# Università degli Studi di Torino

Dipartimento di Matematica Giuseppe Peano

Scuola di Scienze della Natura

Corso di Laurea in Matematica



Tesi di Laurea Triennale

## Applicazione di metodi statistici per la validazione di indici diagnostici predittivi: Il caso delle ospedalizzazioni COVID


Relatore: Prof.ssa Maria Teresa Giraudo
Correlatore: Prof.ssa Chiara Berchialla

Candidato: Alessio Cagnacci


2021/2022

# Indice

# Introduzione

Gli ultimi anni sono stati dominati dalla pandemia SARS COVID-19, che ha colpito la popolazione mondiale, causando una grande quantità di ricoveri ospedaliere ed altrettanto ingente numero di morti.
Ogni giorno i media hanno bombardato la popolazione di dati, numero di ricoveri, numeri dei soggetti positivi e negativizzati, indici di contagio e di mortalità.
Gli ospedali hanno subito un ingente aumento di ricoveri. Sono nati reparti di terapia intensiva COVID, tanti sono stati creati ex novo e gli altri incrementati velocemente in tutto il territorio nazionale ed estero.
Tanto clamore, interesse e rilevanza è stato dato al numero di ricoveri in questi reparti, tanto da diventare la cartina al tornasole dell'aggressività della malattia e spunto per prendere decisioni gestionali nazionali su come affrontare al meglio questo virus.
Ogni ospedale ha raccolto scrupolosamente i dati dei ricoveri, decessi e guarigioni nei reparti di terapia intensiva, sub intensiva, e "ordinari" COVID e comunicati quotidianamente agli organi amministrativi che tramite i media li comunicavano, sotto forma di grafici, schemi, semplice informazione alla popolazione.

Proprio grazie a questi dati, numerosi ricercatori hanno provato a prevedere la gravità della malattia non in generale ma per ogni singola persona, così da potere organizzare le strutture in maniera efficiente e, qualora fosse stato necessario fare una scelta, per mancanza di infrastrutture e materiali, permettere ai medici di intervenire sulle persone con più probabilità di sopravvivenza.

I ricercatori per raggiungere questo risultato hanno elaborato numerosi modelli, il principale dei quali, costruito da LIANG et al. é noto alla società medica e matematica come COVID-GRAM[1].
Quest'ultimo, come molti degli altri elaborati, necessitava di numerosi esami diagnostici e di laboratorio da eseguire su ogni singolo paziente, che purtroppo, per lunghi tempi tecnici di elaborazione non avrebbero fornito un dato immediato.
Basti pensare ad un esame del sangue, che richiede tempi di laboratorio più o meno lunghi o le radiografie o risonanze magnetiche che comportano l'impiego di personale o macchinari, non sempre prontamente disponibili oltre al tempo materiale di esecuzione.

Durante l'emergenza sanitaria però la velocità è essenziale e pertanto trovare una metodologia più rapida avrebbe significato salvare molte più vite.
Numerosi studiosi in campo medico giungono, nel frattempo, alla conclusione che il cosiddetto LUS score, ovvero la gravità di una semplice ecografia toracica, è un test diagnostico molto efficiente. Forti di questo, ad alcuni nasce l'idea di elaborare un altro punteggio da assegnare ad ogni singolo paziente, ricavato unendo alcuni dati del paziente già utilizzati per il COVID-GRAM.[2]
Dapprima unendoli tutti insieme al LUS formando così il GRAM-plus per poi eliminare i predittori più superflui.
Si sceglie quindi di utilizzare il numero di malattie pregresse, la durata dei sintomi prima dell'accesso, il P/F ratio, la presenza di dispnea e la presenza di LUS score superiore a 15.

Nel mio elaborato nei primi due capitoli affronterò i concetti teorici ed i metodi che

sono alla base di questo studio, ovvero la regressione logistica e le curve Roc, e nei successivi, attraverso l'utilizzo del software R, li metterò in praticacon i dati dello studio scritto sopra.

Per rendere di più facile comprensione quanto segue è necessario chiarire che sarà utilizzato il termine:
POSITIVO per indicare le persone con elevato rischio di sviluppare una malattia grave, ovvero il ricovero in terapia intensiva, l'utilizzo di cure palliative o il decesso
NEGATIVO per indicare le persone che contrarranno malattie lievi.

# Capitolo 1

# Regressione logistica

Partendo dalle informazioni di cui siamo in possesso ed avendo a disposizione una variabile qualitativa dicotomica, la *regressione logistica* è il metodo che permetterà di prevedere la gravità della malattia per ciascun paziente.
Affronteremo pertanto in questo capitolo le premesse teoriche necessarie traendo spunto essenzialmente da [3] [4].

## 1.1 Basi per costruzione modello di regressione logistica

Analizziamo per prima la teoria che sta dietro ad una *regressione logistica* e principalmente il metodo *logit*, ovvero il principale metodo di costruzione.

### 1.1.1 Primo approccio al problema

La variabile di risposta del modello, che chiameremo $Y$, è limitata a due valori, 0 e 1, dove $positivo = 1$ e $negativo = 0$.
La probabilità $p = P(Y = 1)$ è il parametro di interesse rappresentando la percentuale dei pazienti che dovrà ricorrere a cure avanzate (terapia intensiva o cure palliative) o addirittura morirà.
La probabilità di $Y = 0$ sarà data quindi da $1 - p$.
Ne calcoliamo *media* e *varianza*:

$$\mathbb{E}[Y] = 0 \cdot (1 - p) + 1 \cdot p = p$$

$$\mathrm{Var}(Y) = 0^2 \cdot (1 - p) + 1^2 \cdot p - p^2 = p \cdot (1 - p)$$

L'idea iniziale è quindi di linearizzare il valore atteso e costruire un modello che lo metterà a confronto con le altre variabili presenti:

$$\mathbb{E}[Y|z] = p = \beta_0 + \beta_1 z + \epsilon$$

Ma si evidenziano subito alcune problematiche:

- I valori previsti della risposta Y potrebbero diventare maggiori di 1 o inferiori a 0 essendo l'espressione lineare per il suo valore atteso illimitata.
- Non è rispettata una delle ipotesi di un'analisi di regressione. La varianza di Y non è costante su tutti i valori della variabile predittiva z.

Pur sembrando quindi questa la procedura più semplice per raggiungere il nostro scopo, abbiamo bisogno di un altro metodo.

### 1.1.2 Modello logit

Abbiamo per prima cosa bisogno di definire due nuove quantità:

**Definizione 1.1.** L' *odds-ratio* è definito come il rapporto tra la probabilità di un evento e quella del suo complementare ovvero:

$$\text{odds}(P(A) = p) = \frac{P(A)}{P(\bar{A})} = \frac{p}{1-p}$$

Notiamo subito che l' *odds-ratio*, a differenza della probabilità, può essere maggiore di 1.
Pur risolvendo in parte uno dei problemi trovati nella prima sezione, ancora non può essere negativo.
Inoltre $\text{odds}(p) = 1$ implica che, l'evento ed il suo complementare, avvengano con la stessa probabilità.
Quindi decidiamo di applicare il logaritmo naturale a questa quantità, che prenderà il nome di *logit*.

**Definizione 1.2.** il *logit* o *log-odds* è definito come:

$$\text{logit}(p) = \log \text{odds}(p)$$

Utilizzando le proprietà del logaritmo abbiamo che:

- $\text{logit}(P(A)) = 0 \rightarrow P(A) = P(\bar{A})$
- $\text{logit}(P(A)) > 0 \rightarrow P(A) > P(\bar{A})$ ed inoltre crescerà più lentamente dell'*odds*.
- $\text{logit}(P(A)) < 0 \rightarrow P(A) < P(\bar{A})$ ed inoltre decrescerà più velocemente dell'*odds* verso 0.

Abbiamo così risolto i problemi riscontrati nella sezione precedente e possiamo quindi assumere che il grafico di *logit* sia una linea retta nella variabile predittiva $z$

$$\text{logit}(p) = \log \text{odds}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 z + \epsilon$$

Il *log-odds* è lineare nelle variabili predittive ed $\epsilon$ rappresenta solo un errore trascurabile.

Dal momento che, è più semplice ragionare in termini di probabilità, cerchiamo di ricavarla dalla formula precedente:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 z \rightarrow \text{odds}(p(z)) = \frac{p(z)}{1-p(z)} = \exp(\beta_0 + \beta_1 z)$$
$$\rightarrow p(z) = \frac{\exp(\beta_0 + \beta_1 z)}{1 + exp(\beta_0 + \beta_1 z)}$$

Quest'ultima equazione prende il nome di *curva logistica*.
Osserviamo che:

- La relazione tra $p$ e $z$ non è lineare ma è un grafico a forma di s
- Il valore $\beta_0$ dà il valore di $p$ quando $z = 0$, ovvero $p = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$
- Il parametro $\beta_1$ determina quanto velocemente $p$ cambia al variare di $z$, senza dimenticare che questa relazione non è di tipo lineare.

