UNIVERSITÀ DEGLI STUDI MILANO-BICOCCA
Scuola di Economia e Statistica
Corso di laurea in
STATISTICA E GESTIONE DELL'INFORMAZIONE

APPLICAZIONE DEI MODELLI DI REGRESSIONE SELF SELECTION E TECNICHE DI MACHINE LEARNING PER L'ANALISI DEI SINTOMI E DETERMINANTI DEL LONG COVID

Relatore: Prof. Pietro Giorgio Lovaglio

Tesi di Laurea di:
Sara Capozio
Matr. N. 853600

Anno Accademico 2021/2022

Indice

Introduzione allo studio

L'Istituto Superiore di Sanità, nel Rapporto ISS CoViD-19 n. 15/2021, definisce Long Covid *"la condizione di persistenza di segni e sintomi che continuano o si sviluppano dopo un'infezione acuta da SARS-CoV-2. Se i sintomi continuano a manifestarsi oltre quattro settimane dall'infezione fino a 12 settimane, si parla di malattia CoViD-19 sintomatica persistente; se i sintomi si prolungano per più di 12 settimane e non possono essere spiegati da nessun'altra condizione, si parla di Sindrome post-CoViD. Il Long-CoViD include entrambe queste condizioni."*

Essendo un fenomeno recente e in continuo sviluppo, la definizione sopra citata è in costante aggiornamento e, in seguito ad un'attenta revisione bibliografica, le formulazioni risultano essere molteplici. Per citarne alcune:

- *"La condizione post COVID-19 si verifica in individui con una storia di infezione da SARS CoV-2 probabile o confermata, di solito 3 mesi dall'inizio del COVID-19 con sintomi e che durano per almeno 2 mesi e non possono essere spiegati da una diagnosi alternativa"* (World Health Organization)
- *"Long-COVID è il termine utilizzato al fine di denotare una persistenza di sintomi in quelle persone che sono state ricoverate per il virus SARS-CoV-2"* (Raveendran et al., 2021).
- *"Segni e sintomi che si sviluppano durante o dopo un'infezione compatibile con COVID-19, continuano per più di 12 settimane e non sono spiegati da una diagnosi alternativa. Di solito si presenta con grappoli di sintomi, spesso sovrapposti, che possono fluttuare e cambiare nel tempo, oltre ad essere in grado di interessare qualsiasi sistema ed apparato dell'organismo".*(National Institute for Health and Clinical Excellence)

Il seguente studio si pone per scopo l'analisi dei determinanti delle sintomatologie Long-Covid, fenomeno che, in questa trattazione, si intenderà generalmente come la persistenza e lo sviluppo dei sintomi oltre il decorso clinico della malattia.

I dati, per affrontare la ricerca, sono stati forniti da un grande ospedale metropolitano specializzato nella cura del Covid-19 e del Long-Covid. Il campione, costituito da 1391 osservazioni, è composto da soggetti che, dopo aver contratto l'infezione da SARSCOV2 ed esserne guariti, hanno aderito all'indagine statistica promossa dalla struttura ospedaliera. Gli individui in analisi hanno contratto l'infezione in due ondate pandemiche

differenti, riguardanti rispettivamente il periodo dal 23/02/2020 al 01/10/2020 e quello dal 01/10/2020 al 26/07/2021. Poiché i soggetti hanno contratto il virus in periodi diversi, i tempi di follow-up risultano differenti e le osservazioni si sono concluse in data 14/10/2021 con l'acquisizione dell'ultimo questionario.

Per perseguire l'obiettivo in studio, sono state costituite delle Macro-Sintomatologie, espresse in termini di durate mediante la tecnica della Componenti Principali. Successivamente, lo studio dei determinanti è stato affrontato tramite l'applicazione di modelli di regressione lineari di Self Selection, supportati dall'utilizzo di metodologie di Machine Learning.

La seguente trattazione si aprirà con la descrizione delle variabili che costituiscono il campione in studio, suddivise in sintomatologie post-Covid (variabili dipendenti) e caratteristiche clinico-anagrafiche dei pazienti (variabili esplicative).

In seguito ad una fase di pulizia dei dati, il secondo capitolo si concluderà con l'aggregazione dei sintomi in macro-gruppi tramite la tecnica delle componenti principali. Poiché uno degli scopi dell'analisi consiste nell'identificare i determinanti Long-Covid, nel terzo capitolo, verranno eseguite delle analisi esplorative in grado di evidenziare ciò che ha contraddistinto ogni ondata infettiva.

La seconda parte della trattazione si concentrerà sulle fasi di analisi. Il quarto capitolo, infatti, è dedicato allo studio dei modelli di Self Selection; in particolare verrà descritto il metodo di stima, le sue proprietà, i vantaggi e i limiti d'applicazione.

Lo studio degli effetti delle covariate sui macro-sintomi suddivisi per ondate, riportato nel sesto capitolo, sarà preceduto da una fase preliminare di selezione delle variabili d'interesse in modo da esplicitare al meglio i nessi esistenti tra le caratteristiche dei pazienti e le sintomatologie post-Covid.

Infine, lo studio si concluderà con l'esposizione dei risultati ottenuti, eventuali limiti e alcuni spunti di riflessione per eventuali analisi future.

Capitolo 1

Descrizione dei dati

Le informazioni relative ai soggetti in studio, rilevate dall'istituto ospedaliero, sono di diversa natura e risultano ripartite nelle seguenti macro-aree:

- Anagrafiche dei pazienti e informazioni personali.
- Farmaci e terapie assunte prima dell'infezione acuta da SARSCOV2.
- Patologie pregresse dei pazienti riscontrate prima dell'infezione da Covid-19.
- Terapie assunte in corso di infezione.
- Sintomi Post-Covid riscontrati dai pazienti nel tempo e oggettivati dai clinici (target in analisi).

Il campione, essendo costituito da soggetti che hanno aderito spontaneamente all'indagine, risulta essere auto selezionato anziché casuale. Potrebbe, quindi, soffrire di un errore sistematico, generato dal processo di reclutamento, denominato "selection bias" che, se non opportunamente controllato, distorcerebbe i risultati dello studio.

Il "selection bias", infatti, noto come errore da selezione, si verifica quando vi è una differenza sistematica tra le caratteristiche dei soggetti selezionati per lo studio e quelle di coloro che non sono stati inclusi, compromettendo così la rappresentatività del campione. Nel corso dell'indagine, la risoluzione di tale problematica è stata raggiunta con l'applicazione di metodologie di regressione apposite, quali i modelli di Self Selection.

Nelle pagine seguenti, con il termine paziente si intenderà un soggetto che ha contratto l'infezione da SARSCOV2 e con paziente asintomatico un individuo che, nonostante la contrazione del virus, non ha sviluppato alcuna sintomatologia Long-Covid.

1.1 Descrizione variabili esplicative

La storia clinica dei pazienti è decritta dalle prime quattro macro-aree sopra citate che riassumono cinquantuno variabili esplicative.

Le informazioni personali comprendono le anagrafiche del paziente: età, sesso, peso, altezza e BMI, indice di massa corporea. Oltre a queste, sono riportate informazioni sull'intensità dell'infezione Covid-19, espressa dall'indice del Word Health Organization (WHO) e sull'eventuale ospedalizzazione del soggetto, rappresentata dalla variabile "Ex Ricovero".

Il secondo macro-gruppo, riguarda le patologie di cui il paziente soffriva prima di sviluppare l'infezione. Alcuni casi hanno presentato comorbidità, ossia l'insorgenza simultanea di più entità patologiche, tra le quali:

- PATOLOGIE POLMONARI: affezioni dei polmoni capaci di compromettere il loro corretto funzionamento.
- PATOLOGIE CARDIACHE: malattie e disturbi che interessano il cuore o alcune sue componenti.
- PATOLOGIE METABOLICHE: patologie che causano un'alterazione del metabolismo dei nutrienti.
- DIABETE: patologia metabolica in cui si verifica un aumento della concentrazione di zuccheri nel sangue.
- PATOLOGIE RENALI: affezioni dei reni capaci di compromettere il loro corretto funzionamento.
- NEOPLASIE: note anche come tumori; indicano patologie in cui si verifica una crescita di cellule neoplastiche nell'organismo.
- PATOLOGIE IMMUNOLOGICHE/AUTOIMMUNI: patologie caratterizzate da un disfunzionamento del sistema immunitario che induce l'organismo ad attaccare i propri tessuti non riconoscendoli come tali.
- EPATOPATIA: affezione del fegato capace di compromettere il suo corretto funzionamento. Si manifesta con steatosi, epatite alcolica o cirrosi epatica.

A seconda della patologia sofferta i pazienti, prima dell'infezione acuta da SARSCOV2, hanno dichiarato di assumere farmaci tra i quali:

- TAO/NAO: insieme di farmaci anticoagulanti orali in grado di ridurre il rischio della formazione di coaguli (trombi).

- ANTIAGGREGANTI (cardioaspirina): ulteriore tipologia di farmaci anticoagulanti orali.
- IPOLIPEMIZZANTE/STATINA: farmaci in grado di contrastare e ridurre il colesterolo al fine di diminuire il rischio di malattie coronariche.
- IPOGLICEMIZZANTI: farmaci che permettono la diminuzione della glicemia, concentrazione di glucosio nel sangue, consentendo il corretto funzionamento dell'ormone insulina. Comprendono antidiabetici orali e insulina.
- CALCIO ANTAGONISTI: classe di farmaci antipertensivi che agiscono bloccando i canali di calcio presenti sulla parete dei vasi arteriosi e del miocardio, riducendo la pressione arteriosa.
- BETA-BLOCCANTI: classe di farmaci antipertensivi.
- ACE INIBITORI/SARTANI: farmaci antipertensivi che agiscono sull'ormone dell'angiotensina II, provocando un conseguente effetto ipotensivo.
- DIURETICI: farmaci antipertensivi che agiscono sul bilancio sodio–acqua, aumentando il volume della diuresi e riducendo così la volemia, quantità complessiva di sangue presente all'interno del sistema circolatorio.
- ANTIARITMICI: farmaci in grado di prevenire o interrompere le aritmie, alterazioni della normale regolarità di contrazione del cuore.
- STEROIDI IMMUNOSOPPRESSORI: farmaci utilizzati per ridurre l'attività del sistema immunitario in particolari condizioni come trapianti d'organo o in caso di malattie autoimmuni.
- STEROIDI INALATORI: steroidi immunosoppressori cortisonici spray utilizzati in caso di reazioni allergiche come asma, orticaria o shock anafilattici oppure per malattie da broncopneumopatia cronico ostruttiva (BPCO).
- ANTIRETROVIRALI: farmaci utilizzati per il trattamento di infezioni da virus HIV.
- DMARDs: farmaci che rallentano la progressione dell'artrite reumatoide, patologia autoimmune infiammatoria cronica che attacca i tessuti articolari.
- ANTITUMORALI: farmaci antineoplastici utilizzati per il trattamento dei tumori. Ne esistono di diversi tipi, classificabili in base al loro meccanismo di azione e alla loro struttura chimica.
- NEURO PSICOATTIVI: farmaci che si dividono in antiepilettici e benzodiazepine. I primi usati per il trattamento dell'epilessia, mentre i secondi

sono una classe di farmaci ansiolitici con effetti ipnotici, anticonvulsivanti e anestetici.

Questi farmaci sono spesso utilizzati in modo combinato e, poiché potrebbero risultare fortemente associati tra loro e/o con la patologia per la quale sono stati prescritti, si riterrà opportuno, prima di applicare i metodi di regressione, condurre un'analisi della collinearità in modo da depurare i risultati da possibili distorsioni.

