Previsione di una serie temporale

Editor: Adnan Sardi

Abstract

Questo report cerca di dare una previsione quantitativa di una serie storica con l'utilizzo del software R studio. La previsone di una serie storica è un processo che richiede lo studio tramite strumenti statistici e matematici di una serie di dati storici per poter poi effettuare una previsione futura. In questo report analizzeremo le azioni di Walmart dal 2001 al 2020.

1. Presentation data

La prima parte prevede di raccogliere i dati temporali della società. Ci sono molti siti che concedono tale possibilità, in questo abbiamo utilizzato la piattaforma *Investing.com* esportando i dati in forma .csv.
Nella cartella *Datas* appartenente dalla directory *Forecast-Walmart-s-stock-with-ARMA-model* possiamo trovare due file .csv. Il primo *WMT Cronologia Dati_unprocessed.csv* è il dataset non ancora processato per l'analisi. Al suo interno possiamo trovare più colonne che andiamo a spiegare qui sotto:

- **Data**, indica la data a cui fa riferimento ogni valore. Si è scelto di prendere sempre il primo giorno di ogni mese.
- **Ultimo**, indica il valore della singola azione alla fine della giornata di quotazione in borsa.
- **Apertura**, indica il valore della singola azione all'apertura del mercato.
- **Massimo**, indica il valore massimo che ha raggiungo la singola azione durante tutta la giornata.
- **Minimo**, indica il valore minimo che ha raggiungo la singola azione durante tutta la giornata.
- **Vol.**, inidca il numero di titoli o contratti scambiati in quella giornata.
- **Var.%**, indica la variazione del titolo dall'ultimo valore che esso possedeva, in questo caso la variazione da mese in mese.

Inoltre della directory oltre a questo dataset possiamo trovare *WMT Cronologia Dati_clearnly.csv* cioè quello che poi è stato effettivamente caricato su R per l'analisi.

2. Data cleaning

La fase di data cleaning è stata suddivisa in due parti.
La prima è stata effettuata su *Excel*, un operazione di preparazione del dataset affinchè R riuscisse a leggere i dati csv in ordine. La seconda invece è stata effettuata su R dove principalmente abbiamo creato una tabella extra che poi è stata chiamata *Value reverse* dove non è altro che il valore delle azioni ma ordinate cronologicamente al contrario. Il motivo di questa scelta è stata riportanta nei paragrafi sucessivi a questo.
Per gli scopi della nostra analisi non abbiamo preso in considerazione molti dati estrapolati ma ci siamo limitati ad analizzare la serie temporale con il solo valore di chiusura (**Ultimo**).
Ovviamente le operazioni effettuate sono disponibili nella sezione **Code**. Mentre per la parte di *Excel* si è usato principalmente l'opzione *Filtro*.

