Statistica (corso progredito)
Richiami sul modello lineare

Leonardo Egidi

Università di Trieste
Corso di laurea magistrale in Scienze Statistiche ed Attuariali

Trieste, a.a. 2021/2022

Il modello di regressione lineare

- I modelli lineari (LM) vengono proposti per studiare le relazioni fra una *variabile risposta* y quantitativa e una o più *variabili esplicative o covariate* $x_1, x_2, \ldots, x_{p-1}$ $(p \geq 2)$ osservate per un campione di n unità.
- Si vuole valutare l'impatto delle covariate sulla media μ_i della variabile y_i per l'-i-esima unità che è rappresentabile tramite la relazione

$$E(y_i|x_{i1}, x_{i2}, \ldots, x_{ip-1}) = \mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1}$$

La risposta y_i può essere quindi descritta come

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{p-1} x_{ip-1} + \epsilon_i \ ,$$

e il modello può essere scritto compattamente per l'insieme delle n unità utilizzando la notazione matriciale,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \ .$$

- $X\boldsymbol{\beta}$ è la componente sistematica del modello.
- $\boldsymbol{\epsilon}$ è la componente stocastica del modello.

In particolare, si ha

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ \vdots & & & \vdots \\ 1 & x_{n1} & \cdots & x_{np-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

con:

- $\boldsymbol{y}$ = vettore $(n \times 1)$ delle risposte;
- X = matrice $(n \times p)$ di regressione contenente i valori delle variabili esplicative, è detta anche *matrice disegno*;
- $\boldsymbol{\beta}$ = vettore $(p \times 1)$ dei parametri (coefficienti) di regressione;
- $\boldsymbol{\epsilon}$ = vettore $(n \times 1)$ delle componenti della variabile errore.

Il modello per l'i-esima unità può essere espresso come

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \epsilon_i,$$

ove $\boldsymbol{x}_i^{\mathrm{T}}$ rappresenta l'i-esima riga della matrice disegno.

Il modello di regressione lineare I

Nel modello lineare

- La variabile y è una variabile quantitativa (continua).
- Le variabili esplicative x possono essere:
  - $\mapsto$ variabili quantitative (numeriche);
  - $\mapsto$ variabili categoriali (fattori).
- Si assume che i valori di X siano delle costanti. X è non stocastica
- La matrice disegno X si assume che sia di pieno rango. Poiché in generale si ha $n >> p$ questo significa che il rango di X è assunto essere pari a p ovvero che le colonne di X sono vettori linearmente indipendenti.

Le assunzioni sulla componente stocastica

Il modello lineare è completato da opportune assunzioni sulle componenti stocastiche ϵ_i

1. $E(\epsilon_i) = 0, \quad i = 1, \ldots, n$ o in termini matriciali $E(\boldsymbol{\epsilon}) = \mathbf{0}$
2. $Var(\epsilon_i) = \sigma^2$ condizione di omoschedasticità
3. $E(\epsilon_i, \epsilon_j) = 0$ per $i \neq j$ condizione di incorrelazione. Le ultime due condizioni possono essere espresse in termini matriciali come
   $Cov(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \sigma^2 \boldsymbol{I}_n$
   dove $Cov(\boldsymbol{\epsilon})$ denota la matrice di varianze e covarianze del vettore aleatorio $\boldsymbol{\epsilon}$.
   Le assunzioni 1-3 costituiscono le assunzioni del secondo ordine. Ad esse di solito si aggiunge
4. $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \Sigma)$
   In notazione matriciale si ha quindi $E(\boldsymbol{y}) = X\boldsymbol{\beta}$ e quindi
   $\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$

