Raport nr 2

Szeregi klasy ARMA

Piotr Mikler 236895

Spis treści

1 Wstęp **2**

2 Analiza wstępna **2**

3 Modele ARMA **3**

4 Dopasowanie modelu **3**

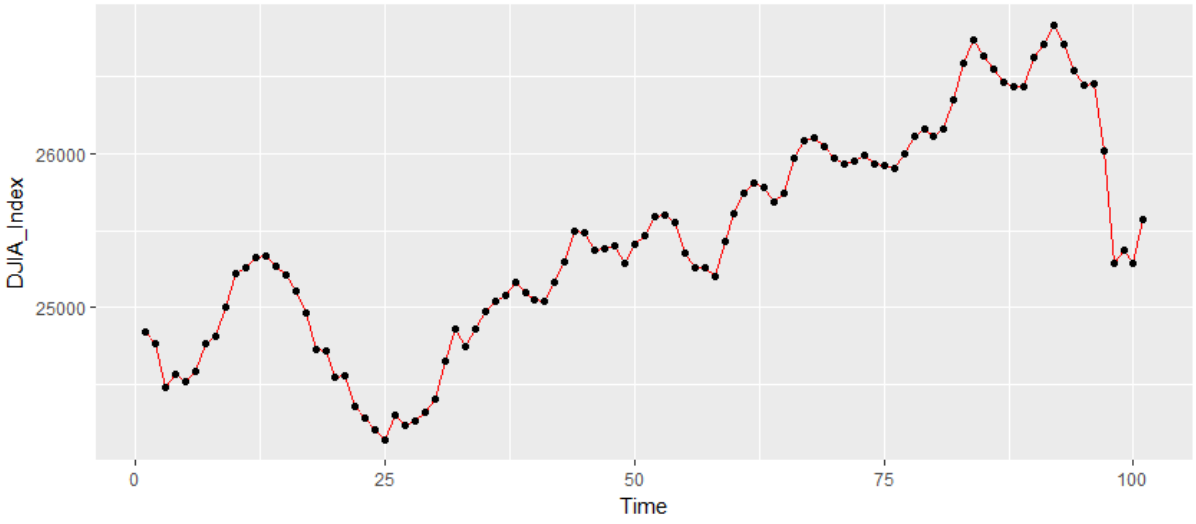
4.1 Dekompozycja Wolda i różnicowanie 3

4.2 Parametry ARMA(p,q) 6

4.3 Analiza reziduów 7

4.4 Predykcja 9

5 Wnioski **10**



Rysunek 1: Srednia wartość indeksu DJIA w okresie od 24.05.2018 do 16.10.2018

1 Wstęp

Praca zawiera analizę średniej wartości indeksu Dow Jones Industrial Average w okresie od 24.05.2018 do 16.10.2018. Przez wartość średnią rozumiem w tu arytmetyczną średnią wartości indeksu na otwarcie i na zamknięcie danego dnia. Wybrany przeze mnie interwał czasowy motywowałem obserwacją wartości indeksu pod kątem prawdopodobnej stacjonarności. Łącznie otrzymałem w ten sposób 100 obserwacji, zachowując sobie dodatkowe 30 w celach sprawdzenia predykcji. Kolejne rozdziały pracy traktują o wstępnej analizie danych, dobieraniu odpowiedniego rzędu modelu ARMA(p,q), analizie reziduów, oraz predykcji wartości.

2 Analiza wstępna

Próbka została pobrana ze strony *stooq.pl*. Dane to historyczne wartości indeksu Dow Jones Industrial Average, czyli jednego z najstarszych amerykańskich indeksów giełdowych, bazujący na mieszanym portfelu akcji złożonym z czołowych przedstawicieli amerykańskiego przemysłu. Skala czasowa dla wygody rozpatrywana jest w formie liczb całkowitych, gdzie 24.05.2018 stanowi moment 0. Wartości indeksu można zobaczyć na wykresie (1). Indeks DJIA charakteryzuje się niskim volatility. Zapewnia to gładki przebieg krzywej, niską wariancję obserwacji wokół kroczącej średniej. Można dopatrzeć się w danych rosnącego trendu, startującego od około obserwacji nr 25. Krzywa wygląda też na posiadającą periodyczny komponent, co objawia się w dość regularnych górkach na wykresie. Dane wyglądają zatem obiecująco w perspektywie

dopasowywania modelu ARMA.

