Università di Pisa

Corso di Laurea in

Artificial Intelligence and Data Engineering

Regressione lineare

per l'analisi dei nuovi casi di COVID-19

Candidato

Franco Terranova

ANNO ACCADEMICO 2021/2022

INDICE

INTRODUZIONE

Si ipotizza che l'analisi sia stata commissionata dal ministero della salute italiano con lo scopo di costruire un modello di regressione per la previsione del numero medio di casi di COVID-19 giornalieri che lo stato debba aspettarsi, esaminando l'andamento della pandemia, delle restrizioni e delle vaccinazioni negli altri Stati del mondo.

1. DATASET

L'analisi è stata svolta sul dataset COVID-19 di Our World in Data, scaricabile al seguente link https://covid.ourworldindata.org/data/owid-covid-data.csv oppure esaminabile attraverso il browser con il seguente Data Explorer https://ourworldindata.org/explorers/coronavirus-data-explorer.

1.1. CONTENUTO DELLA TABELLA

A partire dal dataset indicato è stata effettuata una prima operazione di pre-processing, come descritto nell'Appendice.
Successivamente, l'analisi è stata svolta sui seguenti fattori:

- stringency_index ($\in [0, 100]$): misura che riassume la risposta dei governi, ottenuta considerando diversi indicatori, tra cui chiusure di scuole, chiusure di posti di lavoro e divieti di viaggio, ridimensionati a un valore da 0 a 100 (100 = massima severità). Se le politiche variano a livello regionale, l'indice mostra la risposta della regione più rigorosa.
- total_cases ($\in \mathbf{Z}^+$): numero totale di casi di COVID-19 riscontrati finora.
- total_deaths ($\in \mathbf{Z}^+$): numero totale di morti da COVID-19 riscontrati finora.
- people_fully_vaccinated ($\in \mathbf{Z}^+$): numero totale di persone che hanno completato il ciclo vaccinale (due dosi).
- people_vaccinated ($\in \mathbf{Z}^+$): numero totale di persone che hanno ricevuto almeno una dose di vaccino.
- population ($\in \mathbf{Z}^+$): numero totale di persone che vivono nello stato.
- reproduction_rate ($\in \mathbf{Z}^+$): quante persone vengono contagiate in media da una persona infetta da COVID-19.
- new_cases ($\in \mathbf{Z}^+$): numero di nuovi casi giornalieri.

Sono stati utilizzati i dati relativi all'inizio di settembre, per i quali sono presenti un numero minore di osservazioni con valori mancanti.

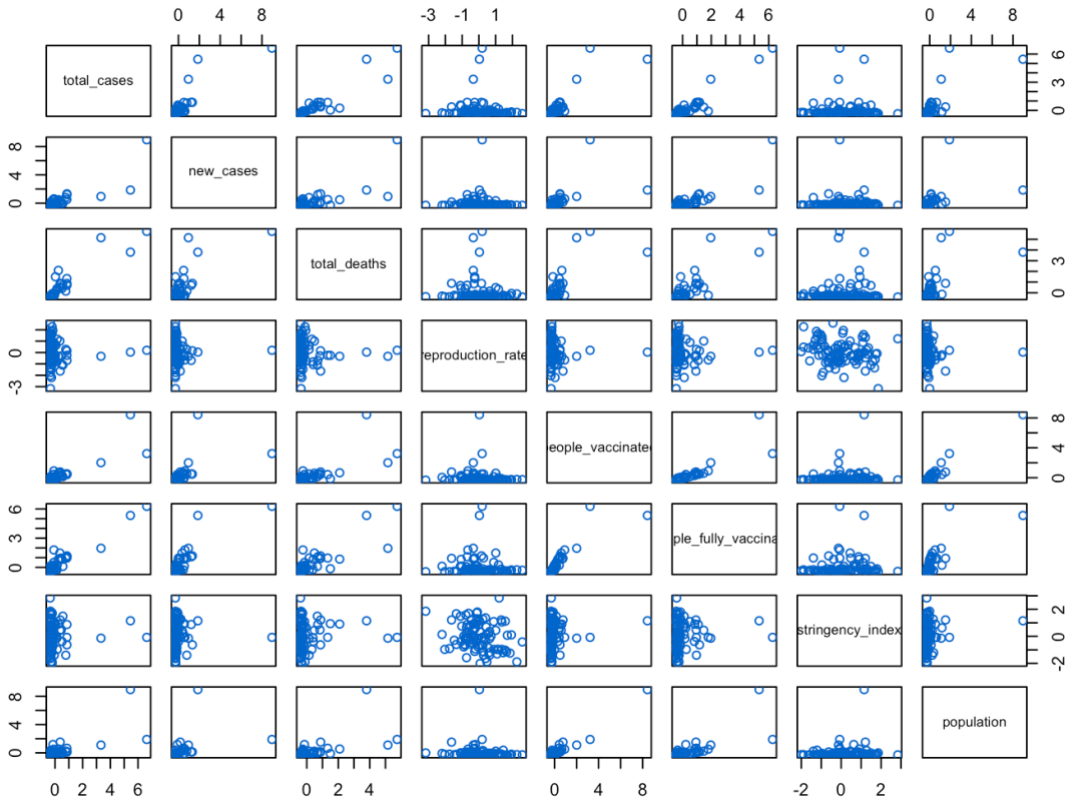
La tabella risultante contiene 8 colonne per un totale di 96 osservazioni.

