Appunti di Statistical Modelling

Prof. Vittadini

Leonardo Alchieri, BSc

First printing, March 2020

Contents

Informazioni sul corso

Il corso è diviso in 3 parti, ovvero modello lineare, modello multi-variato e modello *multi-level*. L'approccio è operativo, con l'obiettivo di vedere le applicazioni su `R` e `SAS`.

Per il corso basta sapere solamente uno dei due strumenti, sebbene sarebbe comodo imparare entrambi.

Alle lezioni si affianco le esercitazioni, in cui si vede come usare i due strumenti per il corso.

L'esame è composto da 3 domande scritte, di cui 2 teoriche e 1 applicata. Quelle teoriche saranno prese tra 15 domande presenti sul sito, mentre quella pratica sarà in esercizio, in cui noi dovremo processare con `R` o `SAS` commentando.

Su `e-learning` sono presenti tutte le informazioni sul corso.

1. Modello Lineare

1.1 Introduzione

Modello Statistico

Cerchiamo di capire che cosa sono i modelli in statistica. Per poter costruire un modello, o statistico o matematico, si deve avere **adattamento ai dati** e **parsimonia**: ovvero le variabili esplicative spiegano in parte il comportamento di una variabile dipendente. Più ci sono parametri, più risulta difficile cercare di spiegare la variabile: questo è il motivo per cui non sempre dei modelli di *Machine Learning* sono migliori di quelli statistici.

Come disse l'epistemologo della scienza **Popper**, ogni modello è intrinsecamente falso: si può sempre migliorare la sua capacità interpretativa.

I **modelli matematici**, che si muovono in campo teoretico/astratto, interpretano perfettamente i dati. Quello **statistico** invece ha meno forza interpretativa, in quanto dominato dall'errore: non riesce a interpretare perfettamente i dati. Nel *residuo* si concentrano una serie di concetti importanti, prima tra tutti la variabilità casuale del fenomeno oppure a un errore sistematico del modello che si possiede, e.g. mancanza di abbastanza variabili. Un altro aspetto importante può essere quello della misura degli errori, in particolare in fisica e chimica nelle misurazioni di pianeti o atomi. Infine si ha difficilmente a disposizione i dati di tutta la popolazione, fatto che induce la presenza di errore nel modello che si utilizza.

Si potrebbe dire che lo studio dei modelli statistici è lo studio dei motivi di errore: esso è fondamentale per analizzare i fenomeni reali. Le quattro ragioni spiegate sopra dicono che cosa è un modello statistico; da questo punto di vista, ci si può poi chiedere come si venga a costruire.

1.2 Costruzione di un modello statistico

Costruire un modello statistico

In termini generali, si deve scegliere di costruire modelli o di spiegazione o di causa/effetto: la differenza tra modelli empirici e modelli causa-effetto. Nel corso ci limiteremo solamente ai primi, in quanto troppo avanzati per i nostri scopi.

Si dovrà quindi individuare le variabili indipendenti e capire quali introdurre in prima istanza.

Fatto questo, si dovrà raccogliere i dati, e cercare di capire come farlo, che esso sia diretto o indiretto.

Scelto il modello si dovrà trovare un modo per **stimare i parametri**, dopo la quale si andrà a eseguire la **verifica del modello**. Infine si andrà a utilizzarlo per **interpretare** e **prevenire**.

Modello di regressione

Di base, noi tratteremo il modello di regressione, in cui, date alcune variabili dipendenti, si usano delle relazioni "note" (fissate) per prevederle tramite delle **variabili esplicative**. Le caratteristiche del modello è di essere deterministico, con la presenza di errori, soprattutto dati dall'ignoranza della relazione reale.

Si potrebbero avere dei modelli lineari o non, a seconda che si costruiscano in termini di come vengono espresse le relazioni tra le variabili.

Nel criterio della **stima**, ci concentreremo i vettori che rendano minime le somme dei quadrati degli errori: questo aumenta al massimo la *parsimonia*. Si noti che si prendono i quadrati per evitare effetti dovuti a somma tra positivi e negativi – difatti la somma degli errori sarebbe nulla.

1.2.1 Modello Lineare Multiplo

Minimi quadrati

Si dimostra che, se la matrice è di rango pieno, il vettore dei parametri $\mathbf{b}$ si trova come:[1]

$$\hat{\mathbf{b}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} \tag{1.1}$$

Questa è la **soluzione dei minimi quadrati**.

Verifica

Nella verifica, si pone che la soluzione è quella dei minimi quadrati con aggiunta degli errori:

$$\mathbf{y} = \mathbf{X}\hat{\mathbf{b}} + \hat{\mathbf{e}} = \hat{\mathbf{y}}$$

Si può inoltre calcolare due misure di bontà, ovvero **varianza residua** e **varianza spiegata**: maggiore è quella spiegata, e minore quella residua, migliore è il risultato. Per questo, si può anche calcolare un indice, detto R^2, compresa tra 0 e 1 (valore che non si raggiunge quasi mai).

In particolare, identificando con $\sigma^2_{mss} = \mathbf{y}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}\mathbf{y}$ la **varianza spiegata**, che identifica la somma dei quadrati dovuta alla regressione, e con $\sigma^2_{ssr} = \hat{\mathbf{e}}^\top \hat{\mathbf{e}} = \mathbf{y}^\top \left[\mathbb{1} - \mathbf{X}\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}\right] \mathbf{y}$ la **varianza residua**, dovuta la somma del quadrato degli scarti tra valori osservati e interpolati, si ottiene:

$$R^2 = \frac{\sigma^2_{mss}}{\sigma^2_{mss} + \sigma^2_{ssr}} \tag{1.2}$$

dove la somma delle due, a denominatore, coincide con la devianza totale $\sigma_{tss} = \mathbf{y}^\top \mathbf{y}$.

Standardizzazione

Si noti che si può standardizzare a priori le variabili, in maniera tale da togliere gli effetti dovuti a differenti ordini di grandezza. In particolare, si identifica con $\mathbf{D_X}$ la matrice diagonale i cui elementi sulla diagonale sono le varianze delle variabili $\mathbf{X}$, ovvero:

$$\mathbf{D_X} = \begin{bmatrix} \sigma_1^2 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \sigma_k^2 \end{bmatrix}$$

[1] Il "cappellino" identifica il valore assunto in presenza del minimo del quadrato degli errori.

dove k indica il numero di variabili esplicative.[2] Si può quindi standardizzare le variabili esplicative e dipendenti come:

$$\mathbf{X}^* = \mathbf{X}\mathbf{D}^{-\frac{1}{2}}$$
$$\mathbf{y}^* = \frac{\mathbf{y}}{\sigma_\mathbf{y}}$$

Scelta del parametro

Nella maggior parte dei casi, quando ci si trova davanti a un dataset non si ha una conoscenza sulla "variabile da predire": spetta a me decidere quale. È inoltre buona pratica quella di, in presenza di un elevato numero di variabile possibilmente esplicative, provare a iniziare da una e aggiungerne altre, per vedere quale porta a un fit migliore.

R quadro corretto

All'aumentare dei parametri per eseguire una regressione, ci si potrebbe aspettare che il comportamento possa cambiare. Di conseguenza, spesso, al posto di mostrare il semplice R^2, si preferisce l'uso di un R^2 **corretto**, o $\bar{R}^2$. Questo tiene conto della numerosità sia delle osservazioni, n, sia dei parametri usati per stimare la regressione, k. Si calcola come:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{\sigma^2_{ssr}}{\sigma^2_{mss} + \sigma^2_{ssr}} \tag{1.3}$$

dove ovviamente la somma $\sigma^2_y = \sigma^2_{mss} + \sigma^2_{ssr}$ è la devianza totale della variabile dipendente.

1.2.2 Modello Lineare Classico

Ipotesi

Il modello lineare classico si basa su 6 ipotesi principali:

1. **Linearità.**
   Dice che sia le variabili esplicative $\mathbf{X}$ sia i parametri $\mathbf{b}$ sono lineari. Ovvero vale la relazione:

   $$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \tag{1.4}$$

   Questo è un semplice punto di partenza per la costruzione di un modello
2. **Non sistematici degli errori.**
   Quest'ipotesi è fondamentale: se ci fossero errori che non hanno una distribuzione che è 0 (ovvero la somma diversa da nullo), non sarebbero residui ma avrebbero un carattere sistematico. Di conseguenza, la parte sistematica, non affetta da errore, è considerata all'interno dell'interpolante.
3. **Sfericità degli errori.**
   Anche questa serve per semplificare: dice che gli errori sono **omoschedastici** e **incorrelati**. Ovvero non vi è correlazioni tra i risultati delle osservazioni (*incorrelati*), ipotesi non sempre molto forte rispetto alla realtà, e che tutti gli errori hanno stessa varianza (*omoschedasticità*). Di conseguenza, la matrice degli errori (o delle correlazioni) delle osservazioni sarà diagonale, con tutti gli elementi uguali:

   $$\Sigma_{\sigma^2_\mathbf{y}} = \begin{bmatrix} \sigma^2 & \dots & 0 \\ 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix}$$

[2] Si ricordi che la matrice $\mathbf{X}$ è quelle delle osservazioni, ovvero quella in cui a ogni variabile k-esima si associano n osservazioni, per ottenere una matrice $n \times k$

4. **Non stocasticità delle variabili esplicative.**
   Difatti, dividendo la variabile $\mathbf{X} = \mathbf{X}_f + \mathbf{X}_s$ in una parte stocastica e in una fissa, si può attribuire la parte stocastica all'errore. Ovvero, si tratta come:

$$\mathbf{y} = \mathbf{X}_f\mathbf{b} + (\mathbf{e} + \mathbf{X}_s\mathbf{b}) = \mathbf{X}_f\mathbf{b} + \mathbf{e}_{new} \tag{1.5}$$

   In tutto il corso si andrà a considerare come valida questa ipotesi.
   Di conseguenza, ipotizziamo che $\mathbf{X}_f$, la parte fissa, sia **incorrelata** con l'errore $\mathbf{e}_{new}$.
5. **Non collinearità delle variabili esplicative.**
   Se il rango di una matrice è inferiore all'ordine (non ha rango pieno), non è invertibile. Di conseguenza, la condizione che viene posta è che le variabili in $\mathbf{X}$ siamo indipendenti, in maniera tale che essa abbia **rango pieno** e, di conseguenza, la matrice $\mathbf{X}^\top\mathbf{X}$ è **non singolare**, e dunque possiede solamente **una soluzione**.
6. **Numerosità della popolazione.**

   #### Necessarie
   Il numero dei casi che si vanno a prendere deve essere maggiore del numero delle variabili, più il termine noto. Ovvero $n > k+1$. Anche in questo caso, se questo non valesse non si potrebbe avere un'unica soluzione alla matrice $\mathbf{X}^\top\mathbf{X}$.

La parte più importante è capire come vengono distribuiti gli errori.

Le ipotesi che devono essere sempre valide sono:

- Non sistematicità degli errori.
- Non stocasticità delle variabili esplicative. In termini generali la si può avere perché attribuisce la parte stocastica delle $\mathbf{X}$ all'errore.
- Non collinearità delle variabili esplicative.
- $n > k+1$.
  Questa, con la precedente, servono per poter permettere l'inversione univoca della matrice $\mathbf{X}^\top\mathbf{X}$.

Viceversa, le altre due ipotesi, *linearità* e *sfericità degli errori* potranno saltare in modelli successivi, come vedremo. Difatti, queste ipotesi non sono sempre valide nella realtà, e una loro omissione può portare alla formulazione di modello più complessi.

Distribuzione degli errori
Una settima ipotesi può essere considerata quella che gli **errori** abbiano tutti una **distribuzione normale**, ovvero $\varepsilon_i \sim N(0, \sigma^2)$. Senza questa non si potrebbero trovare molte informazioni legate ai risultati.

1.2.3 Modello Lineare Campionario

Spesso, si deve scegliere se considerare o la popolazione, che prende *tutti* gli "esperimenti" fatti in k laboratori, o un campione, spesso da un punto di vista pratico più semplice, nei quali si considera solo un sottoinsieme. Il campione può essere scelto o non casuale o casuale, quest'ultimo preferito per poter avere dei risultati più robusti.

Campione
Ogni modello creato su un campione sarà leggermente diverso da quello di altri campione o di tutta la popolazione: al variare del campione, i risultati variano. Da un punto di vista pratico, si possiede ovviamente solamente uno di questi risultati.

Tolti i termini stocastici, la parte $\hat{\mathbf{b}} = \left(\mathbf{X}^\top\mathbf{X}\right)\mathbf{X}^\top\mathbf{y}$ varia solamente al variare della $\mathbf{y}$, e quindi cambia da campione a campione.

Stimatore stocastico

Sul campione si avrà quindi una parte stocastica del modello, e si cerca di quantificare questo errore ε. Si genera quindi un distribuzione campionaria di tipo stocastico dello stimatore $\hat{\mathbf{b}}$: non avendo a disposizione il "risultato vero", si andrà a studiare questa **distribuzione**.

Uno dei metodi da applicare è quello di studiare le **proprietà dello stimatore**, in maniera tale che si conoscano i legami con la popolazione.

Inoltre, siccome l'errore tende a essere di tipo normale, la stessa distribuzione verrà seguita dai parametri della regressione.

Proprietà dello stimatore

- **Correttezza**
  Al variare della distribuzione, deve essere uno stimatore che abbia come valore atteso quello vero della popolazione, ovvero $E[\mathbf{B}] = \mathbf{b}$, dove $\mathbf{B}$ è lo stimatore e $\mathbf{b}$ è il valore assunto dalla popolazione.
- **Efficienza**
  Tra gli stimatori corretti, si sceglie quello che è più efficiente. Ovvero si sceglie lo stimatore con minore varianza.
- **Consistenza**
  Questa proprietà dice che, maggiore è il campione, più "vicino" sarà la stima con il valore vero. Ovvero $\lim_{n\to\infty} P(|\mathbf{B}-\mathbf{b}| < k) = 1$.

Queste 3 proprietà permettono di avere un'approssimazione ragionevole della popolazione.

Stimatori dei minimi quadrati

Si può dimostrare gli stimatori dei minimi quadrati sono sempre corretti. Infatti, per la proprietà di linearità, non collinearità e numerosità della popolazione, si ha che:

$$E[\hat{\mathbf{B}}] = E\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}\right] = E\left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top(\mathbf{X}\mathbf{b}+\mathbf{e})\right] = \mathbf{b} \tag{1.6}$$

Si noti ovviamente che il valore di aspettazione dell'errore è nullo, e il prodotto di una matrice e della sua inversa dà come risultato la matrice identità.

Inoltre, secondo il **teorema di Gauss-Markov**, gli stimatori dei minimi quadrati sono i più efficienti tra tutti gli stimatori lineari possibili. Spesso si usa il termine BLUE, *best linear unbiased estimator*, per identificare gli stimatori del modello OLS, *ordinary least squares*.