3 Modele ARMA

Modele ARMA to modele z dziedziny szeregów czasowych. Pozwalają one na zależności pomiędzy wartościami procesu w różnych odstępach czasu, zniekształcone przez szum biały. Ogólnie proces X_t jest procesem ARMA, jeśli jest stacjonarnym rozwiązaniem równania

$$\phi(B)X_t = \theta(B)Z_t$$

W którym $\{Z_t\}_{t\in\mathbb{N}_+}$ to biały szum, a B to operator przesunięcia wstecz. Wielomiany $\theta(z)$ oraz $\phi(z)$ nazywane są za to kolejno wielomianem średniej ruchomej, oraz autoregresyjnym. One decydują o rzędzie modelu ARMA, a więc o tym w jakim stopniu obserwacje będą zależeć od siebie nawzajem.
W następnych rozdziałach przejdziemy kolejno przez etapy dopasowywania modelu ARMA do danych, używając pakietu R który okazał się bezcennym narzędziem do tych analiz.

4 Dopasowanie modelu

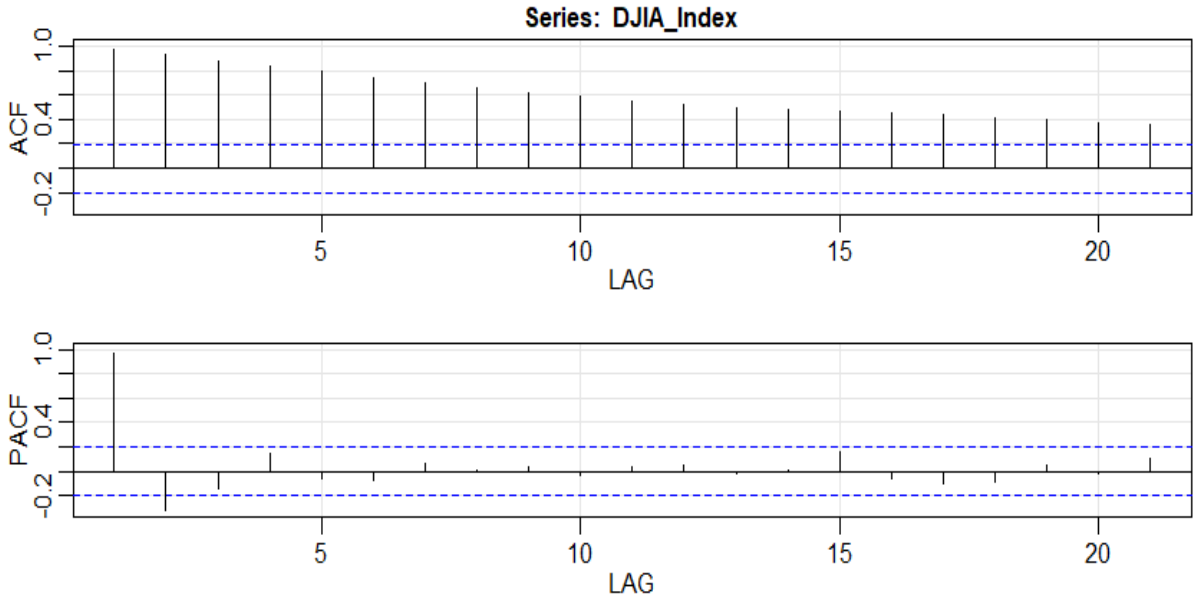
4.1 Dekompozycja Wolda i różnicowanie

Dekompozycja Wolda i różnicowanie to dwa często stosowane podejścia mające na celu sprowadzenie naszych danych do postaci którą można modelować szeregami klasy ARMA.
Dekompozycja Wolda zakłada, że badany szereg czasowy Y_t da się rozłożyć na sumę

$$Y_t = m_t + s_t + X_t$$

Gdzie m_t jest trendem- deterministyczną, nieokresową funkcją, s_t jest komponentem sezonowym- deterministyczną funkcją okresową, natomiast szereg X_t jest szeregiem słabostacjonarnym. Do niego musimy się dostać, ponieważ to właśnie jego będziemy starać się modelować szeregiem ARMA.
Dekompozycja jest pierwszą z możliwości, lecz ja skupię się na drugiej metodzie uzyskiwania słabostacjonarnego procesu, czyli na różnicowaniu danych. W tym podejściu, z zadanego szeregu Y_t tworzymy nowy, definiowany jako

$$X_t = Y_t - Y_{t-h}$$



Rysunek 2: Wykresy autokorelacji i częściowej autokorelacji dla wartości średnich indeksu DJIA

Takie podejście jest często wykorzystywane w danych finansowych, ponieważ w ten sposób modelujemy nie same dane, ale przyrosty między kolejnymi chwilami w czasie. Przyjmować możemy różne wartości parametru h w zależności od tego jak wygląda sezonowość w naszych danych. Moim pierwszym krokiem było stworzenie wykresów autokorelacji i częściowej autokorelacji (rys. (2)). Funkcję autokorelacji $\rho(h)$ estymujemy wzorem (1):