3. Exploration analysis

Per l'esplorazione del data frame R studio mette a disposizione molte funzioni come *view, head, tail, class* (queste quelle maggiormente utilizzate in questa analisi).
Proprio con la funzione class otteniamo l' output *"data.frame"*, che potrebbe darci problemi qual ora vogliamo effettuare un'analisi di una serie temporale. Per evitare questo abbiamo utilizzato la funzione chiave **ts()** che converte il nostro dataframe in una time series.
Questo però non è stato immediato poichè nel farlo si creavano problemi con l'assegnazione del valore nel periodo temporale esatto, per questo motivo si è deciso di creare una colonna che invertisse i valori di chiusura. Nel codice si è lasciato un piccolo frame che indica un altro modo di creare una serei storica dato un data frame.
La time series ottenuta è la seguente:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov
2001	56.80	50.09	50.50	51.74	51.75	48.80	55.90	48.05	49.50	51.40	55.15
2002	59.98	62.01	61.30	55.86	54.10	55.01	49.18	53.48	49.24	53.55	53.90
2003	47.80	48.06	52.03	56.32	52.61	53.67	55.91	59.17	55.85	58.95	55.64
2004	53.85	59.56	59.69	57.00	55.73	52.76	53.01	52.67	53.20	53.92	52.06
2005	52.40	51.61	50.11	47.14	47.23	48.20	49.35	44.96	43.82	47.31	48.56
2006	46.11	45.36	47.24	45.03	48.45	48.17	44.50	44.72	49.32	49.28	46.10
2007	47.69	48.31	46.95	47.92	47.60	48.11	45.95	43.63	43.65	45.21	47.90
2008	50.74	49.59	52.68	57.98	57.74	56.20	58.62	59.07	59.89	55.81	55.88
2009	47.12	49.24	52.10	50.40	49.74	48.44	49.88	50.87	49.09	49.68	54.55
2010	53.43	54.07	55.60	53.64	50.56	48.07	51.19	50.14	53.32	54.17	54.09
2011	56.07	51.98	52.05	54.98	55.22	53.14	52.71	53.19	51.90	56.72	58.90
2012	61.36	59.08	61.20	58.91	65.82	69.72	74.43	72.60	73.80	73.02	72.02
2013	69.95	70.78	74.83	77.72	74.84	74.49	77.94	72.98	73.96	76.75	81.01
2014	74.68	74.70	76.43	79.71	76.77	75.07	73.58	75.50	76.47	76.27	87.54
2015	84.98	83.93	82.25	78.05	74.27	70.93	71.98	64.73	64.84	57.24	58.84
2016	66.36	66.34	68.49	66.87	70.78	73.02	72.97	71.44	72.12	70.02	70.43
2017	66.74	70.93	72.08	75.18	78.60	75.68	79.99	78.07	78.14	87.31	97.23
2018	106.60	90.01	88.97	88.46	82.54	85.65	89.23	95.86	93.91	100.28	97.65
2019	95.83	98.99	97.53	102.84	101.44	110.49	110.38	114.26	118.68	117.26	119.09
2020	114.49	107.68	113.62	121.55	124.06	119.78	129.40	138.85	139.91	138.75	152.79
2021	140.49										

Figure 1: Serie storica di Walmart [2001-2021]

Vedendo i dati in possesso si vede subito come con il tempo il valore di Walmart sia sempre più cresciuto.
Plottando il grafico otteniamo un'altra conferma di ciò.

Dal grafico soprastante si vede come il valore di Walmart è stato costante nel periodo 2001-2012 per poi crescere fino a fine 2015, anno in cui il valore di Walmart è sceso del

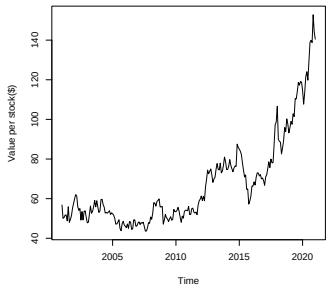


Figure 2: Serie storica di Walmart [2001-2021]

30%. Dopo la decrescita durante il periodo 2015-2016 la sua crescita è stata esponenziale fino ai giorni d'oggi. Per avere una visione più chiara dell'andamento possiamo utilizzare la funzione *aggregate* che aggrega i dati a livello annuale, inoltre si aggiunge un *boxplot* che riepiloga i valori mensilmente.

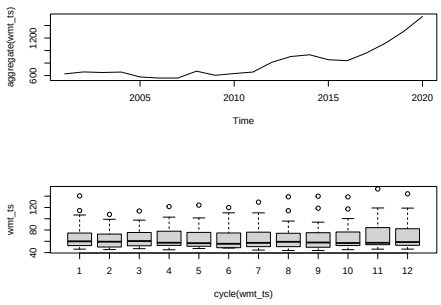


Figure 3: Trend Walmart

Dal primo grafico confermiamo le analisi fatte precedentemente mentre con il grafico box plot possiamo notare come nel mese di Novembre e Dicembre abbiamo più volatilità, questo potrebbe essere causato da una spesa maggiore per via delle festività di quel periodo.

4. Test and trasformation

Una serie storica può essere scritta come:

$$Y_{\mathrm{t}} = f(t) + u_{\mathrm{t}} \tag{1}$$

Dove f(t) rappresenta la parte deterministica della serie storica composta da trend, stagionalità e ciclo mentre u(t) è la componente stocastica. Non sempre la parte deterministica riesce a darci informazioni rilevanti perciò si è deciso di focalizzare la nostra analisi sulla componente stocastica. Effettuiamo una decomposizione della serie con la funzione *decompose* e poi plottiamo i grafici ottenuti.