Inoltre, si dimostra anche che questi stimatori sono **consistenti**.

1.2.4 Test di ipotesi

Distribuzione stimatori

Usando la caratteristica che gli **errori sono normali**, si può dimostrare che anche gli stimatori avranno una distribuzione di tipo normale. Si mostra la dimostrazione per una regressione lineare semplice, ovvero con un solo stimatore, ma si può estende anche in notazione matriciale, ovvero in presenza di più variabili esplicative.

Partendo da una stima, si possono verificare delle ipotesi sui parametri. In particolare, si va a verificare che il **parametro sia nullo**. Ovvero si verifica se la variabile esplicativa ha legami o meno con la variabile dipendenti.

Se il parametro campionario cade nella regione di rifiuto, si rifiuta H_0; viceversa, si accetta. Dalla distribuzione, dalla gran parte dei risultati si ha che il parametro è uguale a 0. Dal punto di vista dell'equilibrio, prima di dire che un parametro sia diverso da 0, si devono ottenere dei risultati nell'aria di rifiuto.

La logica del test è quella di mostrare con la maggior certezza possibile che il parametro possa non essere nullo: si minimizza la possibilità che si prenda un parametro diverso da 0 quando in realtà lo è.

Varianza nota

Nel qual caso sia nota la varianza del parametro sia nota, so che la distribuzione campionaria avrà una distribuzione normale, con valore atteso l'oggetto di stima.

La distribuzione normale sarà del tipo $\beta_j \sim N\left(b_j, \frac{\sigma^2}{n\sigma_{jj}^{-1}}\right)$, con l'ipotesi che $H_0 : b_j = 0$. [3]

Il test viene eseguito come:

$$P\left(-Z_{\frac{\alpha}{2}} < \frac{\beta_j}{\frac{\sigma}{\sqrt{n\sigma_{jj}^{-1}}}} < Z_{\frac{\alpha}{2}}\right) = 1-\alpha \tag{1.7}$$

Ovvero se il parametro ricade nell'intervallo identificato, non si rifiuta; viceversa, se è fuori si rifiuta.

Varianza ignota

Nel qual caso la varianza non sia nota, si eseguo il test con la **t di Student**. Questa distribuzione si costruisce come rapporto tra una normale e un chi-quadro, con $n-k-1$ gradi di libertà.

In particolare, la gaussiana è data dalla distribuzione $\frac{\beta_j}{\frac{\sigma}{\sqrt{n\sigma_{jj}^{-1}}}} \sim N(0,1)$, mentre il chi-quadro è $\frac{s}{\sigma}$, dove s corrisponde alla stima della deviazione standard del parametro. Si dimostra che le due funzioni risultano indipendenti, la seguente è una **distribuzione t**:

$$t \sim \frac{\beta_j}{\frac{s}{\sqrt{n\sigma_{jj}^{-1}}}} \tag{1.8}$$

Su cui si potrà quindi eseguire un **test t**, del tipo:

$$P\left[-t_{\frac{\alpha}{2}} < \frac{\beta_j}{\frac{s}{\sqrt{n\sigma_{jj}^{-1}}}} < t_{\frac{\alpha}{2}}\right] = 1-\alpha \tag{1.9}$$

Test F per il modello

La variabile R^2, rapporto tra *varianza spiegata* e *varianza totale*, come in Eq. 1.2, è un rapporto tra due **chi-quadri**, che ovviamente non sono indipendenti. Per poter avere informazione statistica, si utilizza in letteratura il rapporto tra **varianza spiegata** e **varianza residua**, ovvero

$$F = \frac{\frac{\sigma^2_{mss}}{k}}{\frac{\sigma^2_{ssr}}{n-k-1}}$$

Di conseguenza, questa variabile si distribuisce come una **F di Snedecor**.

In questo caso, la distribuzione è a *una coda*, per cui si deve avere una R^2 maggiore di un certo valore, ovvero all'interno della regione di rifiuto. Sarà di conseguenza:

$$P\left[F < F_{\alpha,(k,n-k-1)}\right] = 1-\alpha \tag{1.10}$$

[3] Con una notazione corretta, si avrebbe che la varianza della matrice degli stimatori β sarebbe $\sigma^2[\beta] = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$. Chiamando questa come Σ, quando si prende il j-esimo stimatore, si dovrebbe prendere il termine diagonale associato (si noti che la matrice è diagonale, per la diagonalità di $\mathbf{X}^\top\mathbf{X}$), ovvero chiama Σ_{jj}.

La notazione riportata sopra, quella usata dal prof, mi risulta di dubbia interpretazione.

Test F su uno o più parametri

Questo test può essere usato per verificare l'ipotesi nulla di uno o più parametri, ovvero del tipo $H_0 : b_1, \ldots, b_q = 0$. In questo caso, si prende un test che prenda in considerazione queste variabili.

Si costruisce la **varianza spiegata delle $k-q$ variabili $\sigma^2_{mss,q}$** e **varianza residua** delle stesse $\sigma^2_{rss,q}$. Ovvero, si ragione come se si avessero solamente le ultime $k-q$ variabili:

$$F_{1-\alpha,(k-q,n-k-1)} = \frac{\frac{\sigma^2_{mss} - \sigma^2_{mss,q}}{k-q}}{\frac{\sigma^2_{ssr}}{n-k-1}} \tag{1.11}$$

Essendo questo un rapporto tra χ^2 indipendenti, si può eseguire un test come *F di Snedecor*, come in Eq. 1.9. Rispetto al caso precedente, si considerano meno parametri.

Intervalli di confidenza per i parametri

In alcuni casi, una volta determinato se il parametro è diverso da 0, può essere utile sapere quale sia l'intervallo in cui si trova il valore vero.

In questo caso, si farà un'ipotesi diversa da zero per il test di ipotesi:

$$P\left[\beta_j - Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n\sigma^{-1}_{jj}}} < b_j < \beta_j + Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n\sigma^{-1}_{jj}}}\right] = 1 - \alpha \tag{1.12}$$

dove l'intervallo identificato dall'equazione è l'**intervallo di confidenza**. Leggendo al contrario, ho la probabilità $1-\alpha$ che il valore vero sia all'interno di esso.

Questo test, come mostrato su normale in Eq. 1.12, può essere eseguito anche come *t test*.

Relazione tra testi di ipotesi e intervallo di confidenza

Si cerca sempre di minimizzare l'**errore di primo tipo**, ovvero ritenere che un parametro sia significativo quando non lo è. Dall'altra parte, non ci si cautela dal rischio opposto, ovvero **errore di secondo tipo**.

Il problema è selezionare solo quelle variabili che sono effettivamente in grado di interpretare la variabile indipendente. Scelgo solo quelle che lo siamo molto chiaramente, per cui il parametro ha dei valori estremi.

Una volta trovate quali sono le **variabili significative** nell'interpretare la variabile dipendente, si usa l'intervallo di confidenza: si va quindi a capire dentro quale intervallo esse si trovino. Non si può essere certi, ma si cerca di dare una probabilità alta del risultato per l'**intervallo di confidenza**.

Se si vuole aumentare la probabilità che il vero parametro sia in esso, si dovrà allargare tale intervallo: si ha sempre un *tradeoff* tra intervallo e probabilità.

1.2.5 Metodo di massima verosomiglianza

Si cerca di usare una funzione di massima verosomiglianza che utilizza il fatto che la distribuzione dei parametri sia normali.

Inoltre, questo metodo arriva allo stesso risultato dello stimatori dei minimi quadrati. Sono inoltre sia i **migliori stimatori lineari**, e sono anche gli stimatori a minima varianza tra **tutti gli stimatori**, proprietà nota come **VUE**.

Si cerca difatti di risolvere un problema del tipo:

$$max_\beta \beta L\left(y_i - \mathbf{X}_i^\top \beta\right), \quad \forall i = 1, \ldots, n \tag{1.13}$$

1.2.6 Variabili esplicative qualitative

Le variabili esplicative categoriche sono definite su valori definiti da attributi, e.g. *sesso*, *religione* etc. Esse posso essere sia **nominali** sia **ordinali**.

Si possono usare delle ***dummy variables***, variabili che non hanno significato quantitativo ma che distinguono tra categorie, e.g. 0 per *femmine* e 1 per *maschi*.

L'esempio più semplice è quello in cui compaiono solamente *dummy*, senza una spiegazione. Per esempio, nel caso *maschi-femmine*, si avrebbero due regressioni separate. In questo caso, la differenza tra le intercette da informazione sulla distrinzione tra i due modelli, con valori diversi della dummy. Interpretativamente, si usano le *dummy* some se fosse una qualunque delle variabili esplicative normali.

Dummy con quantitative

La variabile *dummy* pone differenza che si ha un nuovo coefficiente angolare, in presenza di altre variabili quantitative.

In termini interpretativi, cambia solamente il coefficiente angolare tra le due rette di regressione, una con un valore della *dummy* e una con un altro. Anche in questo caso bisogna stare attenti: ogni coefficiente di regressione è **parziale**, si ha una modificazione di tutte a seconda che ci sia o meno la *dummy*. Inoltre, è difficile interpretare i risultati, che a volte possono essere controintuitivi.[4]

[4] Vedi esempio su cinture di sicurezza.

1.3 Violazioni modello lineare classico

Andiamo qua a vedere che cosa succede se vengono persi alcuni degli assunti nel modello lineare classico, e come comportarsi in tali casi.

1.3.1 Errori eteroschedastici

Nel modello campionario della i-esima soluzione, con soluzione $\hat{\mathbf{B}}$ e i suoi errori. Se si va a vedere la distribuzione dei vari individui e la si paragona con quella di altri, non si può affermare che questi abbiano la stessa varianza. Si noti che si parla della varianza delle **distribuzione**, e non dei singoli campioni: si confrontano le varianza campionarie.

Per l'osservazione j-esima, si avrà che ogni campione porterà con un **errore**, ε_{jk}, dove k indica il k-esimo campione in considerazione. Si può dire che questi errori appartengono a una distribuzione, di **variabile casuale** E_j e varianaz σ_j^2. Oltre a questo, si possono ovviamente calcolare le **correlazioni** tra le diverse *variabili casuale errore* E_j^*.[5]

Se si considera un modello incorrellato ma eteroschedastico, si avrà che $\sigma_i^2 \neq \sigma_j^2, \quad \forall j \neq i \in [1,n]$. Ogni osservazione ha una distribuzione casuale diversa da quella degli altri.

Perdere omoschedasticità

Avere delle osservazioni eteroschedastiche, is ha che si mantiene sia linearità, che correttezza degli stimatori **OLS** (minimi quadrati). Infatti:

$$E[\mathbf{B}^*] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E[\mathbf{y}] = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top E[\mathbf{X}\mathbf{b}+\mathbf{e}^*] = \mathbf{b} \tag{1.14}$$

Quello che però accade è la perdita di **efficienza**, ovvero non sono più *BLUE*.

$$var[\mathbf{B}^*] = E\left[\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}-\mathbf{b}\right)\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}-\mathbf{b}\right)^\top\right] = \\ = \left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)E[\mathbf{e}^*\mathbf{e}^{*\top}]\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\right)^\top \tag{1.15}$$

Ovvero non si ha più la minor varianza.

Nel caso della **t di Student**, si ha che la regione di accettazione muta al variare del campione. Si dimostra inoltre che questa varianza sottostima il valore della varianza vera. Si ha quindi che non si può contare su test come questi con chiarezza: dipendono dal diverso campione e si ha una regione di rifiuto più grande.

Analogo vale anche il **test F**.

1.3.2 Individuare errori eteroschedastici graficamente

Scatter

Il modo per individuare questi errori, si possono usare rappresentazioni grafiche. Si usano dei semplici **scatter plot** per cercare delle anomalie nella distribuzione delle variabili indipendenti con quelle dipendenti. Se si hanno i punti che non sono distribuiti in maniera uniforme sulla retta di regressione, significa che si ha eteroschedasticità.

Residui

Un altro modo è confrontare i residui stimati con quelli predetti, anche in questo caso usando una rappresentazione grafica. In particolare, si se ha omoschedasticità, i valori sono disposti in maniera omogenea, mentre nella seconda tendono a essere disposti a caso, spesso formando dei *cluster*.

[5] Si denota con un * ad apice il fatto che la variabile sia di tipo eteroschedastico.

Residui al quadrato

Un altro modo, è confrontare i residui al quadrato su quelli predetti. Anche in questo caso, si dovrebbe notare una uniformità se gli errori sono omoschedastici.

Residui su regressori

Ancora, in questo caso si confrontano i residui sulle variabili indipendenti, ovvero i **regressori**, quelle variabili usati per predirre il valore dipendente.

Anche qui, ci si aspetterebbe una distribuzione uniforme dei valori sul grafico, spesso rispetto all'asse delle x (i residui sono sempre distribuiti intorno allo 0).

1.3.3 Test per errori eteroschedastici

Test dt White

In questo, l'ipotesi nulla è che gli errori siano omoschedastici.

Data la regressione campionaria delle **y**, con errori **e**, si va a eseguire una regressione OLS sugli errori al quadrato, con variabili esplicative i **regressori**, i **regressori al quadrato** e i loro prodotti.

$$\hat{\mathbf{e}}\hat{\mathbf{e}}^\top = \gamma\mathbf{X} + \delta\mathbf{X}^\top\mathbf{X} + \mathbf{u} \tag{1.16}$$

Sto difatto andando e vedere se si possa spiegare l'eteroschedasticità come cambiamento al variare delle $\mathbf{X}$.

Difatti, se R^2 è elevato, si ha eteroschedasticità; viceversa, no. Per fare questo, si esegui stetistica test sulla variabile $LM = nR^2$.

Il nostro obiettico è quindi quello di avere un *p-value* alto, in modo tale da non respingere l'ipotesi di omoschedaticità.

Test di Breusch-Pagan

Questo test è simile, solo che divide i valori degli errori per la varianza campionaria. Ovvero si fa una regressione del tipo:

$$\frac{\hat{e_i}^2}{s^2} = f(\mathbf{X}_i) + w_i \tag{1.17}$$

In particolare, non viene definita una forma funzionale, che noi prendiamo lineare, e inoltre si dimostra che si può studiare la regione di accettazione con un **F di Snedecor**, data dal rapporto tra somma quadrata $\gamma\mathbf{X}$ e somma quadrata degli scarti w_i.

1.3.4 Errori autocorrelati

Dato un modello campionario con soluzione OLS $\hat{y}_j = \mathbf{X}_j\hat{\mathbf{B}} + \hat{e}_j$, può capitare che più osservazioni siano correlati tra loro. Per esempio, l'acquisto di un'auto da parte di una famiglia quest'anno è sicuramente influenzato dall'acquisto l'anno precedente.

