UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Case Study SMS2

Anno Accademico 2020/2021

- ***Alessandro Belotti*** ***1066721***
- ***Stefano Cattaneo*** ***1068794***
- ***Valter Prifti*** ***1064688***
- ***Matteo Vedovati*** ***1064586***

Descrizione del dataset

Il dataset che abbiamo scelto contiene numerosi dati, raccolti con frequenza giornaliera, riguardo allo sviluppo della pandemia di **Covid-19**. Nel dataset ci sono molte colonne correlate tra loro, ad esempio le colonne *ricoverati con sintomi* e *terapia intensiva* sono strettamente legate a *totale ospitalizzati*.
Abbiamo quindi deciso di escludere dal nostro studio le colonne correlate tra loro e altre informazioni non rilevanti oppure con dati mancanti.
Inoltre, ci siamo resi conto che il valore dei deceduti iniziali presentava un'anomalia: a causa di un **errore di rilevazione** veniva rappresentato un numero negativo di variazione decessi. Abbiamo quindi corretto settando un valore pari a 30 alla linea 121 della tabella nella colonna decessi. Il dato è stato preso analizzando varie fonti che riportavano come decessi giornalieri un numero compreso tra 25 e 35. Abbiamo poi creato nuove tabelle contenenti le **variazioni giornaliere** di tamponi, dimessi e deceduti.

Descrizione dei quesiti

Un parametro fondamentale per il monitoraggio dell'andamento della pandemia è la saturazione delle **terapie intensive**, che si deve mantenere inferiore al 30% di occupazione. Abbiamo quindi concentrato il nostro studio sulle dipendenze che ci sono tra questo indice e gli altri dati a nostra disposizione, da cosa dipende il numero di ricoverati in terapia intensiva?
Dal nostro studio abbiamo notato però che i dati relativi ai *totale positivi* **è falsato** durante i primi due mesi di pandemia.
Abbiamo quindi voluto dare **peso diverso** alle prime osservazioni, ossia quelli dei primi 150 giorni di pandemia.
Fatto ciò abbiamo ricercato i migliore modello per descrivere i ricoveri in terapia intensiva.

Metodi statistici utilizzati

Abbiamo iniziato costruendo dei grafici per poter svolgere **studio bivariato** per cercare di capire il legame tra i vari regressori.
Il metodo di regressione da noi inizialmente usato è stato il metodo **LS**, Least Squares, che è uno strumento che ci permette di trovare una funzione lineare che si avvicini il meglio possibile al nostro insieme di osservazioni, **minimizzando** gli errori.
Per prima cosa abbiamo deciso di costruire un modello di regressione di partenza, considerando 5 regressori, che poi abbiamo valutato osservando la loro significatività, tramite l'uso degli intervalli di confidenza sui coefficienti.
Dopodiché, abbiamo svolto diverse prove, modificando il modello, cercando di considerare regressori più rilevanti, fino a giungere ad un modello migliore.
Osservando i grafici, e notando la scarsa quantità di tamponi durante la prima ondata di contagi della scorsa primavera, abbiamo considerato il fatto che, per svolgere uno studio più interessante (e più corretto), avremmo potuto adottare il metodo **WLS** (Weighted Least Squares), che ci consente attraverso un vettore di pesi, di considerare in modo diverso un gruppo di dati rispetto ad altri.
Per trovare il peso ottimale da attribuire ai dati della prima ondata, abbiamo deciso di usare la **Cross-Validazione**, dividendo il dataset in 8 partizioni, usando 7 partizioni per il training e 1 per il test.

Dopodiché, in un ciclo “for” iterativo, abbiamo definito un contatore, usato per far variare il peso da attribuire ai primi 150 campioni (con la formula $2^{\text{contatore} - 1}$) e abbiamo cercato il migliore, valutandolo in base all’**EQM medio** (media degli Errori Quadratici Medi delle 8 partizioni).