Adesso abbiamo la nostra serie storica composta solo dalla parte stocastica, cioè il grafico con il valore **random** sull'asse delle ordinate. Un metodo per studiare la parte stocastica

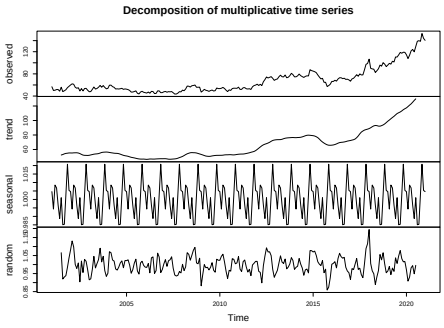
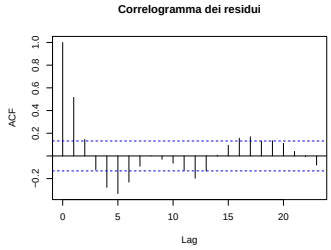
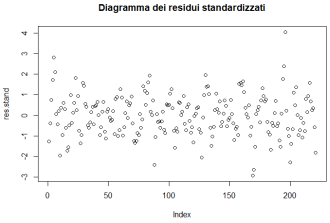


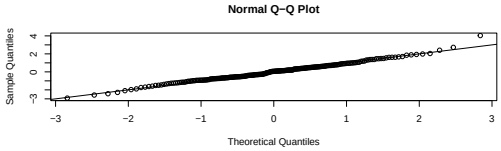
Figure 4: Decomposizione moltiplicativa

è quella di assumerla come un processo a componenti correlati, per questo utilizziamo la funzione ACF e plottiamo il grafico.

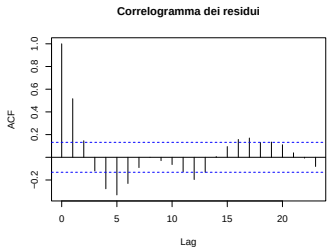


Possiamo effettuare delle prime analisi sul correlogramma ottenuto. L'asse delle ordinate indica quanto una componente k-esima è in relazione con la k-esima + 1 mentre sull'asse delle ascisse abbiamo i pedici di ogni serie storica u(t). Nel grafico si può notare due linee tratteggiate di colore blu che danno un indicazione sul valore di riferimento per capire se due valori di u(t) sono correlati o meno.
I valori non sembrerebbero essere troppo correlati tra loro, infatti solo 8 valori su circa 23 superano la linea tratteggiata.
Si può comunque notare come i valori che superano la banda di riferimento ritornano con una frequenza di k = k + 6/7. Verifichiamo ora se i nostri dati si distribuiscono normalmente. Usiamo i loro residui standardizzati. Qui di seguito vengono riportati i grafici della distribuzione dei residui e il QQ-plot, il codice utilizzato è disponibile sempre nella sezione *Code* della directory principale.





Entrambi i grafici non presentano difformità rilevabili per quanto riguarda la normalità. Per confermare la normalità possiamo utilizzare il test di **Shapiro-Wilk**. Il valore ottenuto da tale test è un **p-value = 0.1724**.
Il valore ottenuto è maggiore rispetto a quello di riferimento di 0.05 quindi ciò ci fa propendere per l'ipotesi nulla ovvero la normalità degli errori e quindi confermano le ipotesi dedotte dai due grafici precedenti.
Plottiamo ora il correlogramma dei residui per poi effettuare altre analisi con test statistici.

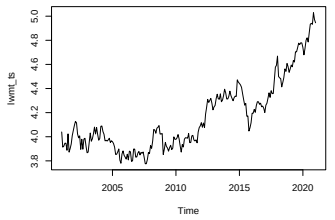


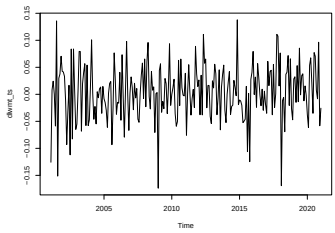
Vediamo subito come l'andamento è pressochè simile al correlogramma precedente. Ora possiamo utilizzare il test di **Ljung-Box** e di **Box-Pierce** per verificare l'assenza di correlazione. I valori ottenuti sono i seguenti

- p-value **Ljung-Box** = 2.2e-16
- p-value **Box-Pierce** = 2.2e-16

Valori bassi del p-value ci consente di rifiutare l'ipotesi nulla cioè di avere residui non autocorrelati.