Quindi possiamo riassumere i concetti principali appena espressi con una definizione:

**Definizione 1.3.** La *curva logistica* è la curva che descrive un modello logit per la regressione logistica:

$$p(z) = \frac{\exp(\beta_0 + \beta_1 z)}{1 + exp(\beta_0 + \beta_1 z)} \qquad oppure \qquad p(z) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 z)}$$

Un esempio di curva logistica è rappresentata dalla figura seguente.

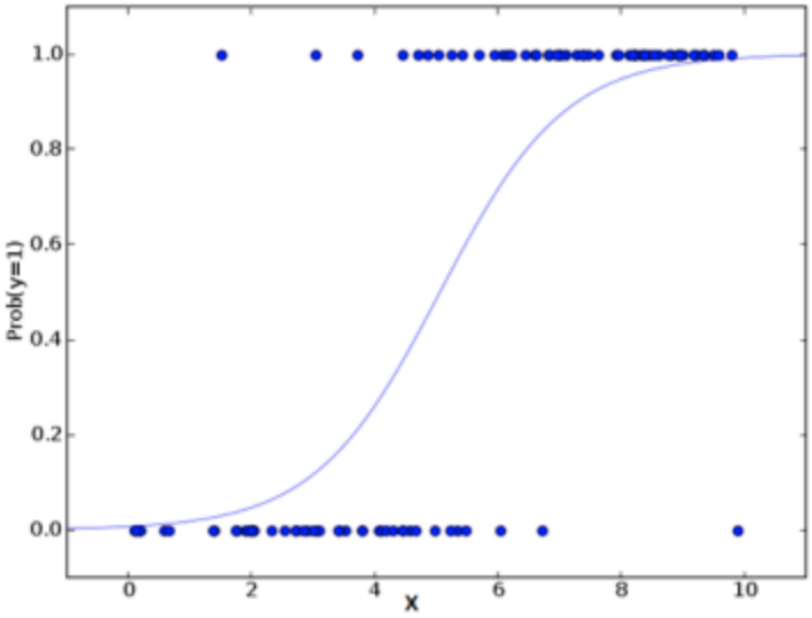



Figura 1.1: Grafico di un modello logit

### 1.1.3 Generalizzazione a più variabili predittive

Fino ad adesso abbiamo esaminato il caso in cui sia presente una sola variabile predittiva. A questo punto generalizziamo il modello al caso in cui ci siano $r$ variabili predittive per la $j$-esima osservazione.

Poniamo per semplicità $\mathbf{z_j} = [1, z_{j1}, z_{j2}, \ldots, z_{jr}]'$ e assumiamo che l'osservazione $Y_j$ sia di *Bernoulli* con probabilità di successo $p(\mathbf{z_j})$ dipendente dal valore delle variabili predittive.
Pertanto:

$$P(Y_j = y_1) = p^{y_j}(\mathbf{z_j})(1 - p(\mathbf{z_j}))^{1-y_j} \qquad per \qquad y_j = 0; 1$$

$$\mathbb{E}[Y_j] = p(\mathbf{z_j}) \qquad e \qquad \mathrm{Var}(Y_j) = p(\mathbf{z_j})(1 - p(\mathbf{z_j}))$$

Per gli stessi ragionamenti già applicati, al posto di usare il valore atteso, si linearizzerà il *logit*; ottenendo quindi:

$$\log \frac{p(\mathbf{z_j})}{1 - p(\mathbf{z_j})} = \beta_0 + \beta_1 \mathbf{z_1} + \ldots + \beta_r \mathbf{z_r} = \boldsymbol{\beta}' \mathbf{z_j} \qquad dove \qquad \boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_r]'$$

La *curva logistica* quindi sarà descritta da:

$$p(\mathbf{z}) = \frac{\exp(\boldsymbol{\beta}' \mathbf{z_j})}{1 + \exp(\boldsymbol{\beta}' \mathbf{z_j})}$$

## 1.2 Analisi di una regressione logistica semplice

In questa sezione introdurremo i metodi utilizzabili per studiare una regressione logistica e stimare gli elementi del vettore $\boldsymbol{\beta}$, al fine di ottenere una stima più corretta possibile del valore dell' *odds-ratio* e di conseguenza del valore della probabilità dell'evento d'interesse.

### 1.2.1 Metodo di massima verosimiglianza

Per prima cosa definiamo cosa è la *verosimiglianza*[5]

**Definizione 1.4.** La funzione di *versosimiglianza* si indica con $\mathcal{L}$ (likelihood) ed è definita come la distribuzione congiunta del campione casuale $X = (X_1, \ldots, X_n)$ di cui sono già note le osservazioni $x = (x_1, \ldots, x_n)$.
Quindi é una funzione nella sola incognita $\theta$, ovvero:

$$\mathcal{L}(x, \theta) = f(x, \theta) = f(x_1, \theta) f(x_2, \theta) \cdots f(x_n, \theta) = \mathcal{L}(\theta)$$

Nel caso in cui $\mathcal{L} > 0$ si usa di solito la *log-verosiglianza* che si indica con L ed è data da $L(\theta) = \log \mathcal{L}(\theta)$.

Una stima degli elementi del vettore $\boldsymbol{\beta}$ può essere ottenuta con il metodo di *massima verosimiglianza* ovvero massimizzando la funzione appena definita.
Nel nostro caso sarà:

$$\begin{aligned}\mathcal{L}(b_0, b_1, \ldots, b_r) &= \prod_{j=1}^{n} p^{y_j}(\mathbf{z_j})(1 - p(\mathbf{z_j}))^{1-y_j} \\ &= \frac{\prod_{j=1}^{n} e^{y_j(b_0 + b_1 z_{j1} + \ldots + b_r z_{jr})}}{\prod_{j=1}^{n} 1 + e^{(b_0 + b_1 z_{j1} + \ldots + b_r z_{jr})}}\end{aligned}$$

I valori dei parametri che massimizzano questa funzione non possono essere espressi in una forma semplice, per trovarli sono necessari metodi di iterazione numerica che, partendo da un valore dato, forniscono questi risultati, nello specifico il metodo dei *minimi quadrati riponderati*.
Denotiamo questi valori ottenuti numericamente con il vettore $\hat{\boldsymbol{\beta}}$.

### 1.2.2 Intervallo di confidenza per parametri

**Definizione 1.5.** Definiamo *intervallo di confidenza* [5] per $\boldsymbol{\theta}$ con coefficiente di confidenza $1 - \alpha$ qualsiasi sottoinsieme S tale che:

$$P(\boldsymbol{\theta} \in S) = 1 - \alpha$$

Quando il campione è grande, $\hat{\boldsymbol{\beta}}$ è approssimativamente una normale con media $\boldsymbol{\beta}$, e l'approssimazione della matrice di covarianza è:

$$\hat{\Sigma}(\boldsymbol{\beta}) \approx \left[ \sum_{j=1}^{n} \hat{p}(z_j)(1 - \hat{p}(z_j)) z_j z_j^{'} \right]^{-1}$$

Le radici quadrate degli elementi sulla diagonale di questa matrice sono le *deviazioni standard* stimate del campione o, l' *errore standard (SE)* degli stimatori $\hat{b}_k$.

Ad esempio l'intervallo di confidenza del 95% per $\hat{b}_k$ è:

$$\hat{b}_k \pm 1,96 SE(\hat{b}_k)$$

In questo caso gli intervalli di confidenza possono essere utilizzati per giudicare l'influenza dei singoli termini per la costruzione del modello *logit*, in modo da potere eliminare quelli insignificanti.

### 1.2.3 Test sul rapporto di verosimiglianza

Indichiamo il valore di *massima verosimiglianza* con $\mathcal{L}_{max} = \mathcal{L}(\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_r)$.
Supponiamo adesso che l'ipotesi nulla sia $H_0 : \beta_k = 0$ e consideriamo il modello ridotto, sarà quindi possibile, con il metodo dei *minimi quadrati ponderati*, calcolare gli stimatori di massima verosiglimiglianza e il valore di quest'ultima:

$$\mathcal{L}_{max,Ridotto} = \mathcal{L}(\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_{k-1}, \hat{b}_{k+1}, \ldots, \hat{b}_r)$$

Nei modelli di regressione logistica si è soliti studiare la veridicità di $H_0$ con la formula che prende il nome di *devianza*:

$$-2\log\left(\frac{\mathcal{L}_{max,Ridotto}}{\mathcal{L}_{max}}\right)$$

$H_0$ viene $rifiutata$ se il valore della *devianza* è alto.
La *devianza* è approssimativamente distribuita come una $\chi^2$ con un grado di libertà (definizioni a seguire), se il modello ridotto ha una variabile predittiva in meno.

**Definizione 1.6.** Se $Z_i$ (con $i = 1, 2, \ldots, g$) è una sequenza di variabili casuali indipendenti normali standardizzate allora la variabile casuale *chi-quadrato* ($\chi^2$) è così definita [5]:

$$X = \sum_{i=1}^{g} Z_i^2$$

Inoltre il parametro $g$ è detto *gradi di libertà* della chi-quadrato.