Discussione degli assunti

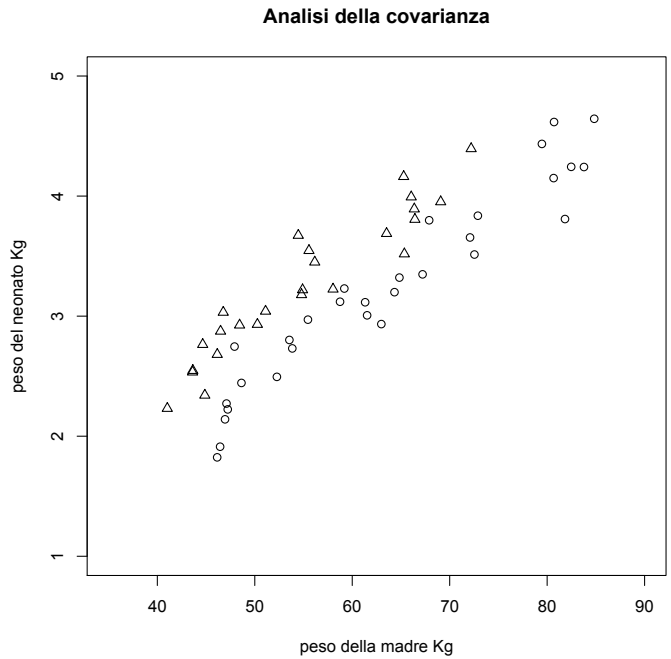
- L'assunzione dell'effetto lineare delle covariate è meno restrittivo di quanto appaia. Si possono introdurre relazioni non lineari trasformando le variabili x. Ad esempio $y_i = \beta_0 + \beta_1 log(z_i) + \epsilon_i$ introduce un effetto di z_i logaritmico. Tuttavia se poniamo $x_i = log(z_i)$ si torna all'usuale modello per la trasformata x. Il modello resta quindi lineare nei parametri.
- **Omoschedasticità degli errori**.
  Si tratta di un'assunzione che può essere verificata con strumenti diagnostici. La conseguenza dell'ignorare l'eventuale eteroschedasticità è una sovrastima della varianza delle stime. Esistono dei rimedi in taluni casi
- **Incorrelazione degli errori**.
  Si tratta di un problema che emerge con dati che sono ordinati (ad esempio temporalmente o spazialmente). Anche in questo caso esistono strumenti per diagnosticare il problema e possibili rimedi.

Covariate continue, fattori, interazioni.

- L'effetto di una variabile esplicativa numerica, se si suppone la linearità, è misurata dal solo parametro che nel modello è associato a tale variabile.
- L'effetto di una variabile esplicativa categoriale (di un fattore) è da leggersi come lo spostamento del livello medio della variabile risposta al netto delle altre variabili nel modello in corrispondenza ai livelli del fattore. In generale, ciascun livello del fattore ha effetto sulla media della variabile risposta: è necessario quindi introdurre un numero di parametri che è legato al numero di livelli del fattore.
- È spesso utile considerare l'interazione fra due fattori o fra un fattore e una variabile quantitativa. Il modello qui considerato è invece poco adeguato per valutare l'interazione fra variabili continue che quindi entrano in forma additiva. È anche possibile valutare l'interazione fra variabili continue (l'interpretazione è meno agevole).

Un esempio

`Peso` alla nascita, Y, in chilogrammi, per un campione di neonati, nate da madre `fumatrice` (F) (nel grafico si usano simboli diversi per fumatrici - il cerchio - e non fumatrici - il triangolino). Per ogni donna è noto il `pesomadre` pre-gravidanza(x)



Com'è ragionevole attendersi il peso del neonato è maggiore se le madri sono non fumatrici. Sembra presente una differenza sistematica fra i due gruppi seppure non cambi la relazione fra peso del neonato e della madre.