Di seguito troviamo un breve riepilogo dei dati contenuti nella tabella.

	total_cases	**new_cases**	**total_deaths**	**reproduction_rate**	**people_vaccinated**	**people_fully_vaccinated**	**stringency_index**	**population**
Minimo	3403	0	14	0.23	23039	20634	23.15	3.825*e^+04
Media	1995392	6545.4	41434	1.0117	19682620	12042212	53.45	4.830*e^+07
Massimo	39543669	202000	642897	1.65	509668131	176481306	96.3	1.393*e^+09

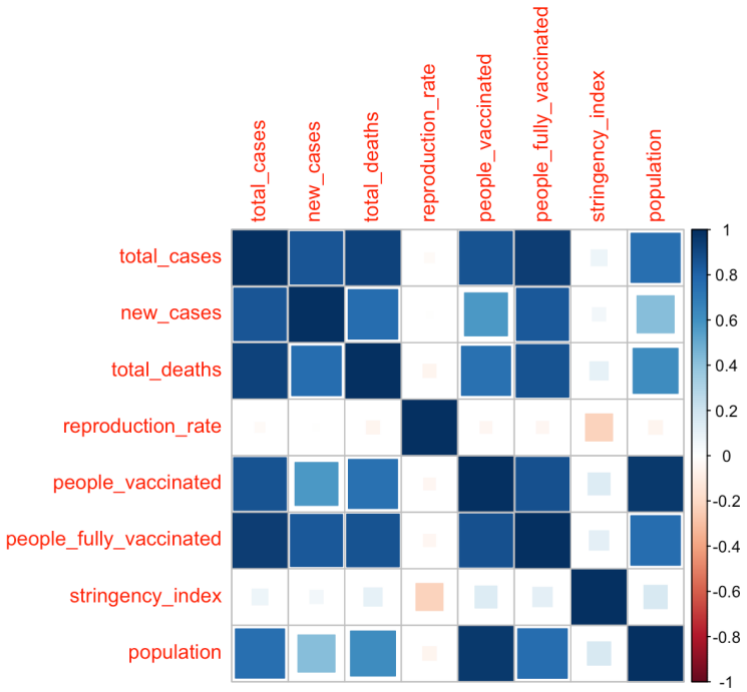
Vista la differenza di scala tra i vari fattori, si è proceduto con la standardizzazione della tabella.

Di seguito troviamo lo scatter plot della tabella, in seguito alla standardizzazione.



1.2. ESPLORAZIONE DEI DATI

Esploriamo innanzitutto le correlazioni tra i vari fattori, al fine di analizzare come essi interagiscano tra di loro.



Dal grafico di correlazione è possibile primariamente notare come stringency_index e reproduction_rate presentano correlazioni molto lievi con tutti gli altri fattori (< 0.25).
Notiamo, invece, una significativa correlazione tra il fattore di uscita scelto con il numero totale dei casi (0.85) e il numero di persone che hanno completato il ciclo vaccinale (0.84).
Dal grafico, inoltre, viene evidenziata la forte correlazione tra il totale dei morti e il totale dei casi (0.93). La correlazione più rilevante tra le presenti è quella tra il numero di persone che hanno ricevuto almeno una dose di vaccino e il numero di abitanti dello Stato (0.97).

