0.99), contengano i valori ignoti $\mathbf{V}_j$; per fare ciò risulta estremamente comodo assumere una distribuzione normale. Nell'immagine sottostante viene mostrato un esempio di come appaiono gli intervalli di confidenza per una serie di test svolti in diversi ospedali; gli intervalli hanno ampiezza differente e sono posizionati secondo un *ranking* crescente del valore osservato; ovviamente ad intervalli di confidenza più piccoli, a parità di α, saranno associate stime più affidabili. La costruzione di intervalli di confidenza fa sì che vengano probabilizzati i valori reali della popolazione.



Ovviamente vi è incertezza riguardo al *ranking* reale: si potrà affermare, con confidenza $1-\alpha$, che un risultato è migliore di un altro solo se il limite inferiore del suo intervallo è maggiore del limite superiore dell'intervallo dell'altro.

Poiché in precedenza è stata mostrata l'incorrelazione tra $\mathbf{v}_j$ e $\boldsymbol{\epsilon}_j$ (derivante da $\text{cor}(\mathbf{A}^{\circ}\hat{\mathbf{v}}^{\circ}, \hat{\boldsymbol{\epsilon}}) = 0$), e assumendo la distribuzione normale, le variabili casuali $\mathbf{V}_j$ e $\mathbf{E}_j$ saranno incorrelate e indipendenti.

Per riassumere le precedenti sezioni:

- **Analisi della Varianza**: serve a determinare se la variabilità viene spiegata meglio dalla varianza tra gruppi (*between*), rispetto a quella nei gruppi (*within*).
- **Analisi della Covarianza ad effetti fissi**: fornisce la stessa informazione, ma depurando dagli effetti dovuti alle caratteristiche individuali.
- **Analisi della Covarianza ad effetti casuali**: fornisce gli stessi risultati ma con informazioni sui singoli gruppi (probabilizzando i risultati mediante impiego di intervalli di confidenza), consentendo entro certi limiti di stilare un *ranking* dei gruppi.

3.3 Modello multilevel

Definite le analisi di varianza e covarianza è ora possibile formulare il modello multilevel; il termine multilivello si riferisce alla duplice realtà (*mixed model*) che tali modelli sono in grado di catturare:

- La regressione sui **Dati Gerarchici** cattura la relazione disaggregata dei gruppi, consentendo quindi di analizzare la varianza nei gruppi (*within*), e quindi le caratteristiche dei singoli individui.
- L'**Analisi dell Varianza** cattura la relazione aggregata, consentendo quindi di descrivere la variabilità tra i gruppi (*between*).

Una volta depurata l'analisi dalle caratteristiche individuali vi sono due possibili fenomeni (prenderemo l'esempio dei risultati agli esami di diverse università):

- La variabilità interna ai gruppi prevale sulla variabilità tra i gruppi: in questo caso i risultati saranno determinati per la gran parte dalle fluttuazioni individuali, ovvero dalle differenze interne degli individui. Considerando le università, in tale scenario, i risultati migliori di una saranno determinati non tanto dalle differenze tra le università stesse (comune, periferica-centro, privata-pubblica,...) ma dalla variabilità nei risultati individuali degli studenti.
- Nel caso opposto la variabilità tra i gruppi prevale su quella nei gruppi: in tale scenario vi sarà un'interpretazione opposta, ovvero i risultati degli studenti saranno per la maggior parte determinati dalle differenze delle università (le quali rappresentano i gruppi).

Quando si verifica il primo caso i risultati non saranno attribuibili all'efficienza del singolo gruppo (ospedale, università, azienda, ...) ma solo ad una sostanziale disomogeneità delle loro distribuzioni; un modello multilevel efficiente deve pertanto ricadere nel secondo caso, ovvero la variabilità nei gruppi necessita sempre di una depurazione. Tale è la ragione de miglioramento che si ha impiegando un modello multilevel invece che un'analisi della varianza, ovvero il superamento del limite dovuto alla disomogeneità nei gruppi; i modelli multilevel sono di conseguenza l'analisi della covarianza ad effetti casuali, e superano le regressioni multilevel le quali si limitano a distinguere i parametri *within* e *between*, senza tuttavia depurarli.