L'ultimo macrogruppo di variabili esplicative descrive le terapie prescritte e somministrate dai medici per contrastare l'infezione acuta da SARSCOV2, quali:

- OSSIGENOTERAPIA: rappresenta la somministrazione di ossigeno al paziente. La somministrazione può avvenire in diversi modi e con diverse intensità.
- STERODI/CORTICOSTEROIDI: terapia in cui, attraverso l'assunzione di medicinali di sintesi, si limita l'azione degli ormoni naturali. Hanno proprietà antinfiammatorie e regolano l'attività di metabolismo e sistema immunitario.
- REMDESEVIR: Terapia nucleotidica che consiste nel bloccare l'enzima RNA polimerasi e l'eccessiva produzione di ATP nei soggetti affetti da COVID-19.
- IDROSSICLOROCHINA (HCQ): terapia indicata per curare le malattie reumatiche e la malaria ma adattata a curare il COVID-19, infatti essa offre un'attività antinfiammatoria e immunomodulatoria.
- INIBITORI DI PROTEASI HIV: terapia che blocca la capacità degli enzimi di metabolizzare le proteine, mantenendole in uno stato inattivo. In questo modo viene limitata la capacità riproduttiva delle cellule COVID-19.
- TERAPIA ANTIBIOTICA: terapia utilizzata per la cura del COVID-19. Presenta effetti benefici per malattie polmonari infiammatorie e ha un'ottima capacità di inibire la replicazione dei batteri patogeni.
- EPARINA: terapia anticoagulante efficace per contrastare le infezioni respiratorie acute dovute dal COVID-19.

1.2 Descrizione variabili dipendenti

Le variabili risposta in studio sono identificate dai sintomi post-Covid riferiti dai pazienti nel tempo e oggettivati dai clinici. I soggetti, infatti, attraverso un colloquio o un questionario scritto, comunicavano il proprio stato di salute permettendone il monitoraggio.

Nel database in esame sono raccolte le informazioni riguardanti la presenza o meno dei sintomi e, nel primo caso, viene specificata la durata con annessa data di inizio ed eventuale data di fine.

I sintomi post-Covid rilevati sono:

- DISPNEA: difficoltà respiratorie.
- TELOGEN EFFLUVIUM: perdita temporanea di capelli.
- ASTENIA: stanchezza.
- MIALGIE: dolori muscolari.
- PALPITAZIONI: anomalie del battito cardiaco.
- MANCANZA OLFATTO
- MANCANZA GUSTO
- AMNESIA: perdite di memoria.
- CEFALEA: dolori alla testa/emicranie.
- ANSIA E PANICO
- INSONNIA: difficoltà o anormale brevità del sonno.
- ALTRO: qualsiasi altro sintomo riscontrato non indicato precedentemente.

La fase di analisi si aprirà, nel successivo capitolo, con la pulizia e la selezione delle informazioni presenti in studio.

Capitolo 2

Pre-Processing variabili dipendenti, esplicative e Componenti Principali

In seguito ad un'analisi esplorativa dei dati, si è osservato come la maggior parte dei soggetti avessero riscontrato molteplici sintomi post-Covid e, per tale ragione, si è deciso di accorparli in macro-categorie. Come metodologia, si è scelto di adottare un approccio data-driven di riduzione della dimensionalità applicato alle durate dei sintomi.
Lavorare con le durate, ha permesso di affrontare l'analisi tenendo in considerazione sia la presenza o assenza della sintomatologia, sia l'importanza e il contributo di ciascun sintomo sul fenomeno del Long Covid.
Prima della creazione dei "macro-sintomi" è stato necessario effettuare una pulizia puntuale delle variabili risposta.

2.1 Pre-Processing variabili dipendenti

Il processo di pulizia delle variabili risposta si è aperto con l'eliminazione della covariata "Altro" e delle sue associate, ovvero durata, data d'inizio e di fine sintomo. Poiché questa, infatti, racchiudeva una serie di sintomatologie specifiche per ogni paziente, non attribuibili con certezza al Long Covid, è stata esclusa dallo studio.
Nei pazienti liberi da sintomi post-Covid, la durata della malattia è risultata essere assente anziché nulla, per cui si è proceduto ad assegnarle valore zero. Questa conversione ha permesso, quindi, di considerare l'intero spettro dei pazienti in studio che, altrimenti, sarebbe risultato incompleto.
Verificando la correttezza della codifica delle durate, tramite il confronto tra la data di inizio e di fine sintomo, si è notato che, per alcuni pazienti, non è stata riportata la data di fine follow-up. Questo ha reso ignota la durata della sintomatologia e, pertanto, in assenza della variabile risposta, i soggetti sono stati esclusi dall'analisi. Al contrario, le durate che presentavano delle incongruenze, sono state corrette.
In seguito alle modifiche apportate, le variabili dipendenti in analisi sono risultate le seguenti: Durata Dispnea, Durata Telogen Effluvium, Durata Astenia, Durata Mialgie,

Durata Palpitazioni, Durata Mancanza Olfatto, Durata Mancanza Gusto, Durata Amnesia, Durata Cefalea, Durata Ansia e Panico e Durata Insonnia.

2.2 Costruzione componenti principali

Per la creazione dei macro-sintomi, è stato adottato un approccio data-driven di riduzione della dimensionalità, in particolare il metodo delle componenti principali.
L'obiettivo è consistito nell'individuare delle aree sintomatologiche comuni del Long Covid, in modo da riassumere i sintomi e studiarne i determinanti.
Queste aree, costituite da variabili fortemente associate tra loro, rappresentano gli indicatori sintetici denominati "macro-sintomi".
Tale approccio ha permesso di ricavare una nuova rappresentazione dei pazienti affetti da disturbi post-Covid passando da uno spazio p-dimensionale, rappresentato dalle durate delle sintomatologie, ad uno spazio inferiore q-dimensionale, rappresentato dalle aree sintomatologiche. La costruzione delle p componenti principali, ottenute dalle combinazioni lineari delle undici variabili di partenza, ha permesso di tener conto del variare complessivo di tutte le covariate, facendo emergere i nessi esistenti tra sintomi e individui.
Per costruzione, l'informazione riprodotta dalle componenti principali, a differenza delle variabili originali, non è equidistribuita. Poiché, infatti, le prime trasformate permettono di riprodurre la quota maggiore del contenuto informativo totale, estraendo le prime q componenti, si spiega gran parte della variabilità degli undici sintomi.
La costruzione delle componenti principali è realizzabile in due modalità: tramite la matrice di varianze e covarianze (Σ), se le variabili originarie sono espresse con la stessa unità di misura e hanno lo stesso ordine di grandezza o in alternativa tramite la matrice di correlazione (R), applicabile in caso di problemi di incommensurabilità. Si è deciso pertanto di svolgere le analisi con la matrice Σ, al fine di esaltare la magnitudine, l'importanza e i contributi di ogni sintomo nello studio del long-Covid. L'applicazione di questo approccio è stata possibile in quanto le dipendenti in studio presentano le medesime unità di misura e ordine di grandezza.

Tabella 1: Percentuale di varianza spiegata dalle prime tre componenti estratte

	1st Principal Component	**2nd Principal Component**	**3rd Principal Component**
Standard Deviation	233.906	106.239	946.530
Proportion of Variance	0.4703	0.0970	0.0770
Cumulative Proportion	0.4703	0.5673	0.6444

Per scegliere il numero ottimale di componenti principali da mantenere in analisi, sono stati utilizzati due criteri: quello della percentuale di varianza cumulativa spiegata e quello dello "scree-plot".

Combinando queste due regole d'arresto si è deciso di estrarre le prime tre trasformate lineari, incorrelate tra loro, in grado di spiegare complessivamente il 64% della variabilità totale.

Per identificare i macro-sintomi, e quindi ottenere una struttura a blocchi, dove ad ogni variabile sia associata una e una sola trasformata, è stato necessario ruotare le componenti.

Sono stati adottati diversi metodi di rotazione, quelli ortogonali "Varimax " e "Quartimax", e uno obliquo, "Oblimin", che in termini interpretativi hanno portato ai medesimi risultati. Successivamente nella trattazione sono stati, quindi, riportati solamente le evidenze ottenute dal metodo "Varimax", in quanto, rispetto alle altre rotazioni, riportava una struttura a blocchi maggiormente definita.

Attraverso la costruzione della matrice di correlazione tra le trasformate ruotate e le variabili dipendenti in esame, è stato possibile valutare l'associazione tra ogni componente e i singoli disturbi al fine di identificare e denominare i macro-sintomi.

Tabella 2: Correlazioni tra variabili e componenti ruotate con metodo Varimax

$\rho(y_i, x_j)$	1st Principal Component	2nd Principal Component	3rd Principal Component
Dispnea	0.7110	-0.1768	-0.5948
Telogen Effluvium	0.4903	-0.2216	-0.5412
Astenia	0.9064	-0.2832	-0.6583
Mialgie	0.8915303	-0.323857	-0.5999
Palpitation	0.599761	-0.2110603	-0.6683
Mancanza Olfatto	0.3058676	-0.945393	-0.2905
Mancanza Gusto	0.2741076	-0.9163021	-0.3450
Amnesia	0.5074905	-0.2177472	-0.8480
Cefalea	0.447186	-0.2311716	-0.6794
Ansia e Panico	0.4785425	-0.2071413	-0.7116
Insonnia	0.53153	-0.2799955	-0.6009327

Osservando la matrice di correlazione tra le componenti ruotate e i singoli sintomi si può notare la presenza di tre gruppi distinti di variabili.

E' stato possibile, quindi, denominare i tre macro-sintomi, identificati dall'associazione tra le variabili e le tre componenti estratte:

- il Sintomo Fisico composto da Dispnea, Astenia e Mialgia
- il Sintomo Sensoriale composto da Mancanza dell'Olfatto e del Gusto
- il Sintomo Psicologico composto da Amnesia, Cefalea, Telogen_Effluvium, Palpitazioni, Insonnia e Ansia e Panico.

La costruzione dei tre macro-sintomi è stata, inoltre, supportata dall'interpretazione di un indicatore statistico di consistenza interna, denominato "Alpha di Crombach". Tale indice è in grado di valutare il grado di coerenza delle variabili che caratterizzano ogni componente. Il suo utilizzo, infatti, ha confermato i risultati della matrice di correlazione riportati nella Tabella 2.

Il metodo delle componenti principali evidenzia una problematica legata al contributo, seppur minimo, delle variabili meno associate nella spiegazione del macro-sintomo. Tale limite impedisce che le tre componenti descrivano soltanto le variabili che le caratterizzano.

Per superare questa problematica, pertanto, è stata eseguita un'ulteriore analisi delle componenti principali. Questa è stata applicata separatamente su ogni macro-sintomo

individuato in precedenza, comprendendo, però, per ciascuna trasformata, solo i sintomi maggiormente correlati alle componenti iniziali (blocchi di variabili in Tabella 2).

Tabella 3: Varianza spiegata dalla prima componente per ogni macro-sintomo

	Componente Fisica	Componente Sensoriale	Componente Psicologica
Cumulative Proportion	0.7467	0.8757	0.4993

In tabella 3 viene riportata la quota di varianza totale spiegata dalla prima componente principale estratta da ognuna delle tre analisi distinte.

La prima, relativa al Macro-Sintomo Psicologico, spiega il 50% della variabilità totale delle sintomatologie: Amnesia, Cefalea, Telogen_Effluvium, Palpitatazioni, Insonnia e Ansia e Panico. Per quanto riguarda il macro-sintomo Sensoriale, la prima componente spiega l'88% della variabilità di Mancanza di Gusto e Mancanza di Olfatto. Infine, quella Fisica spiega il 75% della variabilità dei sintomi Dispnea, Astenia e Mialgia. Complessivamente le prime componenti relative ad ogni macro-sintomo spiegano il 64,07% del contenuto informativo totale.

Una volta ottenuti i punteggi delle componenti su tutti gli individui, poiché essi presentavano valori negativi, è stato necessario traslarli positivamente in modo da renderli logicamente interpretabili in termini di durate. In questo modo, i punteggi delle componenti si sono distribuiti in un range tra valore zero, indicante i soggetti senza sintomi Long Covid, e valori positivi, indicanti il perdurare della sintomatologia.