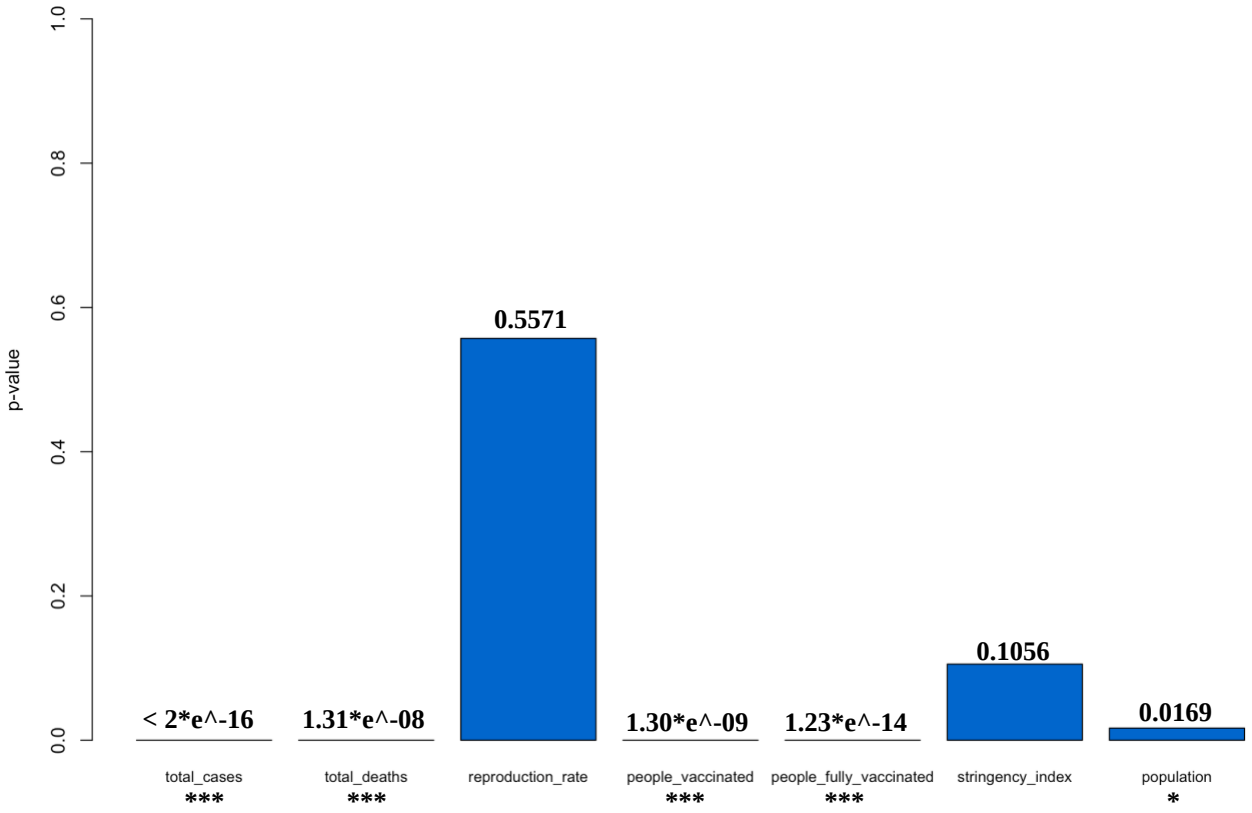
2. REGRESSIONE LINEARE

Per valutare il modello di regressione lineare sono state considerate come metriche la proporzione di varianza spiegata dal modello, la proporzione di varianza spiegata dal modello corretto, il p-value dei coefficienti e quello del modello globale.

```
Residual standard error: 0.2607 on 88 degrees of freedom
Multiple R-squared:  0.9371,    Adjusted R-squared:  0.9321
F-statistic: 187.2 on 7 and 88 DF,  p-value: < 2.2e-16
```

La regressione lineare multivariata ha raggiunto un'elevata proporzione di varianza spiegata, un'elevata proporzione di varianza spiegata del modello corretto e un p-value del modello globale molto basso.

