Analiza cen akcji Netflixa

Michał Wiktorowski

9.02.2023

1 Wstęp

W niniejszej pracy przedstawię zastosowanie modelu ARMA (ang. Autoregressive Moving Average) w analizie szeregów czasowych. Dane które będą rozpatrywane dotyczą cen akcji Netflixa w latach 2002-2022. Plik z danymi można znaleźć tutaj.

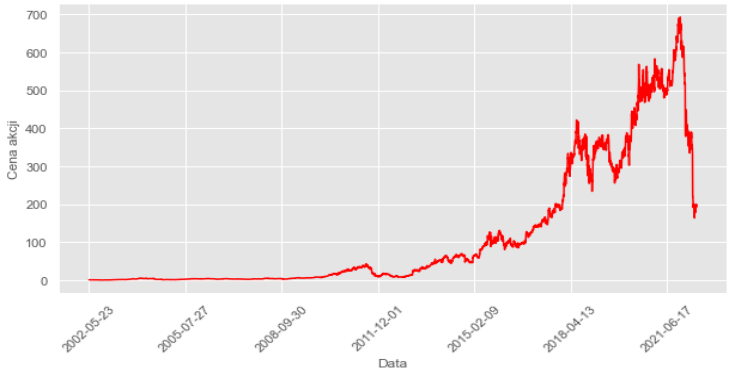
Tabela danych zawierała kilka kolumn, między innymi datę, ceny otwarcia i zamknięcia, a także liczbę wykonanych tranzakcji. Postanowiłem poddać analizie zachowanie ceny akcji od czasu (pod uwagę wziąłem ceny otwarcia).

2 Przygotowanie danych

Szereg $\{X_t\}_{t\in\mathcal{Z}}$ jest szeregiem ARMA(p,q), jeśli spełnia równanie:

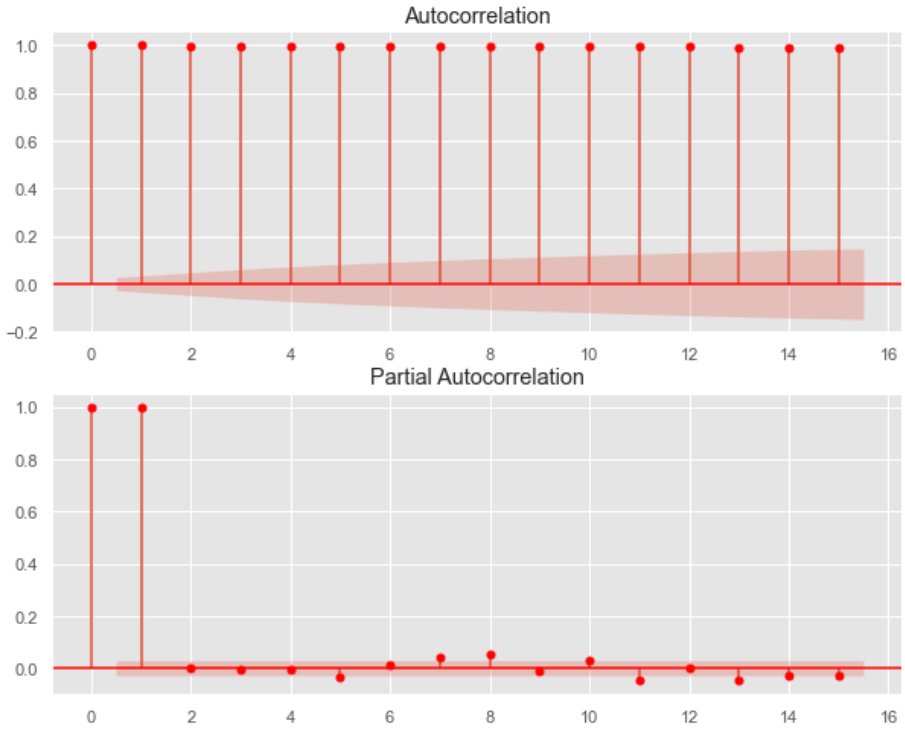
$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = Z_t + \sum_{i=1}^{q} \theta_i Z_{t-i} \qquad \{Z_t\} \sim WN(0, \sigma^2)$$

Bardzo istotnym założeniem o X_t jest słaba stacjonarność, czyli średnia szeregu w czasie musi być stała, a funkcja autokowariancji ma zależeć tylko od "lagu". Bez tego nie możemy mówić o modelu ARMA.