3.3.1 *Empty Model*

Come già esposto in precedenza la costruzione di un modello multilevel parte dalla definizione del modello ANOVA a effetti casuali, detto anche *empty model*. Definendo quindi il modello $y_{ij} = \gamma_{00} + v_j + \epsilon_{ij}$, dove v_j e ϵ_{ij} sono le componenti casuali, manifestazione delle variabili casuali $\mathbf{E}_{ij}$ (i id ad $\mathbf{E}$) e $\mathbf{V}_j$ (iid $\mathbf{V}$), con distribuzioni (sotto ipotesi di omoschedasticità) $\epsilon_{ij} \sim N(0, \sigma^2)$ e $v_j \sim N(0, \tau^2)$; Le variabili casuali $\mathbf{V}_j$ e $\mathbf{E}_{ij}$ sono mutualmente indipendenti ed incorrelate. Possiamo nuovamente definire il coefficiente di correlazione interclasse $\rho^2 = \frac{\tau^2}{\tau^2 \cdot \sigma^2}$ (il quale deriva dalla varianza

costante tra due individui dello stesso gruppo $\text{cov}(y_{ij}, y_{hj}) = \text{var}(\mathbf{V}_j) = \tau^2$) che fornisce la quota di varianza catturata dai gruppi (ovvero la separazione dei gruppi o la loro capacità di spiegare i diversi risultati); nel caso ρ^2 (e quindi τ^2) risulti piccolo (tipicamente minore di 0.1), non è giustificato il passaggio da un modello lineare a uno multilivello, poiché con molta probabilità non esisterebbe una struttura gerarchica nei dati. Per testare la necessità di applicare un modello multilevel si può inoltre utilizzare un test F, sotto $H_0 : \ v_j = v \ \forall j = 1, ..., p$: quando si verifica che l'ipotesi nulla non viene rifiutata, e quindi i v_j sono costanti (e quindi $\tau^2 \to 0$) con confidenza $1 - \alpha$, il passaggio ad un modello multilevel non è giustificato.

3.3.2 *Random Intercept Model* (RIM)

E' possibile inserire nel modello precedentemente descritto una o più variabili esplicative $\mathbf{x}_k$, passando così da un *empty model* a un **Random Intercept Model** (RIM), il quale è un modello misto. La nuova struttura sarà la seguente:

$$\boxed{\text{RIM: } y_{ij} = \gamma_{00} + \beta_1 \cdot x_{ij} + v_j + \epsilon_{ij}}$$

con $\epsilon_{ij} \sim N(0, \sigma^2)$ e $v_j \sim N(0, \tau^2)$, rispettivamente determinazioni delle variabili casuali normali E_{ij} e V_j, le quali risultano indipendenti ed incorrelate. I modello RIM tiene conto sia dell'analisi della varianza, sia della parte di regressione (stimabile attraverso OLS), ed è di conseguenza un modello multilivello a tutti gli effetti (ed in particolare un ANCOVA ad effetti casuali).

L'effetto *shrinkage* dei modelli RIM consente di trattare dati gerarchici anche quando essi presentino gruppi di numerosità differente, senza tuttavia subire effetti distorsivi. Tali modelli contengono due componenti:

- La componente micro (*micro model*), riferita alle caratteristiche dei singoli:

$$y_{ij} = \beta_1 \cdot x_i j + \epsilon_{ij}$$

- La componente macro (*macro model*), riferita alle differenze fra gruppi:

$$\beta_{0j} = \gamma_{00} + v_j$$

E' ovviamente possibile inserire nuove variabili esplicative (*stepwise*) sia riferite alle caratteristiche dei singoli (di primo livello), sia riferite ai gruppi (di secondo livello), sia riferite ad interazioni tra le precedenti; un esempio di tale struttura è fornita dal modello sottostante:

$$\boxed{y_{ij} = \gamma_{00} + \sum_k \beta_k \cdot x_{ik} + \sum_w \lambda_w \cdot z_{jw} + \delta_{wk} \mathbf{z}_{wk} \mathbf{x}_k + v_j + \epsilon_{ij}}$$

con z_{jw} attributo w-esimo riferito al gruppo j-esimo e λ_w il suo parametro.

Esistono diversi metodi per la **stima dei parametri** di modelli multilevel, come il *full Maximum Likelihood* (ML) e il *Restricted Maximum Likelihood* (REML), i quali hanno in comune il fatto che prima stimano la parte di modello lineare e poi

la parte di analisi della varianza. In particolare i REML massimizzano la verosimiglianza dei residui osservati, ottenendo in tal modo le stime del modello lineare (con metodi non ML come OLS e GLS), e successivamente utilizzano tali stime per massimizzare la verosimiglianza dei residui dell'analisi della varianza (eliminando gli effetti misti). Tali metodi hanno diversi algoritmi, tutti caratterizzati dalla natura iterativa la quale converge alle stime di verosimiglianza, come *expectation-maximization*, *Fisher scoring*, IGLS e RIGLS.