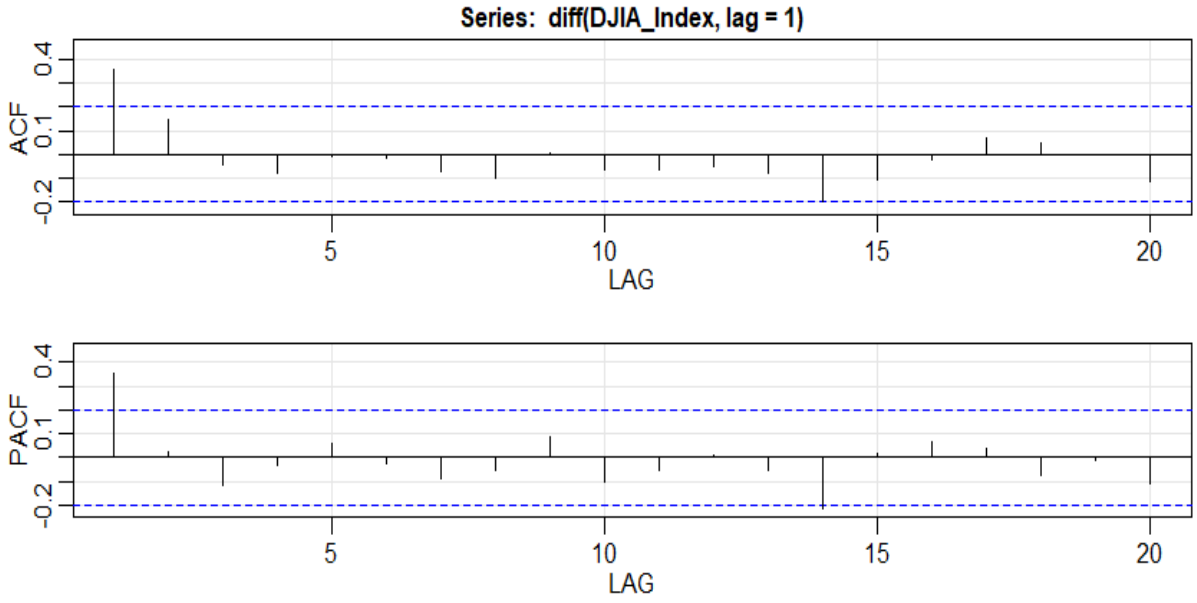
$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_t - \overline{X}_n)(X_{t+h} - \overline{X}_n)$$

$$\hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)} \tag{1}$$

Natomiast funkcja częściowej autokorelacji $\alpha(h)$ jest estymowana w punkcie h jako $\hat{\alpha}(h) \equiv \hat{\phi}_{hh}$, gdzie $\hat{\phi}_{hh}$ to ostatni element wektora $\hat{\phi}_h$ będącego rozwiązaniem następującego macierzowego układu równań (Yule-Walkera)

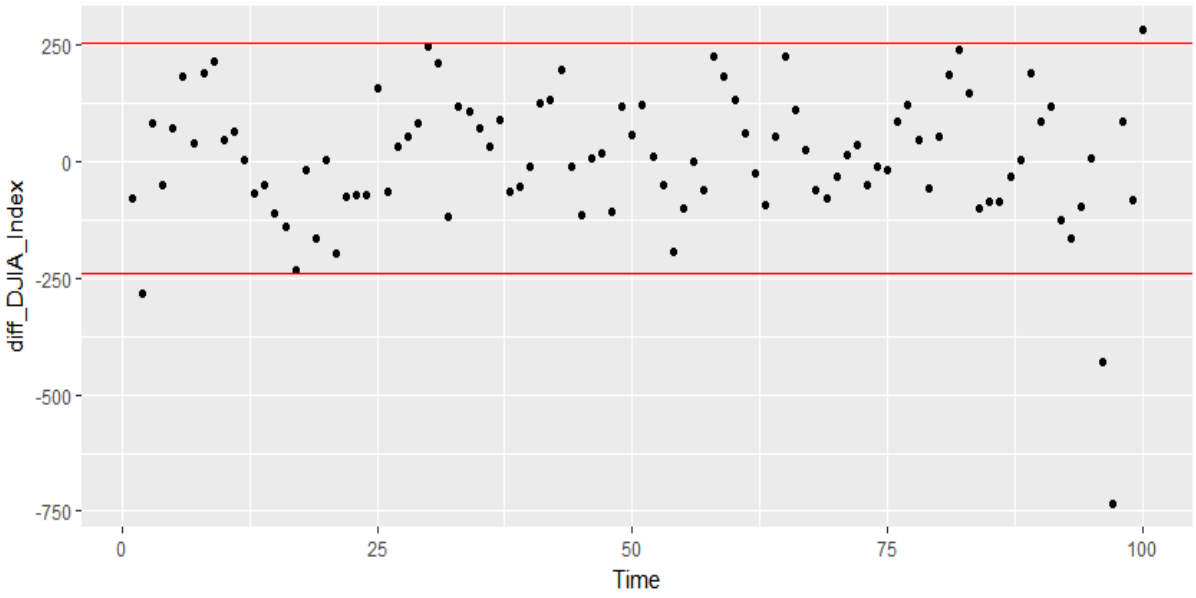
$$\hat{\phi}_h = \hat{\Gamma}_h^{-1} \cdot \hat{\gamma}_h$$