In questo caso, ogni modello per un'osservazione j-esima avrà gli errori distribuiti come una variabile casuale E_j. Nel caso in cui vi sia correlazione tra le osservazioni, come nell'esempio precedente, si ha che ogni variabile casuale calcolata $\hat{E_j^\#}$ sarà correlata con tutte le altre, ovvero vale:

$$cor\left[\hat{E_j^\#}, \hat{E_k^\#}\right] = \frac{1}{n}\left(\hat{E_j^\#}\left(\hat{E_k^\#}\right)^\top\right) = \rho_{jk} \tag{1.18}$$

Possiamo comunque considerare le osservazioni come omoschedastiche, ovvero $var[\hat{E_j^\#}] = \sigma, \quad \forall j \in [1,n]$.

Nel modello lineare multiplo con errori correlati, si può andare a definire il legale di correlazione come un legame di tipo lineare, in cui le variabili al passo $i+1$-esimo solo legate a quelle al passo i-esimo, con l'aggiunta di un effetto aleatorio:

$$\hat{e}^{\#}_{i+1} = \rho \hat{e}^{\#}_{i} + \eta_i \tag{1.19}$$

dove la componente η_i è indipendente e identicamente distribuita come $\sim N(0, \sigma_n)$. In questo caso, per semplicità abbiamo considerato che le correlazioni siano tutte uguali. L'Eq. 1.19 è nota come **processo autoregressivo**.

Autocorrelazione positiva e negativa

L'autocorrelazione si può definire anche come il variare di y al variare di x con lo stesso segno (**positiva**) o alternata (**negativa**). Questo si considera ovviamente tutto oltre un "certo" intervallo di confidenza.

In termini generali, l'autocorrelazione può essere molto più complicato di quello mostrato: si più avere anche di 2^o, 3^o grado e così via. Ovvero significa avere legami oltre che diretti, i–$i+1$, anche più lontano, come i–$i+2$, etc.

Proprietà

Ovviamente valgono ancora **correttezza** e **linearità**, che non dipendon in nessun modo dalla correlazione.

Anche in questo caso, come per l'eteroschedasticità, non si ha più la migliore efficienza per gli stimatori, ovvero *BLUE*:

$$var[\hat{\mathbf{B}}] = \ldots = \left((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\right) E[\mathbf{e}^{\#}\mathbf{e}^{\#\top}] \left((\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top\right)^\top \tag{1.20}$$

Inoltre, non si può usare lo stimatore s^2 per la **t di Student** perché il valore atteso non sarebbe corretto, e.g. con ρ positiva si andrebbe a sottostimare σ^2. Si otterebbe quindi, come nell'omoschedasticità, che si avrebbe una regione di rifiuto più grande di quanto non lo sia. Analogamente anche per il **test F**, che risulterebbe attendibile.

1.3.5 Individuare errori autocorrelati graficamente

Scatter plot di y su x

Anche in questo caso, si vede a occhio un'autocorrelazione se vi sono dei particolari andamenti. Per esempio, la presenza di un "andamento simile a un cardiogramma". Le autocorrelazioni di grado più alto sono difficili da vedere, sebbene si possano comunque notare.

Residui su variabile esplicativa

Anche in questo caso, si notano degli andamenti con oscillazioni.

Residui su residui ritardati

Senza correlazione, gli elementi sarebbe tutti sulla diagonale. Quando questo non accade, ovviamente vi è autocorrelazione.

Correlogramma

Il grafico migliore è **correlogramma**, che permette di individuare il dubbio di individuare le correlazioni di grado superiore. Questo va a sezionare, tramite calcoli complessi, delle correlazioni a tutti i gradi.

1.3.6 Test per l'autocorrelazione

Test di Durbin-Watson

Come ipotesi nulla, si prende che $\rho = 0$. Si prende l'equazione del processo autoregressivo, Eq. 1.19, e si esegue statistica su essa. Si ottiene quindi:

$$DW = \frac{\sum_i \left(e_i^{\#} - e_{i-1}^{\#}\right)^2}{\sum_i (e_i^{\#})^2} = \frac{E[(e_i^{\#})^2] + E[(e_{i-1}^{\#})^2] - 2E[(e_i^{\#}), (e_{i-1}^{\#})]}{E[(e_i^{\#})^2]} \tag{1.21}$$

Considerando che i residui sono omoschedastici, vale che $E[(e_i^{\#})^2] = E[(e_{i-1}^{\#})^2] = \sigma^2$. Inoltre, poiché le correlazioni sono prese tutte uguali, vale anche $E[(e_i^{\#}), (e_{i-1}^{\#})] = E[(\rho e_{i-1}^{\#} + \eta_i), (e_{i-1}^{\#})] = \rho\sigma^2$. Segue quindi:

$$DW = \frac{2\sigma^2 - 2\rho\sigma^2}{\sigma^2} = 2(1-\rho) \tag{1.22}$$

Da questo, si ha che la distribuzione campionaria di DW è centrata su 2. Non sono mai definite delle regioni di accettazione o rifiuto, ma si usa solamente *best practice*: se $DW < 1$ c'è autocorellazione positiva, se $DW > 3$ si considera autocorrelazione negativa.

1.3.7 Regressioni per violazioni

Stimatore WLS

Viene usato in presenza di **eteroschedasticità** e l'obiettivo è quello di tornare in una situazione di **omoschedaticità**. Si andrà quindi a dividere tutti gli elementi per lo scarto quadratico della varianza campionaria:

$$y_i' = \frac{y_i}{s_i} \tag{1.23}$$

Questo vale anche per variabili esplicative che errori. Si ottiene quindi un nuovo modello, del tipo:

$$\mathbf{y}' = \mathbf{X}'\mathbf{B} + \mathbf{e}' \tag{1.24}$$

La varianza del nuovo modello sarà

$$var[e_i'] = \frac{var[e_i]}{var[s_i]} = \frac{s_i}{s_i} \sim const.$$

Si giunge dunque a dei **minimi quadrati pesati**, pesati sulla minima varianza eteroschedastica.

Si deve notare che in alcuni casi, quando il software potrebbe portare a delle varianze negative, si utilizza una funzione esponenziale per fare il fitting.

Stimatore GLS con correlazioni uguali

Nel caso più generale di **errori correlati** e **eteroschedastici**, si usa un'equazione che tiene conto dell'autocorrelazione seriale venne proposta da Durbin, e nota come *General Least Square*.

Studiamo il caso in cui esiste una correlazione fissa, per prima cosa, ovvero $\rho_{ij} = \rho_{ji} = \rho$, $\forall i, j$.

Si costruisce un modello lineare semplice con errori correlati al tempo t (sebbene si possano considerare anche modelli spaziali).

$$y_i = b_0 + b_1 x_i + e_i^{\#} \tag{1.25}$$

dove gli errori seguiranno dalla **stima di correlazione del primo ordine** ρ:

$$e_i^{\#} = a_0 + a_1^{\#} x_1 + \ldots + a_k^{\#} x_k + \rho e_{i-1}^{\#} \tag{1.26}$$

Si definisce poi un'**equazione ritardata** del tipo:

$$\rho y_{i-1} = \rho b_0 + \rho b_1 x_{i-1} + \rho e^{\#}_{i-1} \tag{1.27}$$

Sottraendo alla prima, Eq. 1.25, si ottiene:

$$y_i - \rho y_{i-1} = b_0(1-\rho) + b_1(x_i - \rho x_{i-1}) + w_i$$

dove si è indicato $w_i = e^{\#}_i - \rho e^{\#}_{i-1}$. Questo si può fare poiché gli errori sono una sorta di "cestino", in cui mettere tutte le variabili aleatorie della regressione.

Chiamando $y^{\#}_i = y_i - \rho y_{i-1}$, $b^{\#}_0 = b_0(1-\rho)$ e $x^{\#}_i = (x_i - \rho x_{i-1})$, si ottiene un'**equazione OLS con errori incorrelati**. Infatti:

$$E[w_i] = E[e^{\#}_i - \rho e^{\#}_{i-1}] = 0$$

$$\begin{aligned} cov[w_i, w_{i-1}] &= cov[e^{\#}_i - \rho e^{\#}_{i-1}, e^{\#}_{i-1} - \rho e^{\#}_{i-2}] = \\ &= cov[e^{\#}_i e^{\#}_{i-1} - \rho e^{\#}_i e^{\#}_{i-2} - \rho e^{\#}_{i-1} e^{\#}_{i-1} + \rho^2 e^{\#}_{i-1} e^{\#}_{i-2} = \\ &= \rho - \rho^3 - \rho + \rho^3 = 0 \end{aligned} \tag{1.28}$$

Ovvero si è dimostrato come questi nuovi errori siano incorrelati.

Eseguendo quindi un'equazione dei minimi quadrati, si risolve il problema.

Modello autoregressivo di SAS

Attraverso una procedura computazionale, si introduce nella struttura del modello una parte detta di **errore ritardato**, che tiene conto dell'autocorrelazione di primo ordine, ovvero come:

$$y_i = b_0 + b_1 x_i + AR_i + e_i \tag{1.29}$$

Questo viene calcolato in termini automatici dal modello, e pone come vantaggio l'introduzione anche errori di secondo, terzo ordine e così via.

In presenza di serie storiche con correlazione di ordine più grande, il problema diviene risolvibile solo computazionalmente, come con SAS.

Stimatori GLS per errori non sferici

Il caso più generale è quello di errori che sono eteroschedastici e correlati. Ovvero, la matrice varianza-covarianza degli errori è del tipo:

$$\Sigma_e = \begin{bmatrix} \sigma^2_1 & \rho_{12} & \dots & \rho_{1k} \\ \rho_{12} & \sigma^2_2 & \dots & \rho_{2k} \\ \vdots & & \ddots & \\ \rho_{1k} & \rho_{2k} & \dots & \sigma^2_k \end{bmatrix}$$

Quando ho dei risultati dati da determinazioni campionarie del tipo e^*_i, e^*_{i-1}, se questi sono correlati al primo grado avrò che vale la relazione

$$cov[e^*_i, e^*_{i-1}] = \rho_{i,i-1} \neq \rho_{i-1,i-2} = cov[e^*_{i-1}, e^*_{i-2}] \tag{1.30}$$

Entrambe vengono da correlazioni di primo grado, ma non per forza uguali: è il caso più generale di tutti.

Questo tipo di interpretazione può valere sia per casi *temporali*, sia per casi *spaziali*. Si ricordi che la determinazione degli errori e^*_j è di tipo **campionaria**.

Il modo cerca di eliminare ogni correlazione ed eteroschedasticità. Si cerca quindi di trasformare le variabili, e ottenere un problema OLS. Partendo dal modello campionario

$$\mathbf{y} = \hat{\mathbf{B}}^o \mathbf{X} + \hat{\mathbf{e}}^o \tag{1.31}$$

La matrice di varianza-covarianza si scrive come:

$$S_{\hat{e}^o} = \frac{1}{n}(\hat{\mathbf{e}}^o)(\hat{\mathbf{e}}^o)^\top$$

Si **ipotizza** che esista una matrice per cui vale $S_{\hat{e}^o} = \sigma^2 V V^\top$. Si sta eseguendo una **scomposizione spettrale** della matrice $S_{\hat{e}^o}$ in due matrice identiche e trasposte. Si definisce quindi la matrice degli errori trasformati come:

$$\hat{\mathbf{e}} = V^{-1}\hat{\mathbf{e}}^o \tag{1.32}$$

Si dimostra che questi errori sono omoschedastici e incorrelati:

$$V^{-1} S_{\hat{e}^o} (V^{-1})^\top = V^{-1}\sigma^2 V V^\top (V^{-1})^\top = \sigma^2 \tag{1.33}$$

Si ottiene quindi l'equazione per il modello moltiplicando V^{-1} all'Eq. 1.31

$$V^{-1}\mathbf{y} = \mathbf{y}^o = V^{-1}\mathbf{X}\mathbf{B} + V^{-1}\hat{\mathbf{e}}^o = \mathbf{X}^o\mathbf{B} + \hat{\mathbf{e}} \tag{1.34}$$

La risoluzione di questa porta a trovare lo stimatore

$$\begin{aligned}\hat{\mathbf{B}}^o &= \left[(\mathbf{X}^o)^\top \mathbf{X}^o\right]^{-1} (\mathbf{X}^o)^\top \mathbf{y}^o = \\ &= \left[\mathbf{X}^\top (V^{-1})^\top (V^{-1})\mathbf{X}\right]^{-1} \mathbf{X}^\top (V^{-1})^\top (V^{-1})\mathbf{y} = \\ &= \left[\mathbf{X}^\top (S_{\hat{\mathbf{e}}^o})^{-1}\mathbf{X}\right]^{-1} \mathbf{X}^\top (S_{\hat{\mathbf{e}}^o})^{-1}\mathbf{y}\end{aligned} \tag{1.35}$$

dove le variabili σ^2 si semplificano con le due sostituzioni.

Questo risolve completamente il nostro problema, in quanto l'Eq. 1.35 risolve un problema OLS. Infatti, la nuova matrice di varianza-covarianza risulta:

$$\Sigma_{\mathbf{y}^o} = (\mathbf{B}^o)^\top \Sigma_{\mathbf{X}^o} \mathbf{B}^o + S_{\hat{\mathbf{e}}} = (\mathbf{B}^o)^\top \Sigma_{\mathbf{X}^o} \mathbf{B}^o + \sigma^2 \mathbf{I}_n$$

dove con $\mathbf{I}_n$ si indentifica la matrice identità $n \times n$.

La domanda è se esista effettivamente la matrice V come operato. Per la proprietà della decomposizione spettrale, si può definire $V = \sigma A L^{\frac{1}{2}} A^\top$, dove A è la matrice degli autovettori e L è la matrice diagonale degli autovalori di $S_{\hat{\mathbf{e}}^o}$. Da cui:

$$V V^\top = \sigma A L^{\frac{1}{2}} A^\top (\sigma A L^{\frac{1}{2}} A^\top)^\top = \sigma^2 A L A^\top = S_{\hat{\mathbf{e}}^o} \tag{1.36}$$

per l'ortogonalità degli autovettori, $A^\top A = \mathbf{I}$ e $(L^{\frac{1}{2}}) = L^{\frac{1}{2}}$.[6]

[6] Mi sembra che questo valga, ma non ne sono sicuro.

Proprietà stimatore GLS

Questo stimatore è **corretto**, infatti:

$$\mathbf{B}^o = \left[(\mathbf{X}^o)^\top \mathbf{X}^o\right]^{-1} (\mathbf{X}^o)^\top \mathbf{y}^o = \left[(\mathbf{X}^o)^\top \mathbf{X}^o\right]^{-1} (\mathbf{X}^o)^\top (\mathbf{X}^o \mathbf{b}^o + \mathbf{e}^o) \tag{1.37}$$

$$E[\mathbf{B}^o] = E\left[\left((\mathbf{X}^o)^\top \mathbf{X}^o\right)^{-1} (\mathbf{X}^o)^\top (\mathbf{X}^o \mathbf{b}^o + \mathbf{e}^o)\right] = \mathbf{b}^o + \left((\mathbf{X}^o)^\top \mathbf{X}^o\right)^{-1} (\mathbf{X}^o)^\top E[\mathbf{e}^o] = \mathbf{b}^o \tag{1.38}$$

Si dimostra anche che lo stimatore risulta **consistente**.