Troviamo di seguito rappresentati i p-value dei vari fattori:



2.1. RIDUZIONE DEL MODELLO

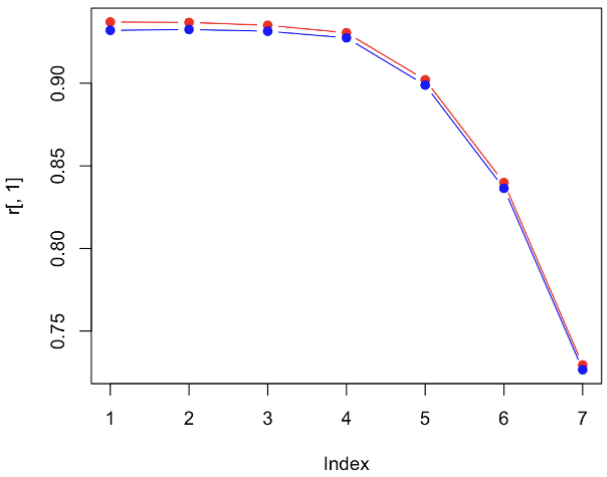
Per quanto riguarda la tabella dei coefficienti, è possibile notare alcuni p-value significativamente più alti degli altri. Si può arrivare allora ad un modello più sintetico eliminando uno ad uno i fattori meno influenti. I fattori eliminati sono stati, in quest'ordine, reproduction_rate, stringency_index e population, giungendo ad un modello con soli 4 fattori e con coefficienti i cui p-value risultano essere tutti inferiori a 3*e^-08.

```
Residual standard error: 0.2693 on 91 degrees of freedom
Multiple R-squared:  0.9305,    Adjusted R-squared:  0.927
F-statistic: 304.8 on 4 and 91 DF,  p-value: < 2.2e-16
```

Per i fattori stringency_index, reproduction_rate e population, l'effetto correttivo della diminuzione del numero di fattori nel modello è più significativo della scarsa diminuzione della proporzione di varianza spiegata e, perciò, ho proceduto con la loro eliminazione.

I p-value dei restanti fattori suggeriscono di fermarci con la riduzione del modello.

Come mostrato nel grafico accanto, se procedessimo con ulteriori eliminazioni dei fattori, otterremmo un calo significativo della proporzione di varianza spiegata (rappresentata in rosso) e della proporzione di varianza spiegata corretta (rappresentata in blu).



Durante la riduzione del modello sono state effettuate le seguenti considerazioni.

Innanzitutto, i fattori total_cases, total_deaths, people_vaccinated e people_fully_vaccinated risultano essere i fattori che catturano meglio la varianza del problema, come era già presumibile dopo l'analisi del grafico di correlazione.

Quest'ultimi raggiungono da soli una proporzione di varianza spiegata e una proporzione di varianza spiegata corretta vicinissime a quelle raggiunte considerando anche gli altri fattori.

Inoltre, come ci si poteva aspettare dalla scarsa correlazione di stringency_index e reproduction_rate con new_cases, i due fattori erano poco influenti al fine dell'analisi.

Infine, il miglioramento del p-value del fattore people_vaccinated dopo l'eliminazione del fattore population (variando da 1.21*e^-09 a un valore minore di 2*e^-16), suggerisce la presenza di una collinearità tra i due. In effetti, la loro correlazione era elevata (pari a 0.97).

2.2. REGRESSIONE NON LINEARE

Si è tentato di effettuare un'analisi tramite un modello di regressione non lineare, in particolare un modello di regressione logaritmico.

