UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Case Study SMS2

Anno Accademico 2020/2021

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

INTRODUZIONE

Nella seconda parte dello studio del nostro dataset vogliamo studiare tramite modelli arima la variabile indipendente scelta nella prima parte del lavoro (**ricoverati in terapia intensiva**).
Dal precedente studio del dataset erano emerse alcune **problematiche** legate alle prime osservazioni dell'andamento pandemico. In particolare lo scarso numero di tamponi fatti aveva restituito un falso numero di positivi, mostrando poi una forte incongruenza con le terapie intensive.
Abbiamo così deciso di tralasciare in questa parte dello studio le **prime 150 osservazioni** (non solo nel regARIMA ma anche nell'arima semplicemente per scopi di confronto dei due modelli).

OBIETTIVO

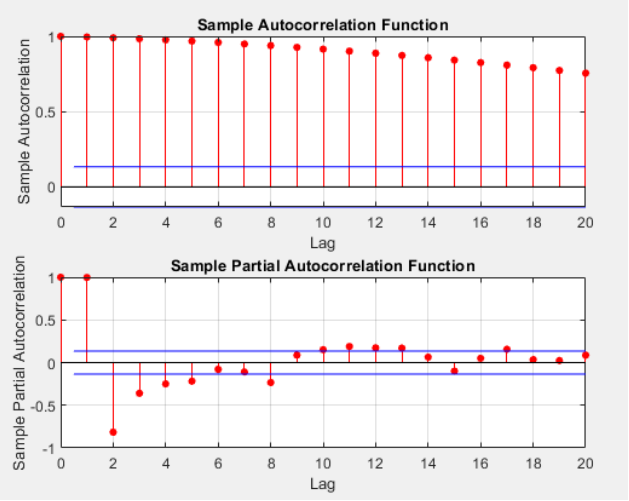
Lo studio svolto in questa seconda parte ha come scopo quello di effettuare una previsione temporale sui ricoveri in terapia intensiva, usando i metodi di modellazione visti a lezione.

ANALISI PRELIMINARE

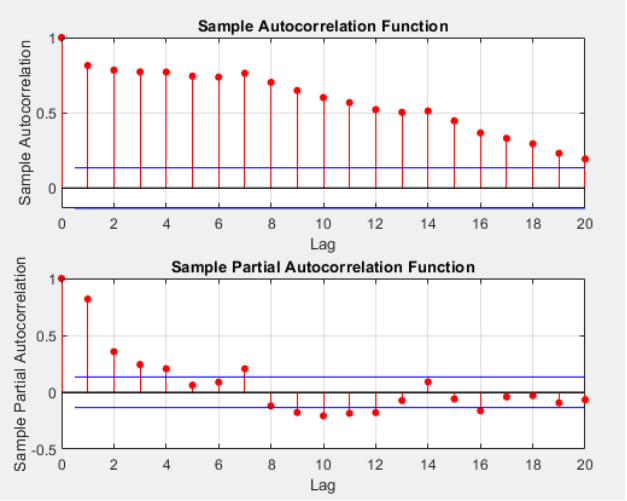
L'andamento totalmente stocastico dei ricoveri in terapia intensiva è modellizzabile tramite particolari tecniche di analisi, in particolare noi abbiamo iniziato lo studio con un semplice modello arima.
Prima di iniziare a spiegare nel dettaglio le tecniche utilizzate specifichiamo che sono state volutamente tralasciate le ultime hf osservazioni, così da effettuare un **training del modello** su una parte del dataset. Le ultime hf osservazioni saranno perciò usate per testare la **capacità previsiva** del modello trovato.
Nota: la parte che riguarda il caricamento del dataset e il suo filtraggio (prime 35 righe dello script) sono già state descritte nella prima parte del report.

ARIMA

Iniziamo con lo studio dell'autocorrelazione, dal grafico notiamo che le y sono **fortemente correlate tra loro**.