Tabella 4: Correlazione tra variabili e la prima componente relativa ad ogni macro-sintomo

	PSICOLOGIA	SENSORIALE	DOLORI_FISICI
Duration_Dispnea	0.5553055	0.2152559	0.7399617
Duration_Telogen_effluvium	0.5878094	0.2062052	0.4522889
Duration_Astenia	0.6497824	0.2906475	0.9184391
Duration_Mialgie	0.6261234	0.2974095	0.8778854
Duration_Palpitation	0.7178715	0.225812	0.55378
Duration_Mancanza_olfatto	0.2866377	0.9461	0.3093979
Duration_Mancanza_gusto	0.3271989	0.9237465	0.2884172
Duration_Amnesia	0.7915428	0.2488035	0.5568533
Duration_cefalea	0.6733272	0.2299677	0.4581546
Duration_Ansia_panico	0.7224656	0.2146631	0.475581
Duration_Insonnia	0.6565268	0.2443891	0.490726

Osservando la matrice di correlazione tra la prima componente relativa ad ogni macro-sintomo e tutte le undici variabili prese in esame (Tabella 4), si può notare come ogni variabile risulti correlata maggiormente al macro-sintomo di appartenenza.
L'esecuzione di tre analisi distinte delle componenti principali ha comportato, però, la caduta dell'assunto di incorrelazione tra le trasformate lineari.

Tabella 5. Correlazioni tra componenti estratte

Correlazioni	Componente Fisica	Componente Sensoriale	Componente Psicologica
Componente Fisica	1	0.32	0.717
Componente Sensoriale		1	0.326
Componente Psicologica			1

Nella Tabella 5, infatti, è possibile osservare tra la componente psicologica-sensoriale e quella sensoriale-fisica delle basse correlazioni rispettivamente pari a 0,32 e 0,326. Al contrario, si evidenzia un'associazione elevata tra la componente psicologica e quella fisica pari a 0,7.
Il valore di correlazione, sopra riportato, sembrerebbe indicare la presenza di un'associazione tra le variabili che compongono la componente fisica e quelle relative alla componente psicologica, suggerendo l'esistenza di una sola componente di natura psico-fisica.
L'approfondimento a supporto di tale ipotesi, attraverso la costruzione del grafico del cerchio delle correlazioni tra le componenti e le singole variabili e quello tra le componenti e i singoli individui hanno permesso di osservare quanto le variabili siano legate alle componenti, quanto siano legate tra di loro e come i soggetti si distribuiscano nello spazio delle trasformate. Tale indagine ha introdotto nuovi quesiti allo studio:

- Esiste un legame tra il sintomo psicologico e quello fisico?
- I soggetti che hanno avuto almeno un sintomo di natura psicologica hanno sviluppato sintomi fisici?
- Sarebbe corretto condurre l'analisi utilizzando una sola componente psico-fisica?

Come scelta metodologica si è deciso, dopo un confronto col personale medico coinvolto, di analizzare i tre macro-sintomi distintamente, ponendo l'attenzione sui quesiti introdotti.

2.3 Pre-Processing variabili esplicative

Dopo aver effettuato il pre-processing e la costruzione delle componenti principali sulle variabili dipendenti, è stato necessario eseguire il processo di pulizia anche sulle variabili esplicative.

La prima operazione è stata la ricodifica della covariata "Colore", indicante il grado di intensità dell'indice Word Health Organization (WHO), convertendo ogni gradazione di colore in un valore numerico da 0 a 3, per ordine crescente di gravità dell'infezione.

Andando ad analizzare la percentuale di valori mancanti su ogni covariata, la colonna "Kg persi durante il ricovero" è stata rimossa in quanto tale variabile presentava una percentuale di missing data superiore al 50%.

Concentrandoci sulle covariate relative alle informazioni personali del paziente, si è deciso di controllare e ricalcolare i valori dell'indice di massa corporea (BMI), sfruttando le informazioni sul peso e l'altezza. Successivamente le covariate "Peso" e "Altezza" sono state eliminate in quanto collineari con la variabile "BMI".

Passando alle variabili riferite ai farmaci, assunti nella fase pre-Covid, e alle terapie, somministrate durante la fase acuta del virus, sono state eliminate le covariate "Nessun Farmaco" e "Terapia durante l'infezione". Queste, infatti, sono risultate ridondanti in quanto esprimibili come combinazioni lineari delle variabili appartenenti ai due rispettivi macro-gruppi. Allo stesso modo la covariata "Almeno un antipertensivo" è apparsa collineare con tutti gli altri farmaci antipertensivi presenti in analisi, in quanto, quasi tutto lo spettro dei medicinali cardiaci, è spiegato dalle variabili appartenenti a tale gruppo. Per questo motivo è stata introdotta una nuova covariata "Altri Antipertensivi" ottenuta dalla differenza fra la variabile "Almeno un antipertensivo" e la somma delle modalità delle altre variabili dello stesso gruppo farmacologico.

Successivamente, si è svolta un'operazione di ricodifica delle variabili "O2 terapia", "O2 terapia (<=2 L/min)", "O2 terapia (6 –2 L/min)" e "O2 terapia(>6 L/min)".

La variabile "O2 terapia", che inizialmente si distribuiva in un range da 1 a 6, dove il valore unitario esprimeva il mancato utilizzo di strumenti per la somministrazione di ossigeno, è stata traslata in un range da 0 a 5. Invece, le variabili indicanti il grado di intensità di somministrazione dell'ossigeno, sono state accorpate in un'unica variabile denominata "O2 intensità" distribuita in un range tra 0 e 3, dove il valore nullo indica la mancata assunzione della terapia. Effettuando un controllo incrociato tra la variabile "O2

intensità" e "O2 terapia", sono stati considerati come missing values i valori sulle due covariate per quei pazienti trattati con somministrazione dell'ossigeno ma con un livello d'intensità nullo e viceversa.

In seguito al controllo delle modalità di tutte le variabili in esame, sono stati corretti i livelli delle covariate "Polmonari", "Ex-ricovero", "Metaboliche", "Patologie Immunologiche", "Remdesevir" e "Cardiache", in quanto presentavano degli errori di imputazione.

La fase di pulizia è terminata con l'analisi della zero-variance che ha portato l'eliminazione della variabile "Salrilumab, Baricitinib" poiché degenere.

Capitolo 3

Analisi descrittive

Le analisi descrittive, eseguite a scopo esplorativo e conoscitivo del campione, sono state utilizzate come supporto nello studio dei potenziali nessi esistenti tra macro-sintomi e possibili determinanti. Il campione descritto è costituito da 1368 soggetti suddivisi per ondate infettive, 575 appartenenti alla prima e 793 alla seconda. Successivamente al processo di pulizia, si sono mantenute in studio 52 variabili, 11 delle quali riportano le sintomatologie post-Covid.

3.1 Analisi delle sintomatologie per ondate

In primo luogo, è stata effettuata un'analisi grafica al fine di studiare la ripartizione dei soggetti nelle diverse aree sintomatologiche individuate.

Grafico1: Partizionamento del campione per sintomatologie prima ondata

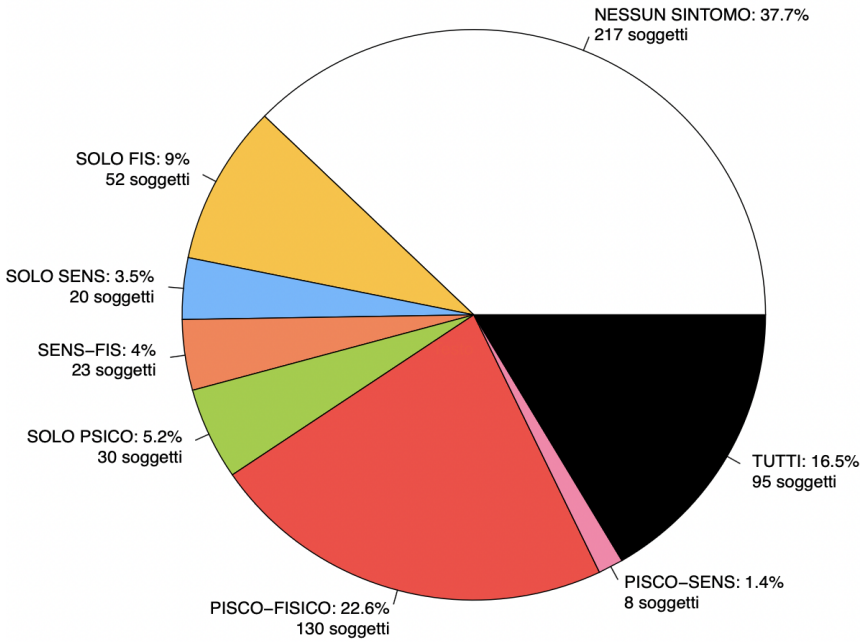
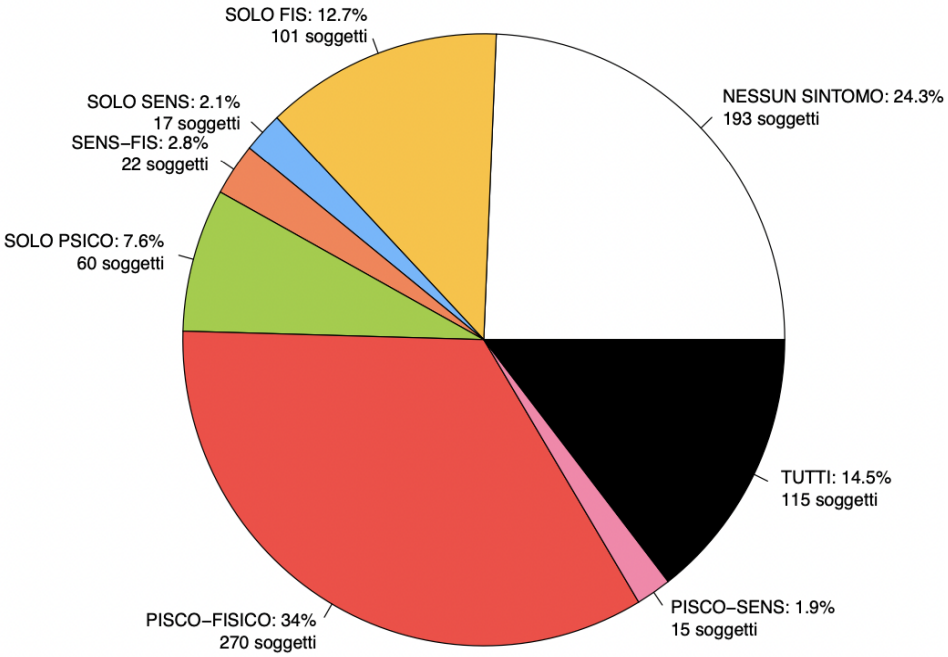


Grafico2: Partizionamento del campione per sintomatologie seconda ondata



Osservando i grafici risulta evidente come il campione in studio sia costituito, in entrambe le ondate, da due sottopopolazioni: coloro che possiedono almeno un sintomo (porzioni colorate) e coloro che non ne hanno mai sviluppato uno (porzione bianca).

Gran parte della popolazione è ripartita in tre macro-gruppi: gli asintomatici, i pazienti che presentano sintomi di natura psico-fisica e coloro che hanno sviluppato tutti i sintomi in esame.

Si osserva come, dalla prima alla seconda ondata, il numero dei soggetti asintomatici sia diminuito del 13% passando dal 37,7% al 24,3% , mentre, la porzione dei soggetti con tutti i sintomi è rimasta pressoché invariata; al contrario, il numero di soggetti con sintomatologie psico-fisiche è aumentato del 11%.