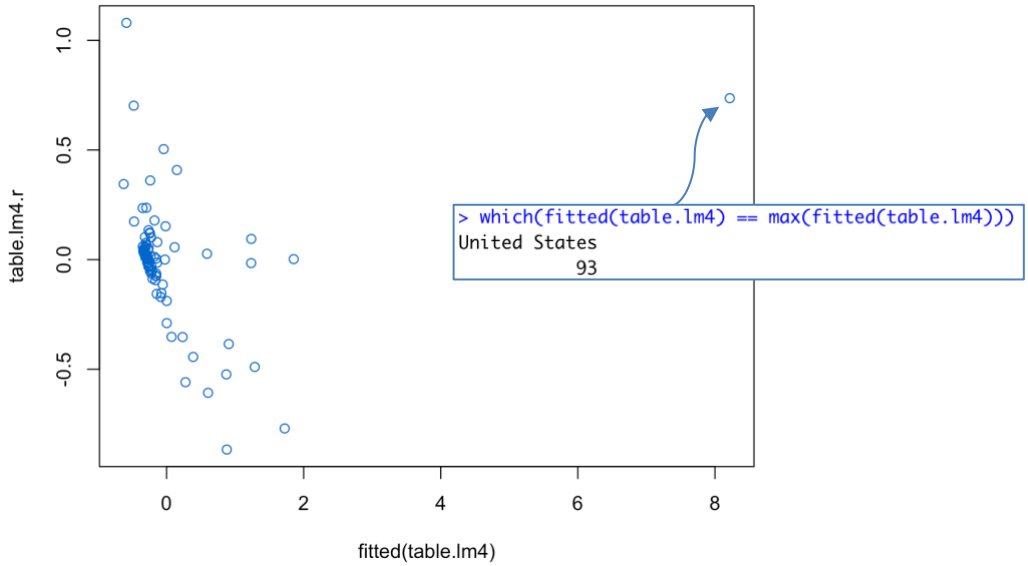
Purtroppo, il modello in questione non ha fornito risultati soddisfacenti, raggiungendo una proporzione di varianza spiegata molto minore rispetto a quella ottenuta considerando il modello di regressione lineare.

Proprio per questo motivo, il resto della relazione verterà solo sull'approfondimento del modello lineare.

```
Residual standard error: 1.508 on 88 degrees of freedom
Multiple R-squared:  0.6608,    Adjusted R-squared:  0.6338
F-statistic: 24.49 on 7 and 88 DF,  p-value: < 2.2e-16
```

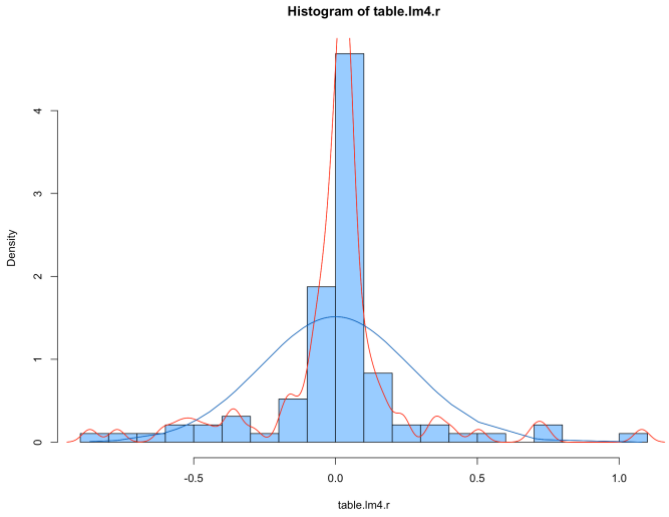
3. ANALISI DEI RESIDUI

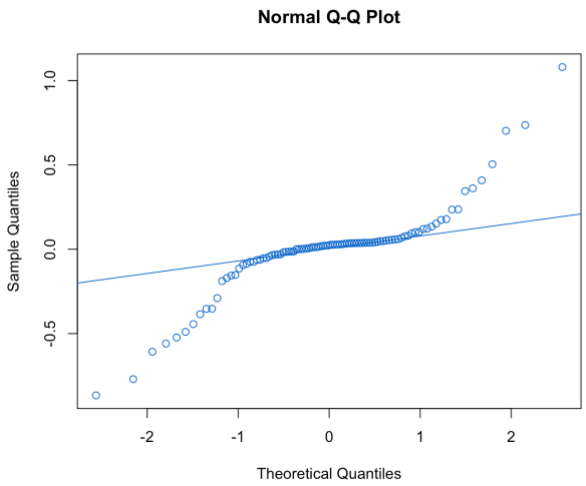
Analizziamo i residui del modello lineare ridotto rispetto ai valori delle previsioni per cogliere eventuali anomalie. Utilizzando il diagramma di dispersione, notiamo la presenza di una figura non omogenea, che indica una possibile struttura dei residui che non è stata riconosciuta.