Rysunek 1: Wykres cen akcji Netflixa od czasu

Na powyższym wykresie wyraźnie widać, że średnia szeregu nie jest stała w czasie. Żeby to potwierdzić, zobaczymy jak wyglądają wykresy autokowariancji i częściowej autokowariancji dla surowych danych. Ponadto skorzystamy z testu Dickey-Fullera (ADF test), który bada obecność pierwiastka jednostkowego w szeregu. Jeśli takowy istnieje, to możemy przyjąć z pewnym prawdopodobieństwem, że szereg jest stacjonarny (domyślnie przyjmiemy prawdopodobieństwo $1 - \alpha = 0,95$). Więcej o teście ADF można przeczytać tutaj.



Rysunek 2: Wykresy autokowariancji i częściowej autokowariancji. Wyraźnie widać silną korelację danych.

Jak możemy zauważyć, wykresy ACF i PACF osiągają wręcz olbrzymie wartości. Ponadto test ADF zaobserwował statystykę, której wartość jest znacznie większa od poziomu istotności. Badany przez nas szereg na pewno nie jest stacjonarny w słabym sensie. Wykonamy teraz odpowiednie kroki, które opisuję poniżej, w celu transformacji naszego szeregu na szereg stacjonarny.

2.1 Stabilizacja wariancji

Na wykresie cen akcji możemy zaobserwować nagłe zmiany wartości. Oznacza to, że ich wariancja zmienia się w chaotyczny sposób. W celu jej stabilizacji, użyjemy na naszych danych transformaty Boxa-Coxa. Ma ona na celu sprowadzenie rozkładu danych do takiej postaci, która jest jak najbliższa rozkładowi normalnemu. Jest ona zadana przekształceniem

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & y \neq 0 \\ \log(y) & y = 0 \end{cases}$$

gdzie λ jest najbardziej optymalnym parametrem przekształcenia. W naszym przypadku $\lambda \approx -0,0214$.

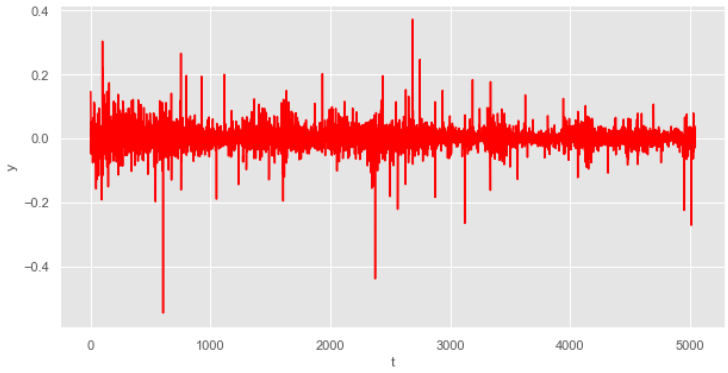
Rysunek 3: Wykres cen akcji po zastosowaniu transformaty Boxa-Coxa

2.2 Dekompozycja trendu

Aby usunąć wyraźny trend skorzystamy z tzw. metody różnicowania trendu. Polega ona na utworzeniu nowego szeregu czasowego według schematu:

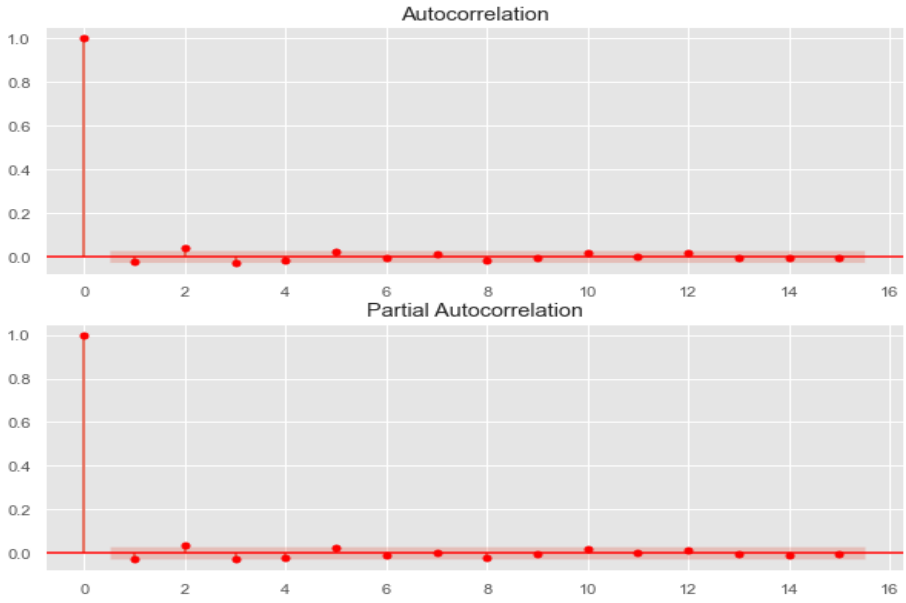
$$Y_t = X_t - X_{t-1}$$

Jest to nowy ciąg zmiennych losowych, którego realizacje są przyrostami w kolejnych punktach czasowych. Rezultat możemy zaobserwować na poniższym wykresie.



Rysunek 4: Dekompozycja trendu. Widać wyraźną zmianę zachowania szeregu typową dla szeregu słabo stacjonarnego

Wynik jest bardzo obiecujący - po dekompozycji trendu, nasz szereg wygląda, jakby spełniał warunki słabej stacjonarności. Potwierdźmy to jeszcze testem Dickey-Fullera, oraz wykresami ACF i PACF.



Rysunek 5: Wykresy ACF i PACF zdekompowanego szeregu czasowego. Widać wyraźną poprawę do stanu sprzed dekompozycji

P-wartość w ADF teście jest równa 0 i jest mniejsza od zadanego poziomu istotności ($\alpha = 0.5$). Możemy zatem przyjąć, że otrzymany szereg jest stacjonarny.

2.3 Dekompozycja sezonowości

Dekompozycja sezonowości wygląda bardzo podobnie do usuwania trendu. Formuła różni się jedynie parametrem przesunięcia d.

$$Y_t = X_t - X_{t-d}$$

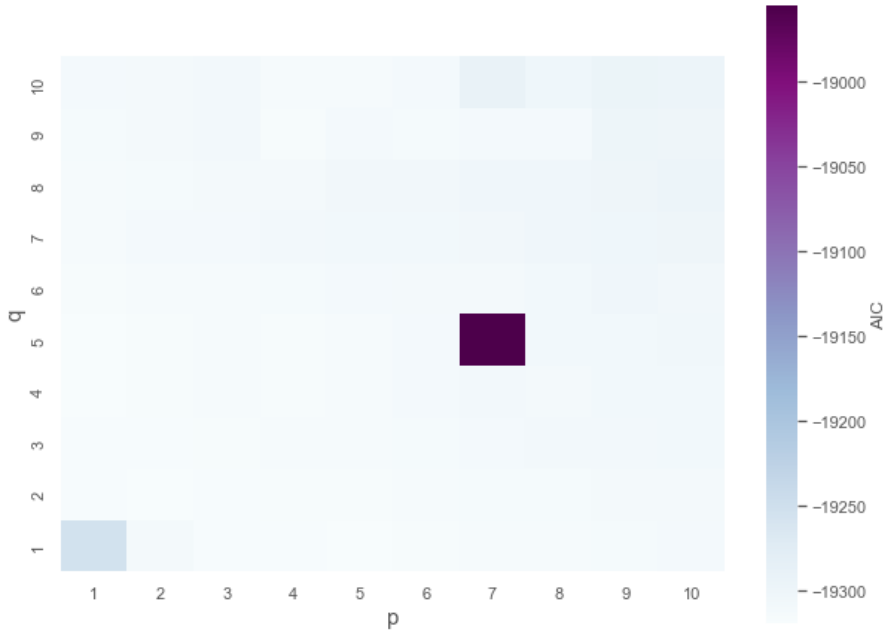
gdzie d jest liczbą danych w sezonie. Ponieważ nasze dane nie charakteryzują się konkretną sezonowością, krok ten pomijamy.

3 Dopasowanie modelu

Po usunięciu z trendu, oraz ustabilizowaniu wariancji, nasze dane są przygotowane do zamodelowania przy użyciu metody ARMA. Aby znaleźć możliwie najlepsze parametry p i q, posłużymy się kryterium informacyjnym Akkaikego (AIC).

$$\mathrm{AIC}(p, q) = -2\log(L) + 2(p + q),$$

gdzie L jest funkcją największej wiarygodności dla residuum naszego modelu. Ustalmy $p_{max} = 10$ i $q_{max} = 10$. Dla każdych $p \in \{1, p_{max}\}$ i $q \in \{1, q_{max}\}$ będziemy generować model ARMA(p, q) (oczywiście bazując na naszych danych). Najlepszym dopasowaniem dla naszych danych będzie ta para liczb, dla której $\mathrm{AIC}(p, q)$ będzie miało najmniejszą wartość.