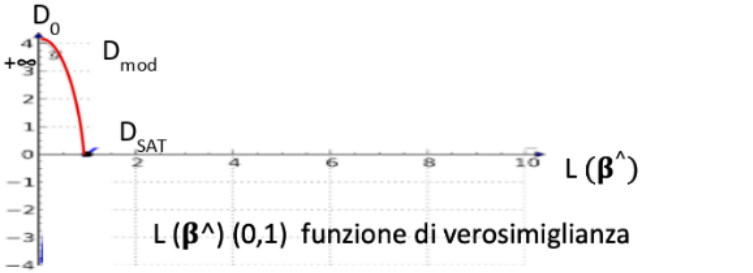
3.3.3 Test d'ipotesi

Dopo il processo di stima è importante verificare la **significatività dei parametri fissi** attraverso appositi test statistici:

Uno dei principali è il **test di Wald**, il quale ha come ipotesi nulla $H_0 : \lambda_w = 0$, e utilizza la statistica test $\mathrm{T}(\lambda_w) = \frac{\lambda\lambda_w}{\mathrm{S.E}(\hat{\lambda}_w)}$, la quale (sotto H_0) si distribuisce approssimativamente come un T-Student con gradi di libertà basati sulla struttura dell'analisi.

Un secondo test è il **Test F**, il quale viene usato per la parte fissa (nullità di r parametri) e della parte random (nullità della devianza spiegata nell'analisi della varianza, *Deviance test*). Attraverso il metodo REML è possibile ottenere iterativamente le stime della devianza residua del modello multiplo $D_M = \mathrm{SSY}_E = -2 \cdot \log L_M$ e la devianza residua dell'analisi della varianza $D_V = \mathrm{SSE} = -2 \cdot \log L_V$, le quali sono misure di bontà di adattamento del modello ai dati.

D_V sotto alcune ipotesi generalmente soddisfatte può essere utilizzato per i test della parte fissa e random, mediante il confronto coi modelli teorici *empty model* (non vi sono variabili esplicative e quindi varianza nei gruppi) e il *saturated model* (un parametro per ogni osservazione, il modello non ha devianza residua né per la regressione, ne per l'analisi della varianza). La costruzione del test segue il seguente schema:



D =(DEVIANZA RESIDUA)= -2 log L(**β**^) (+∞, 0) decrescente

L (0)= 0 Empty model D_0= max= = +∞ ;

L (**β**^) Model D_{Mod}

L ($\beta^{\wedge}_{max}$)=1 Saturated model D_{SAT} =0

$D_0 \geq D_{mod} \geq D_{sat} = 0$ valendo: $L(0) \leq L(\hat{\beta}) \leq L(\hat{\beta}_{max}) = 1$.

Grazie a (-2ln) si ottiene una quantità D (0, +∞)

la cui distribuzione asintotica è del χ^2

Poiché nell'*empty model* sia per la regressione, sia per l'analisi della varianza, le devianze residue assumono valore massimo (coincidono con la SST e distribuendosi come un χ^2_1), idealmente vorrei che il modello fosse il più distante possibile da esso. Dall'altra parte troviamo il *saturated model* il quale ha le devianza spiegate, per entrambe le componenti del modello, pari alla devianza totale, e distribuite come un χ^2_{n-1}. Il modello avrà quindi devianza compresa tra $0 = D_{SAT} \leq D_{MOD} \leq D_0 = +\infty$, e verosimiglianza tra $0 = L(0) \leq L(\hat{\boldsymbol{\beta}}) \leq L(\hat{\boldsymbol{\beta}}_S) = 1$.

Si può ora passare alla definizione del test, il quale utilizza le differenze tra la devianza del modello saturo e quella del modello in esame (sia per regressione che per anova): definendo quindi

$$G_M = (D_0 - D_M)_M = -2 \cdot \log(\frac{L(0)}{L(\hat{\boldsymbol{\beta}})})_M \sim \chi^2_r$$

e

$$G_V = (D_0 - D_M)_V = -2 \cdot \log(\frac{L(0)}{L(\hat{\boldsymbol{\beta}})})_V \sim \chi^2_{p-1}$$

In caso si considerino solo k parametri la distribuzione avrà k gradi di libertà. Se si cade nella regione di accettazione ciò implica che non vi sia nessun miglioramento nel passare dal modello *empty* a quello testato, e di conseguenza i parametri non sono significativamente diversi da 0 (per la parte di regressione), o che la parte di varianza nei gruppi è nulla (per l'anova).