3.2 Analisi profili dei soggetti per ondate

In secondo luogo, si è analizzato il cambiamento dei profili clinici dei pazienti nelle due ondate Covid. Nella tabella sottostante sono riportate le sole variabili che, nel confronto, sono risultate statisticamente significative.

Tabella 6. Analisi profili clinici dei pazienti stratificati per ondate

			ONDATA 1	ONDATA 2	P-VALUE
INFORMAZIONI PERSONALI E DURATE SINTOMI	NUMEROSITA' SOGGETTI	TOTALE	575	793	
	INTENSITÀ	BASSA (0)	139 (24.2%)	209 (26.4%)	0.004
		MEDIA (1)	266 (46.3%)	295 (37.2%)	
		MEDIA-ALTA (2)	57 (9.9%)	94 (11.9%)	
		ALTA (3)	108 (18.8%)	192 (24.2%)	
	ETÀ	MEDIA (SD)	52.5 (15.9)	58.1 (15.0)	< 0.001
		MEDIANA (IQR)	54.0 [4.00, 92.00]	58.0 [10.00, 95.00]	
	CONTEGGIO SINTOMI	MEDIA (SD)	2.61 (2.91)	3.02 (2.69)	0.008
		MEDIANA (IQR)	2.00 [0.00, 11.00]	3.00 [0.00, 11.00]	
	BMI	MEDIA (SD)	25.9 (11.9)	26.3 (4.56)	0.4834
		MEDIANA (IQR)	24.8 [15.4, 263]	26.0 [15.2, 42]	
FARMACI	TAO NAO	0	554 (96.3%)	756 (95.3%)	0.028
		1	11 (1.9%)	33 (4.2%)	
	ANTIAGGREGANTE	0	524 (91.1%)	679 (85.6%)	< 0.001
		1	41 (7.1%)	111 (14.0%)	
	IPOLIPEMIZZANTE STATINA	0	506 (88.0%)	662 (83.5%)	0.004
		1	59 (10.3%)	128 (16.1%)	
	IPOGLICEMIZZANTI	0	528 (91.8%)	712 (89.8%)	0.041
		1	37 (6.4%)	76 (9.6%)	
	ALMENO 1 ANTIPERTENSIVO	0	411 (71.5%)	516 (65.1%)	0.007
		1	154 (26.8%)	274 (34.6%)	
	CA ANTAGONISTI	0	524 (91.1%)	702 (88.5%)	0.019
		1	41 (7.1%)	88 (11.1%)	
	B BLOCCANTI	0	502 (87.3%)	662 (83.5%)	0.009
		1	62 (10.8%)	128 (16.1%)	
	ACE INIBITORI SARTANI	0	454 (79.0%)	587 (74.0%)	0.012
		1	111 (19.3%)	203 (25.6%)	
	DIURETICI	0	528 (91.8%)	699 (88.1%)	0.004
		1	37 (6.4%)	91 (11.5%)	
	STEROIDI IMMUNI	0	552 (96.0%)	783 (98.7%)	0.019
		1	13 (2.3%)	6 (0.8%)	
	ANTIRETROVIRALI	0	551 (95.8%)	753 (95.0%)	0.052
		1	14 (2.4%)	36 (4.5%)	
	NEURO PSICOATTIVI	0	524 (91.1%)	701 (88.4%)	0.016
		1	41 (7.1%)	89 (11.2%)	
PATOLOGIE	CARDIACHE	0	378 (65.7%)	479 (60.4%)	0.033
		1	187 (32.5%)	303 (38.2%)	
	METABOLICHE	0	456 (79.3%)	552 (69.6%)	< 0.001
		1	109 (19.0%)	237 (29.9%)	
	DIABETE	0	521 (90.6%)	701 (88.4%)	0.040
		1	44 (7.7%)	89 (11.2%)	
	N COMORBIDITÀ	MEDIA (SD)	2.18 (1.55)	2.88 (2.02)	
		MEDIANA (IQR)	2.00 [1.00, 11.00]	2.00 [1.00, 11.00]	< 0.001

Tabella 7. Analisi terapie somministrate durante la fase infettiva nelle due ondate

			ONDATA 1	ONDATA 2	P-VALUE
		TOTALE	575	793	
TERAPIE	O2 TERAPIA	0	329 (57.2%)	368 (46.4%)	< 0.001
		NASO CANNULA (1)	68 (11.8%)	116 (14.6%)	
		VENTURI (2)	61 (10.6%)	112 (14.1%)	
		RESERVOIR (3)	15 (2.6%)	12 (1.5%)	
		VENTILAZIONE NON INVASIVA(4)	63 (11.0%)	150 (18.9%)	
		IOT (5)	26 (4.5%)	25 (3.2%)	
	O2 INTENSITÀ	0	325 (56.5%)	367 (46.3%)	< 0.001
		<=2L/min	45 (7.8%)	66 (8.3%)	
		6-2 L/min	29 (5.0%)	58 (7.3%)	
		>6 L/min	155 (27.0%)	287 (36.2%)	
	STEROIDE	0	497 (86.4%)	271 (34.2%)	< 0.001
		1	62 (10.8%)	519 (65.4%)	
	REMDESEVIR	0	518 (90.1%)	663 (83.6%)	< 0.001
		1	39 (6.8%)	127 (16.0%)	
	HCQ	0	340 (59.1%)	785 (99.0%)	< 0.001
		1	217 (37.7%)	5 (0.6%)	
	INIBITORI PROTEASI HIV	0	417 (72.5%)	789 (99.5%)	< 0.001
		LOPINAVIR	134 (23.3%)	0 (0%)	
		LOPINAVIR E DARUNAVIR	4 (0.7%)	0 (0%)	
		DARUNAVIR	2 (0.3%)	0 (0%)	
	TOCILIZUMAB SARILUMAB BARICITINIB	0	524 (91.1%)	788 (99.4%)	< 0.001
		1	32 (5.6%)	2 (0.3%)	
	EPARINA	0	411 (71.5%)	287 (36.2%)	< 0.001
		1	146 (25.4%)	502 (63.3%)	

Tabella 8. Analisi sintomatologie nelle due ondate infettive

		ONDATA 1	ONDATA 2	P-VALUE
	TOTALE	575	793	
DURATA DISPNEA	NUMEROSITA'	172 [29.91 %]	294 [37.07 %]	0.007
	MEDIA (SD)	62.6 (132)	42.8 (72.8)	
	MEDIANA (IQR)	185.50 [60.50, 343.75]	96.50 [62.00, 158.75]	0.1203
DURATA TELOGEN EFFLUVIUM	NUMEROSITA'	129 [22.43 %]	215 [27.11 %]	0.057
	MEDIA (SD)	46.6 (113)	33.4 (65.9)	
	MEDIANA (IQR)	153.00 [91.00, 330.00]	106.00 [69.00, 160.50]	0.2199
DURATA ASTENIA	NUMEROSITA'	263 [45.74 %]	410 [51.70 %]	0.034
	MEDIA (SD)	115 (172)	69.8 (88.6)	
	MEDIANA (IQR)	218.00 [92.00, 387.50]	121.00 [73.50, 192.75]	0.4746
DURATA MIALGIE	NUMEROSITA'	179 [31.13 %]	287 [36.19 %]	0.059
	MEDIA (SD)	89.3 (166)	48.9 (82.2)	
	MEDIANA (IQR)	328.00 [112.00, 438.50]	119.00 [69.50, 192.50]	0.9373
DURATA PALPITAZIONI	NUMEROSITA'	119 [20.70 %]	172 [21.69 %]	0.707
	MEDIA (SD)	54.9 (132)	27.5 (63.3)	
	MEDIANA (IQR)	245.00 [112.00, 390.00]	107.00 [71.00, 167.75]	0.6238
DURATA MANCANZA OLFATTO	NUMEROSITA'	131 [22.78 %]	151 [19.04 %]	0.105
	MEDIA (SD)	46.8 (120)	22.1 (59.4)	
	MEDIANA (IQR)	114.00 [61.00, 370.00]	93.00 [43.50, 182.00]	0.03098
DURATA MANCANZA GUSTO	NUMEROSITA'	122 [21.22 %]	146 [18.41 %]	0.222
	MEDIA (SD)	38.4 (108)	21.5 (61.1)	
	MEDIANA (IQR)	98.50 [39.50, 361.25]	92.00 [43.25, 175.75]	0.1251
DURATA AMNESIA	NUMEROSITA'	124 [21.57 %]	238 [30.01 %]	< 0.001
	MEDIA (SD)	75.4 (160)	42.5 (77.5)	
	MEDIANA (IQR)	367.50 [244.75, 483.50]	127.00 [83.00, 193.75]	0.128
DURATA CEFALEA	NUMEROSITA'	87 [15.13 %]	124 [15.64 %]	0.857
	MEDIA (SD)	39.7 (116)	20.4 (59.8)	
	MEDIANA (IQR)	264.00 [92.00, 394.00]	115.50 [62.00, 186.00]	0.8127
DURATA ANSIA PANICO	NUMEROSITA'	102 [17.74 %]	160 [20.18 %]	0.289
	MEDIA (SD)	48.0 (126)	27.0 (64.4)	
	MEDIANA (IQR)	275.00 [100.25, 426.50]	125.00 [74.00, 184.00]	0.6806
DURATA INSONNIA	NUMEROSITA'	71 [12.35 %]	198 [24.97 %]	< 0.001
	MEDIA (SD)	39.8 (125)	31.5 (67.5)	
	MEDIANA (IQR)	350.00 [145.50, 506.50]	106.50 [65.00, 165.75]	< 0.001

Nelle precedenti analisi descrittive, confrontando le due ondate, si evince come i profili clinici dei soggetti in analisi siano cambiati significativamente a seconda del periodo di infezione del virus.

Nella prima ondata di infezione da SARSCOV2, l'età media dei soggetti che costituiscono il campione risulta essere statisticamente inferiore rispetto a quella dei soggetti della seconda ondata, attestandosi nel primo caso su un'età media pari a 53 anni e nel secondo caso su un'età pari a 58 anni.

Per quanto riguarda le terapie, si evince come il virus sia stato trattato diversamente nelle due ondate. Osservando le differenze in termini di trattamenti e durate, le terapie adottate nella seconda fase sono apparse più efficaci nonostante la maggior fragilità dei soggetti, evidenziata dall'alto numero di farmaci assunti per la cura di patologie cardiache, diabetiche, metaboliche e dall'aumento del numero medio delle comorbidità..

In questa fase, si è verificato un aumento dell'intensità del virus e del numero delle sintomatologie post-Covid, sebbene queste presentino durate ridotte in media di circa la metà. Tale evidenza porta a domandarsi se questi risultati siano dovuti alla maggior cagionevolezza dei pazienti in studio o se le due ondate siano state caratterizzate dalla presenza di un virus di diversa natura.

L'ausilio di strumenti analitici inferenziali permetterà nel seguito della trattazione, di valutare se la diversità del Long-Covid, nelle due ondate, sia stata determinata dalla natura dell'infezione o da caratteristiche proprie dei soggetti.