Vale inoltre un analogo al teorema di Gauss-Markov, noto come **teorema di Aitken**. Ovvero:

$$E[(\mathbf{B}^o - \mathbf{b}^o)(\mathbf{B}^o - \mathbf{b}^o)^\top] = \ldots = \sigma^2 \left((\mathbf{X}^o)^\top \mathbf{X}^o\right)^{-1} = \sigma^2 \left(\mathbf{X}^\top \Sigma_{\mathbf{e}^o}^{-1} \mathbf{X}\right)^{-1} \tag{1.39}$$

ovvero una formula analoga a quella degli stimatoro OLS.

In questo caso, non è però né VUE né BLUE. Però, all'interno è comunque quello caratterizzato a minima varianza, dopo aver eseguito le trasformazioni.

Stimatore FGLS

In questo caso, non si ha la conoscenza della matrice di varianza-covarianza. Ovvero, si devono usare le stime campionare per la matrice, $S_{\mathbf{e}^o} \xrightarrow[N\to+\infty]{} \Sigma_{\mathbf{e}^o}$. Noto come *Feasible General Least Squares*. In questo modo, si può tornare a errori sferici qualsiasi sia la situazione di partenza.

1.4 Multicollinearità

1.4.1 Multicollinerarità-OLS

In questo problema, si va a vedere la **matrice delle correlazioni**, e si va a considerare le variabili con una correlazione maggiore di 0.90. In questo caso, non stando attenti, un normale modello OLS potrebbe dare risultati molto poco attendibili, ed è meglio non considerare tali variabili correlate come *variabili esplicative*.

Definizione

Se una variabile è funzione lineare delle altre, si dice che essa è collineare. In particolar modo, quando una variabile è collineare con più variabili si dice che è **multicollineare**.

Quando si ha anche solo una situazione di "quasi"-multicollinearità, gli elementi diagonali di $(\mathbf{X}^\top \mathbf{X})^{-1}$ diventano molto grandi, e il suo determinante molto piccolo. Aumentano quindi le varianze delle stime dei coefficienti, che di conseguenza diventano instabili.

Inoltre, diventa anche poco attendibile stime sia t che F: aumenta la variabile che il parametro sia statisticamente non significativo, anche se non lo è.

Quando si ha multicollinearità, ovvero con combinazioni di più variabili esplicative, diventa difficile da individuare il fenomeno tramite la matrice delle correlazioni – che invece è utile per la semplice correlazione.

1.4.2 Indici

Indice di tolleranza

Il primo indice che viene usato per vedere la presenza o meno di multicollinearità è nodo come **indice di tollerenza**, compreso tra 0 e 1. Indica il grado di interrelazione di una variabile esplicativa rispetto alle altre.

$$Tol(z_j) = 1 - R^2(x_j | x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p) \tag{1.40}$$

Ovvero si calcola l'R^2 per la regressione della variabile "sospetta" rispetto a tutte le altre variabili. Quando la variabile è multicollineare, si avrà un valore di 0.

Varianza multifattoriale

Questa non è nient'altro che $Vif = \frac{1}{Tol}$, ovvero l'inverso dell'indice di tolleranza. Si considera che si ha multicollinearità quando l'indice è superiore a 10.

Indice di condizione

Un indice più raffinato degli altri due, è dato dal rapporto tra l'autovalore massimo della matrice $\mathbf{X}^\top\mathbf{X}$ e ogni autovalore.[7]

Quando questo indice è maggiore di 10, si indica collinearità.

1.4.3 Verificare Linearità

Ci sono alcune violazione del modello lineare classico molto ricche di conseguenze, come eteroschedasticità, correlazione, che inficiano pesantemente sulla validità del modello – cambio delle distribuzioni t e F, oltre che aumento della regione di accettazione. Nel momento in cui si verificasse che il modello non ha un buon adattamento ai dati, sebbene rispettante tutte le condizioni, potrebbero servire altre cose. Per esempio, l'introduzione di nuove variabili potrebbe cambiare completamente il modello, rendendo dei parametri non significativi, significativi, e viceversa.

Esiste però un fatto intermedio, tra le violazioni e l'introduzione di nuove variabili. Nelle nostre ipotesi abbiamo fatto coincidere la $f(x)$ delle relazione $y = f(x)$ con una funzione lineare. Questo però non è detto che sia valido: sia nei parametri che nelle variabili, ci potrebbe essere un altro tipo di *fit*.

Prima di introdurre nuove variabili, si deve verificare se la relazione sia effettivamente lineare, sia nei parametri che nelle variabili.

- Regressione multipla non lineare nei parametri.

$$y_i = b_0 + b_1 w_i + \log b_2 x_i + \ldots + b_r v_i + e_i$$

- Di grado superiore al 1^o in una variabile esplicativa.

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + \ldots + b_r x_i^r + e_i$$

- Di grado superiore al 1^o in più variabili esplicative.

$$y_i = b_0 + b_1 x_i + b_2 z_i^2 + \ldots + b_r k_i^r + e_i$$

Questi sono solo alcuni dei modelli che si potrebbero costruire, anche spesso con l'uso di altri tipi di funzioni. È evidente che la scelta di quale sia il modello non lineare da adattare non è semplice, con un'infita possibilità di scelta.

Noi andremo a guardare solamente 10–11 delle principali formule, e dei criteri (spesso non intuitivi) per poter decidere quale usare. La statistica moderna fa oggi usi di programmi, e.g. `R` o `SAS`, che riescono a usare un gran numero di simulazioni.

Rappresentazione grafica

Facendo capo a una regressione semplice, con degli **scatterplot** si possono vedere se vi sono degli andamenti non lineari. In particolare, anche in presenza di modelli multipli, è possibile scomporre quest'ultima in tante regressioni semplici (sebbene quella multipla abbia dei parametri leggermente diversi).

Possono essere usati anche grafici come **residui su valori osservati**: in questo caso. una volta interpolata una retta, rimane una struttura che richiama a qualche altra funzione. Si deve quindi osservare una *sistematicità* nei residui della funzione lineare.

[7] Online ho trovato che è il rapporto tra l'autovalore massimo e quello minimo

R^2

Questo parametro, essendo una rappresentazione diretta della bontà del modello, può essere un'ottima modo per vedere se il modello è lineare. Anche se non altissimo, può comunque essere che vi siano altre strutture sottostanti.

F

Anche l'uso di questo parametro può essere utile. Spesso viene usato per l'incongruenza di diversi modelli. In particolar modo, una F non significativa è sicuramente un sintomo di un modello non funzionante.

1.4.4 Funzioni Date

In alcuni casi, si potrebbe già essere a conoscenza, sia per intuitività che per conoscenza, di funzioni da utilizzare.

Funzioni Linearizzabili

Alcune di queste potrebbero essere delle funzioni che, sebbene non prettamente lineari, tramite delle trasformazioni si possono ricondurre alla soluzione di un problema lineare.

Se applicato ai **parametri**, si possono applicare delle diverse funzioni a essi, come per esempio:

$$y_i = b_0 + b_1 x_i + e^{b_2} x_i + \ldots + b_r x_i + e_i$$

in cui si dovrebbe eseguire la semplice trasformazione $b_2' = e^{b_2}$.

Nel qual caso la funzione sia sulle **variabili**, si può operare in maniera analoga, in cui si va a cambiare la funzione. Per esempio:

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + \ldots + b_r x_i^r + e_i = b_0 + b_1 x_i + b_2 k_i + \ldots + b_r w_i + e_i$$

dove sono state eseguite delle semplice sostituizioni.

Funzioni non linearizzabili

In alcuni casi, la funzione data potrebbe non essere intrinsecamente linearizzabile, per cui non esistono delle trasformazioni. Il problema, viene risolto in termini computazionali, con delle approssimazioni. Per esempio:

$$y_i = b_0 + \frac{b_1 x_1}{1 + b_2 x_2 e^{-b_3 x_3}} + e_i$$

in cui si deve approssimare il termine non lineare a diversi gradi, fino a trovare una soluzione. Si cerca in questo caso di **minimizzare la varianza residua** ad ogni passi dell'approssimazione, terminando quando non si hanno delle diminuzioni a gradi successivi. In questo caso, non si applica un metodo analitico ma *iterativo*.

1.4.5 Funzione da individuare

Nella maggior parte dei casi, si deve scegliere quale funzioni usare. Dato il problema come:

$$y_i = f(x_{1i}, \ldots, x_{ki}) + e_i$$

Criterio di scelta

Per poter eseguire la scelta, si va a trovare la funzione che vada a minimizzare lo scarto alla variazione successiva di tutte le variabili esplicative. Ovvero, a ogni ciclo si varia una variabile e si tengono le altre fisse, come:

$$\Delta \mathbf{y} = f(\mathbf{x}_1 + \Delta \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k) - f(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k) + \mathbf{e} \tag{1.41}$$

A seconda di che comportamento si trova, si hanno operativamente delle *best practices*.

Trasformazioni

Esistono due tipi di trasformazioni base:

- Polinomiali in **x**, **y**, in cui o una o entrambe la variabili vengono trasformate con funzioni polinomiali di grado maggiore al primo.
- Trasformazioni logaritmiche, in cui si ha un'approssimazione "percentuale".

Polinomiali in x

In questi casi, si prendono i regressori come potente della unica variabile esplicativa.

Funzioni logaritmiche

Si traformano in logaritmi, e si vede a seconda di caso a caso, in quanto si va a rinunciare al minimo. Tra i diversi casi, sono noti:

- Lineare-log

  $$y_i = b_0 + b_1 \ln(x_i) + e_i$$

  La variazione sarà del tipo:

  $$\Delta y_i \simeq b_1 \frac{\Delta x_i}{x_i}$$

- Log-lineare

  $$\ln(y_i) = b_0 + b_1 x_i + e_i$$

  La variazione è del tipo:

  $$\frac{\Delta y_i}{y_i} \simeq b_1 \Delta x_i$$

- log-log

  $$\ln(y_i) = b_0 + b_1 \ln(x_i) + e_i$$

  Con variazione

  $$\frac{\Delta y_i}{y_i} \simeq b_1 \frac{\Delta x_i}{x_i}$$

In questi casi, spesso nello studio della variazione si usa l'approssimazione $\ln(x+\Delta x) - \ln(x) \simeq \frac{\Delta x}{x}$. Il *coefficiente pendenza* è diverso a seconda dei casi.

1.5 Violazione della normalità

1.5.1 Errori non normali

Sebbene non proprio una "vera" proprietà del modello lineare classico, senza di essa non si potrebbe fare molta della statistica.

Dato un numero finito di campioni della popolazione k, gli errori e_{ij}, dove i indica l'osservazione e j il campione. Ognuno di questi è manifestazione di una variabile casuale E_i. Si hanno quindi n variabili casuali, $E_1, \ldots, E_n$. Se si considera che E_i è normale, ognuno degli errori campionari seguirà una distribuzione Guassiana.

Per campioni grandi, il **teorema del limite centrale** garantisce la normalità. Se però il campione è troppo piccolo, questo non può avvenire.

Conseguenze perdita normalità

Se gli errori non sono normali, i parametri **b** non saranno anch'essi normali, e anche le loro stime **B**.

Di conseguenza, non si possono più fare tutti i test come *t di Student* o *F di Snedecr*. Non si possono inoltre trovare degli intervalli di confidenza per parametri basati sulla variabile standardizzata, e si perde la capacità del campione di rappresentare le popolazione. Le stime OLS non coincidono più a quelle con la *massima verosomiglianza*, e si hanno anche dei problemi computazionali.

Oer gli stimatori OlS non sono più nemmeno corretti e a minima varianza: rimangono corretti e BLUE. Ovvero perdono la proprietà VUE: sono i migliori, ma solo tra i lineari.

1.5.2 Verifica della normalità

Metodi su indici

Sappiamo, dalla letteratura, che quando una distribuzione è normale è **simmetrica**, e di conseguenza **media** e **mediana** coincidono.

Deve anche valere che **simmetria** e **curtosi** siano 0. Infatti, l'indice di simmetria è:

$$S = \frac{E[(\mathbf{X}-\mu)^3]^2}{E[(\mathbf{X}-\mu)^2]^3} \tag{1.42}$$

Il cui valore di aspettazione è $E[S] = 0$. La curtosi è invece:

$$K = \frac{E[(\mathbf{X}-\mu)^4]}{E[(\mathbf{X}-\mu)^2]^2} \tag{1.43}$$

Sotto ipotesi di normalità, si ha che $E[K-3] = 0$.[8]

Boxplot

Un altro metodo è tramite la rappresentazione a **boxplot**. Difatti, si vede graficamente bene se vi è una differenza tra media e mediana, e di conseguenza la presenza di asimettrie.

Distribuzione dei residui

In questo caso, basta vedere se è centrata sullo 0. Ovvero si ha differenza a occhio tra asimettria positiva o negativa.

Distribuzione cumulata sui residui

In questo caso, noi conosciamo già i residui cumulati di una normale. Se la curva empiria coincide con quella "tabulata", tutto okay; altrimento significa che non è presente normalità.

Distribuzione cumulata dei residui su quella delle normale

Si hanno due probabilità, sia su ascisse che ordinate. Se la distribuzione empiria coincide con quella normale, a ogni probabilità coincide la stessa "intensità", o *quantile*, e quindi starebbe sulla diagonale.

Se invece non esiste coincidenza, significa che non vi sarà normalità.

Q-Q plot dei residui

Qua ci sono in ascissa i quantili normali e i residui sulle ordinate. Se i residui dell'empirica sono quelli di una normale, si avranno gli stessi valori delle probabilità cumulate.

Da un punto di vista pratico, è uguale a quello *P-P*, ma con i quantili al posto che le probabilità. Ovvero si avranno solamente diversi assi cartesiani.

[8] Spesso si considera come *curtosi* direttamente $K-3$.

1.5.3 Test non parametrici

Per andare a capire se si è su una distribuzione normale, esistono anche dei test statistici.

Test di Shapiro-Wilk

$$W = \frac{\left(\sum_{i=1}^{n} \beta_i e_{(i)}\right)^2}{\sum_{i=1}^{n} (e_i)^2} \tag{1.44}$$

dove $e_{(i)}$ indica che gli errori sono sommati dal più piccolo al più grande, le β_i sono delle costanti dipendenti da media, varianza e covarianza degli errori. Questi parametri sono costruiti in maniera tale che $W \to 1$ se la distribuzione è normale.

La statistica W è molto asimmetrica e anche valori alti possono portare al rifiuto della normalità.

L'ipotesi nulla H_0 del test è che la distribuzione sia normale. Si deve quindi calcolare il *p-value*, e vedere se sia maggiore o minore del livello di significatività α.

Si noti che in questo test (al contrario tra tutti gli altri proposti), si prende come H_0 che $W = 0$. E quindi si ha che la distribuzione è normale quando si va a *rifiutare* l'ipotesi.