3.3.4 *Random Slope Model* (RSM)

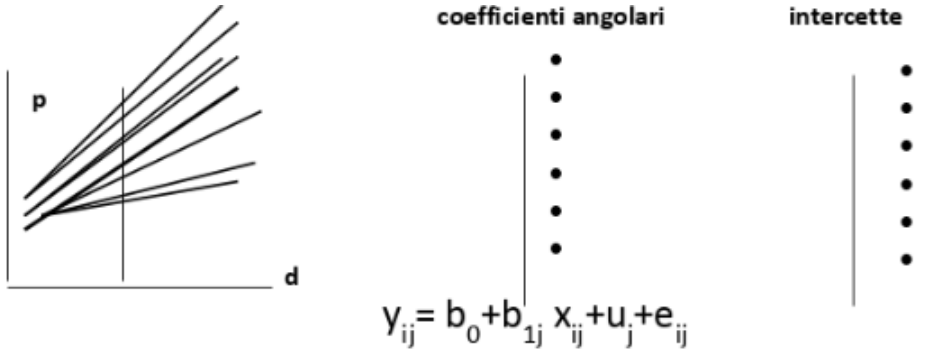
Fin'ora abbiamo presentato modelli multilevel caratterizzati da un'unica parte regressiva per tutti i gruppi; esiste tuttavia un'altra tipologia di modello, il **Random Slope Model** (RSM), il quale utilizza una specifica regressione per ogni gruppo. Tale caratteristica degli RSM li rende maggiormente capaci di rappresentare la realtà rispetto ai RIM poiché assegnano a ciascun gruppo, e per ogni attributo degli individui, un parametro caratteristico, il quale riesce a catturare meglio le differenze dei gruppi.

Riprendendo l'esempio degli ospedali un modello RSM riuscirebbe a tener conto della capacità dei singoli ospedali (che rappresentano i gruppi) di far fronte alle caratteristiche specifiche degli individui curati. Un esempio maggiormente esplicativo riguarda i prezzi delle case: considerando come gruppi tre quartieri residenziali di Milano (uno in periferia, uno in circonvallazione e uno in centro), come variabile risposta il prezzo e come unica variabile esplicativa la dimensione della casa, sarebbe molto più verosimile che i tre gruppi avessero parametri β (ovvero l'incremento nel prezzo a fronte di un incremento unitario nei metri quadrati della casa) differenti a seconda della distanza dal centro, e quindi al gruppo di appartenenza (3 differenti regressioni con tre differenti parametri).

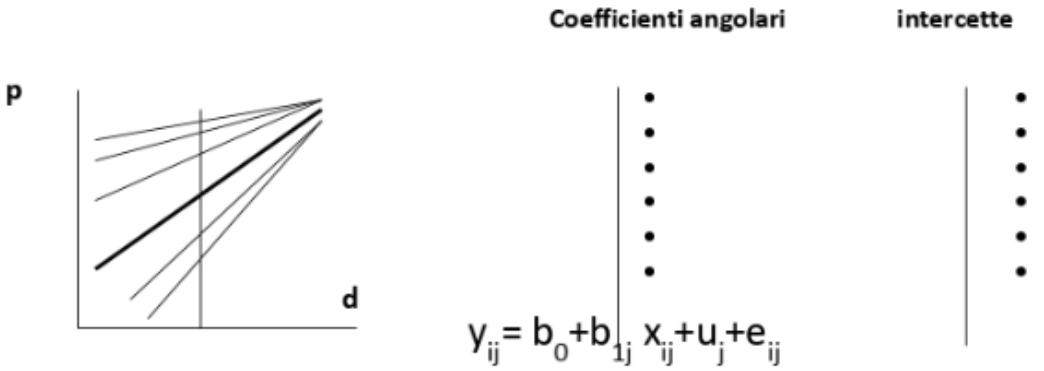
Il **random intercept model** è un particolare tipo di RSM, nel quale ai diversi gruppi vengono date delle intercette differenti, e quindi tengono conto di alcune caratteristiche di partenza del fenomeno, al netto delle variabile esplicative (nell'esempio precedente due case, con le stesse caratteristiche ma zona differente, hanno

prezzi differenti); tuttavia l'aumento della variabile risposta avviene con lo stesso coefficiente angolare per tutto i gruppi.

Una valida alternativa per risolvere tale situazione è consiste nell'impiego di **Random Total Models**, i quali utilizzano, per i diversi gruppi, differenti coefficienti angolari e intercette, come mostrano i grafici sottostanti:



dove viene evidenziato che il parametro b_{1j} varia al variare del gruppo di riferimento. E' ovviamente possibile avere una relazione opposta, ovvero un prezzo di partenza maggiore (intercetta) ma un coefficiente angolare minore; riprendendo l'esempio della case, prendendo come gruppi un quartiere in centro città e uno in campagna, potrebbe verificarsi che il secondo parta con un prezzo maggiore, ma che cresca con una velocità inferiore (poiché in campagna vi è molto spazio, mentre in centro l'aumento della dimensione equivale ad un incremento di prezzo più consistente).