Rysunek 6: Wykres przedstawiający jak zmienia się wartość kryterium Akkaikiego w zależności od doboru parametrów p oraz q.

Okazuje się, że modelem, który najlepiej odzwierciedla ceny akcji Netflixa jest ARMA(1, 4):

$$X_t - \phi X_{t-1} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \theta_3 Z_{t-3} + \theta_4 Z_{t-4},$$

gdzie $\{Z_t\} \sim N(0, \sigma^2)$ są i.i.d. Stąd też wynika, że

$$Z_t + \sum_{i=1}^{4} \theta_i Z_{t-i} \sim \mathcal{N}\left(0, \sigma^2\left(1 + \sum_{i=1}^{4} \theta_i^2\right)\right) = \mathcal{N}(0, \sigma_z^2)$$

Estymatory współczynników ϕ i θ_i możemy wyznaczyć za pomocą metody największej wiarygodności. Oznacza to że szukamy takich współczynników, że funkcja

$$L(x_1, \ldots, x_n, \phi, \theta_1, \ldots, \theta_4, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x_i - \phi x_{i-1})^2}{2\sigma_z^2}\right)$$

przyjmuje wartość najmniejszą. Do wyznaczenia najlepszych parametrów p, q, oraz do wyestymowania współczynników modelu zostały użyte funkcje pythonowe z biblioteki statsmodels.

```
                               SARIMAX Results
==========================================================================================
Dep. Variable:                           y   No. Observations:                  5044
Model:                 SARIMAX(1, 0, 4)   Log Likelihood                9665.668
Date:                  Thu, 09 Feb 2023   AIC                          -19319.335
Time:                          12:05:04   BIC                          -19280.180
Sample:                               0   HQIC                         -19305.618
                                 - 5044
Covariance Type:                    opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.4624      0.360     -1.283      0.199      -1.169       0.244
ma.L1          0.4389      0.360      1.218      0.223      -0.267       1.145
ma.L2          0.0272      0.014      1.931      0.053      -0.000       0.055
ma.L3         -0.0140      0.018     -0.763      0.445      -0.050       0.022
ma.L4         -0.0370      0.013     -2.915      0.004      -0.062      -0.012
sigma2         0.0013   7.27e-06    174.384      0.000       0.001       0.001
==============================================================================
====
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):             13242
0.32
Prob(Q):                              0.92   Prob(JB):
0.00
Heteroskedasticity (H):               0.32   Skew:                            -
0.76
Prob(H) (two-sided):                  0.00   Kurtosis:                         2
8.06
==============================================================================
====
```

Rysunek 7: Informacje dotyczące wybranego modelu. Najważniejsza jest dla nas kolumna "coef", bowiem informuje nas o wartościach estymatorów współczynników oraz wariancji.