Tra i vari residui, notiamo un valore (in alto a destra), tra i tanti, che suggerisce la presenza di una possibile anomalia, la quale risulta essere relativa agli Stati Uniti d'America.

Esaminiamo anche l'istogramma dei residui, il Q-Q plot, gli indicatori di asimmetria, Skewness e Kurtosi, e i risultati del test di Shapiro-Wilk.





Skewness	Kurtosis	p-value (Shapiro-Wilk normality test)
0.1674514	4.369874	3.071*e^-09

I risultati ottenuti ci suggeriscono che la distribuzione dei residui non risulta essere gaussiana.

Una motivazione per cui il risultato non risulta essere troppo soddisfacente è sicuramente la quantità non elevata di osservazioni utilizzate nell'analisi. Oltre, infatti, al numero limitato di Stati mondiali, solo un numero ridotto di Stati presenta valori non nulli per i fattori utilizzati.

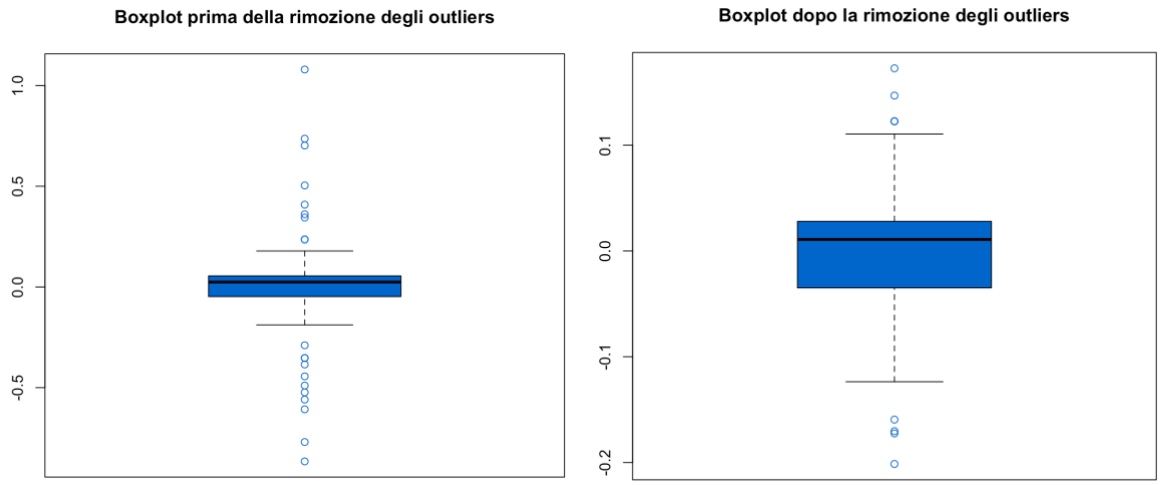
Si è inoltre tentato di rinserire alcuni fattori rimossi durante la riduzione del modello, senza però raggiungere risultati migliori per quanto riguarda i residui.

La presenza di residui anomali può essere inoltre dovuta alla presenza di osservazioni anomale, come gli Stati Uniti d'America, Stato fortemente colpito dalla pandemia che ha raggiunto oltre 45 milioni di casi e quasi 1 milione di decessi, il quale infatti assume il valore massimo di total_cases, new_cases, total_deaths, ma anche di people_fully_vaccinated nel dataset.

Un miglioramento è emerso allora rimuovendo gli Stati relativi ai residui che causavano una forte deviazione da una distribuzione gaussiana, rimuovendo nel complesso 20 stati dal dataset iniziale.

Attraverso un'attenta analisi, è emerso che gli Stati rimossi risultano essere gli Stati che hanno raggiunto i numeri più drastici durante la pandemia.

Di seguito troviamo rispettivamente il boxplot prima della rimozione degli outliers e il boxplot dopo la rimozione degli outliers.



