UNIVERSITÀ DEGLI STUDI DI TRENTO

CORSO DI LAUREA IN ECONOMIA E MANAGEMENT

COVID-19: ANALISI CRITICA ED EMPIRICA DELLA LETALITÀ

IN RELAZIONE ALLE VARIABILI DI ETÀ ANAGRAFICA E GENERE

Tesi di laurea
di Cesare Barbera

Relatore
Ch.mo Prof. Giuseppe Espa

Anno Accademico 2019 / 2020

INTRODUZIONE

Nel mese di novembre dell'anno 2019, il nuovo Sars-CoV-2 inizia a circolare in Cina, in particolare a Wuhan, la città più popolata della parte orientale, perno per il commercio e gli scambi. Tuttavia, non c'era ancora coscienza dell'esistenza del nuovo virus: ciò che inizia ad essere registrato è un certo numero di polmoniti anomale, dalle cause non ascrivibili ad altri patogeni conosciuti.

Soltanto il 9 gennaio avviene la dichiarazione da parte delle autorità cinesi ai media locali che il patogeno responsabile dei casi consiste in un nuovo ceppo di Coronavirus, appartenente alla stessa famiglia dei virus responsabili della Sars e della Mers.

Per quanto concerne i casi, ancora in numero ristretto, questi si concentrano a Wuhan con contagiosità e pericolosità non chiare.

Attraverso i voli internazionali diretti e provenienti dalla Cina che, a causa dell'ignoranza circa le caratteristiche di questo nuovo virus non vengono sospesi, il virus sbarca in altre nazioni.

Inizialmente, i contagi fuori dalla Cina sono circoscritti e limitati, con focolai, per ogni paese colpito, di un ristretto numero di persone. Lo stesso avviene in Italia che, a fine gennaio, conta un numero moderato di contagiati, tutti riguardanti soggetti che erano stati a contatto diretto con il territorio cinese.

Ciò nonostante, nello stesso periodo, cresce la consapevolezza del rischio che l'epidemia si diffonda tanto che, nella serata del 30 gennaio l'Oms dichiara l'"emergenza sanitaria pubblica di interesse internazionale".

Data centrale per la vicenda italiana legata al nuovo coronavirus è venerdì 21 febbraio 2020: emerge un numero considerevole di casi di coronavirus nel lodigiano, in Lombardia. Questo evento rappresenta una realtà allarmante: si tratta di persone non provenienti dalla Cina, indice del fatto che il virus abbia attecchito nel territorio nazionale e che la possibilità di una diffusione incontrollata sia una concreta minaccia.

Nell'arco temporale individuabile fra la fine di febbraio e i primi giorni di marzo 2020, in altri stati europei e non solo, viene registrato un numero sempre crescente di positivi al Covid-19, che acquisisce lo status di epidemia.

L'11 marzo 2020, Tedros Adhanom Ghebreyesus, direttore generale dell'Oms, annuncia nel

briefing da Ginevra sull'epidemia di coronavirus che il Covid-19 rispecchia una situazione pandemica.

Per far fronte all'emergenza in atto sono stati individuati due approcci principali da perseguire ovvero quello della soppressione del virus e quello del suo contenimento.
Il primo si prefigge lo scopo di ridurre il numero di riproduzione del virus al di sotto di 1 contrastando la trasmissione del virus da individuo a individuo fino all'estinzione dell'epidemia. Questo approccio deve essere mantenuto finché il virus è in circolazione o venga trovato un vaccino.
Il secondo approccio, invece, non vuole ottenere un arresto della trasmissione del virus quanto, piuttosto, ridurre l'impatto sulla salute dell'epidemia stessa costituendo quella che è nota come immunità di gregge. Infatti, l'idea sottostante questo approccio è che, attraverso la contrazione del virus da parte di una porzione considerevole della popolazione, si possano sviluppare anticorpi ed immunità nei confronti di un determinato patogeno ottenendo, infine, un rapido declino dei contagiati e della trasmissione del virus. In questo caso, essendo necessario che il virus venga contratto da una parte rilevante della popolazione, il numero di riproduzione di base non deve scendere al di sotto di 1, durante una prima fase, affinché ciò possa avvenire automaticamente in un secondo momento.
Più precisamente, il numero di riproduzione del virus R_0, questo è il simbolo con cui viene indicato il numero di riproduzione di base, stima il numero di soggetti che verranno infettati in un dato momento per ogni soggetto infetto. Questo indicatore è di straordinaria importanza quando si è alle prese con un'epidemia per il suo contenimento perché stabilisce chiaramente quando il numero di infetti diminuirà e l'epidemia tenderà ad estinguersi. La sua interpretazione è molto semplice:

- quando il valore di $R_0 > 1$ siamo di fronte all'evenienza in cui un'epidemia è in espansione poiché ogni infetto contagia più di un altro individuo ed inevitabilmente il numero stesso degli infetti è in aumento.
- Se invece $R_0 < 1$ allora ogni soggetto infetto contagia meno di un altro individuo e l'epidemia è in riduzione con un numero di infetti in diminuzione che diventerà 0 nel futuro, se R_0 resterà inferiore di 1.
- L'ultimo caso rimanente è quello di $R_0 = 1$. Qui per ogni infetto abbiamo un altro singolo individuo che contrae il virus perché contagiato da quest'ultimo e di conseguenza il numero degli infetti rimane stabile nel tempo.

Con l'istituzione del lockdown in Italia si è optato per seguire il primo approccio.

Le ragioni di tale scelta vengono comprese soltanto studiando l'espansione del contagio in relazione alla sua velocità di diffusione. Sebbene molti abbiano portato avanti contro l'attuazione di tale misura, e le conseguenti restrizioni delle libertà individuali di tutti i cittadini, l'argomentazione secondo cui il tasso di mortalità del virus non sia effettivamente superiore di quello di una banale influenza stagionale e che quindi gli effetti reali del virus non giustifichino una misura tanto drastica, non bisogna trascurare il fatto, ed è qui che risiede la vera pericolosità del virus, che l'aumento esponenziale e fuori controllo degli infetti, in un lasso temporale tanto ristretto, implicherebbe una domanda di posti letto in reparti di terapia intensiva e macchinari medici che non sono disponibili in quantità adeguate, comportando il potenziale collasso del sistema sanitario.

Attraverso i dati sui contagi, che sono stati sin da subito resi pubblici, è stato possibile stimare gli effetti di misure restrittive dell'uno o dell'altro tipo e l'impatto di queste sull'evoluzione del numero di contagiati nel tempo. I dati stessi hanno confermato che soltanto misure drastiche come quelle attuate avrebbero abbassato la curva dei contagi in modo da consentire al sistema sanitario di gestire correttamente tutti i casi evitando, dunque, una profonda crisi del sistema stesso.

LO STUDIO

Sin dai primi segnali preoccupanti riportati in Cina da quelle sospettose polmoniti anomale, e ancora di più dopo l'individuazione dell'agente patogeno responsabile, individuato in un nuovo ceppo di coronavirus, sono stati condotti numerosissimi studi ed analisi con lo scopo di pervenire ad una comprensione totale del fenomeno pandemico. Questi hanno riguardato numerosissimi aspetti e tutt'oggi rimangono delle domande che non hanno risposta.

La ricerca costante di un vaccino, parallela a quella incessante di una cura, ha caratterizzato e impegnato totalmente il mondo della ricerca, in un contesto di incertezza globale alimentato da falsa informazione e allarmismo.

Sebbene moltissime delle risorse impiegate alla lotta al coronavirus sono state impiegate per la ricerca di una cura e la sperimentazione del vaccino, un ruolo estremamente rilevante è stato assunto dall'analisi dei dati e dalla ricerca statistica, sia per la comprensione dell'andamento della diffusione del fenomeno epidemiologico sul territorio, sia per quantificare le eventuali relazioni ipotizzate tra il virus e variabili di interesse, sia per testare la validità dei risultati ottenuti in laboratorio.

L'analisi dei dati ha assunto, dunque, un ruolo centrale per la ricerca e lo studio delle caratteristiche del Covid-19 permettendoci di evidenziarne tratti e proprietà di estrema importanza per la salvaguardia della salute pubblica.