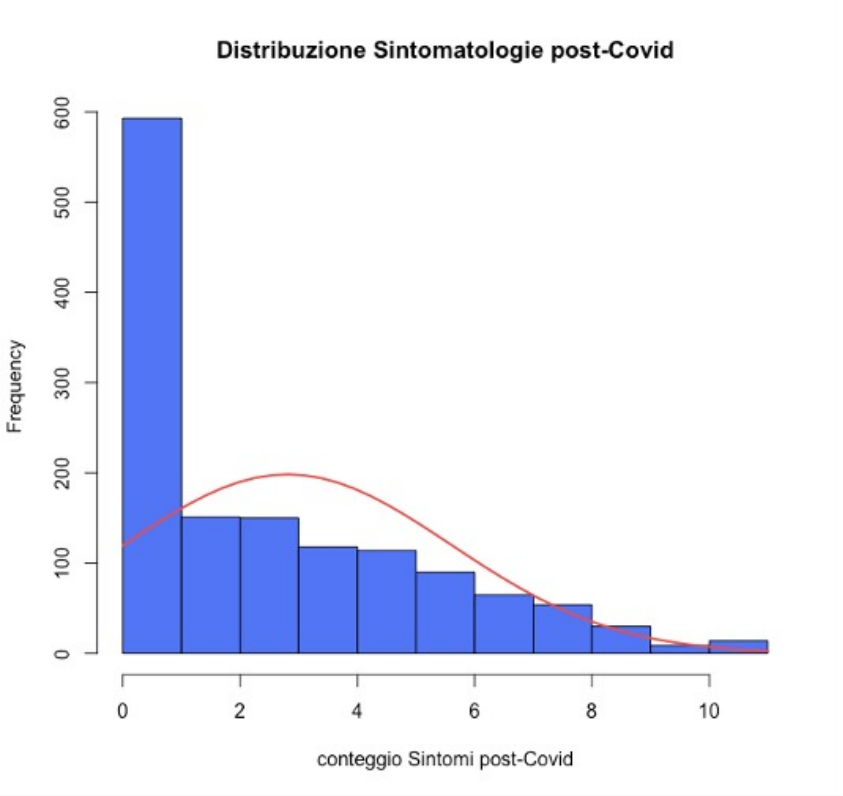
Capitolo 4

Modelli di Self-Selection

4.1 Modelli di Self Selection e Censored Sample

Lo studio dei determinanti che agiscono sul fenomeno del Long-Covid è stato effettuato tramite l'applicazione del modello di regressione statistico Heckit, appartenente alla famiglia dei modelli di Self Selection. Questa metodologia è utilizzata per risolvere problemi legati alla modalità di estrazione del campione in studio, non avvenuta in modo casuale ma tramite auto-selezione campionaria, generando un errore sistematico denominato "bias da selezione".

Grafico 3: Distribuzione dei soggetti appartenenti alla macro-sintomatologia Fisica



Andando ad analizzare la conformazione del campione in studio, rappresentata nel Grafico 3, si nota che siamo in presenza di un campione troncato.

Per campione troncato si intende un insieme di osservazioni la cui variabile dipendente assume due valori differenti: positivo se il paziente presenta il sintomo, nullo se il soggetto è asintomatico. Pertanto, risulta necessario applicare modelli specifici, come il modello Heckit in grado di trattare questa tipologia di campione e di generare stime non distorte e consistenti non ottenibili tramite l'utilizzo di modelli tradizionali.
L'applicazione del modello Heckit consiste nell'esecuzione di una procedura "Two Step" in cui si stimano congiuntamente due equazioni:

1. la select equation, definita dal modello Probit, che gestisce il meccanismo di selezione in grado di studiare l'effetto delle covariate su chi ha avuto o non ha avuto sintomi da Long-Covid .

$$z_i^* = w_i\gamma + u_i$$
$$z_i = \begin{cases} 1 & se\ z_i^* > 0 \\ 0 & se\ z_i^* \leq 0 \end{cases}$$

2. la main equation, definita dal modello di regressione lineare OLS, in grado di studiare l'effetto delle covariate sull'intensità del fenomeno del Long-Covid nei soggetti che hanno sviluppato almeno un sintomo.

$$y_i = \begin{cases} x_i\,\beta + e_i & se\ z_i^* > 0 \\ - & se\ z_i^* \leq 0 \end{cases}$$

Generalmente le due equazioni sopra riportate sono stimate sotto i seguenti assunti preliminari:

$$u_i \sim N(0,1)$$
$$e_i \sim N(0,\sigma^2)$$
$$corr(u_i, e_i) = \rho$$

L'ultima ipotesi di correlazione tra gli errori delle due equazioni permette di valutare se il campione in studio è affetto da bias da selezione, distorsione che determina l'ottenimento di stime inconsistenti.

4.2 Confronto tra modelli Self Selection e modelli tradizionali

La differenza tra i modelli di Self Selection e i modelli tradizionali è possibile osservarla andando ad analizzare il valore atteso condizionato della variabile dipendente:

$$E[y_i|\ y_i \text{ è } osservato] = \ x_i\,\beta + \rho\sigma_e\ \frac{\varphi(w_i\gamma)}{\Phi(w_i\gamma)}$$

Dove:

- ρ rappresenta la correlazione tra u_i e e_i ed il suo valore identifica la presenza/assenza di selection bias
- σ_e la deviazione standard di e_i
- $\frac{\varphi(w_i\gamma)}{\Phi(w_i\gamma)}$ denominato Inverse Mills Ratio, definito dal rapporto tra la funzione di densità e la distribuzione cumulata del modello Probit.

Risulta evidente come l'applicazione del modello Heckit , nel caso di incorrelazione ($\rho=0$) tra i due errori, permette di ottenere dei risultati che si equivalgono a quelli stimati dal modello di regressione lineare OLS.

Al contrario, in presenza di associazione (ρ diverso da zero), e quindi di selection bias, non risulta possibile applicare direttamente il modello lineare in quanto è necessario considerare un termine di correzione aggiuntivo ($\rho\sigma_e\ \frac{\varphi(w_i\gamma)}{\Phi(w_i\gamma)}$) che permette di ottenere delle stime consistenti che tengano conto di tale distorsione.

Nel modello Hekit la selection equation gioca, quindi, un ruolo fondamentale in quanto permette di stimare, tramite il metodo della massima verosimiglianza, l'Inverse Mills Ratio, quantità inserita come variabile esplicativa nell'outcome equation al fine di ottenere stime robuste.

Capitolo 5

Analisi Preliminari e Model Selection

Poiché lo studio ha come duplice obiettivo la valutazione dell'effetto delle covariate sulla presenza/assenza dei sintomi post-Covid e, se presenti, sulla loro intensità, è stato necessario effettuare delle analisi preliminari specifiche in funzione dello scopo considerato.
La trattazione, da questo punto in poi, è stata eseguita in maniera distinta sui macro-sintomi Psicologico e Fisico per le singole ondate. Il macro-sintomo Sensoriale è stato escluso dall'analisi in quanto clinicamente poco rilevante.

5.1 Analisi Preliminari: studio della Collinearità, Separation/Quasi-Separation e Near-Zero-Variance

La fase di analisi si è aperta con lo studio della collinearità, eseguito sia sulle esplicative qualitative che su quelle quantitative. Per quanto riguarda il primo gruppo di variabili, la collinearità è stata verificata mediante la metrica del Chi-Quadro normalizzato. Tale indicatore ha segnalato la presenza di una forte associazione tra le variabili "O2 intensità" e "O2 terapia", determinando l'esclusione di quest'ultima in quanto collineare. Gli indici TOL e VIF hanno permesso di eseguire questa verifica anche per le variabili quantitative, controllo conclusosi con il mantenimento di tutte le covariate in esame.
Per studiare l'effetto dei determinanti sull'intero spettro dei pazienti, sintomatici e non, il controllo della Separation e Quasi-Separation rappresenta un'analisi preventiva fondamentale. Lo studio delle tabelle di contingenza relative alle due ondate per entrambe le macro-sintomatologie ha determinato l'eliminazione delle seguenti covariate:

Tabella 9: Covariate escluse in esame dall'analisi della separation

MACRO SINTOMO PSICOLOGICO		MACRO SINTOMO FISICO	
Prima Ondata	**Seconda Ondata**	**Prima Ondata**	**Seconda Ondata**
TAO NAO Altri Antipertensivi Steroidi Immunologici DMARDs Antitumorali Epatopatia Inibitori HIV	Altri Antipertensivi Steroidi Immunologici DMARDs Antitumorali Epatopatia HCQ Tocilizumab Sarilumab Baricitinib Inibitori HIV	TAO NAO Altri Antipertensivi Antiaritmici Steroidi Immunologici DMARDs Antitumorali Epatopatia Inibitori HIV	Altri Antipertensivi Steroidi Immunologici DMARDs Antitumorali HCQ Tocilizumab Sarilumab Baricitinib Inibitori HIV

Infine, il controllo preliminare della Near-Zero-Variance è risultato essenziale per lo studio dell'intensità dei determinati su coloro che hanno sviluppato almeno una sintomatologia post fase acuta del virus. Questo ha portato all'eliminazione delle seguenti variabili:

Tabella 10: Covariate escluse in esame dall'analisi della Near-Zero-Variance

MACRO SINTOMO PSICOLOGICO		MACRO SINTOMO FISICO	
Prima Ondata	**Seconda Ondata**	**Prima Ondata**	**Seconda Ondata**
TAO NAO Altri Antipertensivi Steroidi Immunologici DMARDs Antitumorali Inibitori HIV	Altri Antipertensivi Steroidi Immunologici DMARDs Antitumorali HCQ Tocilizumab Sarilumab Baricitinib Inibitori HIV	TAO NAO Altri Antipertensivi Steroidi Immunologici DMARDs Antitumorali	Altri Antipertensivi Steroidi Immunologici DMARDs Antitumorali HCQ Tocilizumab Sarilumab Baricitinib Inibitori HIV

5.2 Model Selection

Nella trattazione, l'applicazione dei modelli di regressione di Self Selection è supportata dall'utilizzo di tecniche di Machine Learning per la scelta delle covariate da inserire nelle due equazioni, Probit e OLS, del modello Heckit.

Il processo di Model Selection, eseguito per ogni macro-sintomo e ondata, ha permesso di selezionare le covariate da mantenere in esame in base al loro valore di Importanza, tramite l'applicazione dei seguenti modelli: Alberi Tradizionali, Random Forest e Gradient Boosting. Per Importanza delle covariate si intende la capacità di depurare l'impurità della variabile dipendente passando dal nodo padre ai nodi figli al fine di ottenere dei nodi finali il più possibile omogenei al loro interno.

Essa viene valutata mediante la capacità di massimizzazione dell'Indice di decremento di eterogeneità di Gini, se il target è binario, o della varianza spiegata, in caso di target continuo.

Rispetto ai modelli sopra citati, il metodo Boruta della Random Forest, ha fornito i risultati più robusti, in quanto sfrutta un criterio di Importanza differente basato sul concetto di permutazione. La permutazione è il cambiamento casuale dell'ordinamento delle righe delle colonne del dataset determinando la rottura dell'associazione tra la variabile dipendente e le covariate. Tale strategia misura il grado di importanza di una variabile confrontando il valore medio di Accuracy ottenuto sia sul dataset originale che su quello permutato. Se tale metrica, misurata in entrambe le situazioni, presenta valori simili, allora la covariata non ha alcun potere esplicativo sulla variabile dipendente e l'importanza osservata è imputabile al caso.

