Analisi del prezzo dell'oro

in Relazione con alcuni commodities e indici

Andrea Moressa,1074124
Amin Borqal,1073928
Loris Iacoban,1074130

Anno accademico:
2021-2022

INTRODUZIONE:
Il dataset utilizzato per il progetto è stato preso dal sito Kaggle: **https://www.kaggle.com/code/somyaagarwal69/gold-forecasting-with-regression-arima/data?select=GoldUP.csv**

Il lavoro svolto consiste nel cercare di interpretare, attraverso modelli di regressione e stocastici, l'andamento del prezzo dell'oro,la sua relazione con i regressori presenti nel dataset e una sua previsione.

1. Parametri

Gold_Price : prezzo dell'oro in quel determinato mese

Crude_Oil : prezzo del petrolio in quel determinato mese

Interest_Rate o tasso di interesse : è l'importo che un finanziatore addebita a un mutuatario ed è una percentuale del capitale, l'importo prestato

USD_INR : tasso di cambio dei dollari in rupie

Sensex : si riferisce all'indice di riferimento della BSE (Bombay Stock Exchange) in India; è composto da 30 dei titoli più grandi e scambiati più attivamente e fornisce un indicatore dell'economia indiana

CPI o Consumer price index : dall'inglese, indice dei prezzi al consumo, è un indice che viene calcolato per mezzo di una media ponderata dei prezzi relativi ad un insieme di beni e servizi in un determinato periodo di tempo. Tale insieme è rappresentativo delle abitudini di spesa del consumatore medio. Il CPI è importante in quanto, misurando le variazioni dei prezzi, segnala l'aumento dell'inflazione

USD_Index : l'indice del dollaro; è una misura del valore del dollaro rispetto a un insieme di valute estere

2. Quesiti a cui rispondere

1. Come si comporta il prezzo dell'oro (**Gold_Price**) basandosi sugli altri attributi osservati?
2. Esiste multicollinearità nel modello selezionato al punto precedente (2.1)?
3. Nel modello selezionato al punto (2.1) i residui sono normali? Hanno media pari a zero?
4. Esiste una relazione tra il prezzo del petrolio (**Crude_Oil**) e l'indice del dollaro (**USD_Index**)?
5. È possibile utilizzare il metodo WLS per il modello selezionato al primo punto (2.1)?
6. È possibile fare una previsione dell'andamento del **Gold_Price** in relazione all'indice **CPI**?
7. Qual è il modello nello spazio degli stati migliore per il **Gold_Price**?Possiamo fare inferenza su questo modello?

2.1. Come si comporta il prezzo dell'oro (Gold_Price) basandosi sugli altri attributi osservati?

Per rispondere a questa domanda abbiamo analizzato singolarmente come gli attributi si comportassero con **Gold_Price.**

Regressori	R2	MSE	pValue
Crude_Oil	0.438	8.77e+03	1.78e-31
Interest_Rate	0.0586	1.13e+04	1.58 10^4
USD_INR	0.732	6.05e+03	1.05e-69
Sensex	0.805	5.15e+03	3.17e-86
CPI	0.92	3.31e+03	7.83e-132
USD_Index	0.0176	1.16e+04	0.0402

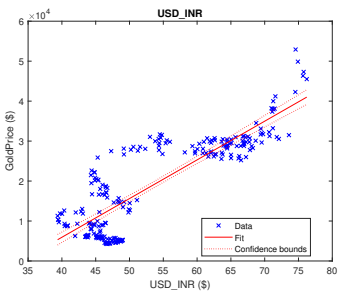
Attraverso la tecnica Stepwise backward selezioniamo il modello migliore che ha come regressori (**USD_INR**, **Sensex**, **CPI**), abbiamo poi applicato un modello di regressione lineare multipla.