Un altro metodo per verificare la veridicità dell'ipotesi $H_0$ prende il nome di *test di Wald*, che utilizza una statistica test $Z = \hat{\beta}_k / SE(\hat{\beta}_k)$ oppure la sua versione $\chi^2$ con un grado di libertà.
In genere però il test sul rapporto di verosimiglianza da valori più precisi.

Si può inoltre generalizzare quest'ultimo test al caso in cui $m$ parametri siano simultaneamente uguali a zero; la *devianza* sopra definita sarà distribuita come una $\chi^2$ con $m$ gradi di libertà.

## 1.3 Regressione logistica con risposta binomiale

In questa sezione generalizziamo ancora i concetti considerando il caso in cui vengono fatte diverse osservazioni per lo stesso valore della covariate $\boldsymbol{z_j}$ e ci siano un totale di $m$ differenti insiemi di dati, nei quali questi predittori sono costanti.
Eseguite $n_j$ prove con la stessa variabile predittiva $\boldsymbol{z_j}$ il risultato $Y_j$ è una distribuzione *binomiale* con probabilità $p(\boldsymbol{z_j}) = p(successo|\boldsymbol{z_j})$.

### 1.3.1 Funzione di massima verosimiglianza e covarianza

Essendo le $Y_j$ indipendenti, la funzione di *massima verosimiglianza* sarà:

$$\mathcal{L}(\beta_0, \beta_1, \ldots, \beta_r) = \prod_{j=1}^{m} \binom{n_j}{y_j} p^{y_j}(\boldsymbol{z_j})(1 - p(\boldsymbol{z_j})^{n_j - y_j})$$

Come nella sezione 1.2.1 lo stimatore di massima verosimiglianza $\hat{\boldsymbol{\beta}}$ si può trovare solo con metodi di iterazione numerica.

Quando il campione è grande la matrice di *covarianza* è:

$$\hat{\Sigma}(\boldsymbol{\beta}) \approx \left[\sum_{j=1}^{m} n_j \hat{p}(\boldsymbol{z_j})(1-\hat{p}(\boldsymbol{z_j}))\boldsymbol{z_j}\boldsymbol{z_j}'\right]^{-1}$$

Inoltre l'i-esimo elemento è una stima per $\hat{\beta}_{i+1}$ e la sua radice quadrata è una stima dell'errore standard sul campione, ovvero $SE(\hat{\beta}_{i+1})$.

Si può inoltre stimare, partendo da questa matrice, la *varianza* di $\hat{p}(\boldsymbol{z_k})$:

$$\mathrm{Var}(\hat{p}(\boldsymbol{z_k})) \approx [\hat{p}(\boldsymbol{z_k})(1-\hat{p}(\boldsymbol{z_k}))]^2 \boldsymbol{z_j}' \left[\sum_{j=1}^{m} n_j \hat{p}(\boldsymbol{z_j})(1-\hat{p}(\boldsymbol{z_j}))\boldsymbol{z_j}\boldsymbol{z_j}'\right]^{-1} \boldsymbol{z_k}$$

## 1.4 Analisi del modello

Dopo avere creato un modello è necessario chiedersi se sia il migliore, ponendoci tre domande:

- C'è qualche deviazione sistematica dal modello scelto?
- Ci sono osservazioni che in maniera anomala non si adattano allo schema generale degli altri dati?
- Ci sono alcune osservazioni che portano grandi cambiamenti nell'analisi statistica se vengono o meno considerati?

. Utilizzeremo i test che seguono per verificarne le risposte e correggere eventuali anomalie.

### 1.4.1 Generalizzazione del concetto devianza

Prima di generalizzare totalmente il concetto cerchiamo di condurci ad un caso più semplice.

Se non ci sono strutture parametriche per il singolo esperimento statistico, la probabilità $p(\hat{\boldsymbol{z_j}})$ può essere stimata usando il numero dei successi $y_j$ in $n_j$ prove.
Sotto questa ipotesi il contributo del caso j-esimo $\hat{\boldsymbol{z_j}}$ alla funzione di verosimiglianza è:

$$\binom{n_j}{y_j} p^{y_j}(\mathbf{z_j})(1-p(\mathbf{z_j})^{n_j-y_j}$$

La funzione è massimizzata dalla scelta $\hat{p}(\mathbf{z_j}) = \frac{y_j}{n_j}$ per ogni $j$ da 1 a $m = \sum n_j$.
Quindi otterremo:

$$-2\log \mathcal{L}_{max,NP} = -2\sum_{j=1}^{m}\left[y_j \log\left(\frac{y_j}{n_j}\right) + (n_j - y_j)\log\left(1-\frac{y_j}{n_j}\right)\right] + 2\log\prod_{j=1}^{m}\binom{n_j}{y_j}$$

Il secondo termine è comune per tutti i modelli possibili.

Possiamo quindi definire la *devianza* (cfr. 1.2.3) fra il modello senza strutture parametriche e il modello adattato ai parametri, avente una costante ed $r-1$ predittori.

Ovvero in formula:

$$G^2 = 2\sum_{j=1}^{m}\left[y_j \log\left(\frac{y_j}{\hat{y_j}}\right) + (n_j - y_j)\log\left(\frac{n_j - y_j}{n_j - \hat{y_j}}\right)\right]$$

Nella quale $\hat{y_j} = n_j\hat{p}(\boldsymbol{z_j})$ è il numero di successi adattato con i parametri.
$G^2$, se il campione di partenza è di grandi dimensioni, è approssimativamente distribuita come una $\chi^2$ con $f$ gradi di libertà, dove $f$ è la differenza tra $m$ (numero di osservazioni) e $\beta$ (numero di parametri stimati).

La *devianza* tra il modello completo e quello ridotto forniscono un contributo per la stima degli altri valori predittori.

$$G^2_{Ridotto} - G^2_{Completo} = -2\log\left(\frac{\mathcal{L}_{max,Ridotto}}{\mathcal{L}_{max}}\right)$$

La suddetta differenza è approssimativamente una $\chi^2$ con $df = f_{Ridotto} - f_{Completo}$ gradi di libertà.

### 1.4.2 Leverage point e influenza delle osservazioni

**Definizione 1.7.** *Leverage* [6] o effetto leva è un termine usato nell'analisi di regressione per indicare una misura di quanto sono lontani i valori delle variabili indipendenti di un'osservazione da quelli delle altre osservazioni.
I *levarage point*, se presenti, sono valori anomali rispetto alle variabili indipendenti. Se vengono cancellati hanno il potenziale di causare grandi cambiamenti nelle stime dei parametri .

Vista la loro influenza è quindi importante individuarli.
Uno dei metodi è l'utilizzo della matrice $\boldsymbol{H}$ costruita come di seguito:

$$\boldsymbol{H = V^{-1/2} Z (Z' V^{-1} Z)^{-1} Z' V^{-1/2}}$$

dove $\boldsymbol{V^{-1}}$ è una matrice diagonale con elementi $n_j\hat{p}(\boldsymbol{z_j}(1 - \hat{p}(\boldsymbol{z_j}))$.
Gli elementi diagonali di $\boldsymbol{H}$ identificano le osservazioni che hanno una forte influenza nel modello di regressione in cui si sta lavorando.

### 1.4.3 Test dei residui e di buona adattabilità

**Definizione 1.8.** I *residui* [5] in un'analisi di regressione sono la differenza (o errore) tra i valori osservati e stimati.
I valori osservati che si trovano al di sopra della curva di regressione hanno un valore residuo positivo e i valori osservati che scendono al di sotto della curva di regressione hanno un valore residuo negativo.

I *residui* possono essere studiati nei casi in cui sembrano esserci problemi di mancanza di adattabilità alla forma del modello *logit* costruito e per la scelta delle variabili predittive.

Ci sono tre modi per definirli:

- Residui devianza: $\mathrm{d}_j = \pm\sqrt{2\left[y_j \log\left(\frac{y_j}{n_j\hat{p}(\boldsymbol{z_j})}\right) + (n_j - y_j)\log\left(\frac{n_j - y_j}{n_j - n_j\hat{p}(\boldsymbol{z_j})}\right)\right]}$
  dove il segno di $d_j$ è lo stesso di $y_j - n_j\hat{p}(\boldsymbol{z_j})$

- Residui di Pearson:
$$r_j = \frac{y_j - n_j\hat{p}(\boldsymbol{z_j})}{\sqrt{n_j\hat{p}(\boldsymbol{z_j})(1-\hat{p}(\boldsymbol{z_j}))}}$$

- Residui di Person standardizzati:
$$r_{sj} = \frac{r_j}{\sqrt{1-\boldsymbol{h_{jj}}}}$$
dove $h_{jj}$ è il j-esimo elemento diagonale della matrice $\boldsymbol{H}$.