Lo studio presentato in questa trattazione ha come tema la letalità del virus. Questa viene definita come: *capacità di provocare la morte; in statistica, quoziente di letalità, il rapporto tra il numero di morti per una data malattia e il numero delle persone affette dalla stessa, relativamente a una data popolazione e a un dato intervallo di tempo.* (Dal dizionario Treccani)

Uno studio sulla letalità è di estrema importanza poiché ci permette di evidenziare i rischi connessi alla contrazione della malattia e l'effetto di variabili di nostro interesse su questo rischio.

La prima parte di questo trattato sarà dedicata alla definizione del modello utilizzato e ad una trattazione teorica degli strumenti implementati per l'ottenimento dei risultati che verranno riportati ed interpretati. in questa sezione verranno definite le ragioni e le ipotesi sottostanti alla scelta del modello e dello stimatore adoperato per pervenire alle stime dei parametri di interesse dello studio.

Il modello individuato per questa trattazione è il modello logit. Saranno esposte la sua derivazione, le sue caratteristiche e verrà approfondito teoricamente il confronto tra un modello grezzo ed uno aggiustato, necessario per l'identificazione di effetti di tipo confondente da parte delle variabili di interesse. All'interno di questo lavoro sarà esposto inoltre il significato delle interazioni tra le variabili esplicative di interesse e la loro interpretazione.

Lo stimatore implementato in questo studio coincide con il metodo di massima verosimiglianza e conseguentemente verrà approfondito e trattato analiticamente il metodo di stima stesso dei parametri, la sua derivazione e le sue proprietà.

Infine, lo scopo dello studio condotto in questo lavoro sarà quello di valutare le relazioni tra età e sesso e la letalità collegata al virus.

Possiamo infatti ritenere che vi sia una forte relazione tra l'età ed il rischio di decesso in quanto le fasce più anziane sono le più deboli e quelle caratterizzate dalla maggior presenza di soggetti cagionevoli o con cartelle cliniche più complicate.

Al contempo è di nostro interesse studiare la relazione stessa tra la letalità del virus ed il sesso del contagiato.

Bisognerà, tuttavia, stare attenti a non individuare come risultato dell'analisi una relazione spuria.

Difatti sappiamo che l'aspettativa di vita per le donne è maggiore di quella degli uomini e, conseguentemente, le fasce più anziane di popolazione saranno composte maggiormente da donne. Risulta, pertanto, fondamentale controllare i risultati per l'età in relazione al sesso così da ottenere l'effetto dovuto all'appartenenza ad una classe rispetto che ad un'altra.

Le variabili di interesse saranno individuate come variabili categoriche in quanto il genere non può essere espresso in termini quantitativi e la variabile età verrà divisa in fasce. Per questo studio sarà, dunque, in primis necessario ottenere i dati relativi al genere e l'età dei contagiati e dei deceduti e elaborarli creando le classi attraverso le quali quantificare gli effetti di nostro interesse.

Una volta effettuata l'operazione di raccolta e elaborazione dei dati, sarà possibile una prima stima separata degli effetti delle variabili genere ed età sulla probabilità di decesso collegata al coronavirus. Si valuterà, dunque, la significatività dei parametri restituiti dal nostro input. I risultati di queste prime stime non sono tuttavia sufficienti a trarre le conclusioni desiderate in

quanto gli effetti così quantificati possono rivelarsi distorti e frutto dell'influenza di fattori non considerati in ognuno dei due modelli sviluppati.

Poiché entrambe le variabili esplicative considerate possono avere un impatto ed alterare l'effetto dell'altra si individuerà un modello congiunto attraverso il quale misurare in che proporzioni cambiano i valori assunti dai parametri una volta che le variabili vengono controllate per l'altra. Anche qui si testerà la validità dei parametri in questo modello per verificare che l'effetto individuato nella prima regressione non sia confondente e quindi sparisca in un modello congiunto.

Successivamente verrà adottato un modello contente le interazioni tra le variabili esplicative. Lo scopo di questa regressione sarà valutare che gli effetti stimati delle variabili siano costanti per i diversi livelli dell'altra variabile. Per testare questa possibilità verrà implementato un test per l'ipotesi che tutti i parametri corrispondenti alle interazioni tra le variabili esplicative siano contemporaneamente nulli.

Conseguentemente ai risultati ottenuti da queste regressioni, si sceglierà il modello preferibile sulla base della significatività assunta dai parametri nei diversi modelli, si costruiranno intervalli di confidenza per i valori assunti da questi e verrà interpretato il loro significato.

L'ultima fase dello studio considererà un'analisi della bontà del fit del modello e del suo adattamento ai dati attraverso un confronto tra i valori predetti e osservati ed il calcolo di statistiche come l'R^2 di McFadden.

I DATI

I dati a nostra disposizione sono estrapolati dal bollettino di sorveglianza integrata aggiornato al 23/06/2020.

“Il bollettino è prodotto dall’Istituto Superiore di Sanità (ISS) ed integra dati microbiologici ed epidemiologici forniti dalle Regioni e dal Laboratorio Nazionale di Riferimento per SARS-CoV-2 dell’ISS. I dati vengono raccolti attraverso una piattaforma web dedicata ed include tutti i casi di COVID-19 diagnosticati dai laboratori di riferimento regionali. I dati vengono aggiornati giornalmente da ciascuna Regione anche se alcune informazioni possono richiedere qualche giorno per il loro inserimento. Per questo motivo, potrebbe non esserci una completa concordanza con quanto riportato attraverso il flusso informativo della Protezione Civile e del Ministero della Salute che riportano dati aggregati.

Il bollettino descrive, con grafici, mappe e tabelle la diffusione, nel tempo e nello spazio, dell’epidemia di COVID-19 in Italia. Fornisce, inoltre, una descrizione delle caratteristiche delle persone affette.”

IL MODELLO

Il modello utilizzato per le nostre analisi si basa sulla regressione logistica. Le variabili che vengono inserite nel modello, età e sesso, vengono proposte sotto forma di variabili dummy. Queste assumono valore 1 quando l'osservazione rispecchia la categoria individuata nella variabile dummy e 0 altrimenti. Nel caso dell'età, essendo di scarso impatto ed interesse l'effetto di un anno in più sulla letalità, queste vengono divise in classi e si valuta l'impatto dell'appartenenza ad ogni classe sulla variabile dipendente. Nel nostro studio vengono inserite variabili dummy per le classi: 55-65; 65-75; 75-85; 85-95; 95+.

Nel nostro modello, studiando l'effetto della sola età sulla letalità, scriveremo, dunque, l'equazione come:

$$y = \beta_0 + \beta_1\delta_1 + \beta_2\delta_2 + \beta_3\delta_3 + \beta_4\delta_4$$

Dove β consiste nel parametro oggetto delle nostre stime e δ rappresenta la variabile dummy.

Si noti che le variabili dummy inserite nell'equazione di stima corrispondono al numero delle classi che rappresentano meno uno perché, rispettando l'ipotesi di rango pieno, necessitiamo n-1 variabili per descrivere la totalità degli effetti considerati nel modello.

Viene da sé che quando tutte le altre variabili sono nulle venga restituito:

$$y = \beta_0$$

che consiste nell'effetto dovuto all'appartenenza alla classe non specificata.

Si noti che questo coincide con l'intercetta e che se volessimo includere una variabile dummy per classe dovremmo escludere l'intercetta dal nostro processo di stima.

Considerando esclusivamente la variabile genere all'interno del modello avremo invece:

$$y = \beta_0 + \beta_1\delta_1.$$

Dove δ_1 assumerà valore 1 nel caso in cui l'osservazione appartenga ad un soggetto di genere maschile e 0 nel caso contrario. Dunque, il modello nel caso di osservazione di genere femminile restituirà:

$$y = \beta_0 .$$

I due modelli così scritti possono essere stimati contemporaneamente in un modello congiunto della forma:

$$y = \beta_0 + \beta_1\delta_{e1} + \beta_2\delta_{e2} + \beta_3\delta_{e3} + \beta_4\delta_{e4} + \beta_5\delta_{s1}$$

dove le variabili dummy con pedice 'e' indicano l'età e quella con pedice 's' il genere.