Test di Kolmogorov-Smirnov

Questo test è difatti un'applicazione matematica del Q-Q plot, dopo aver diviso in classi di frequenze l'intervallo di variazione. Si calcola quindi per ogni intervallo la distanza dal valore aspettato per ogni intervallo.

Si usano delle apposite tavole per vedere il *p-value* D, la somma delle differenze in valore assoluto. s

Test di asimmetria

In questo caso, è basato sul calcolo di S, indice di simmetria, Eq. 1.42.

Si prende come H_0 che $S = 0$. Difatti, si vanno a vedere alcune soglie statistiche, determinando il *p-value*.

Test della curtosi

Analogo a quella della statistica, ma con K. Anche qua, si va a vedere la soglia a cui è associato il *p-value*.

Anche qua, si considera come H_0 che la distribuzione sia normale.

1.5.4 Risolvere violazione normalità

Per un campione maggiore di 25, la violazione di normalità di può risolvere tramite le assunzioni del teorema del limite centrale.

Per campioni più piccoli, si devono operare delle trasformazioni, tra cui:

$$z = \log(y)$$

usata per asimmetria positiva, o quando lo scarto quadratico cresce con y.

$$w = y^2$$

usata per asimmetria negativa, o quando lo scarto quadratico è proporzionale a $E[y]$.

$$k = y^{\frac{1}{2}}$$

quando lo scarto quadratico è proporzionale al valore atteso. E Infine

$$v = \frac{1}{y}$$

se lo scarto quadratico cresce significativamente con y.

1.6 Ouliers e valori anomali

1.6.1 Definizione

L'ultima violazione del modello lineare non concerne né le 6 proprietà, né la normalità, ma se non considerato porta a delle stime non realistiche. Vengono definiti come **valori anomali** quei valori che si discostano dall'andamento generale e **punti influenti** quelli che influenzano in misura rilevante le stime. Con *outliers* si intendono entrambi i termini.

La presenza di *outliers* nel computo di metodi OLS, WLS, GLS o FGLS possono portare a un risultato completamente mutato, difatti portando a uno spostamento della retta di regressione. Bastano uno o pochi valori lontano dai valori medi, affinché cambi la somma dei quadrati.

1.6.2 Identificazione grafica

Boxplot

Il boxplot pone agli estremi delle barre, secondo un'interpretazioni "meccanica", quali sono i valori compatibili con la distribuzione. Fuori da questi vengono identificati gli *outlier*, in maniera automatica.

Scatterplot univariato

In questi, guardando alle regressioni semplici, si possono notare la presenza di punti che inficino alla regolarità. Non è comunque un metodo accurato, in quanto le regressioni multiple sono diverse da quelle singole.

1.6.3 Indicatori

Residui standardizzati

Data la matrice **matrice di proiezione**

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \tag{1.45}$$

che indica quanto il modello sia in grado di predirre $\mathbf{y}$; infatti:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} - (1-\mathbf{H})\mathbf{y}$$

dove sono identificati gli errori come $\mathbf{e} = (1-\mathbf{H})\mathbf{y}$, e quindi $var[\mathbf{e}] = \mathbf{y}^\top(1-\mathbf{H})\mathbf{y}$.

L'elemento diagonale h^*_{ii}, diagonale della matrice di proiezione, prende il nome di **leverage**, e indica l'impatto dell'osservazione i-esima sulla capacità del modello di predire la variabile dipendente.

Si può indicare la varianza dell'osservazione i-esima al variare del campione come:

$$var[e_i] = (1-h^*_{ii})\sigma^2$$

da cui si definiscono i residui per l'osservazione come

$$e^*_i = \frac{e_i}{\sigma\sqrt{1-h^*_{ii}}} \tag{1.46}$$

Questo valore, detto appunti **residuo standardizzato**, indica il residuo al netto della varianza assunto in quel punto.

Si può dimostrare che in un campione distribuito normalmente, il 95% del residui standardizzati sono compresi tra -2 e 2, e al 99% tra -2.5 e 2.5. Dunque, un'osservazione con valore del suo residuo standardizzato, significa che esso è un *outlier*.

In generale, contanto i numeri di *outlier*, si ha un indice di bontà del modello.

Residui studentizzati

Quando un campione non ha elevata numerosità, si va a divedere per lo scarto quadratico medio di **ogni** osservazione, al posto che quello generale, ovvero:

$$e_i^* = \frac{e_i}{s_{e,i}\sqrt{1-h_{ii}^*}} \tag{1.47}$$

Si chiamani inoltre i **residui studentizzati jacknife** in cui si usa $s_{e,(i)}$, ovvero dove si usa la deviazione standard ottenuta eliminando la i-esima osservazione, quindi

$$e_i^* = \frac{e_i}{s_{e,(i)}\sqrt{1-h_{ii}^*}} \tag{1.48}$$

Quando lo scarto tra quelli studentizzati e quelli studentizzato *jacknife* è elevato, significa che l'osservazione i-esima è un *outlier*.

Questo approccio può essere eseguito anche graficamente, su un grafico *standard residuals vs leverage*.

Valori di leverage

Si può usare la stessa matrice $\mathbf{H}$ e i suoi valori diagonali per verificare la presenza di *outlier*. Infatti, valori h_{ii}^* vicini a 1 indicano un effetto molto pesante dell'osservazione sul modello – e se qualcosa influisce troppe, signfica che è un *outlier*.

Si identifica come valore soglia per i *leverage* come $2\frac{k+1}{n}$. Anche in questo caso, si può usare un semplice grafico a barre per identificare i valori troppo elevati.

Covrati

Questi indici identificano quanto varia la matrice delle covarianze all'eliminazione dell'i-esima osservazione. Matematicamente, sono identificati come:

$$covratio_i = \frac{\det\left[s_{(i)}^2\left(\mathbf{X}_{(i)}^\top\mathbf{X}_{(i)}\right)^{-1}\right]}{\det\left[s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\right]} \tag{1.49}$$

Si identifica come valore soglia $1 \pm 3\sqrt{\frac{k+1}{n}}$. Se il cambiamento è maggiore di questo valore, significa che vi è anomalia.

Dffits

In questo caso, si misura l'influenza dell'i-esima osservazione sulla stima dei coefficienti di regressione, ovvero:

$$dffits_i = \frac{\hat{y} - \hat{y}_{(i)}}{s_{e,(i)}\sqrt{h_{ii}^*}} \tag{1.50}$$

Dfbetas

Analogamente a sopra, si misura l'influenza dell'i-esima osservazione sui coefficienti di regressione:

$$dfbetas_i = b - b_{(i)} = \mathbf{X}_{(i)}(\mathbf{X}^\top\mathbf{X})^{-1}\frac{e_i}{1-h_{ii}^*} \tag{1.51}$$

Si prende come valore soglia o 2 o $2\sqrt{n}$, dove n è il numero di osservazioni.

Distanza di Cook

Questo elementi guarda tutti i casi, misurando la l'influenza dell'i-esima osservazione sulla stima dei coefficienti, ma in termini di stima generale del modello. Si prende come soglia il valore di 1.

$$D_i = \frac{(b - b_{(i)})(\mathbf{X}^\top \mathbf{X})^{-1}(b - b_{(i)})}{k\sigma^2_{(i)}} \tag{1.52}$$

2. Modello Lineare Multivariato

2.1 Modello Lineare Multivariato Classico

2.1.1 Definizione

Il modello lineare multivariato è caratterizzato da r variabili esplicative indipendenti e m variabili dipendenti. Esso è costruito come:

$$\begin{array}{lllll} \mathbf{y}_1 & = b_{10} & +b_{11}\mathbf{z}_1+ & \dots & b_{1r}\mathbf{z}_r+\mathbf{e}_1 \\ \vdots & & \vdots & \vdots & \\ \mathbf{y}_m & = b_{m0} & +b_{m1}\mathbf{z}_1+ & \dots & b_{mr}\mathbf{z}_r+\mathbf{e}_m \end{array} \tag{2.1}$$

Ognuna delle m variabili dipendenti è legata alla sua particolare regressione lineare multipla: si hanno m equazioni.

In particolare, date n osservazioni, vediamo in che cosa consista la regressione per una di queste. La i-esima osservazione avrà come variabile dipendente un vettore $(m,1)$ del tipo $\mathbf{y}^i = (y_{i1},\dots,y_{im})^\top$; come variabili esplicative un vettore $(r+1,1)$ del tipo $\mathbf{z}^i = (1,z_{i1},\dots,z_{ir})^\top$; e come matrice dei parametri:

$$\mathbf{B}_{(m,r+1)} = \begin{bmatrix} b_{10} & \dots & b_{1r} \\ \vdots & \ddots & \vdots \\ b_{m0} & \dots & b_{mr} \end{bmatrix}$$

Dunque, indicando la regressione in forma stretta come $\mathbf{y} = \mathbf{Bz}+\mathbf{e}$, dove $\mathbf{y}$ è una matrice (m,n), $\mathbf{B}$ è una matrice $(m,r+1)$, $\mathbf{z}$ una matrice $(r+1,n)$ e $\mathbf{e}$ una del tipo (m,n).

Come nella regressione lineare monovariata, si assume che i parametri siano **lineari**, i **valori attesi degli errori** siano **nulli** e che siano **omoschedastici** e **incorrelati** all'interno sia di ogni equazione, sia tra equazioni diverse. Vale inoltre che le $\mathbf{z}$ variabili esplicative **non sono stocastiche** e sono **non collineare**, con rango massimo, ovvero $(r+1)$. Questo è essenziale affinche la matrice $\mathbf{z}^\top\mathbf{z}$ sia invertibile. Per lo stesso motivo deve essere che $n > r+1$.

Dato che ogni j-esima delle m equazioni è una regressione lineare multipla, a ognuna di queste si può associare una matrice di varianza-covarianza (n,n) del tipo Σ_j, dove n indica il

numero di osservazioni. Di conseguenza, il problema multivariato possiede una **tensore** (n,n,m) di varianza-covarianza.

Dunque, le soluzioni si possono trovare in modo analogo a quelle univariate come:

$$\hat{\mathbf{B}} = \mathbf{y}\mathbf{z}^\top(\mathbf{z}^\top\mathbf{z})^{-1} \tag{2.2}$$

Di conseguenza, i valori predetti sono:

$$\hat{\mathbf{y}} = \hat{\mathbf{B}}\mathbf{z} \tag{2.3}$$

Gli errori si definiscono sempre come differenza tra valori predetti e osservati. In questo caso, si può dimostrare che permangono le condizioni di incorrelazione e ortogonalità. Anche in questo caso, il tensore di varianza-covarianza può essere spezzato in una parte di *spiegate* e una parte di *residue*.

La matrice totale di varianza-covarianza, in cui si considerano le varianze-covarianze delle variabili decisionali j-esime Σ_j, può essere vista come un tensore del tipo (n,n,m,m) quadridimensionale. Si può però rappresentare come matrice (m,m) come avente per elementi delle matrici (n,n). In particolare, se valgono tutte le ipotesi di sfericità, la matrice sarà del tipo:

$$\Sigma = \begin{bmatrix} \sigma^2 I_n & 0 & \dots & 0 \\ \vdots & \ddots & & \vdots \\ 0 & \dots & & \sigma^2 I_n \end{bmatrix} \tag{2.4}$$

Questo vale poiché, per ogni i-esima variabile (tra le m), posso identificare errori tra quella e le altre m, ma anche tra quella e tutti le altri $j \in [1,n]$ osservazioni. Si può inoltre ogni errore $\mathbf{e}_{ij}$ come spezzato in tanti errori, ognuno legato a ogni variabile esplicative, tra le r possibili.

Le proprietà del modello multivariato rimangono le stesse del modello univariato, ovviamente con i giusti accostamenti per il fatto che si lavora con più grossi sistemi.

In questo caso, la normalità degli errori (la “settima" proprietà) è implicata con una **distribuzione normale multivariata**, quindi del tipo $E \sim N(0,\sigma^2 I_{nm})$, dove E è la matrice tale che $\Sigma = E^\top E$, degli errori iid.

2.1.2 Inferenza nella regressione multipla

Il valore atteso degli stimatori è:

$$E[\hat{\mathbf{B}}] = E[\mathbf{y}\mathbf{z}^\top(\mathbf{z}^\top\mathbf{z})^{-1}] = E[(\mathbf{B}\mathbf{z}+\mathbf{e})\mathbf{z}^\top(\mathbf{z}^\top\mathbf{z})^{-1}] = \mathbf{B}$$

In maniera analoga al modello univariato, si dimostra che è consistenza ed efficienza. Quest'ultima, in particolare:

$$E[(\hat{\mathbf{B}}-\mathbf{B})^\top(\hat{\mathbf{B}}-\mathbf{B})] = [\dots] = \sigma^2(\mathbf{z}^\top\mathbf{z})^{-1}$$

Si può inoltre dimostrare che rimane valido il **teorema di Gauss-Markov**, e anche in questo caso gli stimatori sono BLUE. Si può inoltre dimostrare, costruendo uno stimatore di Massima Verosomiglianza, che è anche UNVUE.

Dato il modello del tipo $\mathbf{y} = \mathbf{B}\mathbf{z}+\mathbf{e}$, si può dimostrare che, se il rango della matrice $\mathbf{z}$ è pieno, ovvero maggiore di $r+1$, se gli errori $\mathbf{e}$ sono normalmente distribuiti, lo saranno anche le variabili dipendenti, secondo una **normale multivariata**, ovvero $\mathbf{y} \sim N(\mathbf{B}\mathbf{z},\Sigma_y)$, dove Σ_y è la matrice (n,n) varianza-covarianza delle variabili dipendenti; non è da confondere con Σ, matrice varianza-covarianza degli errori. Vale inoltre che anche lo stimatore $\hat{\mathbf{B}} \sim N(\mathbf{B},\hat{\mathbf{H}})$, dove $\hat{\mathbf{H}}$ rappresenta la matrice (n,n) di **varianza-covarianza spiegata** per la regressione OLS. Questa matrice $\hat{\mathbf{H}}$ si

distribuisce come una **variabile casuale di Wishart** con r gradi di libertà (analogo dei chi-quadri univariati). Anche gli errori **e** si distribuiscono normalmente, per richiesta della regressione.

Si possono inoltre definire le **varianze generalizzate di Wilks**, sia per la varianza-covarianza spiegata sia per quella residua, come il determinante della rispettiva matrice di varianza-covarianza. In particolare, si dimostra che questa varianza generalizzata è 0 se il rango della matrice di varianza-covarianza è minore di m; questo ovviamente accade in presenza di collinearità tra le variabili. Il valore massimo che può essere assunot dalla varianza generalizzata è quando essa è il prodotto delle singole varianze delle variabili; nel caso di quella ridotta, significherebbe che $det[\hat{\Sigma}_e] = \prod_{j=1}^{m} \sigma_j^2$

Viste le distribuzioni di $\mathbf{y}$ e $\hat{\mathbf{B}}$ normali, si possono usare **test** su normale o *t di Student*, per verificare la nullità dei vari parametri. Si può, analogamente al modello univariato, anche la *F di Snedecor*. Non essendoci legami tra le diverse equazioni, per ipotesi, basta verificare che valga per ogni equazione indipendemente; lo stesso, purtroppo, non si può sempre dire per la sfericità.