Modello con un fattore (dicotomico)

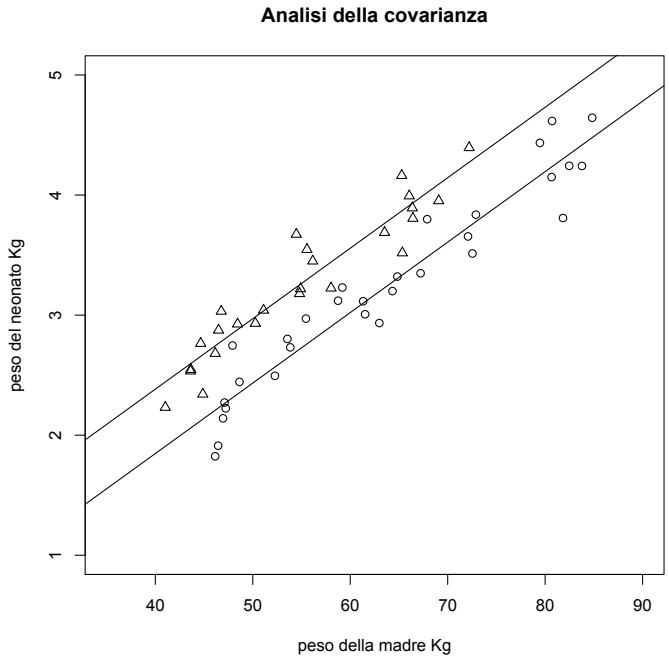
Le due variabili y =peso del neonato e x_1=peso della madre sono variabili numeriche continue e la variabile f_i è un fattore con due livelli (diciamo $F = 1$ se fumatrice, $F = 2$ se non fumatrice); insieme danno luogo al seguente modello

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{(F_i=2)} + \epsilon_i \ .$$

che ha la seguente forma matriciale:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ \vdots & & \vdots \\ 1 & x_k & 1 \\ \vdots & & \vdots \\ 1 & x_n & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \ .$$

Ancora sull'esempio



Nella formulazione data

- il parametro β_1 misura l'effetto del peso della madre sul peso del neonato ed è la pendenza (comune) delle due rette in figura;
- β_0 misura l'intercetta della retta per fumatrici e β_2 misura, a parità di peso della madre, la distanza verticale fra le rette relative a fumatrici e non fumatrici.

Parametrizzazioni alternative

Un modello può avere parametrizzazioni alternative. Ad esempio, il modello precedente può essere scritto nella forma:

$$Y_i = \gamma_1 x_i + \gamma_2 I_{(F_i=1)} + \gamma_3 I_{(F_i=2)} + \epsilon_i \ , \tag{1}$$

ove $I_{(F_i=j)}$ è la variabile indicatore che assume valore 1 se $(F_i = j)$, $j = 1, 2$, e 0 altrimenti.

Il modello è equivalente anche se l'interpretazione dei parametri cambia.

In questo caso il modello in forma matriciale risulta:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 & 0 \\ x_2 & 1 & 0 \\ \vdots & & \vdots \\ x_n & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \ .$$

Interpretazione dei parametri

In tale modello:

- il parametro γ_1 misura l'effetto della variabile quantitativa sul peso del neonato e
- γ_2 e γ_3 misurano i livelli medi del peso del neonato, a parità di peso della madre, per fumatrici e non.
- Le colonne di X della matrice disegno sono ottenute da quelle del modello precedente mediante una combinazione lineare. Il modello resta invariato nella sostanza ma cambia l'interpretazione.

Il modello con interazione

Il modello espresso dalla formula (1) è additivo, nel senso che l'effetto di `peso madre` è lo stesso per ciascun livello di `fumatrice`. Però, ci sono situazioni in cui l'effetto può essere diverso.
In questo caso è necessario assumere un modello con interazioni tra variabili. Ad esempio, la formula esprime il modello di regressione

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 I_{(F_i=2)} + \beta_3 I_{(F_i=2)} x_i + \epsilon_i \ .$$

del fattore `fumatrice`.
In tal caso l'interazione implica che la relazione fra `peso madre` e `peso figlio` sia diversa per fumatrici e non fumatrici.