Si noti che nei tre modelli appena presentati i parametri delle variabili dummy vanno solo a modificare l'intercetta della retta di regressione, che corrisponde all'individuo con età compresa tra i 55 e 65 anni di genere femminile, in quanto, non essendo presente nel modello alcuna variabile continua, la retta di regressione non presenta un coefficiente angolare.

Testare la significatività dei parametri nel modello congiunto è particolarmente importante nel caso specifico di uno studio della letalità di un virus che tiene in considerazione variabili come età e sesso poiché bisogna prestare attenzione che una delle due variabili non sia una variabile confondente. La questione risulta estremamente delicata quando le aspettative di vita di uomo e donna risultino particolarmente diverse.

Considerare il fattore genere per valutare l'impatto che le differenze biologiche fra uomo e donna hanno sulla letalità del Covid-19 è di estrema importanza, ma se questo fattore viene considerato esclusivamente si può incorrere in un errore dovuto al fatto che le aspettative di vita di una donna in Italia sono maggiori rispetto a quelle di un uomo e di conseguenza maggiore mortalità nelle donne sarà dovuta in parte alla loro più veneranda età.

Per questo è particolarmente importante controllare per il fattore età quando si conduce un'analisi di questo tipo ed è importante valutare la significatività dei parametri nel modello che tenga conto anche dell'età così da assicurarsi che l'effetto del sesso sulla letalità non sia dovuto esclusivamente a diverse aspettative di vita di uomo e donna.

Contemporaneamente al modello congiunto, potrebbe essere di nostro interesse valutare un modello che tenga in considerazioni eventuali interazioni tra le variabili considerate per poterne valutare la significatività statistica.

Questo modello, ben più complesso dei precedenti, viene generato scrivendo il modello congiunto e creando le interazioni tra le variabili moltiplicandole tra di loro:

$$y = \beta_0 + \beta_1\delta_{e1} + \beta_2\delta_{e2} + \beta_3\delta_{e3} + \beta_4\delta_{e4} + \beta_5\delta_{s1} + \beta_6(\delta_{e1} * \delta_{s1}) + \beta_7(\delta_{e2} * \delta_{s1}) + \beta_8(\delta_{e3} * \delta_{s1}) + \beta_9(\delta_{e4} * \delta_{s1})$$

Quando si osserva un uomo con età compresa tra 65 e 75 anni il modello restituisce:

$$y = \beta_0 + \beta_1 + \beta_5 + \beta_6$$

Le interazioni tra variabili dummy permettono di alterare il valore corrispondente rispetto al caso in cui le variabili siano considerate singolarmente. In questo specifico caso, nel modello con le interazioni, è come se venisse aggiunto un 'premio' o ridimensionato l'impatto delle due variabili, sulla base del segno del parametro associato al prodotto fra le variabili dummy, per l'appartenenza ad una determinata classe individuata dalle variabili. Si noti che questo effetto si manifesta, nell'esempio, col parametro β_6, che non è presente nel modello congiunto senza interazioni.

Fino ad ora abbiamo trattato le modalità in cui le variabili indipendenti vengono inserite nel modello, ora è il momento di concentrarsi sulla variabile dipendente. Come accennato in precedenza il processo di stima utilizzato per il nostro modello si basa sulla regressione logistica. La probabilità di decesso viene individuata come la proporzione dei deceduti sui soggetti che hanno contratto la malattia. La variabile y che è stata inserita fino ad ora nelle rappresentazioni del modello consiste in una variabile che viene individuata come logaritmo del rischio relativo. Questa variabile dipendente consiste in:

$\ln\left(\frac{p}{1-p}\right)$.

Per una comprensione del significato di questa variabile e della funzione logit, bisogna entrare nel merito dell'odds ratio. Partendo dal caso più semplice di un'unica variabile indipendente categorica, possiamo costruire una tabella che tenga conto delle casistiche corrispondenti a un valore 1 della variabile dipendente e ad un valore 0 della variabile dipendente in base all'esposizione o meno al fattore individuato dalla variabile indipendente.

	y = 1	y = 0	
	a	b	exposed
	c	d	unexposed

dove la riga corrispondente ad exposed individua i soggetti che, nel nostro caso specifico, sono morti per il virus, y = 1, e non sono morti per il virus, y = 0, essendo esposti alla variabile categorica, x = 1. La riga corrispondente ad unexposed indica gli stessi valori ma per i soggetti che non sono individuati dalla variabile indipendente categorica, x = 0.

Nel nostro caso specifico la variabile indipendente che assume valori 0 e 1 è la variabile genere.

Da questa tabella è possibile individuare la probabilità di decesso, dopo aver contratto il virus,

P(y=1), sia per i soggetti esposti sia per quelli non esposti alla variabile indipendente insieme alla probabilità di guarigione, P(y = 0):

$P(y = 1 \mid x = 1) = \frac{a}{a+b}$	$P(y = 0 \mid x = 1) = \frac{b}{a+b}$
$P(y = 1 \mid x = 0) = \frac{c}{c+d}$	$P(y = 1 \mid x = 0) = \frac{d}{c+d}$

Se, e questa è una scelta assolutamente arbitraria, individuiamo come l'appartenenza al genere maschile la categoria exposed, allora la prima individuerà la probabilità di essere deceduti essendo di genere maschile e la seconda quella per il genere femminile.

Possiamo ora individuare separatamente gli odds per la categoria exposed ed unexposed:

$$\text{Odds(exp.)} = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b}$$

$$\text{Odds(unexp.)} = \frac{\frac{c}{c+d}}{\frac{d}{c+d}} = \frac{c}{d}$$

Nel nostro caso, gli odds per un soggetto esposto, individuano il rapporto fra la probabilità che la variabile indipendente sia uno, ovvero un decesso se il soggetto è di genere maschile, e la probabilità che il soggetto guarisca se è dello stesso genere.

Lo stesso viene individuato per i soggetti di genere femminile quando si calcolano gli odds per i soggetti non esposti.

L'odds ratio definito in precedenza è individuato come il rapporto tra gli odds appena identificati.

$$\text{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a*d}{b*c}$$

Un passo ulteriore può essere effettuato definendo come:

$P_1 = P(y = 1 \mid x = 1)$

$P_0 = P(y = 1 \mid x = 0)$

$Q_1 = P(y = 0 \mid x = 1)$

$Q_0 = P(y = 1|x = 0)$

ed il logaritmo dell'*odds ratio*, come:

$\beta = \log(OR) = \text{logit}(P_1) - \text{logit}(P_0)$

Dove la funzione logit consta in $\ln\left(\frac{p}{1-p}\right)$ in cui p rappresenta la probabilità di una casistica rapportata alla casistica complementare.

Essendo P_1 il rischio connesso all'esposizione al fattore individuato da una determinata variabile categorica X Possiamo individuare:

$P_1 = P_x$

$\beta_x = \log(OR) = \text{logit}(P_x) - \text{logit}(P_0)$

$\text{logit}(P_x) = \alpha+\beta x$ dove:

$\alpha = \text{logit}(P_0)$

Da questo segue che:

$P_1 = \frac{\exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}$

$P_0 = \frac{\exp(\alpha)}{1+exp(\alpha)}$

Nel caso specifico di uno studio concernente la letalità di un virus basandosi su età e genere si valuta la probabilità di decesso sulla base dei valori assunti dalle variabili indipendenti, indi per cui occorre individuare le modifiche da apportare quando la variabile indipendente non assuma esclusivamente valori 0 e 1, ma questa è caratterizzata da più di due classi.

Infatti, una variabile come l'età è spesso riportata in termini di classi, poiché l'effetto marginale di un anno in più sulla probabilità P(y =1) è spesso non significativo e quindi di scarso interesse. Chiaramente studi diversi implicheranno campioni e popolazioni diverse che inevitabilmente si differenzieranno per età e di conseguenza non esiste una regola fissa e predeterminata per identificare l'ampiezza e il numero delle classi di età da utilizzare.