In generale il nuovo modello avrà la seguente forma:

$$\boxed{y_{ij} = \gamma_{00} + \beta^*_{1j} \cdot x_{ij} + v_j + \epsilon_{ij}}$$

con $\beta^*_{1j} = \beta^0_1 + \beta_{1j}$ (ovvero somma di una parte comune e una peculiare) parametro del gruppo $j = 1, ..., p$; β^0_1 sarà quindi componente fissa e deterministica, mentre $\beta_{1j} \sim N(\beta^0_1, \nu^2)$ ovvero, al variare del campione, è manifestazione di una v.c. $\mathbf{B}_{1j}$ con distribuzione Normale. A loro volta v_j e ϵ_{ij} saranno, al variare del campione, determinazioni delle variabili causali Normali $\mathbf{V}_j \sim N(\gamma_{00}, \tau^2)$ e $\mathbf{E}_{ij} \sim N(0, \sigma^2)$, tra loro mutualmente incorrelate e indipendenti.

Il coefficiente β_{1j}^* rappresenterà quindi il parametro dell'interazione $(\beta_{1j}^* \cdot x_{ij})$ tra la **parte fissa** $(\gamma_{00} + \beta_1^0 \cdot x_{ij})$ e la **parte random** $(\beta_{1j} \cdot x_{ij} + v_j + \epsilon_{ij})$ del modello; possiamo inoltre suddividere il modello in **parte micro** $(\gamma_{00}+\beta_1^* \cdot x_{ij}+\epsilon_{ij})$, riferita alle caratteristiche individuali, e la **parte macro** $(\gamma_{00} + v_j)$, riferita alle caratteristiche dei gruppi. Le variabili casuali $\mathbf{V}_j$ e $\mathbf{B}_j$ saranno correlate, in quanto intercetta e coefficiente angolare della regressione del gruppo j-esimo, e il loro coefficiente angolare sarà $\text{Cor}(\mathbf{V}_j, \mathbf{B}_j) = \rho$; le coppie di variabili saranno invece incorrelate, indipendenti e identicamente distribuite rispetto agli altri gruppi.

Il **coefficiente intraclasse** assume in questo caso forma:

$$\boxed{\text{VAR}(Y_{ij} \mid x_{ij}) = \tau^2 + 2\rho \cdot x_{ij} + \nu^2 \cdot x_{ij}^2 + \sigma^2}$$

ovvero una forma molto più complessa che nei casi RIM; per i modelli RSM tale indicatore non viene più usato come misura della bontà del modello né come criterio di scelta. Viene tuttavia utilizzato, insieme al parametro, per formulare le seguenti interpretazioni:

ρ positivo con b_j positivo: maggiore è l'intercetta media del valore di **y**

corretto nel gruppo più forte il legame positivo tra **x** e **y**

ρ positivo con b_j negativo: minore è l'intercetta media del valore di **y**

corretto nel gruppo più forte il legame negativo tra **x** e **y**

ρ negativo con b_j positivo: minore è l'intercetta media del valore di **y**

corretto nel gruppo più forte il legame positivo tra **x** e **y**

ρ negativo con b_j negativo: maggiore è intercetta media del valore di **y**

corretto nel gruppo più forte il legame negativo tra **x** e **y**

ρ nullo non c'è legame tra valore intercetta media di **y** corretto nel

gruppo e legame tra **x** e **y**

Dimostrazioni

1. **Correttezza** degli stimatori OLS con errori eteroschedastici:

$$\begin{aligned} E(\mathbf{B}^*) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} + E(\mathbf{e}^*) = \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}E(\mathbf{b}) = \mathbf{b} \end{aligned} \tag{4.1}$$

   Da cui varrà anche la consistenza

2. **Non efficienza** degli stimatori OLS con errori eteroschedastici:

$$\begin{aligned} \text{VAR}(\mathbf{B}^*) &= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{b})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{b})') \\ &= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{B} + \mathbf{e}^*) - \mathbf{b})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{B} + \mathbf{e}^*) - \mathbf{b})') \\ &= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^*)((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^*)') \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{e}^*\mathbf{e}^{*\prime})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}_{e^*}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})' \end{aligned} \tag{4.2}$$