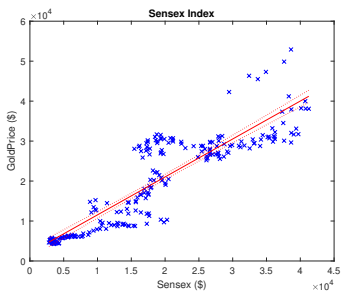
Regressori	R2	MSE	pValue
CPI + Sensex + USD_INR	0.953	2.53e+03	3.65e-156
CPI + Sensex	0.927	3.17e+03	8.96e-135
CPI + USD_INR	0.929	3.13e+03	3.85e-136
USD_INR + Sensex	0.841	4.67e+03	4.31e-95

Dopo l'analisi del coefficiente di determinazione multipla e l'errore quadratico medio, abbiamo deciso di concentrarci sul modello **CPI + Sensex + USD_INR** in quanto il il valore di R2 risulta essere il più alto. Ecco la tabella finale dei parametri $\hat{\beta}$, con statistiche test :

	Regressore	Valore stimato	SE	Statistica T	pValue
β**0**	Intercetta	2437.2	1372.7	1.7755	0.077111
β**1**	CPI	830.46	34.92	23.782	1.741e-64
β**2**	Sensex	-0.69783	0.062574	-11.152	1.8518e-23
β**3**	USD_INR	-583.4	50.383	-11.579	7.8805e-25

Dato che i parametri indipendenti sono tre non è stato possibile il grafico dell'iperpiano di regressione. Sarebbero stati in quattro dimensioni e di conseguenza non rappresentabili. Andiamo quindi a mostrare i grafici parziali per avere un'idea dell'andamento :



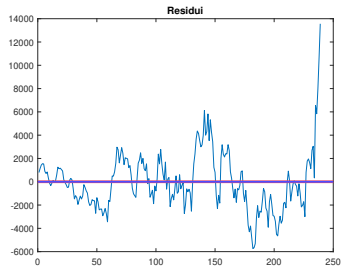


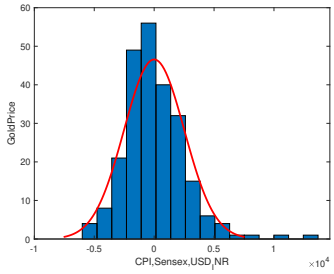
2.2. Esiste multicollinearità nel modello selezionato al punto precedente (2.1)?

La multicollinearità sorge quando c'è un'elevata correlazione tra due o più variabili esplicative, che porta lo stimatore dei minimi quadrati a non dare risultati attendibili. Se c'è una correlazione perfetta tra le due variabili, la matrice (X'X) diventa singolare, ha determinante uguale a 0 circa e perciò non esiste la matrice inversa. Abbiamo quindi verificato che fosse maggiore di 0 per assicurarci che non vi fosse multicollinearità.

2.3. Nel modello selezionato al punto (2.1) i residui sono normali? Hanno media pari a zero?

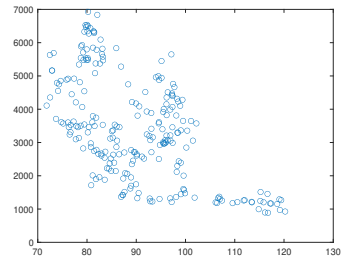
I residui sono la differenza tra i valori osservati e stimati in un'analisi di regressione. L'iperpiano di regressione deve trovarsi lungo il centro dei punti dati. Pertanto la somma dei residui deve essere zero. Abbiamo verificato che i residui avessero una distribuzione normale e media uguale a zero.