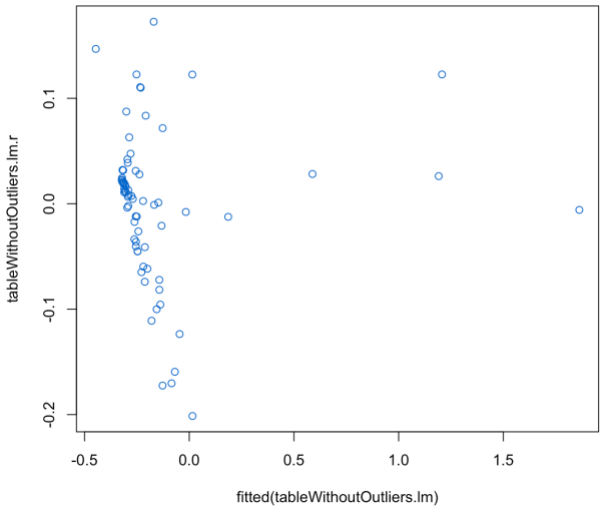

```
> outliers <- boxplot(table.lm4.r, plot=FALSE)$out
> tableWithoutOutliers = table[-which(table.lm4.r %in% outliers),]
```

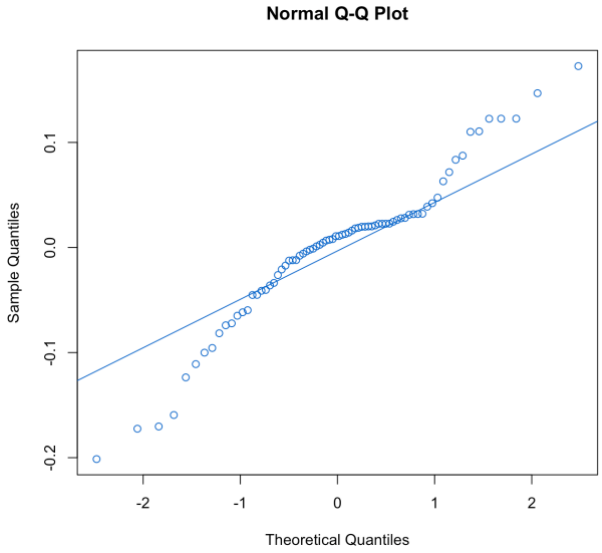
Per quanto riguarda il modello lineare ridotto, in seguito alla rimozione degli outliers notiamo l'aumento della proporzione di varianza spiegata e della proporzione di varianza spiegata corretta di tre punti percentuali.

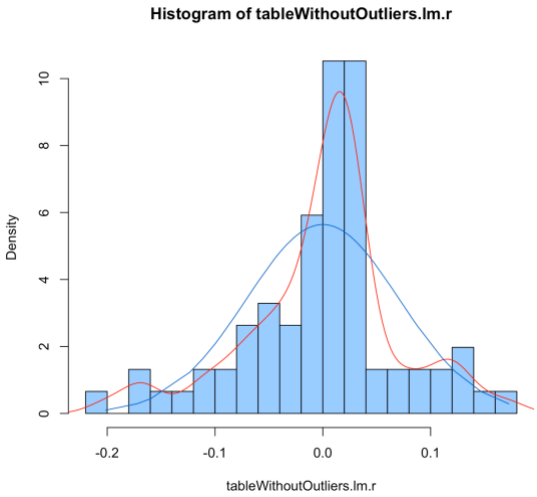
I p-value dei vari coefficienti continuano a rimanere molto bassi.

```
Residual standard error: 1587 on 71 degrees of freedom
Multiple R-squared:  0.9619,    Adjusted R-squared:  0.9598
F-statistic: 448.4 on 4 and 71 DF,  p-value: < 2.2e-16
```

In seguito alla rimozione degli outliers, sono stati ottenuti i seguenti risultati per quanto riguarda i residui:







Skewness	Kurtosis	p-value (Shapiro-Wilk normality test)
-0.4077115	0.8103781	0.004348

È possibile notare il miglioramento del valore della Kurtosi e l'aumento del p-value del test di Shapiro-Wilk di sei ordini di grandezza, nonostante però questi risultati suggeriscano ancora il rigetto dell'ipotesi di gaussianità.