Combinando i criteri di selezione delle variabili sopra enunciati, si sono mantenute in analisi le seguenti covariate:

Tabella 11: Covariate selezionate da inserire nel modello Heckit per il macro-sintomo Fisico

COMPONENTE FISICA			
Prima Ondata		**Seconda Ondata**	
PROBIT	**MODELLO OLS**	**PROBIT**	**MODELLO OLS**
Età	Età	Sesso	Intensità
B_Bloccanti	Sesso	Antibiotico	Età
O2intensità	BMI	Polmonari	Sesso
Intensità	Antiaggregante	Patologie_Autoimmuni	BMI
Antiaggregante	B_Bloccanti	N_Comorbidità	N_Comorbidità
Steroidi_inalatori	ACE_inibitori_Sartani	O2intensità	O2intensità
Polmonari	Steroidi_inalatori	BMI	Antibiotico
Metaboliche	Cardiache	Età	Ipolipemizzante_Statina
HCQ	Metaboliche	Steroide	Steroidi_inalatori
BMI	N_Comorbidità	Intensità	Polmonari
N_Comorbidità	O2intensità	Metaboliche	Metaboliche
Sesso	Eparina	Remdesevir	Steroide
	HCQ		Eparina
	Intensità		
	Steroidi_immunologici		

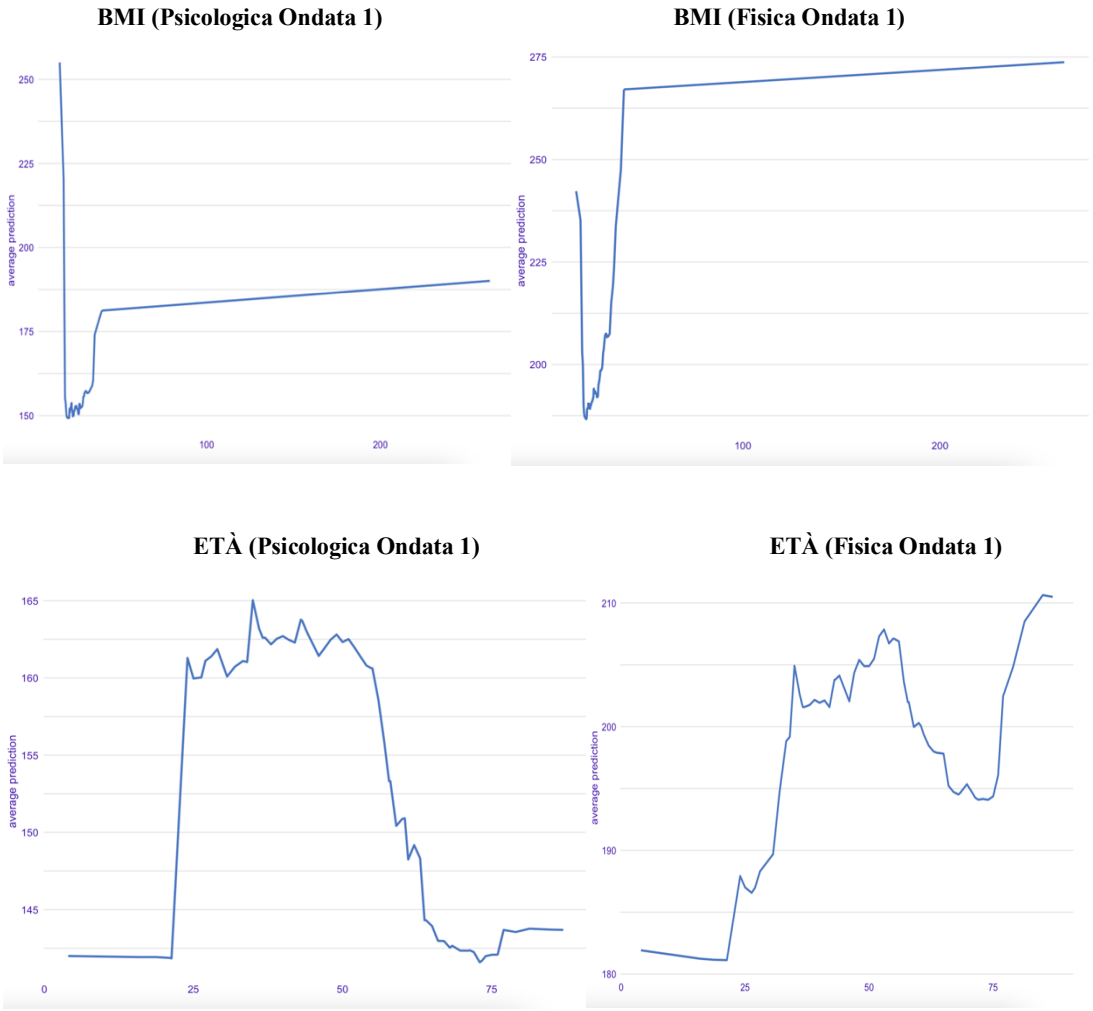
Tabella 12: Covariate selezionate da inserire nel modello Heckit per il macro-sintomo psicologico

COMPONENTE PSICOLOGICA			
Prima Ondata		**Seconda Ondata**	
PROBIT	**MODELLO OLS**	**PROBIT**	**MODELLO OLS**
Sesso	Sesso	Età	Età
Ipolipemizzante_Statina	B_Bloccanti	Sesso	Sesso
Polmonari	Cardiache	Patologie_Immunologiche	Patologie_Immunologiche
N_Comorbidità	N_Comorbidità	Eparina	N_Comorbidità
HCQ	O2intensità	BMI	O2intensità
O2intensità	HCQ	N_Comorbidità	Steroide
BMI	Ipolipemizzante_Statina	Antibiotico	BMI
Età	BMI	Intensità	ACE_Inibitori_Sartani
Intensità	Età	Steroide	Steroidi_inalatori
Metaboliche	Eparina		Polmonari
Antibiotico	Neuro_Psicoattivi		Metaboliche
B_Bloccanti			Antibiotico
Eparina			Eparina
Cardiache			Intensità
ACE_inibitori_Sartani			

Le metodologie sopra enunciate, rispetto ai metodi tradizionali di Model Selection, hanno il vantaggio di considerare e cogliere tutte le possibili forme funzionali che legano le covariate alle variabili dipendenti.

L'analisi dell'Explainability ha quindi permesso di valutare il contributo di ogni variabile esplicativa nella spiegazione del Long-Covid.

Grafico 4: Explainability profilo di alcune variabili quantitative



Nel Grafico 4 sono riportati alcune diagnostiche di applicazione di tale metodologia per macro-sintomi e ondate. Confrontando le due aree sintomatologiche si può, infatti, osservare che il tipo di relazione presente tra le variabili e le durate sintomatologiche è molto simile, evidenza che verrà approfondita nella successiva fase di analisi.

Capitolo 6

Analisi dei risultati

Lo studio dei determinanti del Long-Covid è stato effettuato tramite l'applicazione del modello Heckit alla Macro-Sintomatologia Psicologica e Fisica al fine di valutare eventuali differenze tra le due ondate in esame. All'interno delle equazioni del modello di regressione, oltre alle variabili selezionate dalla Model Selection, sono stati inseriti dei termini d'interazione. Questi sono stati definiti sia da un'analisi effettuata a priori, basata su un attento studio della letteratura scientifica al fine di analizzare l'associazione tra le patologie e i farmaci assunti nella fase Pre-Covid, sia tramite la valutazione dei cambiamenti di significatività andando ad eliminare, all'interno dei modelli, una variabile alla volta.

6.1 Interpretazione dei punteggi dei Macro-Sintomi

L'applicazione dell'analisi delle componenti principali ai sintomi in esame, al fine di costruire delle Macro-Sintomatologie, presenta un limite in termini di interpretabilità. Tale vincolo è determinato dal fatto che le variabili dipendenti, essendo ottenute come combinazione lineare delle variabili che le compongono, non sono esprimibili nell'unità di misura e nell'ordine di grandezza delle covariate originali.

Grafico 5: Scatterplot correlazioni Punteggi del Macro-Sintomo Psicologico e somme delle durate dei singoli sintomi che lo compongono

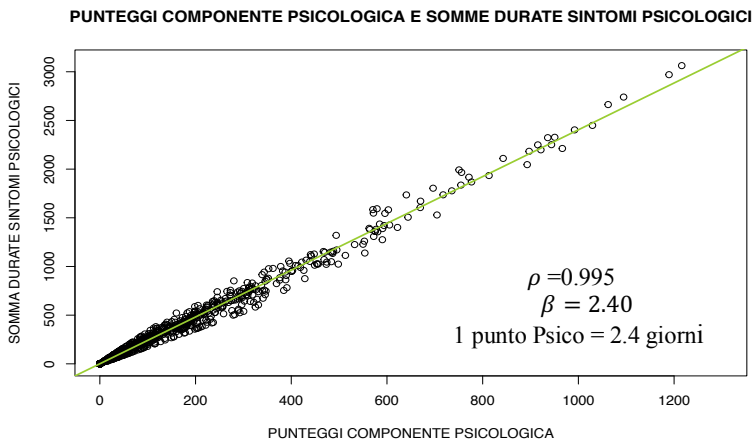
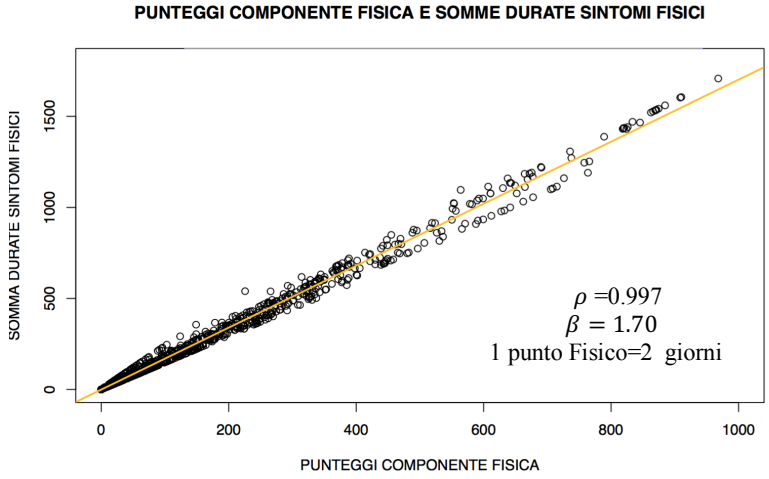


Grafico 6: Scatterplot correlazioni Punteggi del Macro-Sintomo Fisico e somme delle durate dei singoli sintomi che lo compongono



Nello studio tale problematica è stata affrontata tramite l'analisi della correlazione tra i punteggi delle componenti e la somma delle durate dei sintomi appartenenti ad ogni macro-gruppo. La forte correlazione osservata nel Grafico 5 e nel Grafico 6 garantisce l'interpretabilità dei punteggi dei macro-sintomi in termini di durata, unità di misura delle sintomatologie in esame.

Regredendo la somma totale dei sintomi in funzione delle componenti principali, risulta possibile, infatti, stimare il parametro regressivo β, indicatore di conversione in grado di fornire l'equivalente del punteggio della componente in giorni. Questi sono espressi come somma complessiva delle durate sintomatologiche relative ad ogni macro-area.

6.2 Selection e Main equation Macro-Sintomatologia Fisica

Lo studio dell'effetto dei fattori in esame sull'intensità e sullo sviluppo della Macro sintomatologia Fisica è stato effettuato tramite l'applicazione del modello OLS e del modello Logit, metodologia utilizzata, per facilità interpretativa, al posto del modello Probit, precedentemente citato.

I risultati ottenuti dall'analisi sono presentati nella seguente tabella, riportando per le stime dei punteggi ricavati dal modello lineare i rispettivi valori in giorni.