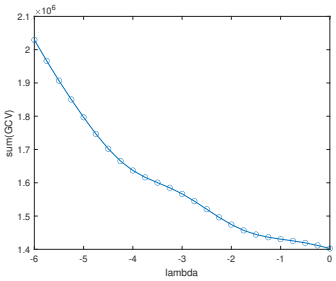




2.4. Esiste una relazione tra il prezzo del petrolio (Crude_Oil) e l'indice del dollaro (USD_Index)?

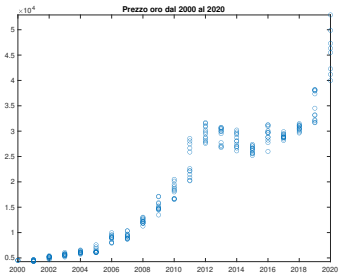
Come si può vedere dal primo grafico, abbiamo una correlazione non lineare tra **Crude_Oil** e **USD_Index** con dati abbastanza dispersi. Per modellizzare questa relazione abbiamo utilizzato una regressione polinomiale ovvvero la regressione spline. Abbiamo utilizzato l'algoritmo di crossvalidazione con k-fold per lo smoothing della funzione, una gestione ottimale del numero di nodi e per trovare il miglior parametro di lisciamento. Abbiamo Calcolato λ e la matrice di penalizzazione Rmat (derivante dalla penalizzazione) per penalizzare il modello, l'ordine è uguale a 6 e nbasis=243. Il lambda ottimale, da noi ottenuto, vale 0.01. Lo smoothing ottenuto tramite crossvalidazione e la funzione stimata le vediamo nei seguenti grafici :

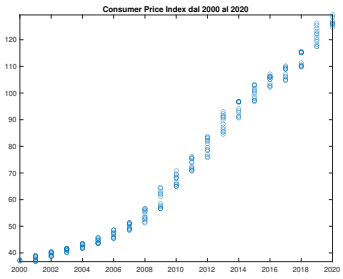


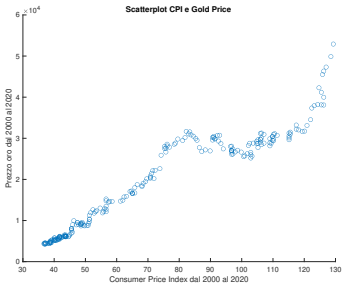


3. Previsione Gold_Price in relazione a CPI mediante l'uso del modello di regressione lineare con errori ARIMA

Come primo passo si costruiamo la regressione lineare semplice con variabile indipendente Yt (**Gold_Price**) ed il regressore in t **CPI** (t = tempo). Successivamente, effettuiamo la modellazione ARIMA(p, q) dei redisui. Abbiamo osservato dal dataset che il **Gold_Price** ed il **CPI** hanno entrambi un trend lineare (primi 2 grafici)e possiamo subito intuire che la media cresce all'aumentare del tempo e questo ci indica che non c'è stazionarietà. Dall'ultimo grafico invece vediamo la forte correlazione tra i due. Vogliamo fare una previsione sull'andamento del trend.





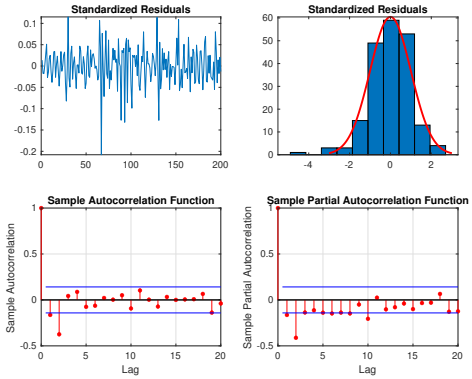


3.1. Regressione con errori ARIMA (regARIMA)

Sfruttiamo le funzioni matlab regARIMA ed estMdl per fare la regressione e stimarne i parametri(tenendo conto anche della differenziazione perché da un'analisi dei redidui preliminare quest'ultimi non sono stazionari). I 3 modelli trovati sono addestrati su 200 osservazioni: AR(1),MA(1),ARMA(1,1).