Difatti, a livello base lo studio del modello multivariato risulta semplicemente lo studio delle soluzioni univariate indipendentemente.

2.1.3 Test di Wilks

Questo test viene definito come:

$$\Lambda = \frac{\det\left[\hat{\Sigma}_e\right]}{\det\left[\hat{\Sigma}_e + \hat{\mathbf{H}}\right]} = \prod_i \frac{1}{1+\lambda_i} \tag{2.5}$$

[1] dove $\lambda_1 \geq \ldots \geq \lambda_n$[2] sono gli autovalori non nulli della matrice $(\hat{\Sigma_e})^{-1}\hat{\mathbf{H}}$, distribuiti come lambda con n, m e r gradi di libertà. Spesso, si usa l'**approssimazione di Bartlett**

$$W = -\frac{n-1}{2(m+r+1)} \ln(\Lambda)$$

che viene distribuita asintoticamente come un χ^2 con m, r gradi di libertà. Si dimostra inoltre che $\frac{1-\Lambda}{\Lambda}$ è asintoticamente una *F di Snedecor*.

Difatti, questo test gioca lo stesso ruolo nel multivariato che veniva preso dal *test F* per il modello univariato.

Il primo modo in cui può essere usato, è per l'ipotesi nulla che $\hat{\mathbf{B}} = 0$, ovvero che nessuna delle variabili **z** sia significativa per la variabile dipendente **y**; in particolar modo, il test fallisce se anche solo una delle variabili esplicative riescano a spiegare la variabile dipendente.

Analogamente si possono costruire test per i valori predetti $\hat{\mathbf{y}}$.

2.1.4 Test su altre ipotesi

Non significatività di alcune variabili indipendenti

Si va a vedere se un gruppo di variabili z_i sono non significative nello spiegare **y**. L'ipotesi *alternativa* è che almeno 1 delle variabili nel gruppo sia significativa, e dunque quella nulla è che nessuna sia significativa. Rispetto a quello precedente, non si vanno a vedere tutte le variabili, ma solamente un sottoinsieme.

Tutte le variabili indipendenti rispetto a un gruppo indipendente

Si va a vedere se tutte le z_i sono non-significative (ipotesi nulla) rispetto a un sottogruppo di variabili dipendenti **y**.

[1] Per far valere l'uguaglianza, si deve ricordare che sui reali vale la relazione $\det A = \prod_i \lambda_i$, dove λ_i sono gli autovalori della matrice A. Inoltre, se le due matrici sono "disgiunte", come lo sono le due varianze, allora vale che si può spezzare l'operatore di det lungo le normale operazioni algebriche.

[2] Non sono sicuro che sia n o m.

Gruppo di dipedenti su gruppo di indipendenti

È a metà strada dei precedenti: si va a vedere se un sottogruppo di esplicative sono significative rispetto a un sottogruppo di dipendenti.

Test dell'uguaglianza di due gruppi

Si va a vedere se un gruppo di variabili esplicative è uguale a un altro.

Test dell'uguaglianza tra stesse variabili rispetto a dipedenti

Tutti questi test vengono tutti costruiti tramite delle Λ di Wilks, ma ridotte ai gruppi considerati. Si può il test come:

$$\Lambda_G = \frac{\det\left[\hat{\Sigma}_{e,G}\right]}{\det\left[\hat{\Sigma}_{e,G} + \hat{\mathbf{H}}_G\right]}$$

Altri test

Esistono altri test, tutti approssimati a chi-quadro ed F, noti come *tracci adi Lawley-Hotelling*, *traccia di Pillai*, *massimo autovalore di Roy-Max*.

2.1.5 Alcune note

La varianza dell'errore della variabile $\mathbf{y}_i$, vettore del tipo $(1,n)$, dove n è la grandezza del campione, è definita come:

$$var[\hat{\mathbf{e}}_i] = \frac{1}{n}\sum_{j=0}^{n} e_{ji}^2 = \frac{1}{n}\sum_{j=0}^{n} \sigma_{ji}^2$$

nella situazione di omoschedasticità, vale ovviamente che $\sigma_{1i} = \ldots = \sigma_{ni} = \sigma^2$, per cui $var[\hat{\mathbf{e}}_i] = \sigma^2$.

Per le correlazioni tra le diverse equazioni, varebbe:

$$cor[\hat{\mathbf{e}}_i, \hat{\mathbf{e}}_k] = \frac{1}{n}\sum_{j=0}^{n}\sum_{l=0}^{n} e_{ji}e_{lk}$$

ovviamente, in incorrelazione, vale che $e_{ji}e_{lk} = 0$ per tutte le combinazioni. Nel caso di errori omoschedastici e incorrelati, vale quindi che $var[\hat{\mathbf{e}}] = \sigma^2 I_n$, dove I_n è la matrice indentità (n,n).

2.2 Modelli multivariati generalizzati

2.2.1 Soluzione dei minini quadrati generalizzati multivariati

In questo caso, si va a esprimere le variabili come "super-vettori", ovvero vettori contententi altri vettori, al posto che usare notazione matriciale. Per esempio:

$$\mathbf{B}^o = (\mathbf{b}_1, \ldots, \mathbf{b}_m)$$

dove $\mathbf{b}_i$ è un vettore del tipo $(1,m)$. Questa notazione si estende di conseguenza anche alla matrice di varianza-covarianza, che sarà una matrice (m,m) contente matrici (n,n) al suo interno.

Si definiscono gli errori trasformati come $\mathbf{e}^o = \mathbf{e}^{o,*}W^{-1}$, si dimostra che si ottengono degli errori trasformati sferici, omoschedastici e incorrelati rispetto a ogni variabile dipendente e fra variabili dipendenti. Applicando questa trasformazione al modello, si ottiene:

$$\mathbf{y}^{0,*} = \mathbf{y}^o W^{-1} = \mathbf{B}^{o,*}\mathbf{z}^{0,*} + \mathbf{e}$$

La soluzione che si ottiene, in termini di "super-vettori", è:

$$\mathbf{b}_j^* = \mathbf{y}_j \Sigma_{\mathbf{e}^*}^{-1} (\mathbf{z}^o)^\top (\mathbf{z}^o \Sigma_{\mathbf{e}^*}^{-1} (\mathbf{z}^o)^\top)^{-1}$$

ogni singola soluzione tiene conto del covariare, e quindi si tengono conto delle differenze anche tra le variabili. Per ottenere W, si esegue una tecnica analoga al caso univariato, ovvero tramite composizione spettrale.

Si ottengono quindi molteplici forme di varianza-covarianza degli errori. In casi sono, **incorrelati e omoschedastici**, tra tutto e tutti. Un caso intermedio può essere il caso in cui gli errori sono **omoschedastici nella stessa equazioni**, ma sono **eteroschedastici tra equazioni diverse**; tutti gli individui si comportano uguali, ma solo rispetto alla stessa variabile dipendente, ovvero si comportano diversamente per ognuna di esse. Questo secondo tende a essere un modello più realistico. Esiste poi il modello più generale, in cui si ha eteroschedasticità sia nella stessa equazione, che tra equazioni diverse. Ovviamente casi più generali possiedono correlazione intra e fra.

Soluzioni FGLS

Spesso, si ha a disposizione solamente la matrice di varianza-covarianza del campione, che tende a quella della popolazione per un campione infinito. Spesso è questo il tipo di soluzione che si possiede.

2.2.2 Modelli SURE

I modelli *seemingly uncorrelated regression equations* sono caratterizzati da variabili esplicative non uguali tra le diverse equazioni. Avere un set completo uguale di variabili esplicative è infatti una visione molto rigida, che spesso non si accosta a quelli che sono i fenomeno reali. Poiché i coefficienti di regressione sono al netto di tutte le altre variabili esplicative, non sempre usare tutte le variabili esplicative è utile.

Per risolvere il problema, si usa come metolodia l'applicazione dei "super-vettori", ovvero si esprimono le variabili dipedenti come un vettore contente vettori:

$$(\mathbf{y}^o)^\top = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_m \end{bmatrix}$$

Per le variabili esplicative, pongo come elementi della matrice $\mathbf{z}^o$ per ogni elemento le variabili esplicative per la m-esima equazione. ovvero:

$$(\mathbf{z}^o)^\top = \begin{bmatrix} \mathbf{z}_1 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \dots & 0 & \mathbf{z}_m \end{bmatrix}$$

dove ogni vettore $\mathbf{z}_i$ contiene tutte le variabili esplicative dell'equazione i-esima. Si avraano quindi in totale $\sum_i r_i$ variabili esplicative, dove r_i è il numero di variabili esplicative per l'equazione i-esima. Di conseguenza, di avranno $\sum_i r_i$ parametri da stimare, ovvero $\mathbf{B}^o$ ha dimensione $(1, \sum_i r_i)$.

Questo modello su può costruire anche con errori omoschedastici e incorrelati, oppure eteroschedastici o anche altre situazioni. Il modello sure più generale usa una matrice di varianza-covarianza per gli errori che è il più generico possibile

2.2.3 Scelta del modello

Per prima cosa, guardando la matrice di varianza-covarianza si sceglie tra un modello classico oppure uno più generalizzato. Viceversa, per ogni equazione si possono togliere le variabili che in un modello non sono significative, andando a usare una stima di tipo SURE. Si cerca poi di migliorare R^2 corretto dal numero di regressori, equazione per equazione.

La scelta di ipotesi, specificazione di modelli e di variabili non è qualcosa di meccanico, ma richiede intuizione ed esperienza, oltre che creatività, fantasia e capacità di leggere la realtà.

3. Regressioni multilevel

3.1 Struttura dei dati gerarchici

Nata poco più di 30 anni, si basa su analisi contestuale, ovvero l'effetto del contesto sociale sul comportamento individuale; e su modelli ad effetti misti, in cui si hanno dei coefficienti di regressione fissi e altri casuali, ovvero la parte stocastica della regressione si ritrova anche nei coefficienti. Rispetto alla tradizionale ricerca, un'osservazione diviene più indipendente e si deve porre importanza su aspetti intermedi, che possono modificare i risultati finali sebbene non si abbia un effetto diretto. È come se andassimo a vedere che cosa significhi appartenere a un certo gruppo, e come questo possa influire sul risultato. Si arriva a studiare i singoli gruppi, con modelli che sono espensione di quelli visti.

Questi scopi sono innestati sul fatto che i dati sono **nidificati**: appartenere a un certo gruppo diventa di fondamentale importanza all'analisi, e non semplicemente come elemento su cui regredire (magari confrontando i gruppi). Si condiziona l'analisi alla presenza dei ordini gerarchici, ovvero gruppi e sottogruppi. A livello descrittivo, ignorando le caratteristiche gerarchiche potrebbe portare e dei risultati completamente diversi; questo tipo di analisi può essere di fondamentale importanza per studi sui medici o sui farmici.

Se si andassero ad aggregare dati micro a livello macro, si avrebbe la cosiddetta **fallacia ecologica**: una correlazione a livello macro non può essere usata per fare asserzioni riguardo a relazione di tipo micro. Viceversa, disaggredando dati macro a livello micro dimenticando la divisione dei gruppi, si avrebbe la **fallacia atomistica**: non si possono usare correlazioni tra variabili a livello micro per fare delle asserzioni a livello macro.

In questi casi, le osservazioni non è detto che abbiano la stessa distribuzione e, quindi, la stessa probabilità di essere estratte tramite *campionamento semplice*; si deve quindi eseguire un campionamento a più stadi, che tenga conto del condizionamento dovuto all'appartenenza a dei gruppi. L'indipendenza della distribuzione può valere dentro i gruppi, in cui si hanno effetti locali; ma non vale per tutto l'universo.

3.1.1 Regressione mutlilevel

Questa regressione lineare tiene in considerazione della presenza di dati gerarchici. Si indica dunque sia l'individuo i-esimo sul campione, sia il gruppo j-esimo.

$$y_{ij} = b_0 + b_1 x_{ij} + e_{ij}$$

ovvero le regressioni hanno un fattore che dipende dall'appartenenza a un certo gruppo.

Si possono avere diversi tipi di relazioni, in questo senso. Una **relazione disaggregata**, in cui si ignora il raggruppamento tra le unità; oppure una **relazione aggregata fra i gruppi**: si prendono le medie dei valori per gruppi, ignorando le unità all'interno dei gruppi. In questo secondo caso, si fa la regressione sulle medie, e quindi:

$$\mu(y)_j = b_0^* + b_1^* \mu(x)_j + e_j \tag{3.1}$$

in cui ovviamente il risultato è dipende dal tipo di regressione scelto, e.g. una normale OLS.

Si può avere però anche la un modello in cui si considerano le **relazioni entro ciascun gruppo**, ovvero si eseguono le regressioni entro ciascun gruppo considerando le deviazioni dalle medie, ovvero

$$y_{ij} - \mu(y)_j = b_1^o (x_{ij} - \mu(x)_j) + e_j^o \tag{3.2}$$

In questo caso ovviamente il termine noto è nullo, in quando la media delle deviazioni è nulla $\mu(y_{ij} - \mu(y)_j) = 0$.

Si può avere poi una **relazione multilelve**, in cui si scompone sia tra i gruppo che entro i gruppi.

$$y_{ij} = b_o^+ + b_1^+ (x_{ji} - \mu(x)_j) + b_2^+ \mu(x)_j + e_j^+ \tag{3.3}$$

In questa, come si può vedere, si considerano entrambi gli effetti. Anche qui, i parametri b_0^+, b_1^+ e $b + 2^+$ sono stimati come OLS. I risultati di questi differenti approcci possono essere, in alcuni casi, completamente diversi tra di loro. Si capisce quindi l'importanza di studiare questi fenomeni con attenzione.

Si definosce in questo senso il **modello di Cronbach**, inc ui si distinguono gli effetti, come:

$$y_{ij} = \alpha + \beta_{within}(x_{ij} - \bar{x}_{.j}) + \beta_{between}\bar{x}_{.j} \tag{3.4}$$

Si definisce **effetto contestuale** come $\delta = \beta_{between} - \beta_{within}$, che identifica quanto i fenomeni micro siano pesanti a livello macro, ovvero l'effetto della variabilità tra di gruppi al netto di quella nei gruppi. La differenza tra le medie delle variabili dipendenti tra i gruppi può essere diversa, a seconda dei fenomeno che si considerano.[1]

Quanto dettagliato sopra, può essere esteso in presenza di più variabili esplicative x_k, ovvero come:

$$y_{ij} = \beta_0 + \sum_k \left[\beta_{k,(within)}(x_{ij,k} - \mu(x)_{j,k}) + \beta_{k,(between)}\mu(x)_{j,k}\right] + e_{ij} \tag{3.5}$$

In termini matriciale, si può esprimere come $\mathbf{y} = \mathbf{x}^@\mathbf{B}^@ + \mathbf{e}$, dove la matrice $\mathbf{x}^@$ è del tipo $(n, 2r+1)$ e $\mathbf{B}^@$ è $(2r+1, 1)$. La solita soluzione dei minimi quadrati sarà in questo causa/effetto:

$$\mathbf{B}^@ = ((\mathbf{x}^@)^\top \mathbf{x}^@)^{-1}(\mathbf{x}^@)^\top \mathbf{y} \tag{3.6}$$

[1] Vedi slide per esempi dettagliati.