Tabella 13: Output modello Logit e OLS sulla macro-sintomatologia Fisica per ondata

MACRO-SINTOMO FISICO												
MODELLO GLM												
PRIMA ONDATA						SECONDA ONDATA						
	Estimate	Std.	t-value	Pr(>|t|)			Estimate	Std.	T value	Pr(>|z|)	
(Intercept)	0.189	0.512	0.370	0.711		(Intercept)	0.419	0.143	2.94	0.003	**
Età	0.022	0.008	2.761	0.006	**	Sesso: Maschio	-0.786	0.168	-4.68	2.91e-06	***
B_Bloccanti1	0.972	0.449	2.166	0.031	*	Polmonari1	0.643	0.279	2.30	0.021	*
Intensità: Media	-0.847	0.314	-2.695	0.007	**	Patologie_Autoimmuni1	1.517	0.619	2.45	0.014	*
Intensità2:Medio-alta	-0.029	0.446	-0.066	0.947		Steroide1	0.612	0.178	3.443	0.001	***
Intensità: Alta	0.015	0.356	0.042	0.967		Remdesevir1	0.493	0.244	2.017	0.044	*
Antiaggregante1	-1.252	0.433	-2.892	0.004	**						
Steroidi_inalatori1	1.721	0.806	2.136	0.033	*						
Metaboliche1	0.923	0.366	2.519	0.012	*						
HCQ1	0.477	0.265	1.798	0.072	.						
N_Comorbidità	-0.205	0.0998	-2.056	0.039	*						
Sesso: Maschio	-0.510	0.225	-2.265	0.024	*						

MAIN EQUATION													
PRIMA ONDATA							SECONDA ONDATA						
Estimate		Days	Std	t value	Pr(>|t|)		Estimate		Days	Std	t value	Pr(>|t|)	
(Intercept)	220.83	37.41	49.86	4.43	0.001	***	(Intercept)	75.72	128.73	45.44	1.67	0,10	.
SessoMaschio	-85.30	-145.01	30.77	-2.77	0.01	**	Età	0.55	0,.3	0.53	1.03	0.30	
Cardiache1	-37.75	-64.18	43.19	-0.87	0.38		Sesso: Maschio	-42.16	-71.67	14.82	-2.85	0.001	**
N_Comorbidità	37.09	63.05	12.73	2.91	0.001	**	BMI	3.53	6.00	1.17	3.01	0.001	**
Eparina1	141.79	241.04	43.75	3.24	0.004	**	N_Comorbidità	-5.46	-9.29	3.04	-1.79	0.07	.
Cardiache1: Eparina1	-164.23	-279.19	67.40	-2.44	0.02	*	O2intensità: <=2L/min	-27.76	-47.19	20.84	-1.33	0.18	
							O2intensità: 2-6L/min	-35.02	-59.53	20.95	-1.67	0.10	.
							O2intensità: >6L/min	-14.20	-24.14	14.96	-0.95	0.34	
							Antibiotico1	131.57	223.68	44.25	2.97	0.00	**
							Età:Antibiotico1	-2.08	-3.54	0.73	-2.87	0.00	**
DIAGNOSTICHE													
Estimate			Std	t value	Pr(>|t|)		Estimate			Std	t value	Pr(>|t|)	
invMillsRatio	77.48		64.07	1.21	0.23		invMillsRatio	-17.84		40.14	-0.45	0.66	
rho	0.31		NA	NA	NA		rho	-0.16		NA	NA	NA	
Multiple R-Squared	0.0703		Adjusted R-Squared		0.0529		Multiple R-Squared	0.0916		Adjusted R-Squared		0.072	

L'esponenziale dei coefficienti ottenuti dal modello Logit ha permesso di ottenere un indice relativo, nominato "Odds Ratio", in grado di misurare l'attitudine o il rischio di contrarre il sintomo in esame. Tale indicatore permette di valutare se sia presente una forma di associazione tra ogni fattore e il rischio di contrarre la sintomatologia (0R ≠1) e quindi di definire se l'esposizione sia stata protettiva (0R<1) o dannosa (OR>1).

Grafico 7: Odds ratio ricavati dalle stime del modello logit per il macro-sintomo Fisico

FISICO PRIMA ONDATA

Variable		N	Odds ratio	p
Età		434	1.02 (1.01, 1.04)	0.006
B_Bloccanti	0	388	Reference	
	1	46	2.64 (1.14, 6.71)	0.030
Intensità	0	102	Reference	
	1	207	0.43 (0.23, 0.79)	0.007
	2	43	0.97 (0.41, 2.39)	0.947
	3	82	1.01 (0.50, 2.05)	0.967
Antiaggregante	0	400	Reference	
	1	34	0.29 (0.12, 0.67)	0.004
Steroidi_inalatori	0	419	Reference	
	1	15	5.59 (1.38, 37.95)	0.033
Metaboliche	0	354	Reference	
	1	80	2.52 (1.25, 5.27)	0.012
HCQ	0	260	Reference	
	1	174	1.61 (0.96, 2.72)	0.072
N_Comorbidità		434	0.81 (0.67, 0.99)	0.040
Sesso	F	209	Reference	
	M	225	0.60 (0.38, 0.93)	0.024

FISICO SECONDA ONDATA

Variable		N	Odds ratio	p
Sesso	F	376	Reference	
	M	362	0.46 (0.33, 0.63)	<0.001
Polmonari	0	654	Reference	
	1	84	1.90 (1.12, 3.37)	0.02
Patologie_Autoimmuni	0	709	Reference	
	1	29	4.56 (1.57, 19.37)	0.01
Steroide	0	258	Reference	
	1	480	1.84 (1.30, 2.62)	<0.001
Remdesevir	0	623	Reference	
	1	115	1.64 (1.02, 2.68)	0.04

Nel Grafico 7, relativo alla macro-sintomatologia Fisica, si sono osservate, tramite un confronto parallelo, delle differenze significative tra le variabili che hanno avuto un impatto sullo sviluppo del sintomo in esame. Si può notare, infatti, che sia nella prima che nella seconda ondata, i soggetti di sesso maschile hanno avuto una minor propensione a sviluppare il sintomo rispetto a quelli di sesso femminile. Differenza osservabile maggiormente nella seconda fase dove le donne hanno presentato un rischio di contrarre la sindrome due volte superiore rispetto agli uomini.

Nel primo periodo di infezione da SARSCOV2, i soggetti con maggior tendenza alla sintomatologia sono stati pazienti con patologie Metaboliche e individui che hanno assunto nella fase pre-covid i farmaci Beta Bloccanti.

Risulta curioso osservare che coloro che hanno manifestato i sintomi del virus in forma più leggera hanno dimostrato un'attitudine due volte maggiore a contrarre disturbi fisici rispetto a chi ha sviluppato un livello medio di infezione. E' importante inoltre sottolineare che tra le terapie, la somministrazione dell'HCQ si è rivelata un fattore che ha favorito il permanere dei sintomi. Infatti, il grafico evidenzia che il rischio di sviluppare disturbi fisici nei pazienti che hanno assunto tale trattamento è circa due volte superiore rispetto a quello di chi non è stato trattato.

Per quanto riguarda la seconda ondata, le conseguenze da Long-Covid si osservano nei pazienti con patologie Polmonari e Autoimmunologiche e nei soggetti trattati con Steroide e Remdesevir.

Lo studio degli effetti delle variabili sui soggetti che hanno contratto il Long-Covid è stato, invece, affrontato tramite l'interpretazione dei coefficienti stimati dall'equazione OLS del modello Heckit.

Dalla Tabella 13 possiamo notare come, passando dalla prima alla seconda ondata, l'aumento o la diminuzione dell'intensità della sintomatologia, per i soggetti che l'hanno sviluppata, sia stata determinata dall'influenza di variabili differenti.

In entrambe i casi punteggi maggiori hanno riguardato il genere femminile rispetto ai soggetti di sesso maschile.

Analizzando la prima ondata si osserva che all'aumentare del numero delle comorbidità è conseguito un prolungamento della macro-sintomatologia Fisica di circa sessantatré giorni; effetto contrario si è verificato nella seconda fase acuta del virus in cui si è registrata una durata inferiore pari a dieci giorni.

Tra le terapie in studio risulta importante sottolineare l'azione benefica della somministrazione dell'Eparina ai soggetti con patologie cardiache ai fini della diminuzione significativa del periodo di prognosi.

Per quanto riguarda la seconda ondata, a livello di profilo clinico, la caratteristica propria del paziente che è risultata associata ad un maggior intensità del disturbo fisico è l'indice di massa corporea. Infatti all'incremento unitario del BMI la durata media della sintomatologia aumenta di circa sei giorni.

Analogamente il perdurare dei sintomi fisici legati al Long-Covid ha interessato pazienti trattati con antibiotico, sebbene all'aumentare dell'età anagrafica tali effetti si riducano.

Effetto contrario si è osservato nei soggetti sottoposti a ossigenoterapia di media intensità.

6.3 Selection e Main equation Macro Sintomatologia Psicologica

Per quanto riguarda lo studio della Macro-Sintomatologia Psicologica, i risultati ottenuti dalle due equazioni del modello Heckit sono presentati nella tabella seguente.

Tabella 14: Output modello Logit e OLS sulla macro-sintomatologia Psicologica per ondata

MACRO-SINTOMO PSICOLOGICO											
MODELLO GLM											
PRIMA ONDATA						SECONDA ONDATA					
	Estimate	Std.	t-value	Pr(>|t|)			Estimate	Std.	T value	Pr(>|z|)	
(Intercept)	0.241	0.169	1.428	0.153		(Intercept)	0.472	0.388	1.219	0.223	
Sesso:Maschio	-0.874	0.209	-4.174	2.99e-05	***	Età	0.001	0.007	0.156	0.876	
Ipolipemizzante Statina1	-1.199	0.388	-3.093	0.002	**	Sesso: Maschio	-1.193	0.169	-7.017	2.26e-12	***
HCQ1	0.839	0.216	3.883	0.0001	***	Patologie Immunologiche 1	-1.288	0.365	-3.532	0.0004	***
Metaboliche1	0.652	0.324	2.016	0.044	*	Antibiotico1	1.716	0.729	2.351	0.019	*
B_Bloccanti1	1.469	0.512	2.872	0.004	**	Steroide1	0.619	0.195	3.183	0.002	**
Metaboliche1: B_Bloccanti1	-1.935	0.788	-2.455	0.014	*	Età:Antibiotico1	-0.024	0.012	-2.011	0.044	*

MAIN EQUATION													
PRIMA ONDATA							SECONDA ONDATA						
Estimate		Days	Std,	t value	Pr(>|t|)		Estimate		Days	Std,	t value	Pr(>|t|)	
(Intercept)	318.00	858.61	52.89	6.01	0.001	***	(Intercept)	156.71	423.12	43.17	3.63	0.001	***
Sesso: Maschio	-79.84	-215.58	41.77	-1.91	0.06	.	Età	-0.72	-1.95	0.45	-1.61	0.11	
Cardiache1	-1.55	-4.19	43.31	-0.04	0.97		Sesso: Maschio	-45.35	-122.45	18.75	-2.42	0.02	*
Eparina1	139.86	377.62	50.30	2.78	0.01	**	O2intensità : <=2L/min	-13.44	-36.29	20.72	-0.65	0.52	
Cardiache1 :Eparina1	-179.80	-485.47	75.75	-2.37	0.02	*	O2intensità : 2-6L/min	-37.04	-100.00	21.82	-1.70	0.09	.
							O2intensità : >6L/min	-7.79	-21.04	15.33	-0.51	0.61	
							BMI	2.30	6.20	1.22	1.88	0.06	.
							ACE_inibitri Sartani1	-25.37	-68.50	13.86	-1.83	0.07	
DIAGNOSTICA													
Estimate			Std,	t value	Pr(>|t|)		Estimate			Std,	t value	Pr(>|t|)	
invMills Ratio	-6.327		77.31	-0.08	0.94		invMills Ratio	-29.01		35.44	-0.82	0.41	
rho	-0.025		NA	NA	NA		rho	-0.26		NA	NA	NA	
Multiple R-Squared		0.067	Adjusted R Squared		0.0467		Multiple R-Squared		0.0954	Adjusted R Squared		0.0782	

Nella macro-sintomatologia Psicologica, come in quella Fisica, si nota che le variabili determinanti del sintomo sono significativamente differenti tra le due ondate e che i potenziali fattori che agiscono su di essa presentano analogie con quelli del sintomo fisico.

Grafico 8: Odds ratio ricavati dalle stime del modello logit per il macro-sintomo Psicologico rispettivamente per prima e seconda ondata

PSICOLOGICO PRIMA ONDATA

Variable	N	Odds ratio	p
Sesso	F 209	Reference	
	M 225	0.42 (0.28, 0.63)	<0.001
Ipolipemizzante_Statina	0 387	Reference	
	1 47	0.30 (0.14, 0.64)	0.002
HCQ	0 260	Reference	
	1 174	2.32 (1.52, 3.56)	<0.001
Metaboliche	0 354	Reference	
	1 80	1.92 (1.03, 3.68)	0.044
B_Bloccanti	0 388	Reference	
	1 46	4.35 (1.70, 12.99)	0.004

PSICOLOGICO SECONDA ONDATA

Variable	N	Odds ratio	p
Sesso	F 376	Reference	
	M 362	0.30 (0.22, 0.42)	<0.001
Patologie_Immunologiche	0 692	Reference	
	1 46	0.28 (0.13, 0.55)	<0.001
Antibiotico	0 492	Reference	
	1 246	5.56 (1.36, 23.86)	0.019
Steroide	0 258	Reference	
	1 480	1.86 (1.27, 2.73)	0.001

Anche in questo caso, in entrambe le ondate, come si osserva nel Grafico 8, i soggetti di sesso maschile risultano aver avuto una tendenza al rischio minore.