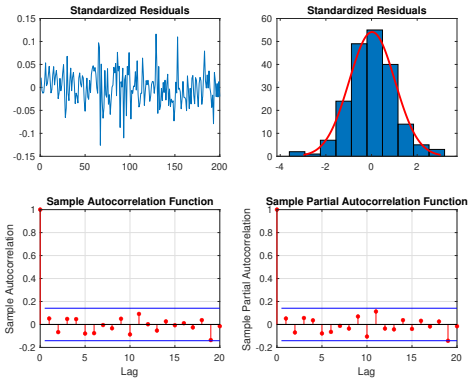
3.1.1 Modello AR(1)

	Valore stimato	SE	Statistica T	pValue
Intercetta	1.1673e-05	0.0022721	0.0051374	0.9959
AR(1)	-0.44061	0.060067	-7.3353	2.2126e-13
β**1**	-0.71563	0.39046	-1.8328	0.066836
Variance	0.0019462	0.00013972	13.929	7.8805e-25



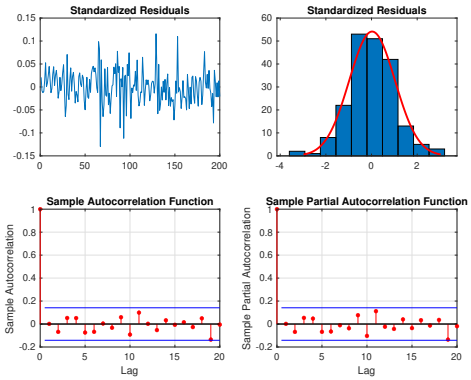
3.1.2 Modello MA(1)

	Valore stimato	SE	Statistica T	pValue
Intercetta	2.2846e-05	0.00015098	0.15132	0.87972
MA(1)	-0.94922	0.021836	-43.471	0
β**1**	-0.34819	0.38337	-0.90825	0.36375
Variance	0.0013427	0.00011997	11.192	4.4749e-29



3.1.3 Modello ARMA(1,1)

	Valore stimato	SE	Statistica T	pValue
Intercetta	2.3595e-05	0.00014933	0.15801	0.87445
AR(1)	0.058927	0.072006	0.81835	0.41315
MA(1)	-0.95366	0.023579	-40.445	0
β**1**	-0.36078	0.38738	-0.93134	0.35168
Variance	0.0013386	0.00011962	11.19	4.542e-29

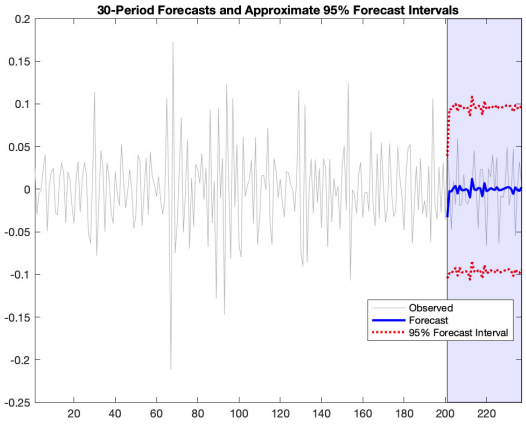


3.2. Analisi dei residui e scelta del modello migliore

I residui hanno media zero,si distribuiscono secondo una Gaussiana e hanno una varianza stabile. Dai grafici di autocorrelazione vediamo che i residui ora hanno un comportamento white noise. I grafici dell'autocorrelazione parziale ci dicono che i residui sono stazionari .

3.3. Forecast

Per scegliere il modello migliore facciamo una stima dell'RMSE di previsione sulle restanti 37 osservazioni e scegliamo il modello con il valore più basso: ARMA(1,1) RMSE=0.8882. Usiamo questo modello per fare la previsione.