Un test generale per l'adattabilità di tutto il modello, specialmente se il campione è piccolo, è ottenuto utilizzando la variabile $\chi^2$ dei residui di Pearson nel modo seguente:

$$\chi^2 = \sum_{j=1}^{m} r_j^2 = \sum_{j=1}^{m} \frac{(y_j - n_j\hat{p}(\boldsymbol{z_j}))^2}{n_j\hat{p}(\boldsymbol{z_j})(1-\hat{p}(\boldsymbol{z_j}))}$$

### 1.4.4 Test di Hosmer - Lemeshow

Uno dei test più utilizzati per verificare la bontà di un modello logistico è quello di *Hosmer - Lemeshow* [7] che si basa sulla divisione in gruppi dei valori di stima della probabilità del modello.
Si suppone che ci siano $n$ stime di probabilità divise in $n$ colonne distinte partendo dal valore più basso per arrivare al valore più alto.
Si decide quindi di divedere queste $n$ colonne in $g$ gruppi, in genere $g = 10$. I gruppi prenderanno il nome di $g_i$ con $i$ che va da 1 a 10, tutti conterranno $n'_i = n/10$ elementi e con l'aumentare del valore di $i$ aumenterà il valore degli elementi. (metodo dei percentili)
In alternativa si può eseguire una divisione per stima di probabilità, inserendo nei gruppi i valori compresi tra date soglie avremmo quindi nel primo i soggetti con stima compresa tra 0 e $0,1$, nel secondo quelli tra $0,1$ e $0,2$ e così via fino al decimo.
In entrambi i casi il test per la bontà statistica di *Hosmer - Lemeshow*, $\hat{C}$, è ottenuto calcolando i valori della variabile $\chi^2$ di Pearson da $g \times 2$ tabelle di osservazioni e stimando le rispettive frequenze. In formule:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k\bar{\pi}_k)^2}{n'_k\bar{\pi}_k(1-\bar{\pi}_k)}$$

dove $c_k$ denota il numero di modelli di covariate presenti nel *k-esimo* decile, $o_k$ rappresenta il numero di risposta attraverso le covariate e $\bar{\pi}_k$ rappresenta una stima di probabilità. In formule:

$$o_k = \sum_{j=1}^{c_k} y_j; \bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j\hat{\pi}_j}{n'_k}$$

Si ha che se abbiamo un buon modello la distribuzione $\hat{C}$ si approssima bene con la distribuzione di una $\chi^2$ con $g-2$ gradi di libertà. Quindi nel test le confrontiamo e riteniamo buono un modello con *p-value* vicina ad uno.

# Capitolo 2

# Curve ROC: Receiver Operating Characteristic

Nel capitolo precedente abbiamo esaminato come costruire il modello di *regressione logistica* più adatto ad analizzare i dati in nostro possesso.
Attraverso le *curve ROC*, adesso, cercheremo di capire come utilizzarlo per prevedere se l'esito sarà positivo o negativo decidendo il valore di soglia migliore. per discriminare tra due classi [8] [9] [10]

## 2.1 Teoria sulle curve ROC

### 2.1.1 Tabella di contingenza

Se si considera, come nel nostro caso un problema di predizione a due classi, scelta una quantità soglia per passare dal risultato negativo a positivo, sono possibili quattro risultati:

**TP**) se il valore della predizione è *positivo* e il valore reale *positivo* (vero positivo)

**FP**) se il valore della predizione è positivo e il valore reale *negativo* (falso positivo)

**TN**) se il valore della predizione è *negativo* e il valore reale *negativo* (vero negativo)

**FN**) se il valore della predizione è *negativo* e il valore reale *positivo* (falso negativo)

I quali possono essere riassunti in una tabella detta di contingenza:

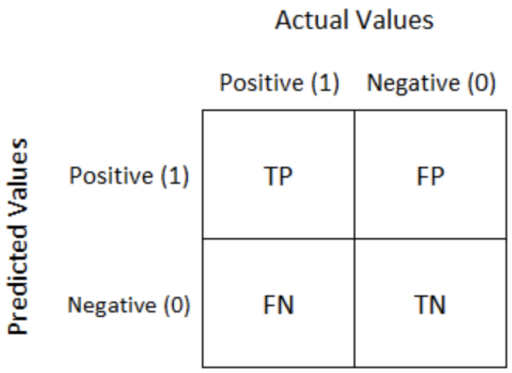



Figura 2.1: Tabella di contingenza

Il modello sarà migliore con un numero basso di risultati falsi.
A tal fine avranno molta importanza tre valori che prendono il nome di *Sensibilità (Se)*, *Specificità (Sp)* e *Precisione*.

Sensibiltà Rappresenta il rapporto tra i veri positivi e i positivi trovati con il modello e si indica con $TPR$.

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

Specificità Rappresenta il rapporto tra veri negativi e i negativi trovati con il modello e si indica con $TNR$.

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

Precisione Rappresenta quanto precisamente il predittore predice i valori positivi e si indica con $PPV$.

$$PPV = \frac{TP}{TP + FP}$$

A volte è utile calcolarla per i valori negativi e prende il nome di $NPV$.

$$NPV = \frac{TN}{TN + FN}$$

I valori $FNR$ e $FPR$ indicano rispettivamente il rapporto tra falsi negativi e negativi e quello tra falsi positivi e positivi.

Partendo da questi concetti analizziamo i valori più utili per cercare di minimizzare gli errori di predizione e trovare la soglia migliore per questi ultimi.
Sarà quindi necessario costruire uno spazio, detto per l'appunto *spazio ROC*.

### 2.1.2 Sensibilità e specificità

Notevole importanza nello studio statistico sia in ambito di diagnostica che di screning è assunta dalla Sensibilità e dalla Specificità.[11]
L'utilità di questi due valori è data dalla loro proprietà di essere strettamente collegate tra di loro; se una è più alta l'altra è bassa e viceversa.
Quindi è necessario sapere in quali casi è preferibile avere l'una o l'altra più elevata.
Un test che rileva in modo affidabile la presenza di una condizione, con conseguente alto numero di veri positivi e un basso numero di falsi negativi, avrà un'alta sensibilità. Ciò è particolarmente utile quando la conseguenza del mancato trattamento della condizione è grave, il trattamento è molto efficace e ha effetti collaterali minimi.
Un test che esclude in modo affidabile gli individui che non hanno la condizione, con conseguente alto numero di veri negativi e un basso numero di falsi positivi, avrà un'alta specificità. Questo è particolarmente importante quando le persone che sono identificate come aventi una condizione possono essere sottoposte a più test e spese.

### 2.1.3 Spazio ROC

Lo *spazio ROC* si forma mettendo la $1 - Specificità$ ($FPR$) sulle $x$ e la $Sensibilità$ sulle $y$, in modo da sfruttare le proprietà espresse in precedenza.
Ogni punto tracciato, quindi, rappresenterà questi due valori per una dato soglia per il predittore.
Il migliore metodo predittivo possibile si troverà nel punto $(0; 1)$, che prende il nome di classificatore perfetto.
La diagonale, che rappresenta i predittori casuali, divide lo spazio in due parti; La parte superiore, dove i predittori sono buoni essendo migliori dei casuali, e la parte inferiore

dove non lo sono.
Partendo da ciò si può notare che, se un classificare è strettamente negativo, attraverso una semplice inversione può diventare ottimo.
Unendo i vari punti tracciati in questo spazio per ogni possibile soglia di predizione, è possibile tracciare la curva, appunto chiamata *Curva ROC* che ci permetterà di capire se il modello predittivo é corretto e la soglia ottimale di predizione.

### 2.1.4 Curva ROC

Poniamo $X$ e $Y$ come due variabili aleatorie indipendenti la prima rappresentante un test diagnostico positivo e l'altra negativo, entrambe saranno definite con un classificatore scelto.
Senza perdere nessuna generalità , e per un valore soglia $c$, il risultato del test sarà positivo se superiore e negativo altrimenti.
Siano $F$ e $G$ le funzioni di distribuzione rispettivamente di $X$ e $Y$.
La *sensitività* del test sarà quindi data da $Se(p) = 1-G(p)$, e la *specificità* sarà definita $Sp(c) = F(c)$.

**Definizione 2.1.** La *curva ROC* viene definita come il grafico di $Se(c)$ su $1 - Sp(c)$ per ogni $c \in [-\infty, \infty]$, oppure equivalentemente in formula:

$$ROC(t) = 1 - G(F^{-1}(1-t))$$

dove $t \in [0,1]$ e $F^{-1}(1-t) = inf\{x \in: F(x) \geq 1-t\}$.

La *curva ROC* così costruita sarà crescente e invariante per trasformazione monotona crescente delle variabili $X$ e $Y$.
Pertanto, il tasso dei veri positivi è dato da:

$$TPR(c) = \int_c^{\infty} F(x)dx$$

Il tasso di falsi positivi è dato da:

$$FPR(c) = \int_c^{\infty} G(x)dx.$$

## 2.2 Utilizzi della curva Roc

In questa sezione vedremo gli utilizzi e le principali proprietà della curva ROC.