Prima della differenziazione



Dopo la differenziazione

Verificando la stazionarietà della nostra variabile indipendente, come era prevedibile osservando il grafico questo test **non viene superato**.
Decidiamo così di differenziare (**d = 1**), facendo di nuovo il test sulla stazionarietà, questa volta viene superato.
Per determinare i parametri p e q migliori utilizziamo un doppio ciclo for calcolando **25 modelli diversi** (p e q da 0 a 4) e per ognuno di questi effettuiamo test sui residui (se IID, normali, omoschedastici) e il calcolo dell'AIC.
Risultano essere tutti non IID, non normali e per arima(0,1,3) e arima(0,1,4) anche omoschedastici.

Tabella dei Test IID(0 -> IID, 1 -> non IID):

	q=0	q=1	q=2	q=3	q=4
p=0	1	1	1	1	1
p=1	1	1	1	1	1
p=2	1	1	1	1	1
p=3	1	1	1	1	1
p=4	1	1	1	1	1

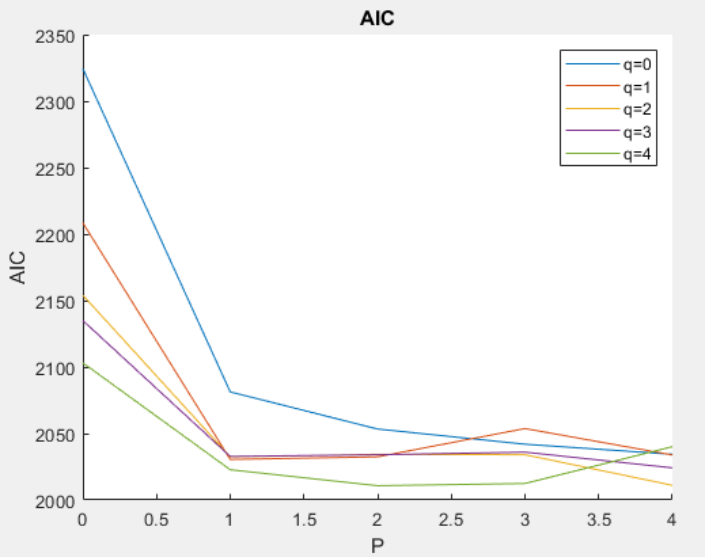
Tabella dei Test normalità(0 -> normale, 1 -> non normale):

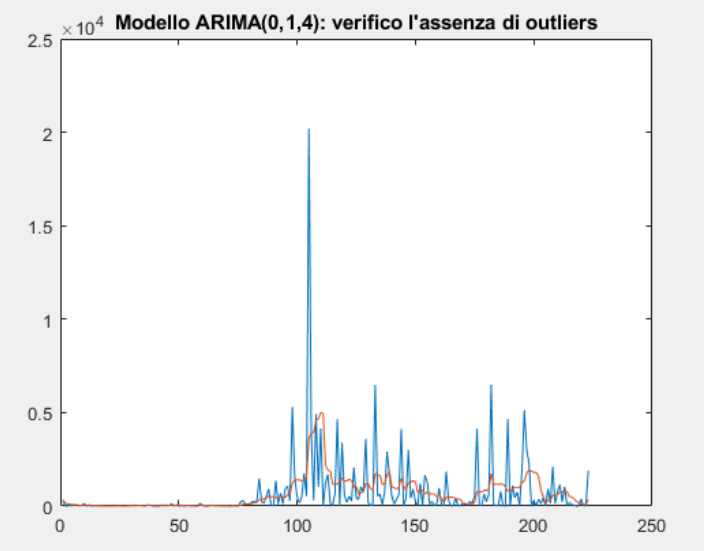
	q=0	q=1	q=2	q=3	q=4
p=0	1	1	1	1	1
p=1	1	1	1	1	1
p=2	1	1	1	1	1
p=3	1	1	1	1	1
p=4	1	1	1	1	1