Qui verrà preso in esame una casistica con variabile indipendente divisa in tre classi in quanto le conseguenze dell'aumento di classi e gli adattamenti che deve seguire il modello sono direttamente conseguenti a quanto verrà esposto.

Identifichiamo una variabile dipendente categorica X ed ipotizziamo che possa assumere tre

diversi valori corrispondenti a tre diversi classi d'età:

X_1 = 1 quando l'età è compresa tra 60 e 70, 0 altrimenti;

X_2 =1 quando l'età è compresa tra 70-80, 0 altrimenti;

X_3 = 1 quando l'età è compresa tra 80-90, 0 altrimenti;

Possiamo individuare nuovamente la tabella che tenga conto delle casistiche corrispondenti a un valore 1 della variabile dipendente e ad un valore 0 della variabile dipendente in base all'esposizione o meno al fattore individuato dalla variabile indipendente:

	y =1	y = 0	
	a	b	X_1 = 1
	c	d	X_2 =1
	e	f	X_3 = 1

Identifichiamo gli Odds:

$$\text{Odds}(X_1 = 1) = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b}$$

$$\text{Odds}(X_2 = 1) = \frac{\frac{c}{c+d}}{\frac{d}{c+d}} = \frac{c}{d}$$

$$\text{Odds}(X_3 = 1) = \frac{\frac{e}{e+f}}{\frac{f}{e+f}} = \frac{e}{f}$$

E due diversi Odds Ratio:

$$OR_1 = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a*d}{b*c}$$

$$OR_2 = \frac{\frac{a}{b}}{\frac{e}{f}} = \frac{a*f}{b*e}$$

Il modello espresso in forma generale per k variabili indipendenti avrà la forma:

$$P(y = 1|x) = \frac{\exp(\alpha + \sum \beta_k x_k)}{1 + \exp(\alpha + \sum \beta_k x_k)}$$

I due modelli identificati singolarmente per le due variabili esplicative, età e genere, sono:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \delta_{s1}$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\delta_{e1} + \beta_2\delta_{e2} + \beta_3\delta_{e3} + \beta_4\delta_{e4}$$

Questi corrispondono ai modelli grezzi del nostro studio.

Si noti che le classi di età individuate nel nostro studio sono 5 e di conseguenza, sempre per l'assunzione di rango pieno, i parametri stimati sono 4 più l'intercetta e, essendovi due singole classi di genere, vi è un singolo parametro stimato più l'intercetta.

Questi vengono definiti grezzi poiché sono riferiti alla popolazione nel suo complesso e osservata in un determinato periodo di tempo.

Contrariamente un modello aggiustato si riferisce alla popolazione classificata in distinti sottoinsiemi e valutata nello stesso periodo.

Con l'intento di assicurarci che gli effetti individuati nei modelli non siano distorti e le variabili considerate non siano confondenti, è nostro interesse costruire un modello congiunto per testare la significatività dei parametri dopo aver controllato per l'altra variabile.

Una metodologia efficace per controllare gli effetti confondenti delle variabili di regressione è la stratificazione della popolazione dello studio, attraverso la combinazione di livelli diversi delle variabili considerate.

Il modello congiunto ha dunque la forma:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\delta_{e1} + \beta_2\delta_{e2} + \beta_3\delta_{e3} + \beta_4\delta_{e4} + \beta_5\delta_{s1}\ .$$

Da cui segue che:

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1\delta_{e1} + \beta_2\delta_{e2} + \beta_3\delta_{e3} + \beta_4\delta_{e4} + \beta_5\delta_{s1})\ .$$

E quindi:

$$p = \frac{\exp(\beta_0+\beta_1\delta_{e1}+\beta_2\delta_{e2}+\beta_3\delta_{e3}+\beta_4\delta_{e4} + \beta_1\delta_{s1})}{1+\exp(\beta_1\delta_{e1}+\beta_2\delta_{e2}+\beta_3\delta_{e3}+\beta_4\delta_{e4} + \beta_1\delta_{s1})}$$

In questo caso p rappresenta la probabilità che la y sia uno, data l'appartenenza di un soggetto alle classi individuate dalle variabili esplicative.

Il modello così esposto tuttavia non considera fattori moltiplicativi fra le variabili esplicative.

Analizzando il caso più semplice, in cui entrambe le variabili esplicative sono individuate in due classi distinte, potremo costruire una tabella che tiene conto delle probabilità corrispondenti a un valore 1 della variabile dipendente in base all'esposizione o meno al fattore individuato dalle

variabili indipendenti:

	y = 1	y = 1	
	$X_2 = 1$	$X_2 = 0$	
$X_1 = 1$	P_{11}	P_{10}	
$X_1 = 0$	P_{01}	P_{00}	

Dove P_{ij} rappresenta la probabilità di che la variabile dipendente assuma valore 1 in base all'esposizione o meno alle variabili indipendenti X_1 e X_2.

Individuando con P_{00} il rischio base, ci sono altri 3 rischi relativi che si possono configurare:

$OR_{10} = \frac{P_{10}Q_{00}}{P_{00}Q_{10}}$: Odds ratio relativo ad un'esposizione al solo fattore X_1

$OR_{01} = \frac{P_{01}Q_{00}}{P_{00}Q_{01}}$: Odds ratio relativo ad un'esposizione al solo fattore X_2

$OR_{11} = \frac{P_{11}Q_{00}}{P_{00}Q_{11}}$: Odds ratio relativo ad un'esposizione sia al fattore X_1 che al fattore X_2

Il modello viene espresso nella forma:

$\text{logit}P(x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$

Di particolare interesse è testare l'ipotesi moltiplicativa: $OR_{11} = OR_{10} * OR_{01}$ sotto la quale il rischio relativo per l'esposizione alla variabile X_1 è indipendente dai livelli di X_2, o, analogamente, il rischio relativo conseguente all'esposizione ad X_2 è indipendente dall'esposizione ad X_1.

Testare questa ipotesi è equivalente a testare l'ipotesi $\beta_3 = 0$.

Avremo, dunque:

$\alpha = \text{logit } P_{00}$

$\beta_1 = \log(OR_{10}) = \text{logit } P_{10} - \text{ logit } P_{00}$

$\beta_2 = \log(OR_{01}) = \text{logit } P_{01} - \text{ logit } P_{00}$

$\beta_3 = \log\left(\frac{OR_{11}}{OR_{10}OR_{01}}\right) = \text{logit } P_{11} - \text{ logit } P_{00} - [\text{ logit } P_{10} - \text{logit } P_{00} + \text{logit } P_{01} - \text{logit } P_{00}] = \text{logit } P_{11} - \text{logit } P_{10} - \text{logit } P_{01} + \text{logit } P_{00}$

Dove $\exp(\beta_3)$ cattura il fattore moltiplicativo per cui il rischio relativo, per la doppia esposizione

ad entrambe le variabili esplicative, differisce dal prodotto dei rischi relativi per le singole esposizioni. Se il parametro è positivo si individua un'interazione positiva, se è negativo avremo un'interazione negativa. Soltanto nell'eventualità in cui il parametro β_3, corrispondente alle interazioni fra le variabili, dovesse essere non significativamente diverso da 0, allora si potrà considerare la possibilità di utilizzare un modello ridotto dei termini di interazione. Si notino anche le differenze di interpretazione cui sono suscettibili i parametri quando si passa da un modello con interazioni ad uno senza, in quanto il significato stesso delle variabili di regressione inserite nel modello dipende da quali variabili vengono incluse nel modello stesso. Nel modello con interazione, o saturato, il parametro β_1 rappresenta il rischio relativo logaritmico per esposizione ad X_1 soltanto per valori assunti da X_2 nulli. Nel modello senza interazioni invece β_1 assume questo significato anche per valori non nulli assunti da X_2. Inoltre, va ricordato che effettuare test ponendo $\beta_1 = 0$ non è di alcun significato in modelli saturati, in quanto un modello con interazioni tra variabili che non presentano effetti singoli statisticamente significativi è di scarsa utilità pratica. Infine, se nessuna delle variabili di regressione costituisce dei termini di interazione con i fattori usati per la stratificazione della popolazione dello studio, ne deriva che il rischio relativo associato ai fattori di rischio inseriti nello studio è costante per le categorie individuate per la stratificazione.