   Di conseguenza OLS con errori eteroschedastici non saranno efficienti, a varianza minima, BLUE o UMVUE (non coincideranno con ML)

3. **Correttezza** degli stimatori OLS con errori autocorrelati:

$$\begin{aligned} E(\mathbf{B}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} + E(\mathbf{e}^{\#}) = \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}E(\mathbf{b}) = \mathbf{b} \end{aligned} \tag{4.3}$$

4. **Non efficienza** degli stimatori OLS con errori autocorrelati:

$$\begin{aligned} \text{VAR}(\mathbf{B}^{\#}) &= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{b})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{b})') \\ &= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{B} + \mathbf{e}^{\#}) - \mathbf{b})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\mathbf{B} + \mathbf{e}^{\#}) - \mathbf{b})') \\ &= E(((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^{\#})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}^{\#})') \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{e}^{\#}\mathbf{e}^{\#\prime})((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}_{e^{\#}}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})' \end{aligned} \tag{4.4}$$

   Di conseguenza OLS con errori autocorrelati non saranno efficienti, a varianza minima, BLUE o UMVUE (non coincideranno con ML)

5. **Incorrelazione degli errori resi incorrelati**:

   - $y_t = \beta_0 + \beta_1 x_t + e_t^{\#}$
   - $e_t^{\#} = a_0 + a_1 x_1 + ... + a_k x_k + \rho e_{t-1}$
   - $y_{t-1} = \rho\beta_0 + \rho\beta_1 x_{t-1} + \rho e_{t-1\#}$
   - $y_t - y_{t-1} = \beta_0(1-\rho) + \beta_1(x_t + \rho x_{t-1}) + w_t \rightarrow y_i^{\#} = \beta^{\#} + \beta_1^{\#} x_t^{\#} + w_t$

   Da cui determino l'incorrelazione:

$$\begin{aligned}\mathrm{Cov}(w_t, w_{t-1}) &= \mathrm{Cov}(e_t - \rho e_{t-1}, e_{t-1}\rho e_{t-2}) = \mathrm{Cov}(e_t e_{t-1} - e_t\rho e_{t-2} - \rho e_{t-1}e_{t-1} + \rho e_{t-1}e_{t-2}) \\ &= \mathrm{Cov}(e_t, e_{t-1}) - \rho\cdot\mathrm{Cov}(e_t, e_{t-2}) - \rho\cdot\mathrm{Cov}(e_{t-1}, e_{t-1}) + \rho\cdot\mathrm{Cov}(e_{t-1}, e_{t-2}) \\ &= \rho - \rho^3 - \rho + \rho^3 = 0\end{aligned} \tag{4.5}$$

   ovvero gli errori sono incorrelati.

6. **Costruzione stimatori sferici** (GLS):

   - partendo dal modello con errori non sferici $\mathbf{y} = \mathbf{XB}^{\circ} + \mathbf{E}^{\circ}$
   - ricavo la matrice di varianze e covarianze campionaria $\mathbf{S}_{E^{\circ}} = \frac{1}{n}(\mathbf{E}^{\circ}\mathbf{E}^{\circ\prime})$
   - ipotizzo ora l'esistenza di una matrice V non singolare tale che $\mathbf{S}_{E^{\circ}} = \sigma^2\mathbf{VV}'$
   - definisco gli errori trasformati (sferici) $\hat{\mathbf{E}} = (\mathbf{V})^{-1}\hat{\mathbf{E}}^{\circ}$
   - dai quali ricavo $(\mathbf{V})^{-1}\mathbf{S}_{E^{\circ}}(\mathbf{V})^{-1\prime} = (\mathbf{V})^{-1}\sigma^2\mathbf{VV}'(\mathbf{V})^{-1} = \sigma^2\mathbf{I}_n$, matrice di varianze e covarianze di errori sferici.
   - Applico la trasformata al modello ottenendo $(\mathbf{V})^{-1}\mathbf{y} = \mathbf{y}^{\circ} = (\mathbf{V})^{-1}\mathbf{XB}^{\circ} + (\mathbf{V})^{-1}\mathbf{E}^{\circ} = \mathbf{X}^{\circ}\mathbf{B} + \mathbf{E}$
   - ricavo gli stimatori GLS $\rightarrow \hat{\mathbf{B}}^{\circ} = (\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{X}^{\circ\prime}\mathbf{y}^{\circ} = (\mathbf{X}'\mathbf{S}_{\mathbf{E}^{\circ}}^{-\mathbf{1}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{S}_{E^{\circ}}^{-1}\mathbf{y}$