Tabella dei Test omoschedasticità(0 -> omoschedastico, 1 -> eteroschedastico):

	q=0	q=1	q=2	q=3	q=4
p=0	1	1	1	0	0
p=1	1	1	1	1	1
p=2	1	1	1	1	1
p=3	1	1	1	1	1
p=4	1	1	1	1	1

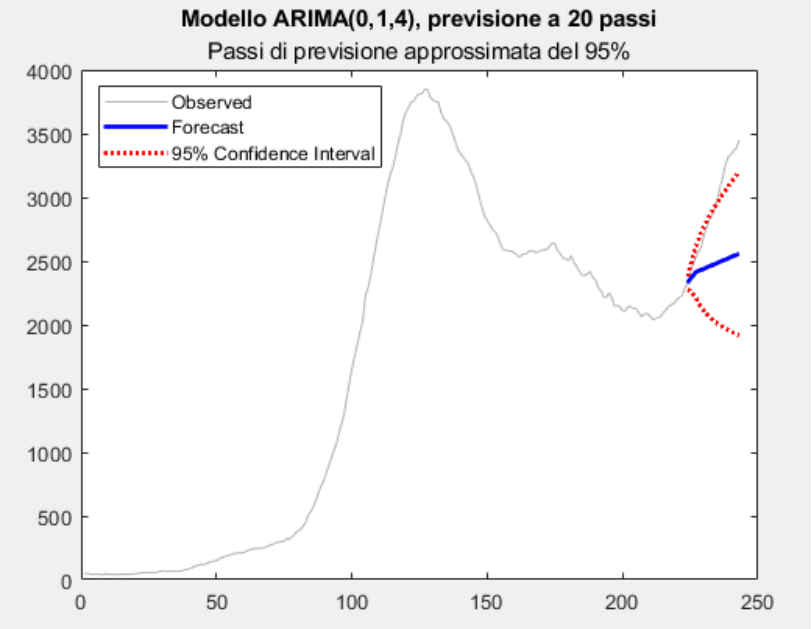
Dal prossimo grafico osserviamo l'andamento dell'AIC in funzione di p e q. Tenendo conto dei risultati ottenuti dai test sui residui e dal minimo valore di AIC scegliamo come modello **migliore arima(0,1,4)**.
Notare come questo non è un modello perfetto, infatti i suoi residui non hanno passato la maggior parte dei test che abbiamo svolto (per quanto riguarda il test sulla normalità il problema potrebbe essere relativo alla scarsità del numero di osservazioni, essendo queste meno di 250).
Effettuando il plot dei residui tramite l'utilizzo di movavg notiamo l'**assenza di outliners**, i valori che si discostano tanto dagli altri non sono infatti tenuti in considerazione nel grafico della media mobile.





Decidiamo poi di fare la previsione con il calcolo del relativo intervallo di confidenza (usando il dataset di training, con le osservazioni (150 : end-hf).
Dal grafico si evince che l'andamento della previsione è crescente come il reale andamento dei ricoveri in terapia intensiva. Questi ultimi sono contenuti nel IC della previsione fino a circa metà delle osservazioni.

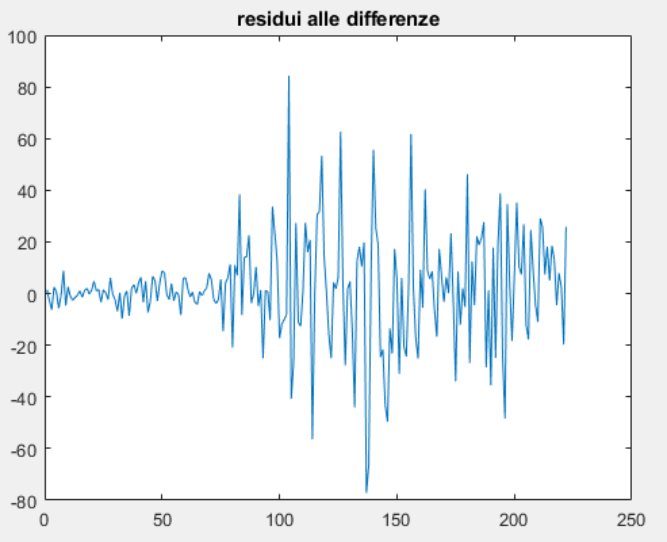
Notiamo come questo modello ha dei grossi **limiti** (soprattutto legati a orizzonti temporali alti).