Le variabili esplicative possono tuttavia assumere valori individuati in più di due categorie, come spesso avviene nel caso dell'età.

Individuiamo così la tabella che evidenzia la probabilità P(y = 1) data l'esposizione o meno ai diversi livelli delle variabili esplicative.

X_2:

Classe1 Classe 2 Classe 3

P_{12}	P_{11}	P_{10}	$X_1 = 1$
P_{02}	P_{01}	P_{00}	$X_1 = 0$

Da questa derivano cinque Odds ratios:

$OR_{ij} = (\frac{P_{ij}\, Q_{00}}{P_{00} Q_{ij}})$

Relativi al rischio base di esito positivo, y = 1, in assenza di esposizione ad entrambi i fattori o assenza di esposizione ad uno ed esposizione alla categoria di valore minore dell'altro.

In termini generali, per un fattore con K categorie sarà necessario esplicitare K-1 variabili esplicative nel modello per descriverne gli effetti.

Il modello, con e senza le interazioni, sarà:

logit $P_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$

logit $P_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Dove il secondo modello verrà adottato solo se non sarò possibile rifiutare l'ipotesi:

$H_0: \beta_4 = \beta_5 = 0$

A favore di:

$H_1: \beta_4 \neq \beta_5 \neq 0$

Si noti che l'ipotesi nulla equivale a testare:

$OR_{ij}=OR_{i0}*OR_{oj}$

Possiamo concentrarci ora sul metodo di stima dei parametri associati alle variabili.

Poiché il modello deriva dalla concezione della variabile dipendente come variabile categorica che assume valore 1 quando il soggetto è deceduto e valore 0 quando il soggetto sopravvive, sotto l'assunzione di una distribuzione di probabilità di tipo logistico per questa eventualità, è possibile trovare i valori dei parametri β che massimizzano la probabilità di aver osservato i dati, attraverso il metodo della massima verosimiglianza ed una funzione di verosimiglianza che viene derivata dalla distribuzione di Bernoulli.

Volendo essere più specifici, l'obiettivo della nostra regressione è stimare la probabilità:

$P(y_i = 1)$

tramite una funzione, scritta in forma matriciale:

$F(X_i{}'\beta)$.

Nulla ci vieta, per esempio, di utilizzare lo stimatore dei minimi quadrati con la semplice variabile dipendente definita come categorica con valori alternativi 0 e 1. Tuttavia, implementare lo stimatore OLS, in questo particolare caso, non è la scelta più popolare in quanto non esclude la possibilità di valori fit non compresi tra 0 e 1, risultato chiaramente incompatibile con la nostra esigenza di pervenire ad una stima di una probabilità inevitabilmente definita tra 0 e 1. In questo senso risulta fondamentale per giungere al risultato stabilito, modellare le variabili esplicative

attraverso una funzione che assuma il range di valori desiderato che può essere individuata in una funzione di densità cumulata. Questo particolare tipo di funzione restituisce la probabilità che una determinata variabile assuma un valore minore o uguale rispetto ad un valore dato.

In questo caso, data una generica funzione di densità cumulata:

$F(X_i{}'\beta)$

la probabilità associata alla variabile dipendente viene individuata come:

$\int_{-\infty}^{X_i{}'\beta} F(X_i{}'\beta)$

Per procedere con la stima dei parametri β si ricorre al metodo della Massima verosimiglianza. Contrariamente al modello OLS, dove si minimizzava il quadrato dei residui della regressione, con questo metodo si vuole ottenere il valore dei parametri che massimizza la probabilità di aver osservato i dati raccolti, sotto l'ipotesi di una determinata distribuzione di probabilità per i dati stessi.

Individuata la probabilità congiunta delle osservazioni campionarie come:

$P(y_1 \dots y_n | \beta) = \prod_{i=1}^{n} P(y_i | \beta)$;

l'approccio della Massima verosimiglianza individua il vettore di valori β che massimizzano questo prodotto. Il primo step per questo processo è dunque l'individuazione di una distribuzione di probabilità per le osservazioni della variabile dipendente Y dati i valori delle variabili indipendenti. Questa funzione viene definita funzione di verosimiglianza in quanto definisce la probabilità di aver ottenuto le osservazioni della variabile dipendente data la probabilità stessa di quelle osservazioni.

La funzione di verosimiglianza è dunque definita come:

$P(y_i | \beta)$.

Nel nostro caso specifico, essendo i valori assunti dalla y solo 0 e 1, può essere individuata nella distribuzione di Bernoulli.

Questa assume la forma:

$f(y_i, p_i) = (p_i)^{y_i}(1 - p_i)^{1-y_i}$

Dove:

p_i è la proporzione di $y_i = 1$ sul totale, ovvero la probabilità che la variabile dipendente assuma

valore 1: $P(y_i = 1)$,

y_i assume valore 1 quando la variabile dipendente avrà assunto quello stesso valore e 0 quando la variabile dipendente avrà assunto valore 0.

Affinché sia possibile stimare i parametri β di nostro interesse, risulta dunque fondamentale poter sostituire p_i con i parametri stessi o una funzione di questi, così da poter massimizzare la funzione di verosimiglianza tramite i parametri ed individuare quest'ultimi con i valori che rendono la probabilità di aver osservato y_i massima.

Per rendere tutto questo possibile, essendo p_i la probabilità di aver osservato y_i, dovremo riscrivere quest'ultima in termini dei parametri tramite:

$$P(y_i = 1) = F({X_i}'\beta)$$

Dove si ricorda che $F({X_i}'\beta)$ consiste in una funzione di densità cumulata.

A questo punto, dovremo fare una scelta, contrariamente a quanto avveniva nello stimatore OLS, ed assumere una funzione di densità cumulata per $P(y_i = 1)$.

Qui introduciamo la funzione logit in quanto l'assunzione nella stima dei parametri tramite regressione logistica è:

$$P(y_i = 1) = \boldsymbol{\Lambda}({X_i}'\beta) = \frac{\exp({X_i}'\beta)}{1+\exp({X_i}'\beta)}$$

Ora possiamo massimizzare la funzione di verosimiglianza sostituendo $\frac{\exp({X_i}'\beta)}{1+\exp({X_i}'\beta)}$ a p_i in quanto entrambi identificano $P(y_i = 1)$.

Otterremo dunque:

$$f(y_i, p_i) = \left(\frac{\exp({X_i}'\beta)}{1+\exp({X_i}'\beta)}\right)^{y_i}\left(1 - \frac{\exp({X_i}'\beta)}{1+\exp({X_i}'\beta)}\right)^{1-y_i}$$

e massimizzeremo la log verosimiglianza, ovvero il logaritmo della funzione di verosimiglianza, così:

$$\ln(l) = \sum_{i=1}^{N}\left[y_i\left(\ln\left(\frac{\exp({X_i}'\beta)}{1+\exp({X_i}'\beta)}\right)\right) + (1 - y_i)\left(\ln\left(1 - \frac{\exp({X_i}'\beta)}{1+\exp({X_i}'\beta)}\right)\right)\right]$$

Le condizioni di massimo implicano che la derivata prima rispetto ai parametri sia posta uguale a 0 e che la derivata seconda sia negativa per determinare la concavità della funzione.

Dunque:

$\frac{\delta}{\delta\beta} l(\beta; y) = \sum_{i=1}^{N} \{ \frac{y_i - \Lambda(X_i'\beta)}{\Lambda(X_i'\beta)[1-\Lambda(X_i'\beta)]} * \lambda(X_i'\beta) \} = 0$;

$\frac{\delta^2}{\delta\beta\delta\beta'} l(\beta; y) < 0$;

Dove λ rappresenta la derivata prima di $\boldsymbol{\Lambda}(X_i'\beta)$.