Stima dei parametri: Lo stimatore dei minimi quadrati

Si può proporre uno stimatore del parametro $\boldsymbol{\beta}$ facendo solo assunzioni su medie e varianze/covarianze di y. In tal caso si può ricorrere al **metodo dei minimi quadrati**:

1. Si sceglie il valore $\hat{\boldsymbol{\beta}}$ di $\boldsymbol{\beta}$ che minimizza

$$LS(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})^2 = (\boldsymbol{Y} - X\boldsymbol{\beta})^{\mathrm{T}} (\boldsymbol{Y} - X\boldsymbol{\beta}) \; ;$$

   da cui

$$LS(\boldsymbol{\beta}) = \boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{\beta}^T X^T \boldsymbol{y} - \boldsymbol{y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$$

2. Derivando $\frac{\partial LS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2X^T \boldsymbol{y} + 2X^T X\boldsymbol{\beta}$ e quindi eguagliando a 0, si ottiene lo stimatore

$$\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} \boldsymbol{y} \; .$$

   Si noti che è necessario assumere che $(X^{\mathrm{T}} X)$ sia non singolare. Questo è assicurato dalla condizione che X sia di pieno rango.

Lo stimatore ai minimi quadrati soddisfa le seguenti proprietà:

1. $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ e $var(\hat{\boldsymbol{\beta}}) = \sigma^2(X^{\mathrm{T}}X)^{-1}$;
2. Asintoticamente
$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 V^{-1})$$
ove $V = \lim_{n\to\infty} X_n^{\mathrm{T}} X_n$, con X_n che è la sequenza di matrici disegno e V è definita positiva;
3. $\hat{\boldsymbol{\beta}}$ è stimatore non distorto con varianza minima fra gli stimatori lineari (Gauss-Markov).

Stima dei parametri: massima verosimiglianza

- Nel caso si faccia l'ipotesi di normalità per la componente stocastica la stima di $\boldsymbol{\beta}$ può essere ricavata con il metodo della *massima verosimiglianza*.
- Si sceglie il valore di $\boldsymbol{\beta}$ che massimizza

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{Y} - X\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - X\boldsymbol{\beta})\right)$$

  è facile verificare (calcolando la log-verosimiglianza e derivando rispetto a $\boldsymbol{\beta}$) che il massimo di questa espressione si ottiene minimizzando la quantità $(\boldsymbol{Y} - X\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - X\boldsymbol{\beta})$.
- Quindi lo stimatore di massima verosimiglianza, nell'ipotesi di gaussianità, equivale a quello dei minimi quadrati.
- Alle proprietà già viste per lo stimatore dei minimi quadrati si aggiunge quella che riguarda la distribuzione esatta di $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X^{\mathrm{T}}X)^{-1})$

La stima della varianza

- La varianza σ^2 viene stimata con la *varianza residua corretta*

$$S^2 = \frac{SSE}{n-p} \ ,$$

ove

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 \\ &= (\boldsymbol{y} - \hat{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{y}}) \\ &= (\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}) \end{aligned}$$

è la *somma dei quadrati dei residui*.

- Sotto ipotesi di normalità, inoltre

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(X^{\mathrm{T}}X)^{-1}) \quad \text{e} \quad \frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p} \ ,$$

con $\hat{\boldsymbol{\beta}}$ e S^2 stocasticamente indipendenti.

Verifiche di ipotesi

I test di interesse più comune sono:

- test per un singolo elemento di $\boldsymbol{\beta}$
  $H_0 : \beta_j = 0 \quad$ contro $\quad H_1 : \beta_j \neq 0$
- test su un sottovettore $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_r)$
  $H_0 : \boldsymbol{\beta}_1 = 0 \quad$ contro $\quad H_1 : \boldsymbol{\beta}_1 \neq 0$
- test di uguaglianza fra due coefficienti
  $H_0 : \beta_j - \beta_r = 0 \quad$ contro $\quad H_1 : \beta_j - \beta_r \neq 0$

Tutte le ipotesi citate sono casi particolari della *ipotesi lineare generale*
$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \quad$ contro $\quad H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}$
dove $\boldsymbol{C}$ è una matrice $r \times p$ di rango $= r \leq p$ e d un vettore $r \times 1$.