7. **Correttezza stimatori GLS**:

$$\begin{aligned}&\mathbf{B}^{\circ} = (\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{X}^{\circ\prime}\mathbf{y}^{\circ} = (\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{X}^{\circ\prime}(\mathbf{X}^{\circ}\mathbf{b}^{\circ} + \mathbf{E}^{\circ}) \\ &E(\mathbf{B}^{\circ}) = E(\mathbf{b}^{\circ} + (\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{XE}^{\circ}) = \mathbf{b}^{\circ}\end{aligned} \tag{4.6}$$

   si può dimostrare anche la loro consistenza

8. **Efficienza degli stimatori GLS**:

$$\begin{aligned}\mathrm{VAR}(\mathbf{B}^*) &= E(((\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{X}^{\circ\prime}\mathbf{y}^{\circ} - \mathbf{b}^{\circ})((\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{X}^{\circ\prime}\mathbf{y}^{\circ} - \mathbf{b}^{\circ})') \\ &= E(((\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{X}^{\circ\prime}(\mathbf{X}^{\circ}\mathbf{B}^{\circ} + \mathbf{E}) - \mathbf{b}^{\circ})((\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{X}^{\circ\prime}(\mathbf{X}^{\circ}\mathbf{B}^{\circ} + \mathbf{E}) - \mathbf{b}^{\circ})') \\ &= (\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{X}^{\circ\prime}E(\mathbf{EE}')((\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\mathbf{X}^{\circ})' \\ &= \sigma^2(\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{E^{\circ}}}{}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{E^{\circ}}}{}^{-1}(\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{E^{\circ}}}{}^{-1}\mathbf{X})^{-1})' \\ &= \sigma^2(\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{E^{\circ}}}{}^{-1}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\circ\prime}\mathbf{X}^{\circ})^{-1}\end{aligned} \tag{4.7}$$

   ovvero efficienti per le variabili trasformate $\mathbf{X}^{\circ}$ e $\mathbf{y}^{\circ}$ (Teorema di Aitken), ma non saranno né Blue né UMVUE

9. **Correttezza degli stimatori OLS** con errori non normali:

$$\begin{aligned} E(\hat{\mathbf{B}}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{b} + E(\mathbf{e}) = \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}E(\mathbf{b}) = \mathbf{b} \end{aligned} \tag{4.8}$$

e poiché vale Gauss-Markov risultano anche efficienti e BLUE; non saranno tuttavia UMVUE, poiché la violazione della normalità non garantisce l'uguaglianza con gli stimatori ML.

10. **Ortogonalità/incorrelazione** tra errori e esplicative (1), e tra errori e valori predetti (2) (Multivariato) :

$$\begin{aligned} &(1)\ \mathbf{E}\mathbf{Z}' = \mathbf{Y}(\mathbf{I} - \mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}\mathbf{Z}') = \mathbf{Y}\mathbf{Z}' - \mathbf{Y}\mathbf{Z}' = 0 \\ &(2)\ \begin{aligned} \hat{\mathbf{Y}}\mathbf{E} &= \mathbf{Y}'\mathbf{Z}(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}\mathbf{Z}'(\mathbf{I} - \mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}\mathbf{Y}') = \\ &= \mathbf{Y}'\mathbf{Z}(\mathbf{I} - \mathbf{Z}(\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}')\mathbf{Y}' = \mathbf{Y}'\mathbf{Z}\mathbf{Y} - \mathbf{Y}'\mathbf{Z}\mathbf{Y} = 0 \end{aligned} \end{aligned} \tag{4.9}$$

11. **Efficienza degli stimatori OLS multivariati**:

$$\begin{aligned} &E((\hat{\mathbf{B}} - \mathbf{B})'(\hat{\mathbf{B}} - \mathbf{B})) = E((\mathbf{Y}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1} - \mathbf{B})'(\mathbf{Y}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1} - \mathbf{B}) = \\ &= E(\mathbf{B}\mathbf{Z}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1} - \mathbf{B} + \mathbf{E}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1})'(\mathbf{B}\mathbf{Z}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1} - \mathbf{B} + \mathbf{E}\mathbf{Z}'(\mathbf{Z}\mathbf{Z}')^{-1})) = \\ &= (\mathbf{Z}\mathbf{Z}')^{-1}\mathbf{Z}E(\hat{\mathbf{E}}'\hat{\mathbf{E}})\mathbf{Z}(\mathbf{Z}\mathbf{Z}')^{-1} = \sigma^2(\mathbf{Z}\mathbf{Z}')^{-1} \end{aligned} \tag{4.10}$$

12. **Costruzione stimatori** in presenza di errori non sferici (GLS multivariato):