### 2.2.1 AUC: area sotto la curva

**Definizione 2.2.** L' *AUC* (area sotto la curva ROC) viene definita come:

$$AUC = \int_0^1 ROC(t)dt$$

Questo valore, che ha molta importanza, è la probabilità che un classificatore, dati due valori a caso uno positivo e l'altro negativo, ci dica con esattezza quale dei due è positivo.
in formule:

$$A = \int_{x=0}^{1} TPR(FPR^{-1}(x))dx = \int_{\infty}^{-\infty} TPR(T)FPR'(T)dT =$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{I}(T' > T)F(T')G(T)dT'dT = P(Y > X)$$

L' *AUC* ha un valore che oscilla tra 0 e 1, dove 1 rappresenta la perfezione e $0,5$ un modello casuale.
In genere si considera buono un predittore se tale valore è più vicino ad 1 rispetto che a $0,5$ e quindi il modello con *AUC* maggiore sarà quasi sicuramente il migliore.
Questo metodo ha il vantaggio di riassumere in un unico valore la qualità del predittore e di conseguenza di ottenere un risultato non dipendente dal valore soglia scelto.
La possibilità di due diverse curve con la stessa area è un limite che queste possono avere. In quest'ultimo caso si dovrà procedere con il confronto di parti di area.
Altro limite potrebbe essere l'avere bisogno di sapere la bontà del predittore per alcuni valori soglia specifici con la conseguente necessità di dovere analizzare altri fattori.

### 2.2.2 Optimal cut point

Uno dei principali utilizzi della *curva ROC* è la ricerca del migliore valore soglia possibile per un predittore, questo punto prende il nome di *Optimal cut point*.

Abbiamo visto che per costruzione la *curva ROC* rappresenta la relazione tra $specificit$à e $1-sensibilit$à a seconda del valore soglia scelto nel predittore e che lo spazio è costruito in modo tale che il predittore perfetto si trova nel punto $(0;1)$.

Il valore soglia migliore sarà la soglia corrispondente al luogo della curva più vicino a tale punto.
I principali metodi per calcolarlo sono:

1. Il metodo basato sull'indice di *Youden* definito come la distanza massima tra la *curva ROC* e la bisettrice, cercando il punto corrispondente e la soglia che rappresenta.
$$J = \max_i(specificit\text{à}(i) + 1 - sensibilit\text{à}(i))$$

2. Il metodo che utilizza la definizione quindi si trova la distanza da $(0;1)$ ad ogni punto della curva e si prende il punto in cui è minore. L'optimal cut point sarà la soglia corrispondente.

3. Il metodo che si basa sull'analisi dei costi benifici calcolondo la pendenza della ROC nelle varie soglie e soppesando quindi i costi per ogni diversa possibilità. Alla fine si scelgono quelle con costo più basso così calcolato:
$$S = \frac{1-p}{p}\frac{C_{FP} - C_{TN}}{C_{FN} - C_{TP}}$$
dove p è la probabilità di avere esito positivo e C rappresenta il costo.

4. Il metodo che si basa su minimizzare l'errore di massima frequenza e prende il nome di Minimax ovvero in formule:
$$\min_c(\max(p(1 - Se(c)), (1 - p)(1 - Sp(c))))$$

Etc Numerosi altri metodi sono riportati nell'articolo [18]

# Capitolo 3

# Introduzione ai dati

In questo capitolo, attraverso l'utilizzo del software R, eseguiamo una prima analisi dei dati in nostro possesso.
Si sceglie di utilizzare R essendo uno dei software più completi per la statistica computazionale e la grafica. Utilizzeremo più precisamente la versione 4.0.2.

## 3.1 Ricavare i dati

In questa sezione vedremo quali dati si sceglie di utilizzare per il nostro studio e si cerca di costruire il database migliore per i nostri obbiettivi.

### Ricerca dei dati

Durante la prima ondata del COVID-19 tra il 26 febbraio 2020 e il 17 maggio dello stesso anno sono stati raccolti i dati di "anamnesi" di tutti i pazienti adulti con una confermata infezione da SARS-CoV-2 nell'ospedale San Giovanni Bosco di Torino.
Sono stati esclusi i pazienti che hanno contatto il COVID in ospedale e che avevano già in precedenza problemi gravi ai polmoni, come ad esempio un infezione batterica molto estesa.
La presenza della malattia veniva confermata attraverso l'uso di un tampone molecolare.
Accertata la positività entro 48 ore venivano raccolte per ogni paziente le caratteristiche demografiche, le malattie, i sintomi e giorni passati dall'inizio dei primi sintomi, vari segni clinici, risultati dei test di laboratorio e radiologici.
Tutti i dati sono infine stati inseriti in un database Excel dedicato.

### Preparazione del database

Una volta ottenuto l'insieme di dati da analizzare è necessario importarlo in R. In questo caso essendo in formato Excel si utilizza il pacchetto *readxl*. [12]
Lo chiameremo per semplicità DB.
In seguito dopo avere utilizzato il comando "attach(DB)" sarà possibile eseguire tutti i controlli e, in caso ci siano dei problemi, eseguire tutti i cambiamenti necessari al pieno funzionamento del nostro database.

## 3.2 Analisi descrittiva dei dati

In questa sezione vedremo come utilizzare R per eseguire un analisi descrittiva e lo applicheremo ai nostri dati.

## Analisi descrittiva in R

Il software R ha diverse funzioni che ci permettono di analizzare il nostro database in modo da avere una descrizione del campione.
Nel caso in cui ci troviamo a lavorare con variabile quantitative, come ad esempio età o LUS score, è possibile attraverso la funzione "summary()" avere un riassunto delle principali analisi statistiche come la media, la mediana e i quantili.
Nel caso in cui si voglia avere informazioni più precise sulla media, essendo il dato probabilmente di maggiore importanza, è possibile utilizzare il cosiddetto "t.test()" il quale ci restituirà il valore e un analisi dettagliata dello stesso che comprende:
La costruzione di una variabile T-student per descrivere la media e quindi un intervallo di confidenza della stessa al 95%.

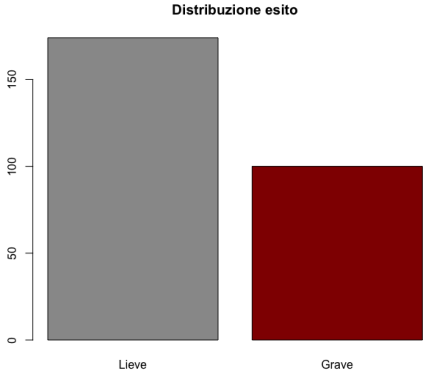
**Definizione 3.1.** La distribuzione $T$ di *Student* con $n$ gradi di libertà è definita posto $Z$ come una variabile *normale* e $G$ una $\chi^2$ con $n$ gradi di libertà come:

$$T_n = \frac{Z}{\sqrt{G/n}}$$

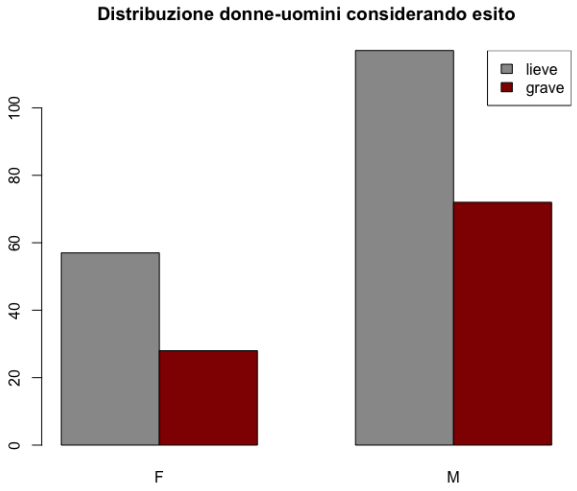
Questo test può essere eseguito anche unendo una variabile qualitativa, come ad esempio l'esito, in modo da avere un analisi dettagliata in entrambi i casi.
R si può pure utilizzare per la rappresentazione grafica in questo caso verrà eseguita utilizzando la funzione "hist()" che restituirà appunto un istogramma della variabile.
Inoltre utilizzando il pacchetto *ggplot2* [13], è possibile eseguire grafici molto più avanzati, come nel nostro caso unire due diversi istogrammi in modo da eseguire uno studio più dettagliato e specifico grazie alla contemporanea valutazione della stessa variabile quantitativa in relazione ai casi della variabile qualitativa.
Nel caso delle variabili qualitative, come ad esempio il sesso, si predilige l'uso dei grafici con la funzione "barplot()".
Anche in questo caso attraverso la creazione di una tabella è possibile confrontare due variabili qualitative per poi eseguire un "barplot()" congiunto delle stesse (come il grafico sul sesso a seguire) o attraverso la funzione "prop.table()" avere una semplice analisi della frequenza delle stesse.
Un altra tipologia di grafico utile sono i diagrammi a mosaico che si ottengono utilizzando la funzione "mosaic.plot" si sceglie di non utilizzarli in quanto non risultano molto utili.