Verifiche di ipotesi

Il modello può essere ristimato sotto la restrizione lineare $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d}$.
I residui di tale modello sono ${}_{H_0}e_i$ e si possono calcolare $SSE_{H_0} = \sum_i^n {}_{H_0}e_i^2$
e la statistica

$$\frac{n-p}{r}\frac{SSE_{H_0} - SSE}{SSE},$$

che, quando H_0 è vera e sotto l'assunzione di normalità si distribuisce come una $F_{r,n-p}$.

Test su un singolo coefficiente

Se si considera il test per β
$H_0 : \beta_j = 0$ contro $H_1 : \beta_j \neq 0$
si può applicare la regola vista sopra (dove C è un vettore di zeri con 1 nella posizione j mentre $d = 0$), e si può mostrare che

$$\frac{\hat{\beta}_j^2}{\widehat{Var(\hat{\beta}_j)}} \sim F_{1,n-p},$$

che è il quadrato di una t di student con $n - p$ gdl e quindi equivalente alla seguente statistica test

$$t_j = \frac{\hat{\beta}_j}{\widehat{Var(\hat{\beta}_j)}^{1/2}}$$

Si può usare lo stesso risultato per ottenere un intervallo di confidenza per β_j al livello $1 - \alpha$ $\hat{\beta}_j \pm t_{n-p,1-\alpha/2}(Var(\hat{\beta}_j))^{1/2}$

Scomposizione di somme di quadrati

- È utile ricordare la scomposizione della devianza dei valori $\boldsymbol{y}$ in:

$$SST = SSR + SSE, \tag{2}$$

ossia, la somma dei quadrati totale (*devianza totale*) in somma dei quadrati di regressione (*devianza spiegata dal modello*) e somma dei quadrati residua (*devianza dei residui*). Lo studio delle componenti della (2) è fondamentale nei LM, in quanto il confronto tra SSR e SST è un indicatore della bontà di adattamento del modello.

- Sia $\mathcal{F}_1$ il modello minimale con la sola intercetta ($p = 1$). Sia $\mathcal{F}_p$ il modello corrente a p parametri e sia $\mathcal{F}_{p_0}$ un modello ridotto, con $1 < p_0 < p$. Allora, la varianza spiegata dal modello corrente $\mathcal{F}_p$ può essere suddivisa per le componenti del modello stesso, come si mostra nella Tabella 1 sul prospetto detto di *analisi della varianza*.

L'analisi della varianza

Tabella 1: Prospetto di analisi della varianza (anova)

Fonte di variabilità	g.l.	SQ	test su miglioramento
totale	n	SST	
costante	1	$n\bar{Y}^2$	
totale corretta	$n-1$	SST_{cor}	
miglioramento con $\mathcal{F}_{p_0}$ rispetto a $\mathcal{F}_1$	p_0-1	SSR_{p_0}	$\frac{SSR_{p_0}/(p_0-1)}{SSE_{p_0}/(n-p_0)} \sim F_{p_0-1,n-p_0}$
miglioramento con $\mathcal{F}_p$ rispetto a $\mathcal{F}_{p_0}$	$p-p_0$	$SSR_p - SSR_{p_0}$	$\frac{(SSR_p-SSR_{p_0})/(p-p_0)}{SSE_p/(n-p)} \sim F_{p-p_0,n-p}$
residui di $\mathcal{F}_p$	$n-p$	SSE_p	