Quando queste condizioni sono rispettate ricaviamo i valori dei parametri che massimizzano la funzione di verosimiglianza e quindi otterremo i valori per cui la probabilità di aver avuto le osservazioni di y è massima.

ANALISI EMPIRICA

Una volta effettuato il processo di raccolta dei dati, è possibile procedere con l'analisi dell'impatto che età e genere hanno sulla letalità del Covid-19.

Preliminarmente è stata costruita una tabella che rappresenta le proporzioni dei soggetti deceduti, rispetto ai totali considerati nell'analisi, sulla base dell'appartenenza ad una determinata classe di età e genere:

	num	den	y	sex	eta
1	889	19800	0.04489899	2	1
2	2590	19019	0.13617961	2	2
3	6177	19472	0.31722473	2	3
4	7605	16763	0.45367774	2	4
5	1947	3869	0.50323081	2	5
6	281	23124	0.01215188	1	1
7	807	12952	0.06230698	1	2
8	2702	14628	0.18471425	1	3
9	6113	24670	0.24779084	1	4
10	4053	15220	0.26629435	1	5

Nella tabella il valore della y corrisponde al rapporto tra i valori riportati. Il valore '2' per la categoria sex corrisponde al genere maschile mentre le classi di età sono suddivise da 1 a 5 dove la prima corrisponde a 55-65 e l'ultima 95+. Ogni classe ha un intervallo di 10 anni.

Una volta individuato il valore delle proporzioni, possiamo costruire i nostri primi due modelli. Uno basato solo sulla regressione della variabile dipendente sul sesso, e l'altro esclusivamente sull'età. Questi due modelli, particolarmente semplici portano i seguenti risultati:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.703184	0.009203	-185.06	<2e-16 ***
sex2	0.568927	0.012390	45.92	<2e-16 ***

per il sesso.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.57479	0.02964	-120.61	<2e-16 ***
eta65	1.44519	0.03476	41.58	<2e-16 ***
eta75	2.53080	0.03211	78.83	<2e-16 ***
eta85	2.87153	0.03143	91.38	<2e-16 ***
eta95	2.79478	0.03349	83.45	<2e-16 ***

per l'età.

Queste tabelle mostrano le stime dei coefficienti attribuiti dallo stimatore e accompagnati dai corrispondenti errori standard e p values. Si noti come la relazione tra stime ed errori standard implichi p values molto piccoli mostrando forte evidenza contro l'ipotesi H_0: $\beta_i = 0$. Tutti i parametri stimati in entrambi i modelli sono infatti significativamente diversi da 0 con valori di p, ovvero la probabilità di avere la data osservazione sotto l'ipotesi che H_0 sia vera, estremamente piccoli.

L'interpretazione dei coefficienti in un modello di regressione logistica, contrariamente a quanto avviene per un modello lineare stimato ad esempio attraverso OLS, non è l'aumento che subisce la variabile dipendente per incremento unitario della variabile esplicativa. In realtà il coefficiente così stimato è di scarsa significatività perché la stima della probabilità avviene nella forma di:

$$p = \frac{\exp(\beta_0+\beta_1\delta_{e1}+\beta_2\delta_{e2}+\beta_3\delta_{e3}+\beta_4\delta_{e4}+\beta_1\delta_{s1})}{1-\exp(\beta_0+\beta_1\delta_{e1}+\beta_2\delta_{e2}+\beta_3\delta_{e3}+\beta_4\delta_{e4}+\beta_1\delta_{s1})}$$

Tuttavia, nel nostro caso specifico con variabili dummy, si ottiene che $[\exp(\beta_1\delta_{e1} + \beta_5\delta_{s1})]$ è l'aumento del rischio relativo rispetto all'osservazione base, ovvero donna di 55-65 anni, per un uomo appartenente alla fascia di età 65-75 anni.

Avremo infatti che il rischio relativo logaritmico per l'osservazione base sarà:

$$\left(\frac{p}{1-p}\right) = \exp(\beta_0)$$

Poiché tutte le variabili dummy sono nulle in questo caso, e:

$$\left(\frac{p}{1-p}\right) = \exp(\beta_0 + \beta_1\delta_{e1} + \beta_5\delta_{s1})$$

Sarà il rischio relativo logaritmico per l'osservazione maschile con età tra 65-75 anni.

Per le proprietà degli esponenti:

$$\exp(\beta_0 + \beta_1\delta_{e1} + \beta_1\delta_{s1}) = \exp(\beta_0) * \exp(\beta_1\delta_{e1} + \beta_5\delta_{s1})$$

e quindi l'aumento del rischio relativo logaritmico dall'osservazione base ad un'altra sarà dato, nel caso specifico, da:

$$[\exp(\beta_1\delta_{e1} + \beta_5\delta_{s1})]$$

E, in generale, da:

$[\exp(\beta_1\delta_{en} + \beta_5\delta_{s1})]$ se è uomo, oppure:

$[\exp(\beta_1\delta_{en})]$ se è donna. Con n =1 ,… ,N che indica la classe di età di appartenenza

Quindi nel caso del modello che considera esclusivamente il fattore genere, passando da un'osservazione femminile ad una maschile, il rischio relativo logaritmico aumenterà: exp(0.568927) = 1.76637, leggermente superiore ad una volta e mezzo.

Nel caso del modello che considera esclusivamente l'età, passando da una classe all'altra, avremo un aumento del rischio relativo logaritmico di:

	parameters	log relative risk increase:
eta65	1.44519	4.24266
eta75	2.53080	12.56355
eta85	2.87153	17.66402
eta95	2.79478	16.35903

Una volta individuati i due modelli separati e valutato singolarmente la significatività di ogni variabile esplicativa, possiamo stimare un modello congiunto. La stima del modello fornisce i seguenti risultati:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.08053	0.03114	-131.03	<2e-16 ***
eta65	1.34651	0.03491	38.57	<2e-16 ***
eta75	2.47828	0.03226	76.83	<2e-16 ***
eta85	2.98450	0.03169	94.18	<2e-16 ***
eta95	3.09928	0.03419	90.64	<2e-16 ***
sex2	0.89421	0.01398	63.95	<2e-16 ***

Si noti che ogni singolo parametro stimato è significativamente diverso da 0. Questo risultato è di estrema importanza perché chiarisce che né genere né sesso sono variabili confondenti. Appariva evidente il rischio di imputare un effetto al genere che fosse in realtà dovuto a fattori biologici dovuti alle differenze di età e aspettative di vita fra i due gruppi, invece che afferenti a differenze dovute all'appartenenza ad un genere piuttosto che ad un altro. Al contempo, gli effetti delle due variabili considerate, stimate attraverso i parametri, passando dal modello ristretto a quello congiunto, subiscono variazioni di entità minima mostrando come gli effetti di ogni singola variabile, non subendo variazioni una volta controllato per l'altra, siano da imputare largamente alle variabili stesse e non vi sia evidenza di una larga distorsione dovuta al loro essere confondente. Questo è evidenziato nella seguente tabella che mette a confronto i valori dei parametri associati alle variabili nel modello ristretto ed in quello congiunto:

	ristretto	congiunto
eta65	1.445187	1.346509
eta75	2.530803	2.478282
eta85	2.871527	2.984501
eta95	2.794779	3.099283
sex2	0.5689271	0.8942054

Una volta stimato il modello congiunto possiamo anche valutare la presenza o meno di interazioni tra le variabili esplicative che siano significativamente diverse da 0. Il modello con le

interazioni presenta i seguenti risultati:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-4.39805	0.06002	-73.275	<2e-16	***
eta65	1.68670	0.07017	24.037	<2e-16	***
eta75	2.91332	0.06369	45.742	<2e-16	***
eta85	3.28762	0.06181	53.193	<2e-16	***
eta95	3.38454	0.06276	53.929	<2e-16	***
sex2	1.34064	0.06914	19.390	<2e-16	***
eta65:sex2	-0.47668	0.08092	-5.891	3.85e-09	***
eta75:sex2	-0.62247	0.07397	-8.415	<2e-16	***
eta85:sex2	-0.41604	0.07238	-5.748	9.02e-09	***
eta95:sex2	-0.31421	0.07842	-4.007	6.16e-05	***