## I risultati

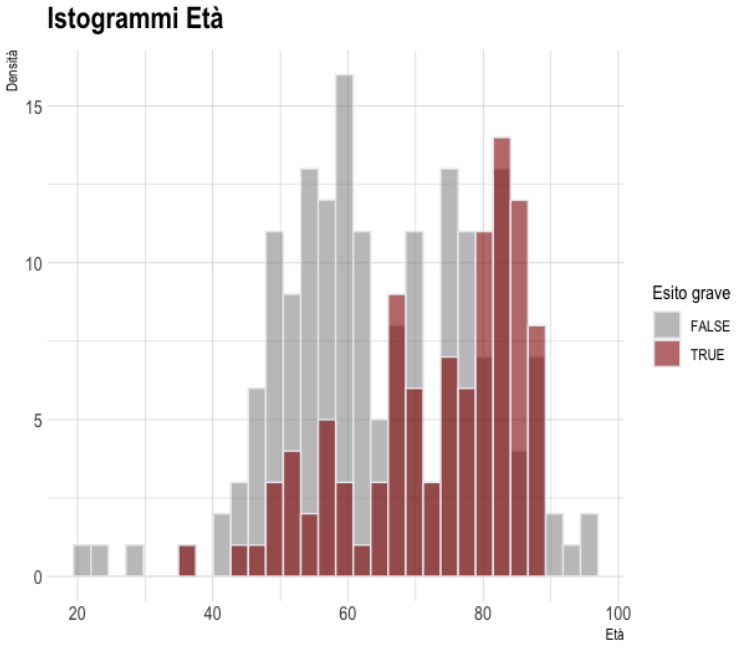
I dati forniti indicano che il numero di pazienti ricoverati durante la prima ondata di Covid nell'ospedale San Giovanni bosco di Torino è 274 di cui 100 $(36,5\%)$ hanno sviluppato una malattia grave o il decesso.



Come si può notare dal grafico a seguire i pazienti erano prevalentemente uomini, circa il 69% totale, dei quali il 38% ha sviluppato una patologia grave.
Le donne invece hanno sviluppato una patologia grave "solo" nel 32% dei casi.
Si osserva quindi che il Covid appariva notevolmente più pericolo per gli individui di sesso maschile.

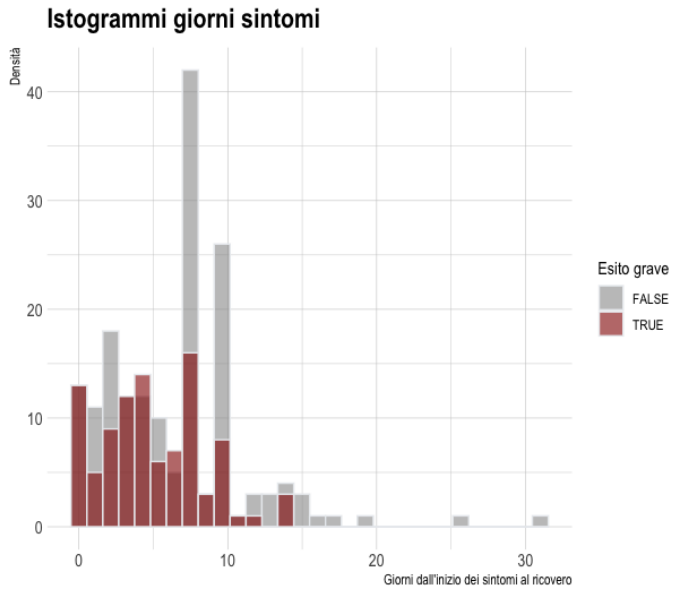



L'età, nel nostro campione, che considera solo persone maggiorenni, ha un range che va dai 21 ai 96 anni.
Nel grafico a seguire, che rappresenta la distribuzione in base all'esito, si nota come questo valore gioca un ruolo molto importante per stabilire la gravità del ricoverato:



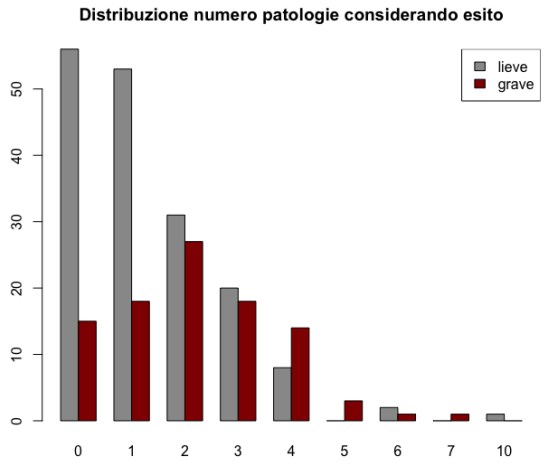
Utilizzando il "t.test" vediamo che l'età media dei ricoverati è di $67,7$ anni e che, nel caso in cui ci riconduciamo solamente ai malati gravi, sale sensibilmente fino ad arrivare a $72,6$.
Abbiamo quindi la che conferma che l'aumentare dell'età ha una grande incidenza sulla pericolosità della malattia.

Come notiamo dal grafico a seguire altro dato di rilievo sono i giorni trascorsi tra la manifestazione del primo sintomo ed il ricovero.




Studiando la media dei giorni trascorsi osserviamo che nel campione totale è di $5,8$ giorni e che scende a $4,8$ considerando solo i casi gravi.
In questo caso ci accorgiamo che un assenza iniziale dei sintomi e un improvvisa presenza acuta degli stessi rende molto più probabile il ricovero in terapia intensiva.

Altro da importante è l'anamnesi delle malattie pregresse dei ricoverati:



Come si vede dal grafico la presenza di un numero maggiore di malattie pregresse aumenta notevolmente la probabilità di andare in contro ad una malattia grave.
Una patologia tumorale a maggior ragione incrementa il rischio di una malattia grave.
Nel gruppo in esame era presente o comunque era stato presente in 20 ricoverati ed ha portato allo sviluppo di una malattia grave in 9 di essi.

Nel $51,7\%$ dei casi totali era presente dispnea al momento dell'accesso ospedaliero e la percentuale si alza a $65\%$ se consideriamo solo i casi poi diventati gravi.
Ci accorgiamo quindi che la presenza di dispnea ha molta importanza nell'aggravare della malattia.

Fra i parametri raccolti nel nostro studio vi è anche il rapporto P/F.[14]

**Definizione 3.2.** Il rapporto P/F rappresenta l'indice della respirazione alveolare ed è così definito:

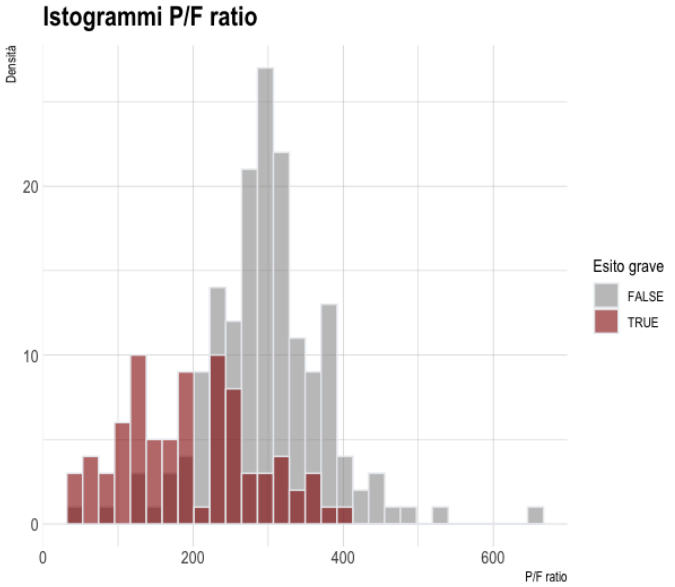
$$P/F = \frac{PaO_2}{FiO_2}$$

dove:

- $PaO_2$ rappresenta la pressione parziale dell'ossigeno nel sangue arterioso, in genere oscilla fra gli 80 e i 100 mmHg, nei casi di problemi respiratori scende.

- $FiO_2$ rappresenta la percentuale espressa in frazione di ossigeno somministrato, nell'aria è al $21\%$ quindi in questo caso sarà $0,21$.

Questo valore è quindi da considerarsi ottimale se superiore a 450 e normale se non inferiore a 300.
Il valore del rapporto P/F al momento del ricovero era in media di $263,9$ che nei casi diventati gravi si abbassa notevolmente a una media di 196.
Abbiamo quindi la conferma che l'abbassarsi acutizza i sintomi come si può notare anche dal grafico a seguire:



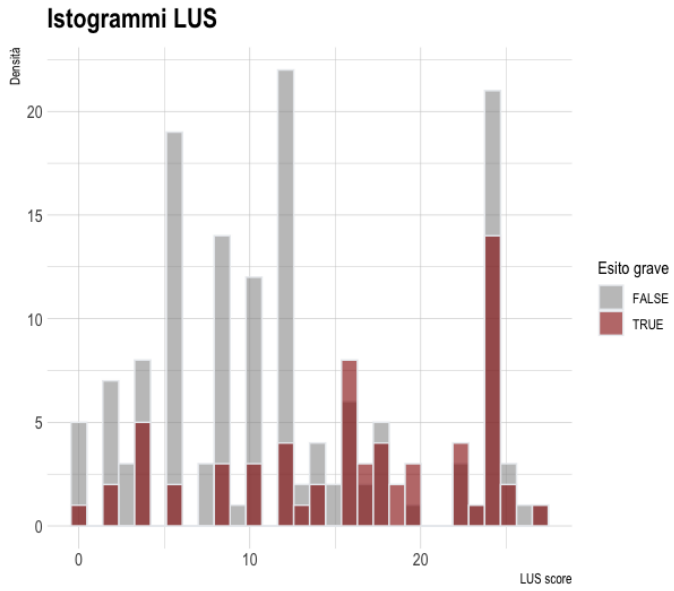
I pazienti ricoverati sottoposti a radiografia, nell'$82,4\%$ presentavano danni dovuti al Covid. Pertanto, a conferma della pericolosità del virus la quasi totalità delle persone ha danni più o meno gravi, la percentuale aumenta a $92,8$ se relativa ai soli casi che risulteranno gravi.
Alla gravità dei danni riscontrati è stato attribuito un punteggio, ovvero il cosiddetto LUS score.[15]

**Definizione 3.3.** Il LUS score è un punteggio che viene dato ad una ecografia polmonare al fine di quantificarne la gravità.
Si calcola eseguendo l'ecografia in 12 punti distinti a ciascuno dei quali viene attribuito un punteggio da 0, che indica nessun danno o quasi, fino a 3 che indica un punto totalmente danneggiato.