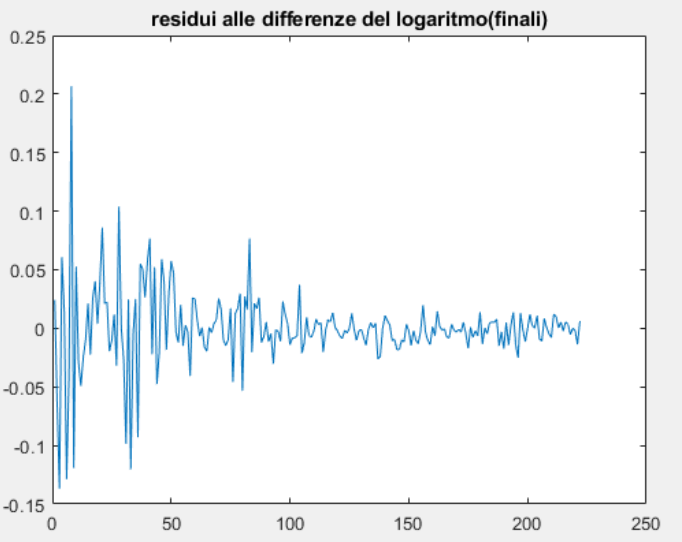


Per questo abbiamo deciso di proseguire il nostro studio modellando l'andamento dei ricoveri in terapia intensiva tramite regArima, usando come **parte deterministica** i regressori scelti nella prima parte dello studio.
Ricordiamo che i regressori (trovati nel nostro precedente studio) che spiegano meglio la variabile dipendente sono *ricoverati_con_sintomi*, *totale_positivi* e *deceduti_giornalieri*.

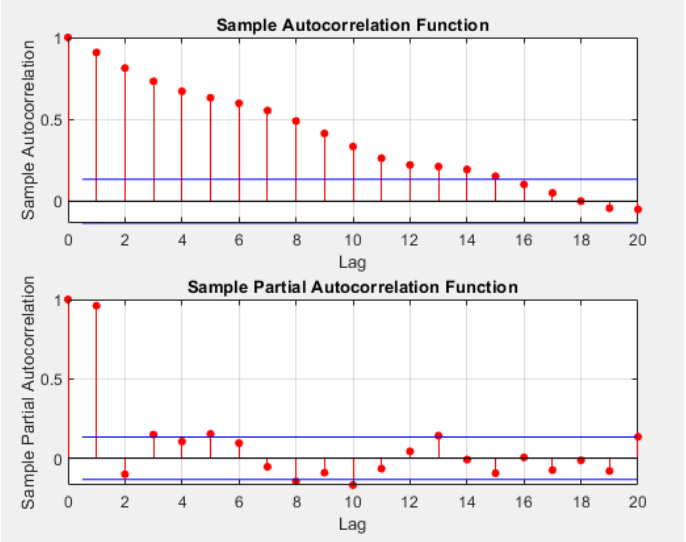
REGARIMA

Iniziamo effettuando la stima dei coefficienti di regressione e la verifica della stazionarietà dei residui di regressione. Notiamo come sia i residui, sia i regressori, sia la y non passano il test. Inoltre dal grafico della correlazione (vedi sotto) vediamo come i residui sono **fortemente correlati** fra loro, ossia c'è ancora una parte del dataset che può essere ancora spiegata.
Differenziando il dataset tutto diventa stazionario, effettuiamo poi di nuovo il calcolo dei residui di regressione con dataset differenziato.
Plottando i residui così ottenuti notiamo forte **eteroschedasticità**, osservazione confermata dall'archtest. Per cercare di risolvere usiamo la **funzione logaritmo** ai nostri dati (pur vedendo nel grafico un miglioramento, il test continua a confermare la loro eteroschedasticità)