Si noti come in presenza di interazioni, i valori dei parametri associati al genere ed alle interazioni, costituite dalle moltiplicazioni delle variabili esplicative tra di loro, non siano tutti contemporaneamente non significativamente diversi da 0. Questo vuol dire che per ognuno di questi parametri possiamo rigettare l'ipotesi nulla, che questi siano pari a 0, perché la probabilità della stima ottenuta, sotto l'ipotesi che l'ipotesi nulla sia vera, è estremamente bassa. In questo caso ci aspettiamo che un test F, per valutare l'ipotesi che tutti questi parametri siano contemporaneamente nulli, ci restituisca un p value piccolo, così da poter rifiutare l'ipotesi nulla e attestare la presenza di interazioni tra le due variabili.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL	9	24057.5				
eta	4	19702.9	5	4354.5	<2.2e-16	***
sex	1	4250.3	4	104.2	<2.2e-16	***
eta:sex	4	104.2	0	0.0	<2.2e-16	***

In effetti questo è il risultato suggerito proprio dal test in questione. Nella superiore tabella sono

riportati i risultati che coincidono esattamente con quanto previsto. I parametri dei termini di interazione tra le variabili risultano infatti tutti contemporaneamente statisticamente significativi come evidenziato da un p value estremamente basso minore di 2.2e-16. Di conseguenza si rifiuta l'ipotesi nulla:

H_0: $\beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$

a favore di:

H_1: $\beta_6 \neq \beta_7 \neq \beta_8 \neq \beta_9 \neq 0$

La significatività nei termini di interazione ci conduce alla conclusione che il modello migliore e preferibile per la stima della letalità del virus sulla base delle variabili genere ed età è il modello congiunto includente le interazioni tra parametri.

Questo risultato implica che l'effetto di età e genere non saranno costanti per i livelli dell'altra variabile. Avremo un'attenuazione della letalità, rispetto al rischio base, per ogni doppia esposizione alle classi dei fattori di rischio in quanto siamo posti di fronte ad interazioni negative.

Il confronto fra i parametri stimati nei modelli grezzi e quello congiunto con le interazioni mostra:

	ristretto	congiunto
eta65	1.445187	1.686696
eta75	2.530803	2.913317
eta85	2.871527	3.287616
eta95	2.794779	3.384540
sex2	0.568927	1.340644

In questo caso, tutti i parametri sono ancora statisticamente significativi indicando che nessuna delle due variabili considerate sia di tipo confondente, ma l'effetto del genere subisce una variazione importante, indicando come, nella stima attraverso il modello grezzo, l'effetto del genere sulla letalità venga sottostimato e come questo effetto sia in parte dovuto a fattori biologici da imputare alle diverse aspettative di vita tra uomo e donna.

Il passo successivo nella nostra analisi consiste nel calcolo degli intervalli di confidenza al 5% per ogni parametro stimato nel modello.

	Upper limit	lower limit	parameters
(Intercept)	-4.5179830	-4.2826165	-4.39805
eta65	1.5505589	1.8257348	1.68670
eta75	2.7904211	3.0401760	2.91332
eta85	3.1685550	3.4109212	3.28762
eta95	3.2635404	3.5096461	3.38454
sex2	1.2066091	1.4777380	1.34064
eta65:sex2	-0.6363199	-0.3190126	-0.47668
eta75:sex2	-0.7688175	-0.4787712	-0.62247
eta85:sex2	-0.5593358	-0.2755244	-0.41604
eta95:sex2	-0.4691500	-0.1616454	-0.31421

I valori compresi tra quelli individuati nella tabella corrispondono all'intervallo di confidenza ed il limite superiore e inferiore evidenziano i valori limite oltre i quali la probabilità di quelle osservazioni risulta inferiore al 5%. Attraverso l'intervallo riportato, infatti, individuiamo l'insieme di valori plausibili per il parametro oggetto delle nostre stime con un livello di confidenza scelto del 95%.

Possiamo ripetere la stessa operazione elevando il numero di Nepero ad ogni parametro:

	parameters	upper limit	lower limit
(Intercept)	0.01230136	0.01091101	0.01380649
eta65	5.40160516	4.71410415	6.20735446
eta75	18.41778333	16.28787652	20.90892251
eta85	26.77894405	23.77310815	30.29313680
eta95	29.50440816	26.14192511	33.43643179
sex2	3.82150196	3.34213253	4.38301993
eta65:sex2	0.62083835	0.52923647	0.72686637
eta75:sex2	0.53661726	0.46356089	0.61954423
eta85:sex2	0.65965653	0.57158861	0.75917390
eta95:sex2	0.73036238	0.62553377	0.85074280

Si ricordi che i valori individuati dalla tabella, con l'esclusione dell'intercetta e della variabile genere e delle interazioni, rappresentano l'aumento del rischio relativo logaritmico per una donna appartenente ad una fascia d'età diversa da quella base, ovvero 55-65 anni. Per gli uomini questo è dato dal valore individuato sotto la colonna parameters, moltiplicato per il valore corrispondente alla variabile genere, moltiplicato per il valore corrispondente alla interazione tra la fascia di età di appartenenza ed il genere.

Infatti:

$[\exp(\beta_1\delta_{en} + \beta_5\delta_{s1} + \beta_{n+5}\delta_{en}\delta_{s1})] = [\exp(\beta_1\delta_{en}) * \exp(\beta_5\delta_{s1}) * \exp(\beta_{n+5}\delta_{en}\delta_{s1})]$ in cui n = 1 per eta 65 fino a n = 4 per eta95.

Una volta individuato l'intervallo di confidenza per ogni parametro possiamo calcolare i valori previsti dal modello e metterli a confronto con quelli osservati. Questo passaggio è particolarmente utile perché è un'evidenza circa la bontà del modello ed offre un riscontro per valutare se le variabili genere e sesso effettivamente abbiano un impatto sulla letalità del virus e se questa stessa può essere spiegata attraverso le differenze all'interno della popolazione in termini delle variabili esplicative considerate.

	sex	eta	num	den	y	fv
1	2	1	889	19800	0.04489899	0.04489899
2	2	2	2590	19019	0.13617961	0.13617960
3	2	3	6177	19472	0.31722473	0.31722471
4	2	4	7605	16763	0.45367774	0.45367768
5	2	5	1947	3869	0.50323081	0.50323075
6	1	1	281	23124	0.01215188	0.01215188
7	1	2	807	12952	0.06230698	0.06230697
8	1	3	2702	14628	0.18471425	0.18471423
9	1	4	6113	24670	0.24779084	0.24779082
10	1	5	4053	15220	0.26629435	0.26629433

Nella tabella vengono riportate le osservazioni di deceduti e contagiati per ogni categoria di genere ed età, individuate nelle prime due colonne. 'Num' e 'den' corrispondono rispettivamente ai deceduti e dai contagiati, mentre la 'y' corrisponde al rapporto tra queste due variabili, ovvero il valore vero delle proporzioni di deceduti su contagiati per ogni categoria. La colonna 'fv' riporta i 'fittted values' ovvero i valori previsti dal modello. Si può subito notare come i valori veri e quelli fit siano quasi identici. Da questo confronto quindi il modello appare uno stimatore preciso delle probabilità di decesso sulla base dell'appartenenza ad una determinata classe di età o genere.

Un ulteriore strumento per valutare la bontà del fit di un modello di massima verosimiglianza, come la regressione logistica, è lo pseudo R2 di McFadden, definito come:

$$R^2 = 1 - \frac{\ln(lc)}{\ln(lnull)}$$

Dove Lc corrisponde alla funzione di verosimiglianza massimizzata per il modello completo e Lnull corrisponde al modello che presenta solo l'intercetta.

L'R2 di McFadden assumerà valori vicini a 0 quando il modello completo sarà molto vicino al modello ristretto implicando che le variabili inserite nel modello completo abbiano scarsa significatività. Questo tenderà ad 1 invece quando Lc sarà 1, ovvero quando il modello restituirà

con probabilità vicina ad 1 un valore di Y = 1 quando questa è 1 ed un valore di Y = 0 quando questa è 0, ovvero la funzione di verosimiglianza è vicina ad 1 e le variabili considerate nel modello sono significative.