3.1.2 Analisi Anova

Per comprendere i modelli multileve, si deve partire dall'analisi della varianza e della covarianza.

Nella sua versione descrittiva e non gerarchica, essa viene espressa come:

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

dove γ_{00} è la media di y su tutta la popolazione e u_j è la differenza tra la media generale e quella del gruppo j-esimo. Di conseguenza, si identifica con $\gamma_{00} + u_j$ la media relativa al gruppo j. e_{ij} risulta il residui relativo alla variazione individuale. Si va a vedere come la variabile dipendente cambi per effetto dell'appartenenza ai diversi gruppi e a fluttuazioni individuali.

Chiamando $n = n_1 + \ldots + n_p$ la numerosità del campione totale e n_j la numerosità dle gruppo j-esimo, posso definire:

$$\mathbf{y}^\top = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_p \end{bmatrix} \qquad \mathbf{u}^\top = \begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix} \qquad \mathbf{E}^\top = \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_p \end{bmatrix}$$

In questo caso, ogni vettore $\mathbf{y}_j$ ($\mathbf{e_j}$) è di dimensione $(n_j, 1)$. Si può definire una matrice di appartenenza $\mathbf{A}$ di dimensione (n, p), costruita tale che valga $\mathbf{A}_j\mathbf{u} = \mathbf{u_j}$, dove $\mathbf{A}_j$ è una matrice del tipo (n_j, p) e $\mathbf{u_j}$ è una matrice $(n_j, 1)$ contente tutti u_j nelle righe. In questo modo, posso scrivere l'equazione delle regressione come:

$$\mathbf{y}_j - \gamma_{00} = \mathbf{A_j u} + \mathbf{e}_j \tag{3.7}$$

Considerando tutti i gruppi messi insieme, si esprime come:

$$\mathbf{y} - \gamma_{00} = \mathbf{Au} + \mathbf{e} \tag{3.8}$$

In cui le matrici $\mathbf{y}$, $\mathbf{Au}$ e $\mathbf{e}$ sono del tipo (n, p).

Con questo artificio, sono riuscito a esprimere in termini di modello lineare l'Eq. **??**. Da questa, si può ottenere la **devianza totale** come:

$$SST = (\mathbf{y} - \gamma_{00})^\top (\mathbf{y} - \gamma_{00}) = \mathbf{u}^\top \mathbf{A}^\top \mathbf{Au} + \mathbf{e}^\top \mathbf{e} \tag{3.9}$$

dove si identificano le due componenti di **devianza tra i gruppi** $SSF = \mathbf{u}^\top \mathbf{A}^\top \mathbf{Au}$ e la **devianza nei gruppi** $SSE = \mathbf{e}^\top \mathbf{e}$. Si ottiene quindi la nota relazione:

$$SST = SSF + SSE \tag{3.10}$$

Nell'ipotesi di **omoschedasticità**, in cui la varianza di ogni errori è pari a σ^2, si può definire $\mathbf{e}^\top \mathbf{e} = n\sigma^2$ e $\mathbf{u}^\top \mathbf{A}^\top \mathbf{Au} = n\tau^2$. Si definisce quindi il **coefficiente di correlazione interclasse** come:

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2} \tag{3.11}$$

che indica quanta varianza, rispetto a quella totale, è dovuto a quella tra i gruppi.

3.1.3 Analisi della Covarianza

Il limite dell'analisi della varianza si vede quando gli appartenenti a gruppi diversi sono differenti tra loro: non si riesce tenere conto di questo effetto. L'analisi della varianza mostrerebbe che c'è una forte variabilità dovuta ai gruppi, ma in realtà è presente omogeneità dentro i gruppi, ma semplicemente sono diversi tra loro.

Si deve quindi fare analisi della varianza, al netto delle caratteristiche degli individui di ogni gruppo. Ovvero l'**analisi della covarianza**, in cui si applica un modello lineare extra:

$$y_{ij} = (\beta_0 + \sum_k \beta_{kj} x_{kij}) + \gamma_{00} + v_j + e_{ij} \tag{3.12}$$

Ho quindi un modello lineare e una normale analisi della varianza.

In termini matriciali,

$$\mathbf{y} - \gamma_{00} = \mathbf{xB} + \mathbf{Av} + \mathbf{e} \tag{3.13}$$

Il residuo del modello lineare è difatto la variabili dipendente dell'analisi della varianza. Sto facendo l'analisi della varianza su ciò che di $\mathbf{y}$ non ho spiegato in termini di modello lineare. Si può esprimere come $\mathbf{y} - \gamma_{00} = \mathbf{xB} + \mathbf{y}_e$. Le soluzioni OLS di questo modello sono:

$$\hat{\mathbf{B}}^\top \mathbf{x}^\top \mathbf{x} \hat{\mathbf{B}} = SSX \tag{3.14}$$

nota come **devianza spiegata di regressione**.

Si può quindi andare a risolvere l'equazione:

$$(\mathbf{y} - \gamma_{00}) - \mathbf{x}\hat{\mathbf{B}} = \hat{\mathbf{y}}_e = \mathbf{Av} + \mathbf{e} \tag{3.15}$$

Purtroppo, non si può applicare in questo caso un normale modello OLS, in quanto non vale che la matrice $\mathbf{A}$ sia di rango pieno. Un unico metodo non esiste, tanto che è ancora attivo dibattito in letteratura; il metodo che noi vediamo è quello di porre dei vincoli sui parametri. In particolare, se il rango è $p-1$ si pone:

$$\sum_j u_j = 0$$

nel caso in cui il rango sia inferiore, ci sono altri vincoli ancora. Difatti, si vanno ad eliminare elementi della matrice $\mathbf{A}$ per trovare una matrice di rango pieno $\mathbf{A}^o$ di dimensione $(n, p-1)$, per cui il modello diviene:

$$\hat{\mathbf{y}}_e = \mathbf{A}^o \mathbf{v}^o + \mathbf{e}$$

che possiede soluzione OLS.

Unendo i risultati fin qui ottenuti, il modello di **varianza-covarianza** è:

$$SST = (\mathbf{y} - \gamma_{00})^\top (\mathbf{y} - \gamma_{00}) = \hat{\mathbf{B}}^\top \mathbf{x}^\top \mathbf{x} \hat{\mathbf{B}} + (\hat{\mathbf{v}}^o)^\top (\mathbf{A}^o)^\top (\mathbf{A}^o)(\hat{\mathbf{v}}^o) + \mathbf{e}^\top \mathbf{e} \tag{3.16}$$

dove si identificano le tre parti $SSX = \hat{\mathbf{B}}^\top \mathbf{x}^\top \mathbf{x} \hat{\mathbf{B}}$, devianaza spiegata di regressione, $SSV = (\hat{\mathbf{v}}^o)^\top (\mathbf{A}^o)^\top (\mathbf{A}^o)(\hat{\mathbf{v}}^o)$, devianza spiegata fra gruppi; e $SSE = \mathbf{e}^\top \mathbf{e}$, devianza residua.

$$SST = SSX + SSV + SSE \tag{3.17}$$

Ho quindi la struttura di variabilità divisa in 3 parti, a differenza del normale modello lineare con l'analisi della varianza, che è solamente in 2.

Nell'ipotesi di omoschedasticità, analogamente a quanto viene fatto per l'analisi della devianza, si può definire:

$$n(\tau^o)^2 = (\mathbf{v}^o)^\top \mathbf{A}^\top \mathbf{A} (\mathbf{v}^o) \qquad n(\sigma^o)^2 = \mathbf{e}^\top \mathbf{e}$$

Il **coefficiente di correlazione interclasse** è definito come:

$$\rho^o = \frac{(\tau^o)^2}{(\tau^o)^2 + (\sigma^o)^2} \tag{3.18}$$

Questa non sarà più inficiata da caratteristiche diverse.

3.1.4 Analisi della covarianza campionaria

Si deve tradurre il modello in termini campionari, come per il modello lineare. Esso sarà analogo, ma in termini di variabili ed errori campionari. L'errore campionario darà vita, al variare del campione, a variabili casuali, determinate da un distribuzione comune (ignorando la distribuzione degli errori tra i gruppi, e considerando il modello generale). Si avranno anche variabili casuali per le $\mathbf{y}$, distribuite normalmente come $N(\gamma_{00}, \sigma_y^2)$, prescindendo dalla struttura dei singoli gruppi.

Sotto l'ipotesi di normalità delle variabili esplicative campionarie $\mathbf{y}$, si può dimostrare che le somme degli errori al quadrato hanno una distribuzione chi-quadro χ^2; analogo anche per tutte le devianze che la compongono:

$$SST \sim \chi^2_{(n-1)} \qquad SSX \sim \chi^2_{(r)} \qquad SSY_e \sim \chi^2_{(n-r-1)} \qquad SSV \sim \chi^2_{(p-1)} \qquad SSE \sim \chi^2_{(n-p-r-1)}$$

Si può inoltre mostrare che $cor(\mathbf{x}\hat{\mathbf{B}}, \mathbf{A}^o\hat{\mathbf{v}}^o) = cor(\mathbf{x}\hat{\mathbf{B}}, \hat{\mathbf{e}})$, in quanto $\hat{\mathbf{y}}_e$ è il residuo della regressione di $\mathbf{y}$ e vale che $\hat{\mathbf{y}}_e = \mathbf{A}^o\hat{\mathbf{v}}^o + \hat{\mathbf{e}}$ e $cor(\mathbf{A}^o\hat{\mathbf{v}}^o, \hat{\mathbf{e}}) = 0$. Da questi fatti, si ha che le grandezze SSX, SSV e SSE sono **incorrelate**, e quindi le loro distribuzioni sono indipendenti. Poiché i rapporti tra χ^2 indipendenti sono delle F di Snedecor, si ha che variabile:

$$F_{r,n-r-1} = \frac{\frac{SSX}{r}}{\frac{SSE_y}{n-r-1}} \qquad F_{p-1,n-r} = \frac{\frac{SSV}{p-1}}{\frac{SSE}{n-p-r}}$$

e si possono quindi eseguire i vari test statistici. Si vuole avere la $F_{r,n-r-1}$ nella regione di rifiuto, e quindi rifiuta l'ipotesi per cui le $\mathbf{x}$ *non* spieghino le $\mathbf{y}$. Analogamente, se $F_{p-1,n-r}$ cade nella regione di rifiuto, si rifiuta l'ipotesi per cui i gruppi *non* spieghino la variabilità di $\mathbf{y}$.

3.1.5 Ancova ad effetti casuali

Si introduce la **covarianza a effetti casuali**. In ognuno dei gruppi (sottopopolazioni), noi non prendiamo tutti gli elementi e prendiamo alcuni campioni, da cui si ricava la variabile casuale E e poi quella $\mathbf{y}$, come detto nella sezione precedente.

Le medie dei gruppi, come definito prima $\gamma_{00} + v_j^o$, non sono quindi dei parametri fissi e si dimentica l'effetto di campionamento. Per dei dati gerarchici, questa definizione non va bene: v_j^o deve essere anch'esso la determinazione di una variabile casuale V_j, variabile casuale media campionaria, una per ogni gruppo. Esso sarà distribuito normalmente con valor medio 0 e varianza τ^2.

Si ha quindi da risolvere il modello:

$$y_{ij} - (\beta_0 + \sum_k \beta_{jk} x_{ijk}) = \gamma_{00} + v_j^o + e_{ij} \tag{3.19}$$

dove, a differenza del modello della covarianza tradizionale, anche la variabile v_j^o è la determinazione di una variabile casuale, generata al cambiare del campione. In questo modo, si sta rispettando la gerarchia del modello, data da *campionamento a due stadi*.

Poiché vale la relazione $cor(A^o\hat{\mathbf{v}}_j^o, \hat{\mathbf{e}}) = 0$, si può affermare che le distribuzioni V_j e E_j sono incorrelate e indipendenti.

A livello stocastico, definire v_j di tipo campionario, non posso avere dei risultati in termini di "valore preciso" della media campionaria. Per poter passare da informazioni campionare a quelle su tutta la popolazione, si deve usare l'**intervallo di confidenza**. Si sta andando a eseguire una **probabilizzazione della gerarchia** fra medie parziali: minore è il loro intervallo di confidenza, maggiore è la forza nel determinare il vero valore di v_j. Siccome questi sono relativi a sottopopolazioni diverse, ognuna di queste variabili casuali è indipendente e incorrelata con le altre.

3.1.6 Modello Multileve: definizione e passaggi risolutivi

Caratterizzato da dati gerarchici, in cui la regressione cattura la relazione disaggregata tra i dati e descrive la varianza nei gruppi. L"analisi della varianza cattura anche la varianza aggregata fra i gruppi.

Si possono definire due macrocategorie di modelli multilevel:

- **Modelli misti**, in cui si fissano i coefficienti della regressione tra **x** e **y** e l'errore di secondo livello, con l'errore individuale visto come errore casuale.
- **Modelli casuali**, in cui si analizza anche la relazione disaggregata tra **x** e **y** come variabile casuale.

Il **modello Multilevel** è la situazione più generale che si può ottenere, di cui la regressione lineare, regressione multilevel, analisi della varianza a effetti fissi o analisi della varianza a effetti casuali sono dei casi particolari.

Il primo modello che si guarda è sempre quello OLS, su cui vado a verificare tutte le proprietà. Faccio poi l'analisi della varianza a effetti casuali, analisi della covarianza a effetti misti e, infine, analisi della covarianza a effetti casuali. Si costruiscono quindi dei modelli a passaggi.

Il primo risultato è quello di un **modello OLS**, con errori omoschedastici e incorrelati. Se questo non vale, vado a correggere tutti questi effetti (come è stato mostrato fino ad ora nel resto del corso), confrontando le violazioni. Per andare avanti, non si vuole che risultati complessi siano inficiati da errori, come *eteroschedasticità*, *normalità* o altro. Si può trasformare questo modello OLS in uno multileve, come un **modello Cronbach**, per avere anche nella prima parte una distinzione tra *between* e *within*; non è comunque esssenziale, in quanto è un passaggio che viene eseguito in fase successiva nelle analisi di varianze e covarianze.