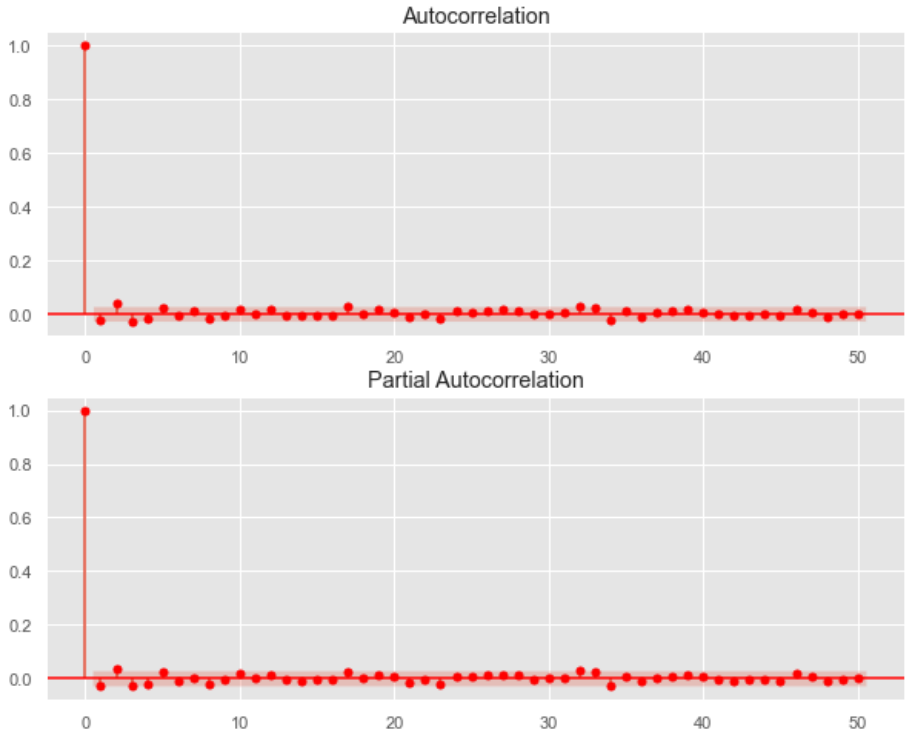
4 Ocena jakości modelu

4.1 Przedziały ufności dla ACF i PACF

Numeryczna konstrukcja przedziałów ufności dla funkcji autokowariancji (ACF) i częściowej autokowariancji (PACF) sprowadza się do symulacji Monte Carlo. Generujemy N (na przykład $N = 1000$, chociaż im więcej, tym lepiej) trajektorii modelu ARMA, a z nich wyznaczamy wartości resztowe. Dla każdego z wygenerowanych wykresów znajdujemy funkcje ACF i PACF w zależności od lagu h.

Następnie dla każdej wartości h wyznaczamy kwantyle rzędu $\frac{\alpha}{2}$ i $1 - \frac{\alpha}{2}$ (na każde h przypada N wartości funkcji ACF i PACF). Zatem przedziały ufności funkcji ACF i PACF dla każdego lagu h możemy wyrazić jako:

$$[q_{\frac{\alpha}{2}}\mathrm{ACF}(h),\ q_{1-\frac{\alpha}{2}}\mathrm{ACF}(h)]$$
$$[q_{\frac{\alpha}{2}}\mathrm{PACF}(h),\ q_{1-\frac{\alpha}{2}}\mathrm{PACF}(h)]$$

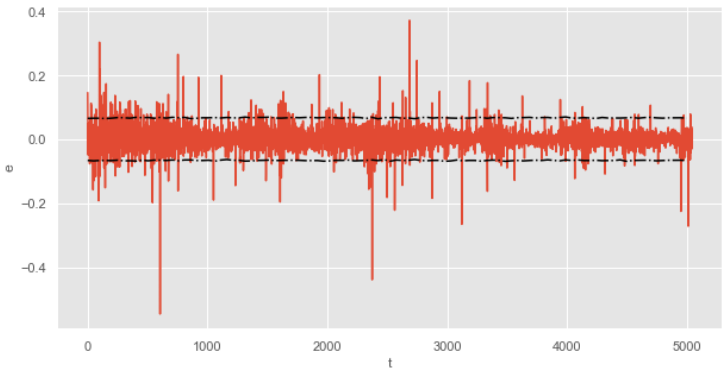


Rysunek 8: Wykresy ACF i PACF

Jak widać na powyższych wykresach, zdecydowana większość wartości znajduje się w ów przedziałach ufności. To dobry znak, ponieważ oznacza on pewną zgodność ustalonego modelu z analizowanymi danymi.

4.2 Porównanie trajektorii z liniami kwantylowymi

Wyznaczymy teraz linie kwantylowe dla modelu ARMA(1,4) metodą Monte Carlo i porównamy je ze zdekomponowanym szeregiem.

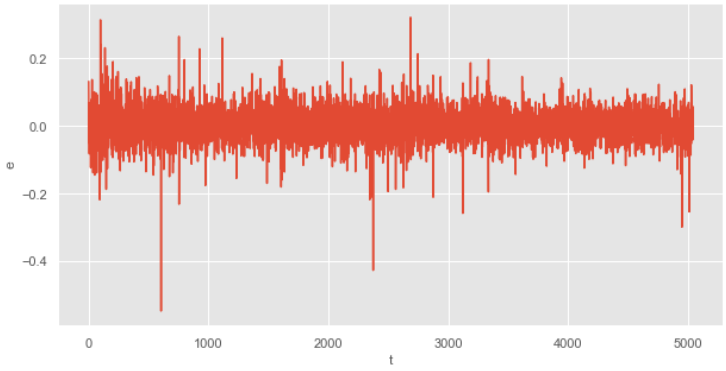


Rysunek 9: Linie kwantylowe dla zdekompowanego szeregu

Z powyższego wykresu możemy stwierdzić, że linie kwantylowe są w miarę dobrze dopasowane do badanego szeregu czasowego. Nie jest to jednak dopasowanie idealne. Może to wynikać z dużej liczby wartości mocno odstających.