La medesima risposta all'esposizione si osserva tra pazienti affetti da patologie Metaboliche e tra coloro che hanno assunto farmaci Beta Bloccanti, mostrando entrambi, nella prima ondata, una maggior propensione allo sviluppo del sintomo in esame.

Per quanto riguarda la seconda ondata, le patologie di natura immunologica hanno rappresentato un fattore protettivo, riducendo il rischio di sviluppare la malattia del 72%.

Si sono dimostrate al contrario potenzialmente dannose alcune terapie, quali la cura steroidea ed in particolare l'antibiotico.

Coloro, infatti, a cui è stata somministrata tale terapia hanno registrato un'attitudine alla sintomatologia sei volte superiore rispetto a chi non è stato trattato.

Le analogie tra i due macro-sintomi sono state evidenziate anche nel modello lineare OLS.

In conclusione, è importante evidenziare che il valore del coefficiente dell'Inverse Mills Ratio, fattore correttivo da considerare in presenza di selection bias, non è significativo

in nessuno dei quattro modelli stimati, Probit e OLS, per le sintomatologie Psicologica e Fisica. Tale risultato è determinato dalla presenza di mancata associazione tra le due equazioni del modello Heckit, smentendo l'ipotesi della presenza di un campione auto-selezionato affetto da distorsione. Medesimi risultati osservati si potrebbero ottenere applicando i modelli di regressione tradizionali.

6.4 Confronto tecniche differenti di Model Selection

La trattazione si conclude con il confronto dei risultati ottenuti applicando due metodologie differenti di Model Selection a supporto dei modelli Heckit.

E' stato eseguito, infatti, un ulteriore studio al fine di valutare eventuali similitudini o differenze tra le variabili selezionate dai metodi descritti nei capitoli precedenti e quelle estratte da metodologie tradizionali, le quali, a differenza dei modelli di Model Selection, si limitano ad osservare solamente la presenza di un'associazione lineare tra le covariate e i macro-sintomi in esame.

Tabella 15: Confronto output modello Heckit con metodi di Model Selection differenti

MAIN EQUATION	Macro-Sintomo Psicologico				Macro-Sintomo Fisico			
	Model Selection Lineare		Model Selection ML		Model Selection Lineare		Model Selection ML	
	Ondata I	Ondata II	Ondata I	Ondata II	Ondata I	Ondata II	Ondata I	Ondata II
Sesso: Maschi	-	-	-	-	-	-	--	--
Età	-	-	-	-				
BMI		+		+		++		++
ACE_inibitori_Sartani1		-		-				
Ipolipemizzante_Statina1					-			
Diuretici1					-			
N_Comorbidità					++	-	++	-
Tocilizumab_Sarilumab_baricitinib1	-*							
Eparina1	+		++		+		++	
O2intensità: <=2L/min								
O2intensità: 2-6 L/min		-		-		-		-
O2intensità: >6 L/min								
Antibiotico1								++
Cardiache*Eparina			-				-	
Età*Antibiotico								--

SELECTION EQUATION	Macro-Sintomo Psicologico				Macro-Sintomo Fisico			
	Model Selection Lineare		Model Selection ML		Model Selection Lineare		Model Selection ML	
	Ondata I	Ondata II	Ondata I	Ondata II	Ondata I	Ondata II	Ondata I	Ondata II
Sesso: Maschi	---	---	---	---	-	---	-	---
Età					++		++	
Intensità: Media					-		--	
Intensità: Media-Alta								
Intensità: Alta								
Steroidi_inalatori1	+	+			+		+	
Diuretici1						+*		
Antiaggregante1					--		--	
Neuro_Psicoattivi1	-*	+*						
Ipolipemizzante_Statina1	--		--					
Ca_antagonisti1	-							
B_Bloccanti1	+		++		+		+	
ACE_inibitori_Sartani1	++							
Diabete1		-						
Patologie_Immunologiche1		---		---				
Renali1	-							
Metaboliche1			+		+		+	
Polmonari1						+		+
Patologie_Autoimmuni1						++		++
N_Comorbidità					-		-	
HCQ1	+++		+++		+		+	
Steroide1		++		++		+++		+++
Remdesevir1		+*				+		+
Antibiotico1		+		+				
O2intensità: <=2L/min								
O2intensità: 2-6 L/min					+*			
O2intensità: >6 L/min								
Età*Antibiotico				-				
Metaboliche*B_Bloccanti			-					
B_Bloccanti1:ACE_inibitori_Sartani1	-*							

Capitolo 7

Conclusioni

La stima del modello Heckit ha permesso di valutare, tramite l'interpretazione del coefficiente di correlazione tra le due equazioni Probit e OLS , che il campione non è affetto da auto selezione e quindi risulta rappresentativo della popolazione.
I risultati conseguiti dallo studio in esame hanno evidenziato, per entrambe le macro-sintomatologie post-Covid, differenze statisticamente significative passando dalla prima alla seconda ondata di infezione.
Lo studio ha evidenziato che il Long Covid ha riguardato soprattutto il genere femminile interessando tra le due ondate soggetti con profili clinici differenti per patologie e trattamenti assunti durante la fase infettiva.
In particolare i trattamenti in esame che hanno determinato una riduzione del permanere delle sintomatologie post-virali sono stati, nella prima ondata, l'eparina, limitatamente ai soggetti cardiaci, mentre nella seconda terapie di somministrazione di ossigeno e, per soggetti d'età avanzata, l'antibiotico.
Al contrario, ad eccezione dell'ossigenoterapia, gli stessi trattamenti si sono rivelati dannosi per soggetti non rispondenti ai profili clinico e anagrafico sopra citati, in quanto associati ad una maggior intensità dei disturbi.
Lo studio ha evidenziato , inoltre, che i trattamenti che hanno agito significativamente sulla durata non sono stati, invece, rilevanti per lo sviluppo della sindrome post-virale.
A tal fine, infatti, si sono rivelate dannose terapie quali il Remdesevir, gli steroidi e gli antibiotici; ed in generale non si sono registrati trattamenti potenzialmente protettivi.
Il confronto dell'effetto delle terapie nelle due ondate in esame porta, quindi, ad ipotizzare che il contagio sia stato determinato da un virus che ha mutato la sua natura tra le due fasi andando a colpire soggetti con anamnesi cliniche differenti.
L'analisi trasversale tra macro-sintomatologie, inoltre, ha evidenziato che i sintomi fisici e psicologici presentano gli stessi determinanti, rafforzando l'ipotesi di una possibile associazione tra i due gruppi e quindi dell'esistenza di un'unica macro-sintomatologia psico-fisica.
La trattazione condotta si è posta l'obiettivo di fungere da analisi preliminare del fenomeno, i cui risultati potrebbero rappresentare il punto di partenza di più approfonditi

studi al fine di rafforzare le ipotesi sopra riportate. Sarebbe, infatti, opportuno, condurli su un campione nazionale al fine di ottenere risultati generalizzabili, non solo alla popolazione residente nell'area dell'ospedale metropolitano come nel caso in esame, ma anche a quella dell'intera penisola.

Un'ulteriore ambito d'indagine potrebbe, inoltre, rivelarsi l'effetto del vaccino sullo sviluppo e sull'intensità del Long Covid, fattore che potrebbe risultare potenzialmente protettivo per i soggetti infettati.

Bibliografia e Sitografia

- *R. Carter Hill, William E. Griffiths, Guay C.Lim, Principle of Econometrics 2014*
- *Martinella, V. (2022, May 9). Long Covid nei malati di tumore: chi rischia di più? CorrieredellaSera.* https://www.corriere.it/salute/sportello_cancro/22_maggio_09/long-covid-malati-tumore-b7aa30ca-c616-11ec-80ae-9f956e43a8f0.shtml
- *Martinella, V. (2022b, July 24). Long Covid, i sintomi: le otto manifestazioni più frequenti che perdurano nel tempo. Corriere della Sera. https://www.corriere.it/salute/malattie_infettive/cards/otto-manifestazioni-piu-frequenti-long-covid-che-perdurano-tempo-cosa-sappiamo-disturbi-piu-comuni/post-covid-19-condition_principale.shtml?viewName=APERTURA_FOTO_FISSA*
- *Marrone, C. (2022, May 12). Long Covid, almeno un sintomo in oltre la metà dei pazienti due anni dopo il ricovero. Corriere della Sera. https://www.corriere.it/salute/neuroscienze/22_maggio_12/long-covid-sintomi-pazienti-due-anni-dopo-d31e56ca-d132-11ec-b465-8b7c23727ee0.shtml*
- *Cervia C, Zurbuchen Y, Taeschler P, et al. Immunoglobulin signature predicts risk of post-acute COVID-19 syndrome. Nat Commun. 2022;13(1):446. Published 2022 Jan 25. doi:10.1038/s41467-021-27797-1*
- *Taquet M, Dercon Q, Luciano S, Geddes JR, Husain M, Harrison PJ. Incidence, co-occurrence, and evolution of long-COVID features: A 6-month retrospective cohort study of 273,618 survivors of COVID-19. PLoS Med. 2021;18(9):e1003773. Published 2021 Sep 28. doi:10.1371/journal.pmed.1003773*
- *Heckman Selection | Model Estimation by Example. Michael Clark. https://m-clark.github.io/models-by-example/heckman-selection.html#*
- *Bogdan Oancea ,Heckman correction technique-a short introduction. Published 2017*
- *The Heckman Sample Selection Model. (2018, September 17). Rob Hicks. https://rlhick.people.wm.edu/stories/econ_407_notes_heckman.html*
- *A Heckman Selection-t Model. (2000). Taylor & Francis. https://www.tandfonline.com/doi/full/10.1080/01621459.2012.656011?casa_token=HBaEH3xDXEIAAAAA%3AQgrYwdUejTnd1E9dur_Y9EzgRMQbh1Mo8VuSL5RY1WQk8mcX7c-kZpPI6FyN04XFVhXDpBP8No*

Ringraziamenti

La prima persona che vorrei ringraziare immensamente per essere stata sempre presente nei momenti più bui e difficili è il mio più grande punto di riferimento, mia nonna Chiara. A te, vorrei dedicare questa tesi in memoria dell'appoggio che mi hai dato fin dall'inizio nell'affrontare questo duro percorso e del forte sostegno che hai sempre rappresentato per me.

La seconda persona che vorrei citare è mia mamma, la mia roccia e il mio rifugio più sicuro, senza la quale non avrei raggiunto tutti i miei più importanti obiettivi.

Ringrazio tutti i miei famigliari, in particolare zia Federica e tutte le mie amiche, Annalisa, Cristina, Eleonora, Beatrice e Martina, compagne di avventure, di ricordi e di crescita grazie alle quali ho scoperto il vero valore dell'amicizia.

Desidero ringraziare, inoltre, il relatore di questa tesi, il professore Pietro Giorgio Lovaglio, per la disponibilità e l'attenzione dimostrate durante lo svolgimento del lavoro.

Ringrazio inoltre tutti i miei compagni di corso con cui ho condiviso questi tre anni di crescita personale e professionale, in particolare Stefano, Francesco, Camilla e Pietro.

Un ringraziamento speciale va ad una persona importante, Luca, compagno di tesi, per avermi sostenuta, sopportata e soprattutto per esserci sempre.