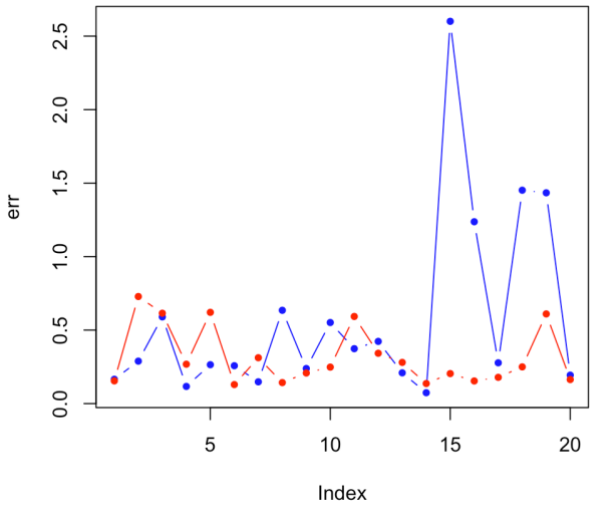
Con la riduzione del numero di osservazioni, tuttavia, potremmo rischiare di avvicinarci ad una situazione di overfitting.

4. PREDIZIONE E AUTOVALUTAZIONE

Considerando i dataset con outliers e senza outliers, dividiamo quest'ultimi in una porzione utile a generare i modelli di regressione lineare ed una porzione di test per verificare la bontà delle previsioni di ogni modello.

Estraendo un solo set di dati abbiamo una variabilità nella stima, introducendo quindi una distorsione se facciamo una sola scelta.

Estraiamo quindi 86 osservazioni per il set di training e 10 osservazioni per il set di test per il modello che considera anche gli outliers, e 68 osservazioni per il set di training e 8 osservazioni per il set di test per il modello che non considera gli outliers ed effettuiamo un numero di prove pari a 20.



	media	**mediana**	**deviazione standard**
Errori del Modello	0.5765557	0.2831227	0.6361377
Errori del Modello privo di Outliers	0.3169362	0.248945	0.197918

La diminuzione del numero di osservazioni utilizzato per la costruzione dei modelli non incide molto sulla proporzione di varianza spiegata, che si attesta comunque intorno alla percentuale precedentemente indicata.

È evidente una differenza nella media, nella mediana ma soprattutto nella deviazione standard degli errori dei due modelli; vi è infatti una forte variabilità nell'errore del modello che prende in considerazione anche gli outliers.

In quest'ultimo è stato possibile osservare che l'errore risulta essere più elevato quando nei set di test si presentano le anomalie precedentemente rilevate, come gli Stati Uniti d'America o il Brasile.

CONCLUSIONI

Dai risultati ottenuti sul modello di regressione lineare, è emerso come le vaccinazioni e il numero totale di casi siano rilevanti nel determinare il numero di nuovi casi giornalieri.
I due valori precedentemente indicati riguardano persone che hanno raggiunto l'immunità, indicando quindi come le vaccinazioni siano uno strumento decisivo per superare la pandemia.
Notiamo inoltre la scarsa influenza dell'indice di restrizione e ciò suggerisce, quindi, che in questa fase della pandemia raggiunta grazie alle vaccinazioni, le restrizioni possano essere un indicatore meno importante.
È emerso inoltre che escludere gli stati colpiti in modo drastico permette di ridurre l'errore medio e raggiungere un modello più generale.
Nonostante sia rimasta una struttura dei residui non catturata, il modello privo di outliers presenta un'elevata proporzione di varianza spiegata, pari al 96.2%, e quindi rimane comunque fortemente valido.
Lo strumento in questione potrebbe trovare utilità nel supporto decisionale del Ministero della Salute, per esempio nello stimare il numero medio di casi di COVID-19 previsti e prendere successivamente decisioni sulle restrizioni da adottare, per esempio.
Un modello più accurato potrebbe tuttavia essere raggiunto con un maggior numero di stati su cui effettuare l'analisi.

APPENDICE

L’operazione di pre-processing ha previsto l’aggregazione delle informazioni per Stato e la scelta di un sotto-insieme di fattori, escludendo quelli con un eccessivo numero di valori mancanti e quelli non rilevanti per la predizione.
Sono stati utilizzati i dati relativi all’inizio di settembre, per i quali sono presenti un numero minore di osservazioni con valori mancanti.
Le righe di codice relative all’operazione di pre-processing sono presenti nel file R commentate, visto che è stata allegata la tabella contenente i dati in seguito all’operazione di pre-processing.