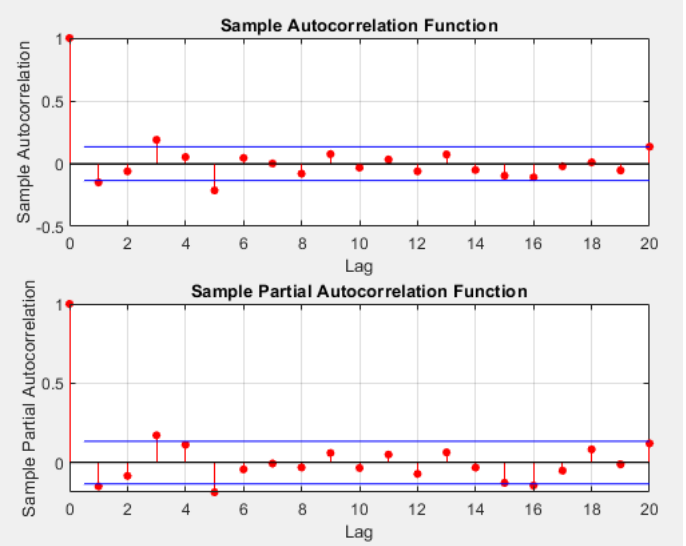




Dai grafici di autocorr e parcorr sui residui finali, prendiamo in considerazione come possibili modelli: arima(3,1,0), arima(5,1,0), arima(0,1,3) e arima(0,1,5) .



Residui iniziali



Residui finali

Stimiamo quindi i parametri di questi 4 modelli e calcoliamone i residui, così da fare inferenza su questi.

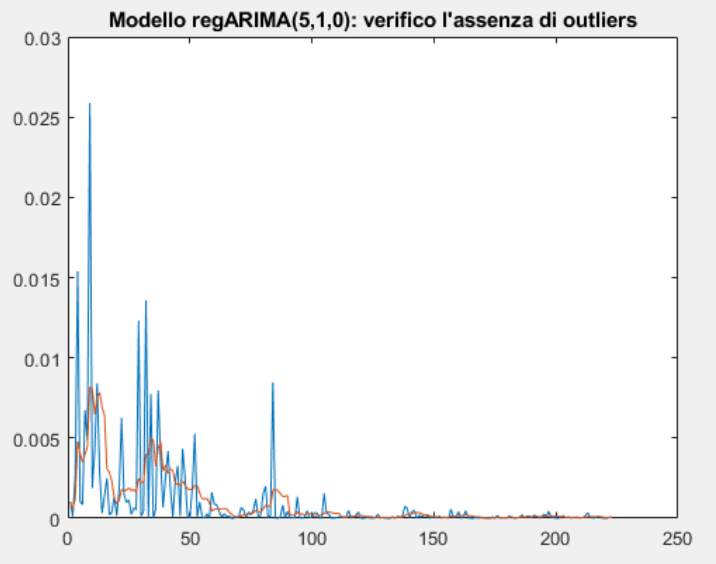
Diagnosi residui dei modelli:

	IID	Normalità	Omoschedasticità
regARIMA(3,1,0)	1	1	1
regARIMA(5,1,0)	0	1	1
regARIMA(0,1,3)	1	1	1
regARIMA(0,1,5)	0	1	1