Questo per il nostro modello equivale: 0.9918435

Un valore estremamente alto che implica che il fit del modello è ottimo.

CONCLUSIONE

Sulla base dei risultati ottenuti dall'analisi empirica esposta, possiamo trarre le conclusioni inerenti all'impatto che le variabili genere ed età hanno avuto sulla probabilità del virus di causare morte.

I dati pubblicati nel bollettino dell'Istituto Superiore di Sanità sono stati classificati per genere ed età, dove la seconda variabile è stata divisa in 5 classi differenti con intervallo di 10 anni, partendo dalla prima classe che include tutti i soggetti con età compresa tra 55 e 65 anni fino all'ultima che comprende i soggetti con più di 95 anni.

Con i dati a nostra disposizione così elaborati, è stata effettuata una prima stima separata degli impatti delle due variabili sulla letalità del virus. Tale analisi ha evidenziato delle relazioni significativamente diverse da 0 con livelli di confidenza pari al 95%. La letalità si è, dunque, rilevata connessa ad entrambe le variabili.

L'impatto delle variabili di interesse, stimato tramite i primi due modelli grezzi, è stato verificato con la costruzione di un modello congiunto. Questo modello presenta contemporaneamente entrambe le variabili esplicative, età e genere. Lo scopo di questa regressione è quello di valutare come variano gli effetti che sono stati imputati alle singole variabili nei modelli grezzi, una volta che ogni variabile viene controllata per l'altra, e se questi effetti sono ancora statisticamente significativi. I risultati di questa regressione evidenziano variazioni nei valori dei parametri associati alla variabile età di piccola entità. Il parametro associato alla variabile genere, invece, assume un valore chiaramente superiore nel modello congiunto rispetto a quello grezzo. Questo risultato ci porta ad affermare che l'effetto di questa variabile sia parzialmente spiegato da differenze nei fattori biologici derivanti dall'età dei soggetti considerati.

Gli errori standard del modello congiunto a loro volta seguono la tendenza evidenziata dai parametri, non subendo aumenti di grandi dimensioni eccetto che per la variabile genere, suggerendo che parte della variabilità della variabile esplicativa genere sia comune all'età, come d'altronde ci si attendeva considerando i valori della variazione dell'effetto stimato passando da un modello all'altro. I risultati della regressione del modello congiunto, sebbene implichino che parte dell'impatto del genere sulla letalità sia da attribuire al fattore età, non offrono comunque alcuna evidenza che le due variabili possano essere di tipo confondente, in quanto tutti i

parametri rimangono statisticamente significativi.

Data l'evidenza che entrambe le variabili non siano di tipo confondente, attraverso la costruzione di un modello che presenti interazioni tra le variabili esplicative del modello, si è valutata l'eventualità in cui gli effetti stimati per ogni variabile non siano costanti per tutti i livelli dell'altra variabile. Per testare questa ipotesi si è dovuto ipotizzare che i parametri associati alle interazioni siano tutti nulli e si è valutata la probabilità di aver osservato le stime dei parametri ottenuti sotto questa ipotesi. I risultati di questo test ci conducono a concludere che è impossibile rigettare l'ipotesi nulla per cui i parametri corrispondenti alle interazioni siano tutti nulli. Questo suggerisce che i parametri delle interazioni siano tutti contemporaneamente statisticamente significativi e che gli effetti stimati concernenti il genere non siano costanti per tutti i livelli d'età e che l'effetto attribuito all'appartenenza ad una determinata classe d'età non sia costante per i due generi.

Da quest'ultima regressioni traiamo come conclusione che il modello preferibile per valutare l'impatto delle variabili esplicative di interesse sia un modello congiunto con interazioni tra le variabili.

I parametri stimati tramite questo modello hanno la seguente interpretazione:

- L'intercetta elevata al numero di Nepero rappresenta il rischio relativo per l'osservazione base, ovvero una donna con età compresa tra 55-65.
- Il rischio relativo di una donna con età individuata in una classe diversa rispetto a 55-65 anni, aumenta, rispetto a quello dell'osservazione base, di un fattore moltiplicativo pari al numero di Nepero elevato al parametro associato alla classe di età di appartenenza.
- Il rischio relativo per un uomo con età compresa tra 55-65 anni aumenta, rispetto a quello dell'osservazione base, per un fattore moltiplicativo pari al numero di Nepero elevato al parametro associato alla variabile sesso.
- Il rischio relativo per un uomo con classe di età diversa da 55-65 anni, aumenta, rispetto a quello dell'osservazione base, di un fattore moltiplicativo pari al prodotto tra:
    1. Il numero di Nepero elevato al parametro corrispondente alla classe di età di appartenenza;
    2. Il numero di Nepero elevato al parametro associato alla variabile genere;
    3. Il numero di Nepero elevato al parametro stimato per l'interazione tra la variabile genere e la classe d'età di appartenenza.

All’esito del lavoro si è verificata la bontà del fit del modello.

In prima istanza, con l’indice R^2 di McFadden. Questa statistica è un indice della bontà di adattamento del modello ai dati e viene usata per modelli non lineari, in sostituzione del meglio noto R^2, poiché è una misura più generica e adeguata. Questo assume ugualmente valori compresi tra 0 e 1, dove i risultati vicino a 0 indicano un fit particolarmente scarso e quelli vicini ad 1 un livello di adattamento ottimo, e fa parte della famiglia degli pseudo-R^2, ovvero le statistiche per modelli non lineari che indicano la bontà del fit di un modello. L’R^2 di McFadden per la nostra regressione assume un valore estremamente alto e vicino ad 1 a conferma della bontà di adattamento del modello congiunto con interazioni ai dati.

In seconda istanza, è stata presentata una tabella contente i valori stimati tramite il modello ed i veri valori assunti dalla variabile dipendente. Questi sono molto simili, a conferma di quanto già evidenziato dalla statistica R^2 di McFadden, e evidenziano ulteriormente la capacità delle variabili esplicative di spiegare con precisione la variabilità della variabile dipendente.

BIBLIOGRAFIA

Breslow N.E., Day N.E., *Statistical Methods in cancer research Volume 1: The Analysys of case-control studies*, IARC Scientific Publication n. 32

Riani M, *Il modello di regressione logistica, introduzione e inferenza*, UNI.NOVA, Parma, 2013

Bollettino dell'Istituto Superiore di Sanità (ISS), *Epidemia COVID-19*, Roma, 23 giugno 2020 https://www.epicentro.iss.it/coronavirus/bollettino/Bollettino-sorveglianza-integrata-COVID-19_23-giugno-2020.pdf

TABELLE

DISTRIBUZIONE DEI CASI DIAGNOSTICATI DAI LABORATORI DI RIFERIMENTO REGIONALE (N=239.709) E DEI DECESSI SEGNALATI (N= 33.542) PER

FASCIA DI ETÀ E SESSO

Soggetti di sesso maschile

classe di età (anni)	numero casi	numero deceduti
0-9	1135	1
10-19	1937	0
20-29	6117	12
30-39	8520	43
40-49	13127	211
50-59	19800	889
60-69	19019	2590
70-79	19472	6177
80-89	16763	7605
90+	3869	1947
età non nota	14	0
Totale	109.773	19475

Soggetti di sesso femminile

classe di età (anni)	numero casi	numero deceduti
0-9	1021	3
10-19	1908	0
20-29	7734	4
30-39	10441	23
40-49	18188	81
50-59	23124	281
60-69	12952	807
70-79	14628	2702
80-89	24670	6113
90+	15220	4053
età non nota	24	0
Totale	129.910	14067

Casi totali

classe di età (anni)	numero casi	numero deceduti
0-9	2156	4
10-19	3845	0
20-29	13851	16
30-39	18961	66
40-49	31315	292
50-59	42924	1170
60-69	31971	3397
70-79	34100	8879
80-89	41433	13718
90+	19089	6000
età non nota	38	0
Totale	239.683	33542

SOMMARIO