3.1.7 Modello con medie incondizionate

Noto anche come modello **vuoto**, si tratta di un modello ANOVA ad effetti casuali del tipo:

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

in cui sia e_{ij} (errori) che v_j (media campionaria) sono distribuiti casualmente con distribuzioni normale. Si potrebbe anche ipotizzare che sia gli errori che le medie campionarie non siano omoschedastici, sebbene questo è al di fuori degli scopi di questo corso. Il campionamento all'interno di un gruppo j è completamente diverso da quello degli altri gruppi: **campionamento a due stadi**; nel primo si scelgono i gruppi, nel secondo gli elementi dei gruppi. Gli errori e le medie campionarie sono quindi indipendenti e incorrelate.

L'ipotesi che si può allargare è quella di incorrelazione fra gli errori dentro ogni sotto-popolazione, in maniera analoga in quanto fatto nel modello OLS. Gli altri casi richiederebbero l'uso di modelli multilevel più complessi.

Dunque, come formulazione si può scomporre come:

$$var[\mathbf{y}] = \sigma^2 + \tau^2 \tag{3.20}$$

ovvero fra e tra i gruppi, dove $tau^2 = cov[y_{ij}, y_{i'j}]$, ovvero la parte della variabilità generale condivisa dagli elementi all'interno di ogni gruppo. Si definisce, come precedentemente accennato, il **coefficiente di correlazione interclasse** come:

$$\frac{\tau^2}{\tau^2 + \sigma^2} \tag{3.21}$$

che indica quanto della varianza della variabile esplicativa si può attribuire alla devianza tra i gruppi. Nel qual caso questo coefficiente sia basso, si possono usare delle regressioni tradizionali, che non tengano conto di dati gerarchici. Viceversa, per valore maggiori di 0.10, si deve considerare il modello avendo variabili gerarchiche.

In termini di studio statistici, si può usare il **test F** per vedere se le intercette casuali v_j siano in complesso equivalenti, come già mostrato per l'analisi della covarianza. Nel modello multilevel, si ha interesse quando questo tipo ti test rimanda il rifiuto dell'ipotesi nulla. La parte più specifica del modello è quella di rimandare risultati in termini dell'intervallo di confidenza della variabile casuale V_j, in quanto media campionaria che cambia a variare del campione. Come buona pratica, spesso si confrontano le code (di tutte le medie).

3.1.8 Random Intercept Model

Questo modello prende anche il nome di **mixed model**, presenta anche una parte dovuta alla regressione lineare, senza intercetta:

$$y_{ij} = \gamma_{00} + \beta_{ij} x_{ij} + v_i + e_{ij} \tag{3.22}$$

Anche in questo caso, si considerano gli errori e le medie campionarie come indipendenti e incorrelate.

In questo modello, si avranno 3 effetti da considerare: l'effetto della stima del parametro $\hat{\mathbf{B}}$, l'effetto della differenza fra i gruppi e della differenza nei gruppi.

Si possono avere anche **campioni non bilanciati**, ovvero con numerosità differente tra i gruppi. La particolarità del modello Multilevel è di usare l'effetto *shrinkage*, ovvero nel risultato finale vanno a pesare i campioni con meno elementi.

Il modello può essere visto, a livello di microdati, come *micromodel*, oppure come *macromodel* in termini di analisi della varianza. Il primo modello è fisso, mentre il secondo è casuale. Si possono introdurre variabili esplicative di secondo livello, oltre che interazioni tra variabili di primo e secondo livello: tutti modi per trovare soluzioni sempre più complesse e raffinate. A livello pratico, si dovrebbe sempre introdurre la regressione multilevel anche nella parte lineare OLS.

3.1.9 Stima e test di ipotesi

Anche termini di stima, si hanno due soluzioni: prima il modello lineare, poi l'analisi della varianza. I metodi di **full maximum likelihood** (FML) e **residual maximum likelihood** (REML) servono per stimare tutti i parametri calcolati, ovvero i coefficienti di regressione γ e le componenti della varianza σ^2 e τ^2. Sono uno alternativo all'altro: essi vanno prima a stimare la parte lineare, e poi quella della varianza.

Il secondo è più preciso e, con metodi come *OLS* o *GLS*, stima i parametri del modello lineare, e usa queste stime per massimizzare la verosomiglianza dei residui per ottenere le stime dei parametri di varianza. L'algoritmo è in grado di eseguire le correzioni per l'allargamento alle ipotesi di eteroschedasticità e correlazione. In maniera pratica, si usano alcuni algoritmi, come *EM*, *Fischer scoring*, *IGLS* o *RIGLS*, che convergono alle stime GML o REML.

Questi metodi, in prima istanza vanno usare metodi di *non likelihood*, come OLS o GLS (sono in grado di distinguere) per minimizzare la *devianza residua*

$$D_M = \sum_{ij} ((y_{ij} - \gamma_{00}) - x_{ij} b)^2 = SSY_e = -2 \ln L_M \tag{3.23}$$

dove si identifica con L_M la funzione di massima verosomiglianza.

Come passaggio successivo, si massimizza la verosomiglianza dei *residui* del modello lineare y_e, ovvero si minimizza la devianza:

$$D_V = \sum_{ij} \left((\hat{y_e})_i - A^o v^o_{ij} \right)^2 = SSE = -2 \ln L_V \tag{3.24}$$

Data la non linearità di questi modelli, vengono stimati in maniera iterativa, con convergenza alle stime finali quando non si hanno dei miglioramenti. La **non linearità** non è dovuto alla

presenza di parametri lineari (lo sono), ma dalla **struttura gerarchica dei dati**. Gli algoritmi citati precedentemente fanno esattamente questi passaggi a "tentativi".

Parametri fissi

Sui **parametri fissi del modello**, si prende come ipotesi nulla che siano tutti nulli, ovvero $H_0: \gamma_h = 0$, dove γ_h indica tutti i parametri fissi, come γ_{00} o β. Il test è noto come **Test di Wald**, definito come:

$$T(\gamma_h) = \frac{\hat{\gamma}_h}{s.e.(\hat{\gamma}_h)} \tag{3.25}$$

Si comporta circa come un t test, sebbene sia molto più complesso. Il risultato è comunque riconducibile per vedere la significatività dei parametri.

Test F

Si può comunque usare un test F per verificare nullità di tutti i parametri nel modello lineare e anche per la nullità della devianza fra i gruppi.

Deviance test

Per verificare la parte di **analisi della varianza** e le ipotesi di **uguaglianza di più parametri** si usa un test noto come **deviance test**. Questa verifica è valida anche per modello più complessi di quello lineare, come le *regressioni logistiche*. Data la **funzione di massima verosomiglianza del modello** L, si definisce la devianza residua come:

$$D = -2\ln(L)$$

la cui distribuzione di probabilità è asintotica a un χ^2. Dal modello REML, si ottengono una devianza residua per il modello lineare multiplo, D_M, e una devianza residua per l'analisi della varianza D_V.

Si può utilizzare questo approccio in 3 differenti casi:

- **empty model**, modello in cui non ci sono variabili esplicative, e non vi è devianza fra i gruppi. Questo è quello che si vede quando si studia l'analisi della varianza, senza parte di modello lineare.
- **saturated model**, in cui nella parte regressiva il modello non possiede SSY_e, ovvero non esiste devianza residua e passa perfettamente per i punti; e analogo per la parte di varianza.

Il limite superiore della devianza residua è ∞, in situazione di empty model. Viceversa, nel saturated model, se la funzione di massima verosomiglianza vale 1, significa che la devianza è 0. Ricapitolando, nell'empty model si ha $D = \infty$, mentre il quello saturato $D = 0$. Difatti, essendo questi due modello agli estremi, tutti i valori di devianza di un modello reale si troveranno tra questi due estremi, in maniera decrescente. Ancora, si dimostra che questa funzione D, decrescente tra ∞ e 0, è asinotica a un χ^2, grazie alla presenza di $-2\ln$.

Nell'**empty model**, si ha devianza massima, si avrà:

$$D_{0,M} = SSY_e = -2\ln L(0)_M = \max = SST$$

$$D_{0,V} = SSE = -2\ln L(0)_V = \max = SSY_e$$

queste di distribuiscono come dei χ^2_1, con 1 grado di libertà. Il parametro in entrambi i casi è solamente 1: una sola intercetta per entrambi, dato che non si hanno variabili esplicative. In questo caso, le devianze residue coincidono con quelle totali, ovvero:

$$SSY_e = SST \qquad SSE = SSY_e$$

Nel **saturated model**, c'è interpolazione perfetta dei dati, e quindi:

$$D_{sat,M} = SSY_e = -2\ln(\hat{\beta}_m ax)_M = 0$$

$$D_{sat,B} = SSE = -2\ln(\hat{\beta}_m ax)_V = 0$$

per cui le devianze spiegate coincidono con quelle totali, ovvero:

$$SSX = SST \qquad SSF = SSY_e$$

distribuite come χ^2_{n-1} con $n-1$ gradi di libertà.

Modello proposto

Il caso del modello proposto è quello del modello che sta a metà strada tra i due estremi visti prima. In questo caso, i parametri si distribuiscono nella parte regressiva come χ^2_{r+1} con $r+1$ gradi di libertà e χ^2_p con p gradi di libertà per la parte di analisi della varianza. Da queste quantità si può costruire un test, definito come la differenza con il modello *empty*:

$$G_M = (D_0 - D_{mod})_M = -\ln\left(\frac{L(0)}{L(\hat{\beta})}\right)_M \qquad G_V = (D_0 - D_{mod})_V = -\ln\left(\frac{L(0)}{L(\hat{\beta})}\right)_V \tag{3.26}$$

con distribuzioni, rispettivamente χ^2_r e χ^2_{p-1}. La parte fissa può riguardare tutti gli $r+1$ parametri, anche solo $k < r$ parametri oppure anche uno solo. Ovviamente vengono presi con H_0 di nullità dei parametri.

3.1.10 Random Slope Model

Usando un modello lineare per dei dati gerarchici, si potrebbe entrare in errore. Al fine di questo, si devono usare dei modelli di regressione in grado di pesare le differenze tra i gruppi della popolazione. Si potrebbe avere eterogenerità tra le regressioni lineari quando eseguite su gruppi differenti: si parla in questo caso di *interazione gruppo-varianza*. In modello **random slope**, depuro dalle differenze tra i gruppi e come questi possano inferire in maniera diversa.

Eseguiamo tutto il percoso per i modelli multilevel. Il punto di partenza è il modello OLS, in cui non si apporta differenza tra i gruppi. Usare un modello **random intercepts**, come usato fin'ora, considera solamente intercette differenti, ma tutti gli stessi coefficienti angolari. Spesso però, non è detto che le rette di regressione debbano essere le stesse per tutti i gruppi. Si ha quindi un modello del tipo:

$$y_{ij} = b_0 + b_{1,j}x_{ij} + u_{ij} + e_{ij}$$

noto anche come **random total model**. Si avrà una differenza tra i modello totali se si ha covarianza positiva (le rette si allontanano) oppure negativa (si avvicinano) o, ancora, se è zero (non hanno un legame tra loro). La covarianza indica il legame tra i differenti gruppi, e.g. al variare tra tra i gruppi, in ordine, si ha aumento dei coefficienti angolare delle regressioni.

La rappresentazione analitica di questo modello è del tipo:

$$y_{ij} = \gamma_{00} + (\beta_1^0 + \beta_{1,j}^*)x_{ij} + v_j + e_{ij} \tag{3.27}$$

in cui la parte β_1^0 è la parte comune a tutti i gruppi, fissa; e la parte $\beta_{1,j}$ è determinazione di una variabile casuale al variare del campione. Anch'essa è ipotizzata normale del tipo $N(\beta_1^0, \mathbf{v}^2)$.[2]

La parte fissa del modello è $\gamma_{00} + \beta_1^0 x_{ij}$ mentre quella random il resto. Si può anche distinguere tra parte micro, $\gamma_{00} + (\beta_1^0 + \beta_{1,j})x_{ij} + e_{ij}$, e parte macro, $\gamma_{00} + v_j$.

[2] Sono confuso da che cosa sia questo $\mathbf{v}^2$

Le variabili casuali V_j e E_j sono incorrelate e indipendenti fra i gruppi e dentro i gruppi. Invece, per quanto detto, V_j e B_j all'interno di un gruppo sono legati, secondo la relazione $cov(V_j, B_j) = \rho$, che può essere positivo, negativo o nullo, come detto sopra. Ovviamente tra coppie diverse si prendono indipendenti e incorrelate, oltre che identicamente distribuite.

Il calcolo del **coefficiente di interclasse** diviene può compless

$$var(y_{ij}|x_{ij}) = \tau^2 + 2\rho x_{ij} + \mathbf{v}^2 x_{ij}^2 + \sigma^2 \tag{3.28}$$

questo si usa molto meno e non è un buon criterio di scelta del tipo di modello.[3]

[3] Vedere le varie situazioni di legame tra ρ e $\beta_{1,j}$ e i loro effetti.

Domande

- Come mai posso considerare la R^2 come il rapporto tra due χ^2 indipendenti? Sbaglio, o la varianza totale contiene anche quella spiegata? A casa mia non mi sembrano indipendenti.

  Risposta
  Avevo sbagliato a capire la slide. Era riferito alla variabile F, rapporto tra devianza spiegata dalla regressione e devianza intrinseca.
- Non ho capito come dovrebbe essere la funzione esponenziale che si usa quando si ha varianza negativa, per il metodo **WLS**.
- Nell'Eq. 1.28, intuisco come mai $cov[e_i^{\#}, e_{i-2}^{\#}] = \rho^2$, ma mi sfugge il diretto passaggio algebrico.
- Quando costruisco la matrice di varianza-covarianza per gli errori non sferici, come mai la matrice è simmetrica? Ovvero, come mai vale sempre la relazione $cov[e_i, e_j] = cov[e_j, e_i]$?

- **Domanda**
  Online ho trovato che l'indice di condizione è visto come il rapporto tra il massimo e il minimo autovalore della matrice di *design*.

  Risposta
  Sì e no. Difatti, si mettono gli autovalori in ordine crescente, e si fanno i rapporti successivi per trovare gli indici di condizione.

- **Domanda**
  Nel test di Shapiro-Wilk, che tipo di distribuzione possiede W? Stessa domanda per tutti gli altri test.

  Risposta
  Sono delle particolari distribuzioni asimmetriche, di cui non andiamo a studiare alcunché.

- **Domanda**
  Come mai assumo, nell'**Ancova ad effetti casuali**, che le variabili v_j siano indipendenti e incorrelate tra di loro? A me sembra qualcosa un po' tirato che non è detto avere sempre

validità: i gruppi non sono degli Universi isolati.

Risposta
nella desrittiva si hanno medie campionarie, campionante in maniera indipendente. Sono popolazione tra di loro indipendenti, e quindi l'estrazione non ha a che fare con le altre. Non ho il legame nell'estrazione. È questo il passaggio fondamentale. È indipendenza più forte anche del campionamento di tutta la popolazione, in quanto vado a prendere da urne diverse al posto che solamente un'unica urna.

- **Domanda**