5 Weryfikacja założeń dotyczących szumu

Sprawdźmy jeszcze czy nasze dane spełniają założenia dotyczące zachowania residuów. Zbadamy m.in. zachowanie ich średniej i wariancji, a także przetestujemy je pod kątem niezależności i normalności rozkładu.



Rysunek 10: Wykres przykładowej trajektorii residuów

5.1 Testowanie średniej szumu

W pierwszej kolejności zbadamy, czy średnia wartości resztowych jest stała. W tym celu posłużymy się testem t. Przedstawię poniżej definicję statystyki testowej testu t:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

gdzie x_1, x_2 - średnie, n_1, n_2 - długości prób, a s_p to tzw. skombinowane odchylenie standardowe (ang. pooled standard deviation) zdefiniowane wzorem:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Jeśli p-wartość poszczególnej statystyki jest większa od poziomu istotności (w naszym przypadku $\alpha = 0,05$) to nie ma podstaw do odrzucenia hipotezy zerowej.

W celu możliwie najlepszej aproksymacji p-wartości testu, skorzystamy z symulacji Monte Carlo. W każdym kroku będziemy generować dwie realizacje ARMA(1, 4), a następnie poddamy testowi t ich szumy. Dla 1000 kroków, p-wartość wynosi około $p \approx 0,63$ co jest wynikiem znacznie większym od α. Wobec tego możemy przyjąć, że średnia szumu jest statystycznie stała.

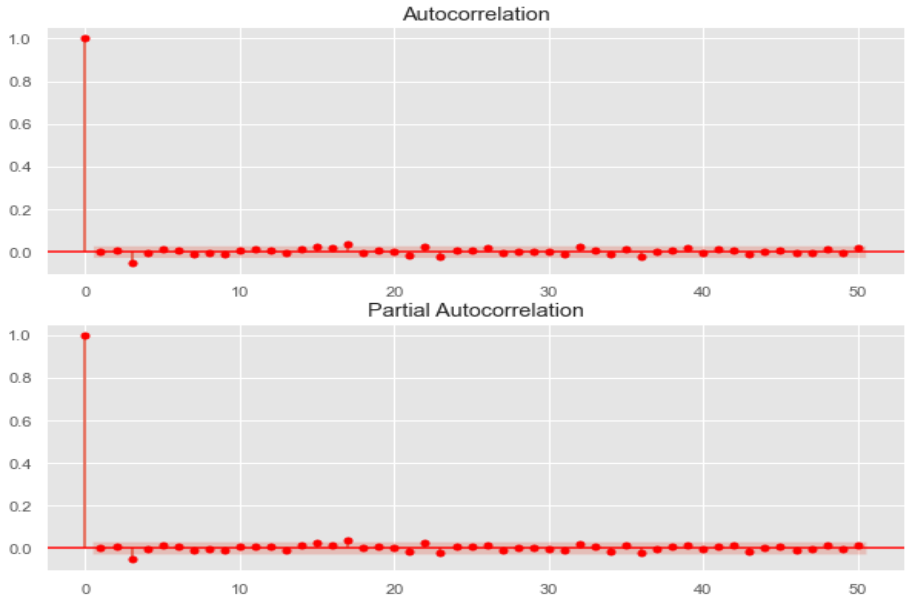
5.2 Testowanie wariancji szumu

Następnym krokiem będzie przetestowanie residuów modelu pod kątem wariancji. Sprawdzimy czy występuje tzw. efekt ARCH. W najprostszych słowach jest to stan, w którym wariancja szeregu zmienia się wraz z upływem czasu. Skorzystamy tutaj testu ARCH-LM (Lagrange-Multiplier test). Aby więcej się dowiedzieć o samej idei testu, można przeczytać tutaj. W przeprowadzonych symulacjach Monte-Carlo, otrzymaliśmy p-wartość $p \approx 0,00093$ co jest wynikiem

znacznie mniejszym od poziomu istotności $\alpha = 0,05$. Możemy zatem odrzucić hipotezę o obecności efektu ARCH w szumie.

5.3 Testowanie niezależności

Ważnym założeniem w modelach typu ARMA jest niezależność residuów. Sprawdźmy najpierw, jak wyglądają wykresy ACF, oraz PACF dla wartości resztowych.