• La perdita di bontà di adattamento del modello $\mathcal{F}_{p_0}$ rispetto a $\mathcal{F}_p$ può essere valutata attraverso la statistica

$$F = \frac{(SSE_{p_0} - SSE_p)/(p-p_0)}{SSE_p/(n-p)} \sim F_{p-p_0,n-p} \cdot$$

Coefficiente di determinazione

- Il coefficiente di determinazione R^2 è definito come la % di varianza spiegata dal modello di regressione ed è una misura di bontà dell'adattamento del modello e si basa sulla scomposizione appena vista ovvero
  $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$
- $R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSE}{SST}$
- È compreso fra 0 e 1 e più vicino esso è a 1 e maggiore è la bontà dell'adattamento del modello
- È pari al coefficiente di correlazione al quadrato fra $\boldsymbol{y}$ e $\hat{\boldsymbol{y}}$
- In modelli nidificati R^2 è sempre maggiore per il modello più ampio
- Il coefficiente di determinazione corretto è $R_c^2 = 1 - \frac{n-1}{n-p}(1 - R^2)$. Esso tiene conto del numero di variabili esplicative inserite nel modello.

Previsioni e residui

- La media (condizionata) di $\boldsymbol{y}$ può essere stimata da $E(\hat{\boldsymbol{y}}) = \hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$ pertanto $\hat{\boldsymbol{y}} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{y} = Hy$
- $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ è una matrice quadrata di dimensione n ed è detta matrice *cappello* o di proiezione e ha le seguenti proprietà:
  1. è simmetrica e idempotente
  2. rango(H)=traccia(H)=p
  3. h_{ii} è compreso fra $1/n$ e 1
  4. la matrice $I - H$ è pure simmetrica e idempotente con rango pari a $(n - p)$
- i residui del modello sono pari a $\boldsymbol{e} = (I - H)\boldsymbol{y}$
- sotto l'ipotesi di normalità $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2(I - H))$

L'analisi dei residui I

- Per una verifica della bontà di adattamento di un modello è necessario effettuare un controllo empirico delllo stesso tramite un'analisi dei residui definiti come differenza fra valori osservati e previsti della variabile dipendente.
- Nei LM i residui vengono opportunamente standardizzati e si ha

$$r_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{S^2(1 - h_{ii})}} \ , \tag{3}$$

  ove h_{ii} è l'i-esimo elemento diagonale di $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.
- Al fine di identificare gli outliers si introducono inoltre i residui studentizzati $e_i^* = \frac{e_i}{S_{(i)}\sqrt{1-h_{ii}}} = e_i(\frac{n-p-1}{n-p-e_i^2})^{1/2}$ dove $S_{(i)}$ è la varianza dei residui stimata omettendo l'osservazione i,

L'analisi dei residui II

- e le distanze di Cook $\frac{1}{p} r_i^2 \frac{h_{ii}}{1-h_{ii}}$
  che presentano valori elevati per le osservazioni che se escluse modificano in modo sostanziale i parametri del modello.
  - $\mapsto$ Strumenti grafici: grafici dei residui contro le variabili esplicative e contro i valori stimati dal modello; diagramma Q-Q normale dei residui, diagramma dei valori h_{ii} e delle distanze di Cook.
  - $\mapsto$ Verifica dell'ipotesi di normalità anche mediante appropriati test (basati su valutazioni di simmetria e curtosi)

Estensioni del modello nel caso di varianza non costante

- Le assunzioni sugli errori potrebbero essere in alcuni casi non soddisfatte. Ad esempio gli errori potrebbero essere eteroschedastici o correlati
- L'assunzione $Cov(\epsilon) = \sigma^2 I$ va allora rimpiazzata con $Cov(\epsilon) = \sigma^2 W^{-1}$ ove W è una matrice definita positiva
- Se utilizziamo lo stimatore visto in precedenza, ignorando l'eteroschdasticità e la correlazione, esso continua ad essere corretto ma la matrice di varianza e covarianza che stimiamo non è quella corretta.
- I test o gli intervalli di confidenza che ricaveremo porteranno quindi a conclusioni errate