- dato il modello di partenza $\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}^\circ$ (con $\mathbf{E}^\circ$ errori non sferici)
- si ipotizza l'esistenza di una matrice $\mathbf{W}$, non singolare, tale che: $\mathbf{\Sigma}_{E^\circ} = \sigma^2\mathbf{W}\mathbf{W}'$
- da cui deduco $\mathbf{E}^\circ = \mathbf{E}(\mathbf{W})^{-1}$
- e da cui ricavo la matrice di varianze e covarianze degli errori: $\mathbf{\Sigma}_{E^\circ} = (\mathbf{W})^{-1}E(\mathbf{E}^{\circ\prime}\mathbf{E}^\circ)\mathbf{W}'\mathbf{W}(\mathbf{W})^{-1} = \sigma^2\mathbf{I_{nm}}$ ovvero errori sferici.
- applico la trasformata al modello: $(\mathbf{W})^{-1}\mathbf{y} = \mathbf{y}^\circ = (\mathbf{W})^{-1}\mathbf{X}\mathbf{B}^\circ + (\mathbf{W})^{-1}\mathbf{E}^\circ = \mathbf{X}^\circ\mathbf{B} + \mathbf{E}$
- e infine ricavo lo stimatore OLS del modello così trasformato (GLS): $\hat{\mathbf{B}}^\circ = (\mathbf{Z}^{\circ\prime}\mathbf{Z}^\circ)^{-1}\mathbf{Z}^{\circ\prime}\mathbf{y}^\circ = (\mathbf{Z}'\mathbf{\Sigma}^{-1}_{\mathbf{E}^\circ}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Sigma}^{-1}_{E^\circ}\mathbf{y}$
- la matrice $\mathbf{W}$ sfrutterà le proprietà degli autovettori/autovalori attraverso la decomposizione spettrale di $\mathbf{\Sigma}_{E^\circ}$: $\mathbf{W} = \sigma\mathbf{K}\mathbf{L}^{\frac{1}{2}}\mathbf{K}'$ con $\mathbf{K}$ matrice autovettorie $\mathbf{L}$ matrice diagonale degli autovalori di $\mathbf{\Sigma}_{E^\circ}$

13. **Analisi della Covarianza** (ANCOVA):

- definisco, a partire dall'ANOVA, un modello che tenga conto delle caratterostiche della popolazione: $y_{ij} - (\beta_0 + \sum_k \beta_{kj} x_{ijk}) = \gamma_{00} + v_j + e_{ij}$, o in termini matriciali $(\mathbf{y} - \boldsymbol{\gamma}_{00}) - \mathbf{XB} = \mathbf{Av} + \mathbf{e} = \mathbf{y}_e$
- ricavo gli stimatori OLS $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} - \boldsymbol{\gamma}_{00})$ da cui deduco la devianza spiegata dalla regressione SSX= $\hat{\mathbf{b}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{b}}$
- pongo dei vincoli sui parametri per rendere la matrice $\mathbf{A}$ di rango pieno, ottenendo una matrice $\mathbf{A}^\circ$ di dimensione $(n, p - s)$ (s numero vincoli).
- il modello diviene ora $\mathbf{y}_e = \mathbf{A}^\circ$ + e, la cui soluzione OLS è data da: $\hat{\mathbf{v}}^\circ = (\mathbf{A}^{\circ\prime}\mathbf{A}^\circ)^{-1}\mathbf{A}^{\circ\prime}\mathbf{y}_e$, da cui deduco la devianza spiegata dall'analisi della covarianza (o *between*) SSV=$\hat{\mathbf{v}}^{\circ\prime}\mathbf{A}^{\circ\prime}\mathbf{A}^\circ\hat{\mathbf{v}}^\circ$
- le devianze seguiranno tale relazione: S
  - SSX+SSV=SST
  - SSE=e'e (devianza *within*)
  - SST-SSX=SSV+SSE=SSY_e ; con SSY_e devianza residua di regressione.
- sotto ipotesi di omoschedasticità $\frac{SSV}{n} = \hat{\mathbf{v}}^{\circ\prime}\mathbf{A}^{\circ\prime}\mathbf{A}^\circ\hat{\mathbf{v}}^{/} = \tau^2$ e $\frac{SSE}{n} = \mathbf{e}'\mathbf{e}/n = \sigma^2$
- definisco il coefficiente di correlazione intraclasse$\rightarrow \rho = \frac{\tau^2}{\tau^2+\sigma^2}$, ovvero la quota di varianza totale dovuta ai grupp, al netto delle caratteristiche di ciascun gruppo