4. Previsione mediante l'utilizzo di un Modello state-space

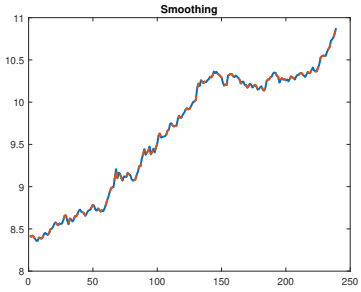
Lo scopo dell'analisi state space è stimare le proprietà rilevanti delle variabili partendo dalla conoscenza delle osservazioni. Lo stato non lo osserviamo direttamente, ma indirettamente tramite un altro vettore che indicheremo con Yt che in generale può avere dimensione diversa dal vettore dello stato. Abbiamo due equazioni: una ci descrive lo stato del sistema l'altra ciò che osserviamo di quel sistema, o meglio, come osserviamo. Idealmente vogliamo conoscere la x, ma la osserviamo tramite la y. Abbiamo costruito 3 diversi modelli in MatLab:Modello locale a livello stocastico,modello locale deterministico e modello di sjdhdhfuwuf. In base all'analisi dei residui ed all'AIC abbiamo scelto come modello migliore quello locale a livello stocastico .

4.1. Smoothing:

L'operazione di smoothing consente di stimare attraverso l'applicazione ricorsiva del Kalman Filter i vettori di stato , la varianza di stato , gli errori delle osservazioni e gli errori di stato. Pertanto, l'operazione di smoothing consente di stimare tutte le variabili incognite del modello state space.

$$x1(t) = x1(t-1) + (0.04)u1(t) \tag{1}$$

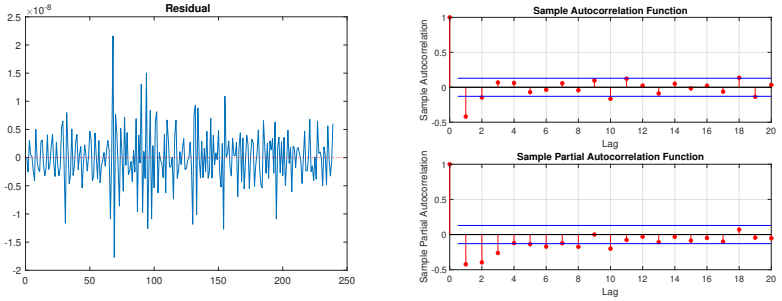
$$y1(t) = x1(t) + (1.17e-05)e1(t) \tag{2}$$



ci ha dato un grafico più "allisciato" perché tiene conto del fatto che c'è un errore di osservazione.

4.2. Analisi dei residui:

I residui hanno media zero e varianza stabile ed inoltre dal grafico si può intuire che sono stazionari ed effettivamente lo sono da quanto abbiamo riscontrato dal kpss test. Il grafico di autocorrelazione indica che dal lag 1 i residui assumono un comportamento white noise mentre quello di autocorrelazione parziale ci conferma la stazionarietà.



5. Conclusioni:

Nella prima parte del report con delle tecniche di regressione lineare multipla abbiamo verificato la forte correlazione esistente tra il **Gold_price** e gli indici/commodities **CPI, Sensex, USD_INR**. Abbiamo deciso inoltre di applicare la regressione spline con tecniche di crossvalidazione tra il **Gold_price** e **USD_Index** per evitare overfitting e trovare un modello che "spiegasse" il piú possibile una correlazione non lineare. Infine abbiamo cercato di migliorare il modello di regressione lineare multipla cercando un iperpiano di regressione che "spiegasse più accuratamente" i dati con l'utilizzo dei minimi quadrati ponderati. Nella seconda parte del report i 2 modelli stocastici che abbiamo usato ci hanno consentito di fare una previsione del **Gold_Price** usando un modello di regressione lineare semplice con errori ARIMA. Il grafico di previsione ottenuto è frutto della combinazione della parte regressiva con la parte ARIMA. Nell'ultimo punto che riguarda lo state space model abbiamo applicato lo smoothing per ottenere delle stime degli errori e con queste abbiamo ottenuto un grafico del **Gold_Price** attenuato dai rumori.