Presentazione e discussione dei risultati

Il modello di partenza da noi utilizzato ha dato come risultati:

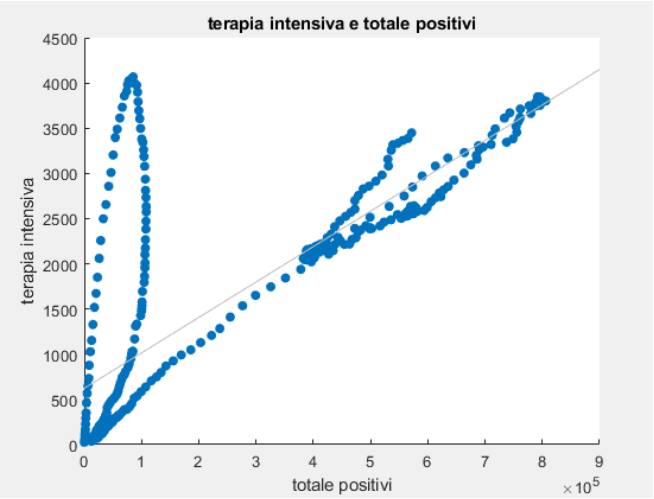
	Beta_hat	tStat	IC inf	IC sup
intercetta	-110.28	-5.3007	-151.19	-69.378
ricoverati_con_sintomi	0.09558	26.06	0.088369	0.10279
totale_positivi	-0.00018605	-1.2698	-0.00047412	0.00010202
dimessi_giornalieri	-0.013032	-3.7472	-0.01987	-0.0061944
deceduti_giornalieri	1.3827	10.375	1.1206	1.6447
tamponi_giornalieri	0.00079288	4.5514	0.00045036	0.0011354

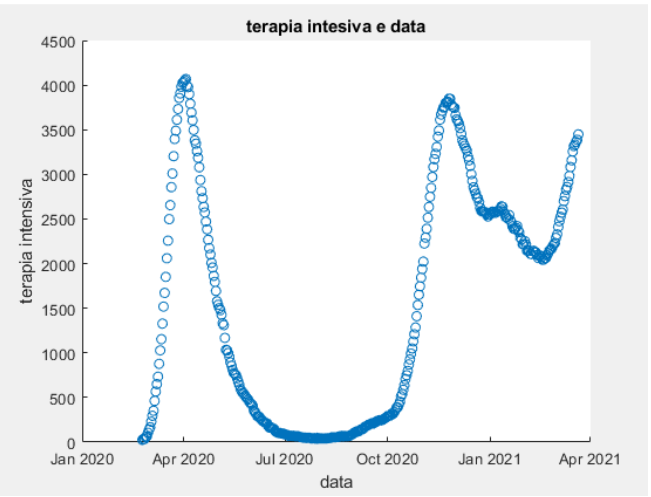
L'indice di correlazione ($\mathbf{R^2}$) risulta essere **0.97389**, mentre il **RMSE** (Root Mean Square Error) risulta **215.74**. Notare come il *totale positivi* ha il valore 0 compreso nell’IC, ossia è un regressore non significativo.
Il modello da noi scelto ha dato come risultati:

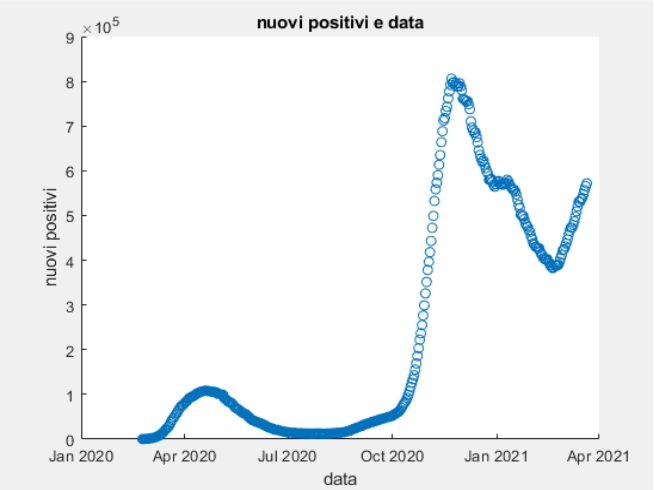
	Beta_hat	tStat	IC inf	IC sup
intercetta	-66.716	-3.5806	-103.35	-30.083
ricoverati_con_sintomi	0.10267	29.619	0.095859	0.10949
totale_positivi	-0.00040267	-5.1544	-0.00055627	-0.00024908
deceduti_giornalieri	1.0625	8.6625	0.82132	1.3036