Il punteggio quindi varia da 0 a 36 ed è possibile dividere questo intervallo in modo da distinguere la gravità totale in 4 fasce:

- $0 =$ ecografia normale
- $[1;5]$ coinvolgimento lieve dei polmoni
- $[6;15]$ coinvolgimento moderato dei polmoni
- $> 15$ coinvolgimento grave dei polmoni

La media nel nostro gruppo di ricerca è di $13,4$.
Usando il "t.test" notiamo però una grande differenza tra la media nei casi rimasti lievi per i quali è di $12,4$ e quella dei casi diventati gravi in cui addirittura sale fino a $16,2$.
Questa grande differenza si può vedere meglio dal grafico a seguire nel quale troviamo la distribuzione per i nostri pazienti sia nel caso lieve che in quello grave.



# Capitolo 4

# Analisi statistica dei dati

In questo capitolo esaminiamo come mettere in pratica, sempre attraverso l'utilizzo del software R, gli argomenti teorici espressi nei capitoli precedenti.

## 4.1 Regressione logistica

La regressione logistica nel software R si ottiene utilizzando il comando "glm()".
Inoltre si può analizzare il risultato ottenuto attraverso l'utilizzo della funzione "summary()" la quale restituisce i coefficienti e altre analisi importanti come la devianza di cui abbiamo parlato nella sezione 1.2.3.
Infine con il pacchetto *ResourceSelection* [16] e più precisamente il test di *Hosmer-Lemeshow* (cfr. 1.4.4) valuteremo numericamente la bontà del modello per prevedere se la malattia di un ricoverato sarà o meno grave.

### LUS

Nel capitolo precedente abbiamo osservato che una delle variabili più incidenti nelle gravità di un ricovero è il LUS score che rappresenta la quantità di danni riscontrabili con un ecografia polmonare.
L'utilizzo del LUS score come variabile predittiva viene proposto come prima ipotesi nell'articolo sul COWS-score [2] e partendo da questo, utilizzandolo da solo in una regressione logistica possiamo provare a prevedere la gravità della malattia.
Per fare ciò si utilizza il comando di R glm().
Otteniamo attraverso "summary(lus)" che ovviamente sia il LUS che l'intercetta sono molto significativi.
Il test di Hosmer-Lemeshow ci restituisce invece un $p - value$ di $0,2917$.
Il valore è elevato ma non troppo quindi ci aspettiamo che anche lo studio delle curve ROC ci dia un risultato non troppo soddisfacente.

### GRAM-score

L'intero articolo di Liang et [1] si basa sulla costruzione di questo modello di regressione.
Utilizzando dei metodi di tipo LASSO riescono, partendo da 72 variabili predittivi, a trovare le dieci più utili e con queste decidono di costruire il modello migliore possibile.
Le variabili utilizzate sono:
la presenza di una radiografia anormale, di emottisi, di dispnea, di incoscienza, il numero di malattie, la storicità tumorale, il rapporto tra i neutrofili e i linfociti, la lattate deidrogenasi e la bilirubina diretta.
Come abbiamo fatto con il LUS, utilizziamo il software per ottenere il modello e valutare la sua bontà.

Attraverso il "summary" notiamo che i valori più significativi sono il numero delle malattie, il valore dell'LDH e della bilirubina.
L'incoscienza e l'emottisi non vengono considerate correttamente a causa di un grande squilibrio di questi dati all'interno del nostro dataset infatti abbiamo solo 2 veri e tutti gli altri sono falsi.
Eseguiamo poi infine il test di Hosmer ed otteniamo un valore di $p - value$ uguale a 0.7561 che ci conferma la grande bontà del modello costruito da Liang.

### COWS-score

L'articolo intitolato per l'appunto COWS etc. [2] nasce per la creazione di un modello migliore rispetto a quelli già noti.
Per fare questo, come abbiamo visto nell'introduzione, si parte dal GRAM e, utilizzando vari metodi, si trova che le variabili necessarie e sufficienti da utilizzare per predire la gravità della malattia sono cinque:
Il numero di malattie pregresse, la durata dei sintomi prima dell'accesso, il P/F ratio, la presenza di dispnea e la presenza di LUS score superiore a 15.
Partendo da questi predittori si crea un modello di regressione attraverso R e con l'utilizzo del summary eseguiamo una prima analisi.
Notiamo quindi che pure in questo caso il numero delle malattie è molto significativo come anche il p/f ratio.
Pure in questo caso infine eseguiamo il test di Hosmer-Lemeshow ottenendo un $p-value$ di 0.7857 ovvero di poco superiore al GRAM.
Questo valore ci fa quindi verificare che il modello COWS è ottimale e migliore rispetto ai modelli precedentemente costruiti.
Al fine di dimostrare quest'ultima affermazione si sceglie di utilizzare le curve ROC.

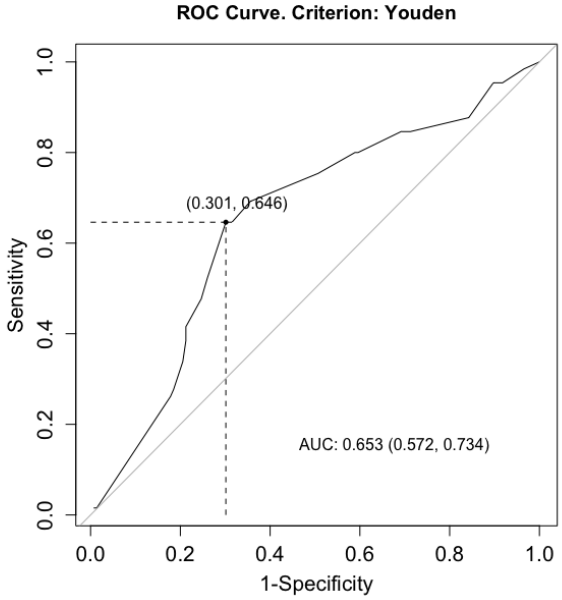
## 4.2 Costruzione e analisi delle curve ROC

Una volta calcolato il modello di regressione logistica come visto nella sezione precedente sarà possibile analizzarlo utilizzando le curve ROC, per fare ciò è necessario il pacchetto di R *pROC* . [17]
Per prima cosa creeremo la curva con il comando "roc(responce=x , predictor=y )" per poi disegnarla con "plot.roc()" e analizzarla con "smooth()".
Sarà inoltre possibile confrontare 2 curve (A e B) attraverso il comando "roc.test(A, B)".
Il pacchetto *OptimalCutpoints* [18] sarà infine utile per calcolare la soglia migliore per ogni curva ROC trovata.

### LUS

Per prima tracciamo la curva ROC del cosiddetto LUS-score.
Attraverso una veloce analisi notiamo che l'area sotto la curva è di 0,6385, quindi il LUS è un buon predittore ma non ottimo. Infatti il valore è superiore a $0,5$ ma comunque non vicino ad uno.
Al fine di individuare il valore di soglia migliore, si sceglie di utilizzare tutti i metodi della sezione 2.2.2 ed otteniamo valori di 16 per i metodi di Youden e di minima distanza, 18 per il metodo di Minimax e 27 per il metodo costi/benefici.
La soglia migliore fra questi nel nostro caso è di 16, infatti in questo punto i valori di $Sp$ e $Se$ sono entrambi elevati, ovvero $Se = 0.6462$ e $Sp = 0.6986$.

Nella figura che segue è rappresentata la curva e questo punto con relative stime migliorate di AUC, Sp e Se:




Si può inoltre notare che questa soglia equivale a 16 che è proprio il valore scelto per la creazione del COWS.