Eteroschedasticità

- Illustreremo il caso della presenza di eteroschedasticità. In tal caso $Cov(\epsilon) = \sigma^2 W^{-1}$ con $W^{-1} = diag(1/w_1, 1/w_2, \ldots, 1/w_n)$
- Se si moltiplica ϵ_i per $\sqrt{w_i}$ si ottengono gli errori trasformati $\epsilon_i^* = \sqrt{w_i}\epsilon_i$ che hanno varianza costante
- $var(\epsilon_i^*) = var(\sqrt{w_i}\epsilon_i) = \sigma^2$ e gli errori sono omoschedastici.
- Il modello non cambia se trasformiamo anche la variabile risposta e le covariate (inclusa l'intercetta) in modo analogo.
- Otteniamo quindi $y_i^* = \sqrt{w_i}y_i$ e $x_{ij}^* = \sqrt{w_i}x_{ij}$ per ognuna delle p variabili esplicative (intercetta inclusa) e quindi il modello

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \cdots + \beta_{p-1} x_{i(p-1)}^* + \epsilon_i^*$$

  ha errori omoschedastici e valgono per esso le stesse assunzioni viste, inclusa l'omoschedasticità.
- Le trasformazioni fatte, espresse in termini matriciali, corrispondono a pre moltiplicare i termini del modello $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ per $W^{1/2}$

Minimi quadrati pesati

- ponendo quindi $\boldsymbol{y}^* = W^{1/2}\boldsymbol{y}$, $\boldsymbol{X}^* = W^{1/2}\boldsymbol{X}$ e $\boldsymbol{\epsilon}^* = W^{1/2}\boldsymbol{\epsilon}$ otteniamo

$$\boldsymbol{y}^* = X^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$$

  e siamo ora nel caso un modello lineare omoschedastico per cui la stima dei parametri si ottiene con

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (X^{*\mathrm{T}}X^*)^{-1}X^{*\mathrm{T}}\boldsymbol{y}^* \\ &= (X^{\mathrm{T}}W^{1/2}W^{1/2}X)^{-1}X^{\mathrm{T}}W^{1/2}W^{1/2}\boldsymbol{y} \\ &= (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\boldsymbol{y}\end{aligned}$$

- Tale stimatore è detto dei **minimi quadrati pesati**
- Si noti che i pesi sono inversamente proporzionali alle varianze (diverse) per ogni ϵ
- Osservazioni che sono soggette a maggiore errore casuale pesano quindi meno

Selezione delle variabili esplicative e scelta del modello

In numerose applicazioni la lista di variabili esplicative candidate a entrare nel modello è lunga. Come selezionarle?
Un approccio grezzo è il seguente:

Stimare il modello più complesso che include tutte le potenziali covariate (ed eventualmente le interazioni). Poi rimuovere tutte quelle non significative dal modello. (selezione all'indietro - backward)
Si tratta di una strategia non consigliabile per diversi motivi:

- Il modello che risulta potrebbe comunque essere ridondante
- Con numerose covariate è maggiore il rischio di multi-collinarità (covariate correlate)
- Ci sono molti modelli con essenzialmente la stessa performance ma differente interpretazione sostanziale. Non si ha alcuna garanzia che le variabili che restano siano realmente le più rilevanti dal punto di vista interpretativo.

Selezione delle variabili esplicative

Altri criteri (altrettanto grezzi) per selezionare le variabili sono i seguenti:

- Selezione fra tutti i sottoinsiemi possibili di covariate (si sceglie il migliore fra $\sum_{j=1}^{p} \binom{p}{j}$ modelli possibili)
- Selezione in avanti (forward)
- Selezione a passi - stepwise (una combinazione fra quella in avanti e quella all'indietro)

Uno dei principii da considerare nella costruzione del modello è il cosidetto *rasoio di Occam*, che consiglia di selezionare fra modelli che hanno essenzialmente la stessa performance quelli che sono meno complessi.