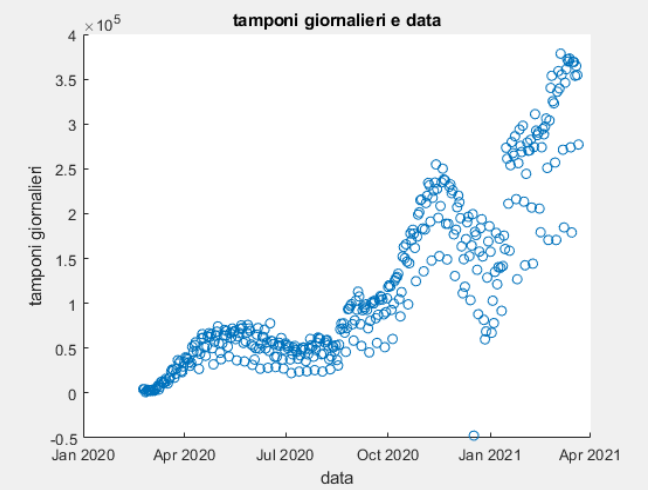
L'indice di correlazione ($\mathbf{R^2}$) vale **0.97182**, mentre il **RMSE** (Root Mean Square Error) risulta **223.55**.
Il secondo modello è stato ottenuto togliendo i *dimessi giornalieri* e i *tamponi giornalieri*, questo al fine di rendere significativo il numero di *totale positivi*.

Svolgendo lo studio bivariato abbiamo notato che per la prima ondata, nel periodo Febbraio-Luglio (corrispondenti ai primi 150 campioni) , c’è stato un **picco** di saturazione delle terapie intensive nonostante i nuovi positivi fossero **molto pochi rispetto a quelli dei mesi successivi**. Sempre grazie allo studio bivariato, abbiamo poi capito che questo è dovuto all’avere effettuato pochi tamponi nel periodo in questione.



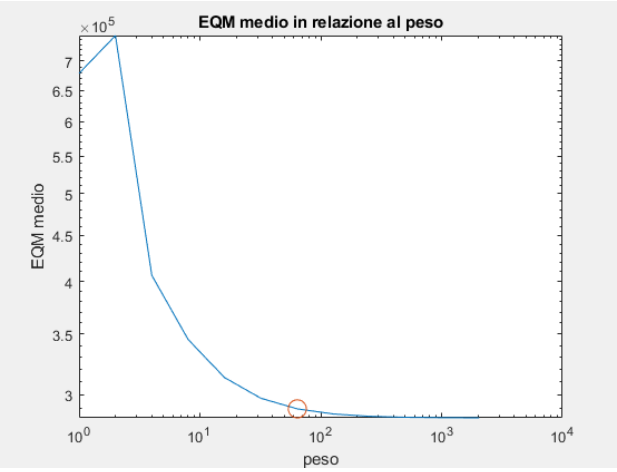






Come si vede dai grafici sovrastanti i posti occupati di terapia intensiva nella prima ondata sono comparabili con quelli delle successive ondate mentre i positivi sono significativamente inferiori ed in linea con i tamponi giornalieri.

Svolgendo il **Metodo dei Minimi Quadrati Ponderati**, assegnando pesi differenti ai primi 150 campioni, abbiamo infine trovato che il modello migliore trovato è quello avente **peso** uguale a **64** poiché abbiamo notato che da questo valore l'**EQM medio** decresce di un valore trascurabile.



WLS:

EQM Medio	Peso
6.7856e+05	1
7.4694e+05	2
4.0632e+05	4
3.4516e+05	8
3.1341e+05	16
2.9739e+05	32
2.895e+05	64
2.8574e+05	128
2.8406e+05	256
2.8336e+05	512
2.8309e+05	1024
2.8298e+05	2048

In conclusione il modello migliore che abbiamo trovato è quello con i beta stimati tramite WLS, dove le prime 150 osservazioni sono pesate 64, valore massimo da noi definito. Questo ci permette, noti i *ricoverati con sintomi*, il *totale positivi* e i *deceduti giornalieri*, di ottenere una buona stima dei ricoveri in terapia intensiva.