Rysunek 11: Wykresy ACF i PACF dla przykładowej trajektorii residuów

Wykresy zdają się być obiecujące - wartości są zawarte, albo są bardzo blisko przedziałów ufności. Do zbadania problemu skorzystamy jeszcze z testu Ljunga-Boxa, który poddaje testowi niezależność residuów. Definiujemy następującą statystykę testową:

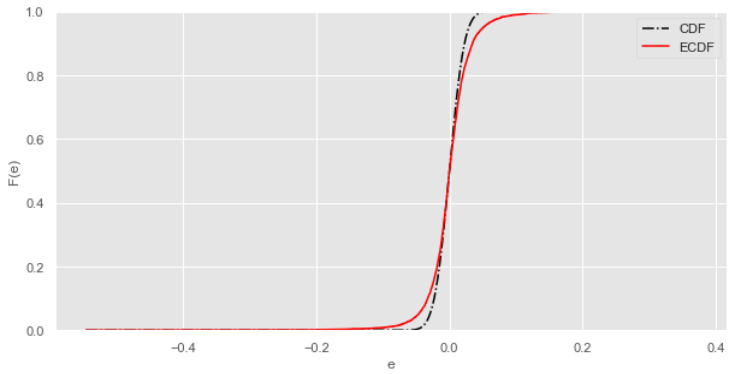
$$Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{n-k},$$

gdzie n jest długością próby, h liczbą lagów. Jeśli p-wartość jest większa od zadanego poziomu istotności, to nie ma podstaw do odrzucenia hipotezy zerowej. Więcej o teście Ljunga-Boxa można przeczytać tutaj. Wykonamy test Ljunga-Boxa dla pierwszych 100 lagów. Przyjmujemy za poziom istotności $\alpha = 0.05$. Wartość statystyki $Q \approx 112$, a p-wartość $p \approx 0,1914$. Ponieważ p-wartość jest znacznie większa od α to nie mamy podstaw do odrzucenia hipotezy zerowej. Możemy zatem przyjąć, że badane dane są niezależne.

5.4 Testowanie normalności rozkładu

5.4.1 Dystrybuanta

W pierwszej kolejności porównamy dystrybuanty empiryczną i teoretyczną.

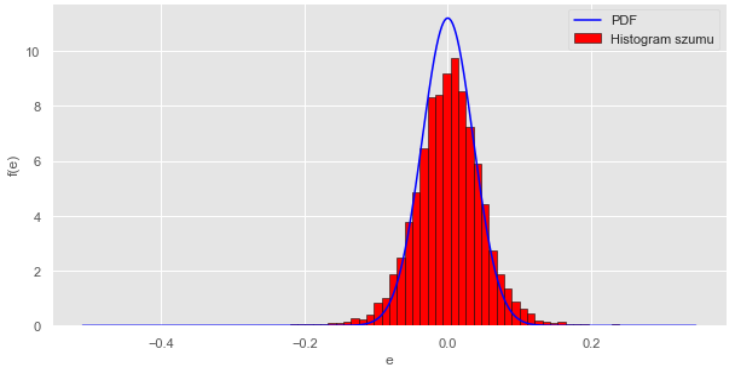


Rysunek 12: Dystrybuanta empiryczna i teoretyczna rozkładu residuów

Jak możemy zauważyć, dystrybuanta teoretyczna rośnie gwałtowniej od wykresu empirycznego - szczególnie jest to widoczne na początku i końcu dystrybuanty. Jednak wartości środkowe są niemal identyczne. W celu głębszej analizy zbadajmy wykres gęstości residuów.

5.4.2 Gęstość rozkładu

Sprawdźmy czy teoretyczna gęstość rozkładu wiernie odzwierciedla histogram gęstości residuów naszych oryginalnych danych. Przypomnijmy, że nasz wyestymowany rozkład jest rozkładem normalnym $\mathcal{N}(0, \sigma^2) = \mathcal{N}(0, 0.0013)$.