Criteri per la scelta del modello

Con riferimento ai LM abbiamo già visto alcuni criteri

- R^2 e R^2 corretto
- test F (per modelli innestati)

ma se ne possono considerare altri, fra questi

- il criterio C_p di Mallows

$$C_p = \frac{\sum_i^n (y_i - \hat{y}_{iM})^2}{\hat{\sigma}^2} - n + M$$

dove M è il numero di covariate nel modello e $\hat{y}_{iM}$ sono i valori previsti con tali M covariate. Il modello da preferire è quello con il più basso C_p.

- Akaike Information Criteria (AIC)

$$AIC = -2l(\hat{\beta}_M, \hat{\sigma}^2) + 2(M+1)$$

Modelli con miglior adattamento sono quelli con AIC più piccolo.
Per un modello lineare con errori gaussiani e p parametri AIC può essere espresso come $AIC = n\log(\hat{\sigma}^2) + 2(p+1)$ ove $\hat{\sigma}^2$ è pari a SSE diviso per n.

- Bayesian Information Criteria (BIC)

$$BIC = -2l(\hat{\beta}_M, \hat{\sigma}^2) + \log(n)(M+1)$$

Evitare la collinearità

Una diagnosi di collinerarità può essere fatta calcolando il *fattore di inflazione della varianza* (VIF) asssociato al j-mo predittore

$$VIF_j = \frac{1}{1 - R_j^2}$$

dove R_j^2 è il coefficiente di determinazione quando x_j viene regredito sulle restanti covariate. ($VIF_j > 10$ è di solito un sintomo che quella variabile potrebbe essere la causa della collinearità).
Possibili soluzioni sono:

- l'omissione di covariate
- l'uso delle componenti principali ottenute dai regressori (o altre combinazioni dei regressori)
- regressione a cresta (regressione Ridge)

Ridge regression

- La Regressione Ridge è stata introdotta per evitare la collinearità adottando un criterio alternativo a quello dei minimi quadrati

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda \boldsymbol{I})^{-1}\mathbf{X}^T\boldsymbol{y}$$

dove λ è un parametro di regolazione scelto con opportuni metodi.

- Tuttavia la regressione Ridge è un esempio di minimi quadrati penalizzati. Si può infatti anche mostrare che corrisponde a introdurre la seguente penalità

$$\hat{\boldsymbol{\beta}}_{PLS} = \arg\min_{\beta}[(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \text{pen}(\boldsymbol{\beta})]$$

ove

$$\text{pen}(\boldsymbol{\beta}) = \sum_{j=1}^{p} \beta_j^2 = \boldsymbol{\beta}^T\boldsymbol{\beta}$$

Se λ è grande il termine di penalizzazione domina e tutti (o quasi tutti) i coefficienti vengono schiacciati verso lo 0

- pen($\boldsymbol{\beta}$) può anche essere inteso come un termine che misura la complessità del modello. Una penalità che è grande se molti $\boldsymbol{\beta}$ sono grandi e diversi da 0.
- $\lambda \geq 0$ è un parametro di regolazione che riflette quanto peso si dà al termine di complessità. che può essere scelto ad esempio usando metodi di cross validation.

Il LASSO (Least Absolute Shrinkage and Selection Operator)

Anche il LASSO corrisponde a una penalizzazione per la funzione di perdita dei mininmi quadrati

$$\hat{\boldsymbol{\beta}}_{LASSO} = \arg\min_{\beta}[(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|]$$

- La penalizzazione scelta con il LASSO tende anch'essa a schiacciare alcuni valori dei coefficienti verso 0. Coefficienti che hanno valori più piccoli vengono schiacciati più decisamente verso lo 0 rispetto a quanto faccia la regressione Ridge.
- Si tratta di un metodo in cui si sceglie un compromesso fra il fit del modello e la sua potenzialmente eccessiva complessità
- Non esiste una soluzione in forma chiusa del problema di ottimo per cui occorre ricorrere a opportuni metodi numerici.
- I metodi come il LASSO o la regressione Ridge, anche per la loro capacità di comporre complessità del modello e adattamento, vengono detti metodi di regolarizzazione.