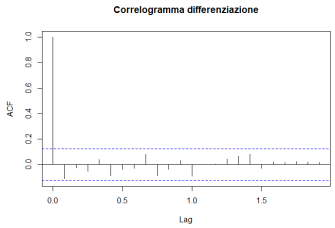
Proviamo ad effettuare delle manipolazioni sui dati per eliminare la correlazione tra i residui. Prima effettuiamo una trasformazione logaritmica per ridurre la varianza poi effettuiamo un operazione di differenziazione per eliminare il trend. I risultati sono i seguenti.





I valori di p-value dei due test risultano ora uguali a 0.3467 e 0.3784, maggiori rispetto ai valori precedenti mentre il correlogramma associato non presenta valori che superano il livello di confidenza del 95%.

Altri due test che possono essere effettuati sono quelli di **Dickey-Fuller** e di **Phillips-Perron**.



5. Model predict

Ad una serie storica è possibile stimare un modello che può essere a media mobile, auto regressivo oppure entrambi. Dopo aver individuato più modelli possiamo utilizzare la funzione AIC, dove il miglior modello è quello che minimizza proprio l'AIC. Nel nostro caso il miglior modello è un modello ARMA(1,0). Analizziamo ora il qq-plot dei residui del nostro modello.

I residui sembrerebbero distribuirsi bene se non per la coda inferiore. Ripetendo i test sulla normalità otteniamo un p-value di 0.1443 per il test di Shapiro-Wilk mentre 0.03698 per quello di Jarque Bera. Quest'ultimo risultato ci interroga sulla normalità del nostro modello e di come la parte iniziale del nostro qq-plot non sia dovuto al caso.
Possiamo utilizzare ora la funzione *predict* per poter fare una predizione del titolo in base al modello trovato.

Si evice subito come la predizione è influenzata negativamente da qualche fattore non considerato. Il titolo dovrebbe decrescere fino a rimanere costante, cosa che non è successa.

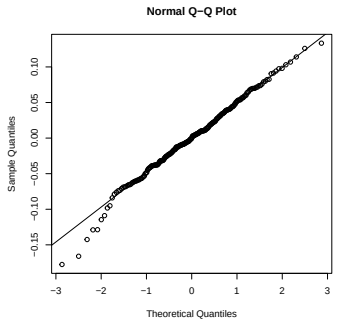


Figure 5: QQ plot ARMA(1,0)

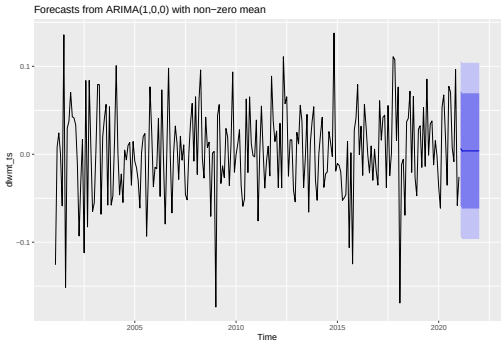


Figure 6: Predizione azioni Walmart

6. Conclusion

Lo scopo della nostra analisi era quella di capire come poter utilizzare R studio per lo studio di una serie temporale per darne una predizione finale. Il modello non sembra avere grosse problematiche se non per la parte finale dove la predizione è completamente errata.
Alcuni accorgimenti utili per il miglioramento del modello possono essere una maggior acquisizione di dati, questo perchè le serie storiche funzionano meglio se abbiamo un dataset corposo. Altro accorgimento che è saltato subito all'occhio è la distribuzione dei residui del miglior modello che non è ottimale. Una seconda differenziazione come la prima potrebbe essere una soluzione.
Per stimare la bontà del nostro modello può essere utilizzato l *RSME*, nel codice è stato annotato come commento.
Per concludere abbiamo costruito un modello che può essere una base per un suo futuro miglioramento, miglioramento che deve tener anche conto delle oscillazioni del titolo dovuto a fattori esterni come avvenimenti geopolitici o notizie macroeconomiche. Il fatto che il titolo abbia avuto un crollo a fine 2015 e poi un boom nel 2020 sono dovuti appunto a fattori esterni che non sono stati tenuti conto.