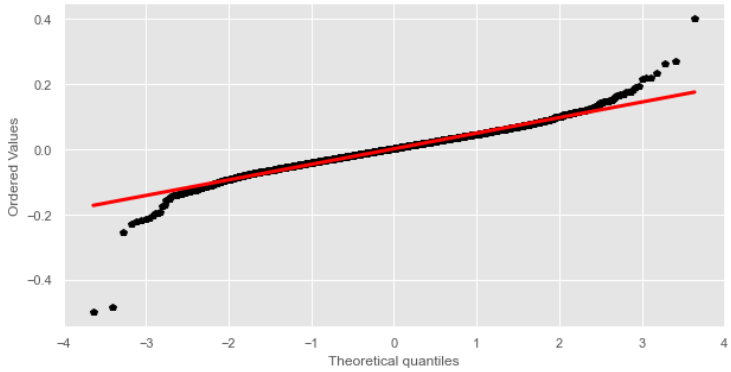


Rysunek 13: Histogram residuów porównany z teoretyczną gęstością

Jak możemy zauważyć, teoretyczny wykres gęstości ma bardzo podobny kształt do histogramu - rzeczywisty rozkład szumu jest nieco bardziej mezokurtyczny, ponieważ ma cięższe ogony, a mniej wartości jest skupionych wokół średniej. Jest to jednak drobna różnica, a samo dopasowanie jest bardzo dokładne.

5.4.3 Wykres kwantylowy

Spójrzmy teraz jak się prezentuje wykres kwantylowy zastosowany do analizowanego szumu:



Rysunek 14: Wykres kwantylowy residuów

Podobnie jak w przypadku gęstości, kwantyle teoretyczne przyjmują zbliżone wartości do empirycznych. Nieco większe odchylenie możemy zauważyć dla wartości krańcowych - ponownie ma to związek z cięższymi ogonami w rzeczywistym rozkładzie szumu.

5.4.4 Test Andersona-Darlinga

Ostatnim narzędziem, którym posłużymy się do oceny normalności rozkładu, będzie test Andersona-Darlinga. Charakteryzuje się on stosunkowo dużą mocą (większą od testu Kołmogorowa-Smirnova) i w przeciwieństwie do testu Shapiro-Wilka nie ma górnego ograniczenia jeśli chodzi o długość próby. Naszą hipotezą zerową jest zachowana normalność rozkładu. Testujemy ją przeciwko hipotezie alternatywnej, która mówi przeciwnie. Statystyka testowa prezentuje się następująco:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\log F(X_i) + \log(1 - F(X_{n-i+1})],$$

gdzie n jest długością próby, $F(X)$ jest dystrybuantą naszego rozkładu, a X_i jest i-tą realizacją posortowanej próby. W przypadku nieznanych parametrów

średniej i wariancji stosuje się poprawkę:

$$A^{2*} = A^2 \left(1 + \frac{0,75}{n} + \frac{2,25}{n^2}\right)$$

Dla przykładowej trajektorii residuów, wartość statystyki testowej $A^2 = 11,358$ co jest niebotycznym wynikiem. Warunkiem przyjęcia hipotezy zerowej jest większa wartość statystyki od wartości krytycznej testu. Nawet na poziomie ufności $\alpha = 0,01$ punkt krytyczny jest równy $1,091$ co jest wyraźnie mniejszym rezultatem od wartości statystyki. Zatem bez problemów możemy założyć, że badany rozkład jest normalny.

6 Wnioski

Jak się okazuje, za pomocą szeregów czasowych typu ARMA możemy przeprowadzić obszerną analizę danych, która pozwoli nam na znalezienie podstawowego modelu, który najlepiej będzie odzwierciedlał nasze dane. Istnieją jednak jeszcze dokładniejsze metody analizy, jak chociażby badanie wariancji przy użyciu modelu GARCH. Pozwala on na sprawdzenie zjawiska tzw. heteroskedantyczności czyli "sposobowi" w jaki zmienia się wariancja wartości resztowych w czasie. Pozwala nam to na jeszcze dokładniejszą optymalizacje w doborze modelu.

Jak spojrzymy na zachowanie cen akcji Netflixa, to możemy zauważyć duże i nagłe zmiany wartości, a pod sam koniec notujemy olbrzymi spadek. Z pewnością to utrudniło analizę - w szeregu czasowym, nawet po dekompozycji, występowały liczne wahania, które osiągały mocno odstające wartości. Mimo to udało nam się dobrać taki model, który przede wszystkim spełnia warunki zachowania residuów, jest stacjonarny i niezależny. Ciężko tutaj jednak o dobrą predykcję, ze względu na wcześniej wspomniany, znaczący spadek cen. Nie ma jednak co się dziwić - przez wiele lat Netflix był wyborem numer jeden spośród serwisów filmowych. Z biegiem czasu powstały kolejne wypożyczalnie jak chociażby Amazon Prime, czy Disney+ które bardzo szybko zyskały popularność. A to wpłynęło na ostateczny koniec dominacji Netflixa na rynku.