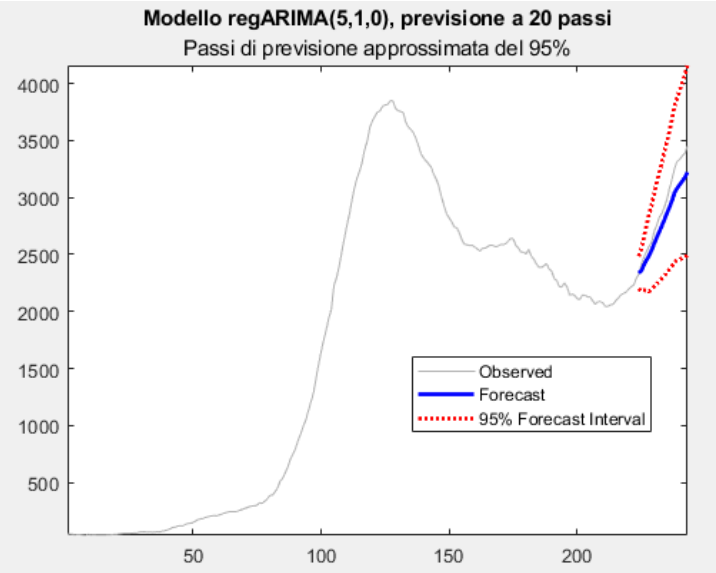
AIC sui modelli:

	AIC
regARIMA(3,1,0)	-892.6
regARIMA(5,1,0)	-902.07
regARIMA(0,1,3)	-892.91
regARIMA(0,1,5)	-899.09

In base ai risultati dei test e dal minimo valore di AIC scegliamo come modello arima(5,1,0). Notiamo inoltre come il modello scelto non presenta outliners significativi, come mostrato dal grafico moveavg.



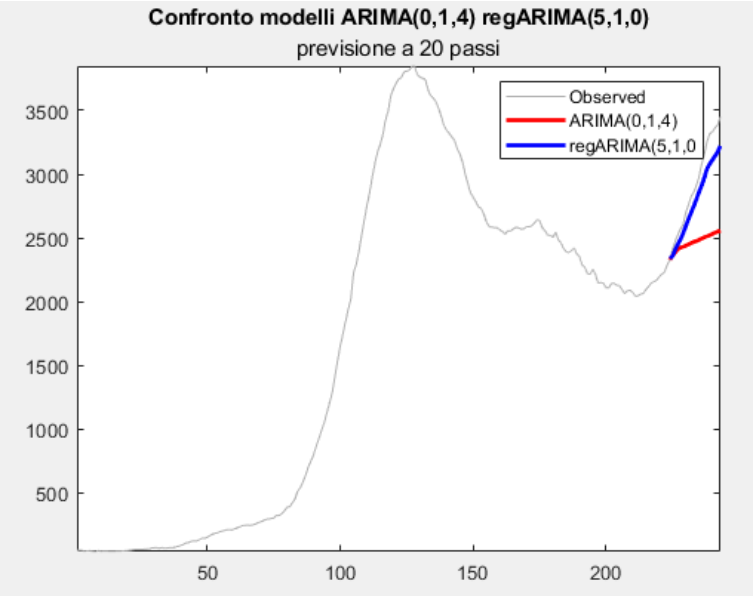
Effettuiamo poi la previsione usando questo modello.



Il comando forecast è stato applicato ai dati in formato logaritmico, passando come predittori previsivi gli ultimi hf valori delle nostre covariate (dataset di test).
Dal grafico notiamo l'alta precisione della previsione, che è molto vicina alle reali osservazioni.

CONFRONTO TRA PREVISIONE ARIMA E REGARIMA

Il confronto grafico fra gli andamenti previsivi trovati dai due modelli mostrano una significativa **differenza per lag ampi**. Il modello regArima mantiene l'andamento molto simile alla vera traiettoria lungo tutto l'orizzonte previsivo hf.
Questo non accade con il modello arima, dove la pendenza varia quasi subito, **discostandosi** molto dalle vere osservazioni dei ricoveri in terapia intensiva.



Possiamo concludere che il modello contenente la parte esogena/**regressiva** è risultato **migliore** durante il nostro studio. Per quanto riguarda i test sui residui il modello arima scelto ha residui omoschedastici (ma non IID e non normali), mentre il modello regArima scelto ha residui IID (ma non normali ed eteroschedastici).
La scarsa qualità dei residui è quindi un limite dei nostri modelli, che potrebbe essere superato usando modelli più complessi e specifici.