## GRAM-score

Proseguiamo tracciando la curva ROC per il GRAM-score.
L'area sotto la curva in questo caso è di 0.8405, quindi notevolmente migliore rispetto a quella del LUS.
Anche in questo caso cerchiamo l'optimal cut point utilizzando tutti i metodi sopra espressi.
Il migliore in questo caso è lo Youden che fornisce una soglia di 0.3194, $Sensibilit$à di 0.7705 e $Specificit$à uguale a 0.7883.
A seguire il grafico nella figura:

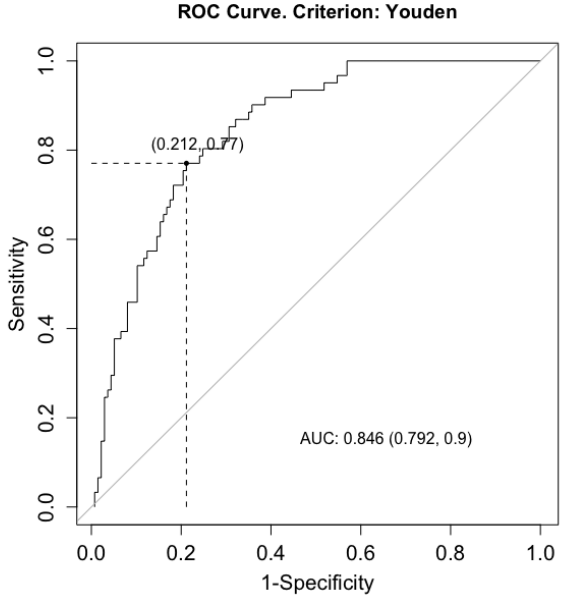
## COWS-score

Infine tracciamo la curva ROC per il COWS, l'area sotto di essa è di 0.8439 quindi di pochissimo superiore a quella per il GRAM ma, considerando che sono sufficienti un numero molto inferiori di parametri, lo rende il modello migliore tra quelli visti.
Questo fatto ci viene confermato dai test di confronto fra curve ROC attraverso l'utilizzo del comando roc.test(GRAM, COWS). Si sceglie di utilizzare il metodo Bootstrap. [17]

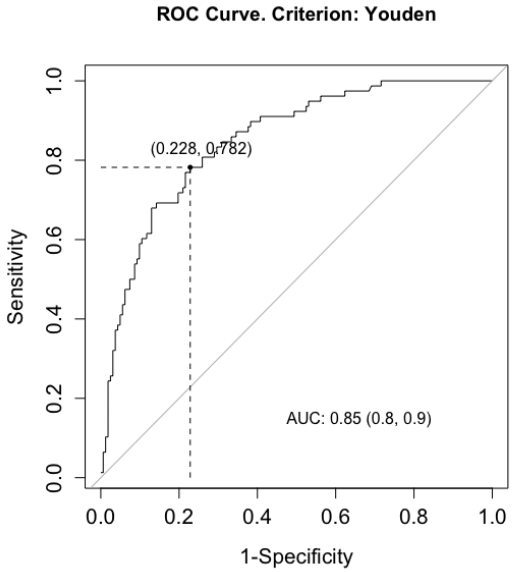
**Definizione 4.1.** Il metodo Bootstrap si basa sull'utilizzo di questa formula:

$$D = \frac{AUC1 - AUC2}{s}$$

dove s è la deviazione standard della differenza di Bootstrap, AUC1 è l'area della prima curva e AUC2 quella della seconda.
D infine viene confrontata con una normale standard che esprime l'alternativa che cerchiamo e ci restituisce un p-value, dandoci quindi una stima della probabilità che ciò sia vera.




Il test ci restituisce $D = -0.1242$ e, impostando l'alternativa "less", che la probabilità che l'area di COWS sia maggiore di quelli di GRAM è di 0.4506.
Il risultato del test ci conferma nuovamente che il punteggio COWS è di poco migliore; nell'articolo infatti è stato maggiormente affinato in modo da rendere i risultati ancora più attendibili.
Cerchiamo adesso il valore soglia ottimale come visto nei punti precedenti.
Utilizziamo tutti i metodi pure in questo caso scegliamo quello trovato con Youden (0.3106) essendo migliore per i valore di $Sp = 0.7821$ e $Se = 0.7840$ .
Data l'importanza del modello sono da segnalare $PPV = 0.6354$ e $NPV = 0.8819$.
Di seguito viene riportato il grafico nella figura:



# Conclusione

Come espresso nel capitolo di introduzione l'intento che mi sono posto nell'elaborare la mia tesi è quello di, partendo dai dati pervenuti dall'ospedale San Giovanni Bosco di Torino, analizzare quanto avvenuto, elaborato ed utilizzato durante l'emergenza COVID.
Ho esaminato e rielaborato i modelli creati da vari ricercatori, e soprattutto compreso perché sia stato utilizzato, nella maggior parte dei casi, il modello studiato da Boero et.
Ho a tal fine ricreato ed analizzato, anche con tecniche differenti, modelli di regressione logistica, validati attraverso apposite tecniche ed infine giunto alla conclusione di quale, visti i risultati ottenuti, fosse il migliore.
Per prima cosa, avendo a disposizione numerose variabili, ho cercato di capire quali fossero quelle più utili al nostro scopo, ovvero più efficaci per prevedere la gravità della malattia e con queste ho quindi creato tre modelli:
LUS, GRAM, COWS.
Le ho poi validate con il test di Hosmer-Lemeshow e lo studio delle curve ROC, cercando infine la soglia migliore per dividere i positivi dai negativi.
Sono giunto alla conclusione che sia il modello GRAM che il COWS sono ottimi per prevedere tale esito, dal momento che entrambi hanno attendibilità molto simili.
Nella realtà, considerata l'emergenza epidemiologica e la velocità necessaria per decidere sulle modalità di procedere su ogni singolo paziente, era necessario però avere a disposizione un modello che fosse semplice e veloce da calcolare, e soprattutto che non necessitasse di troppi esami diagnostici. Di conseguenza il COWS score diventa sicuramente il modello migliore necessitando solamente di una ecografia, esame veloce da effettuare e non invasivo.
Questo non fa che confermare l'importanza della studio svolto da Boero et. dimostrando che il loro modello era il migliore creato.

In ambito di biostatistica, ovvero della statistica applicata alla medicina, l'uso di metodi analoghi a quello descritto potrebbe essere molto utile in ambito di prevenzione.
Alla luce dell'aumento di malattie oncologiche e difficilmente curabili nella popolazione, avere dei mezzi che ci forniscano una previsione di esito in base a fattori determinati è di grande aiuto sia in ambito di prevenzione che di cura.
La conoscenza e lo studio di questi fattori permette di ottimizzare le risorse a disposizione indirizzando la prevenzione su quegli esami diagnostici, quelle abitudini di vita e comportamenti in genere che permettono di avere una vita migliore.

Il loro utilizzo comunque a mio avviso potrebbe non fermarsi solo alla medicina, ma anche a settori completamente differenti come l' economia e il management, per lo studio dei mercati finanziari e la gestione del rischio.

# Bibliografia


[1] Liang et al. (2020), *Develepment and Validation of a Clinical Risk Score to predict the Occurence of Critical Illness in Hospitalized Patients With COVID-19*, JAMA Internal Medicine 2020;180(8), pagine 1081-189.

[2] Boero et al. (2020), *The COVID-19 Worsenig Score (COWS) - a predictive bedside tool for critical illness*, Wiley Echocardiography 2021;00, pagine 1-10.

[3] Johnson e Wichern (1982), *Applied Mutivariate Statistical Analysis*, Sesta edizione, sezione 11.7.

[4] Hastie, Tibshirani e Friedman (2001), *The Elements of Statistical Learning*, Seconda edizione, sezione 4.4.

[5] Piccolo, Domenico(1998), *Statistica*, Terza edizione.

[6] Wikipedia, *Leverage(Statistics)*, `https://en.wikipedia.org/wiki/Leverage_(statistics)`

[7] Hosmer e Lemeshow, *Applied Logistic Regression*, seconda edizione, sezione 5.2.2, pagine 147-155

[8] Fawcett, Tom(2006), *An introducion to ROC analysis*, Elsevier, Pattern Recognition Letters; 27(8) pagine 861–874.

[9] Gonçalvez et al. (2014), *Roc curve estimation: An overview*, Revstat -Statistical journal marzo 2014; 12(1), pagine 1-20

[10] Wikipedia, *Receiver operating characteristic*, `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`

[11] Wikipedia, *Sensitivity and specificity*, `https://en.wikipedia.org/wiki/Sensitivity_and_specificity`

[12] Package, *readxl*, `https://cran.r-project.org/web/packages/readxl/readxl.pdf`

[13] Package, *ggplolt2*, `https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf`

[14] DimensioneInfermiere, *Guida semplice all'interpretazione e al calcolo del rapporto P/F*, `https://www.dimensioneinfermiere.it/guida-semplice-calcolo-del-rapporto-pf-pao2fio2/`

[15] Esaote, *COVID-19 Pneumonia Lung Ultrasound (LUS) assessment of severity of involvement*, `https://www.esaote.com/fileadmin/user_upload/clinical-solutions/Esaote-poster-LUS-Covid-A1-160000293-V1-0321-LR.pdf`

[16] Package, *ResourceSelection*, `https://search.r-project.org/CRAN/refmans/ResourceSelection/html/hoslem.test.html`

[17] Package, *pROC*, `https://cran.r-project.org/web/packages/pROC/pROC.pdf`

[18] Package, *OptimalCutpoints*, `https://cran.r-project.org/web/packages/OptimalCutpoints/OptimalCutpoints.pdf`