W powyższym równaniu $\hat{\Gamma}_h = [\hat{\gamma}(i-j)]_{i,j=1}^{h}$, natomiast $\hat{\gamma}_h = [\hat{\gamma}(1), \hat{\gamma}(2), \ldots, \hat{\gamma}(h)]'$.
Wykresy tych dwóch funkcji mówią nam wiele o tym jak wyglądają zależności między kolejnymi obserwacjami. Jeśli dla wybranej wartości *lag* otrzymujemy wartość ACF lub PACF wystającą ponad zaznaczony na niebiesko poziom, możemy wnioskować, że istnieje zależność między wartościami szeregu o indeksach odległych o ten *lag*. Wartości ACF (autocorrelation function) bardzo powoli zanikają, natomiast wartość PACF (partial autocorrelation function)



Rysunek 3: Wykresy autokorelacji i częściowej autokorelacji dla zróżnicowanego szeregu wartości średnich DJIA

jest znacząca jedynie dla $lag = 1$. Sugeruje to autoregresyjny charakter danych, który często da się zneutralizować różnicując dane o (w tym przypadku) $h = 1$. Spróbowałem właśnie tego podejścia i po raz kolejny narysowałem ACF i PACF, tym razem dla zróżnicowanych danych (rys. (3)). Zróżnicowanie szeregu w tym przypadku zdziałało cuda. Widzimy zdecydowanie, że zarówno ACF jak i PACF mają wartości znaczące tylko dla $h = 1$. Jest to dla nas bardzo dobry znak, ponieważ prawdopodobnie dostaniemy dobre dopasowanie modelu ARMA już przy użyciu niewielu parametrów. Chciałbym zwrócić również uwagę na fakt, że wartości próbkowych ACF jak i PACF obie praktycznie dotykają niebieskich linii dla $lag = 14$. Biorąc pod uwagę fakt, że jest to odległość czasowa dokładnie dwu tygodni, wydało mi się to podejrzane. Sytuacja nie powtarza się jednak dla $lag \in \{21, 28, 35\}$, więc odrzuciłem hipotezę periodycznej zależności o okresie 2 tygodni. Zróżnicowanie danych ma dodatkowo tę zaletę, że zazwyczaj usuwa trend i sezonowość z danych. Jeśli spojrzymy na powstały szereg Y_t =diff(DJIA_Index), widoczny na rysunku (4) zobaczymy, że istotnie nie ucieka on od zera i ma dość stały rozrzut (stałą wariancję). Widzimy, że istnieją cztery obserwacje które znacząco odbiegają od pozostałych. Gdyby nie były to dane giełdowe, uznałbym je za odstające i usunąłbym z próby. W tej sytuacji jednak, myśląc o charakterze danych giełdowych nie chciałbym przypadkowo zaniżyć wariancji danych, a tym samym osłabić modelu. Z tego powodu nie usuwam tych obserwacji. Średnia zróżnicowanego szeregu to $\overline{X} \approx 7.3$, a wariancja w próbie to $\sigma^2 \approx 21054.3$. Otrzymany w dro-



Rysunek 4: Wykres zróżnicowanych średnich wartości indeksu DJIA. Kolorem czerwonym zaznaczone poziomy $q_{0.5} \pm 1.5 \cdot IQR$

dze różnicowania szereg X_t powinien wykazywać właściwości szeregu słabostacjonarnego. Aby to sprawdzić, użyłem testu adf (augmented Dickey-Fuller test). Bada on stacjonarność przy pomocy regresji wielorakiej. Test odrzucił hipotezę niestacjonarności szeregu X_t na rzecz hipotezy alternatywnej (stacjonarności) z p-wartością na poziomie 0.016, zatem w sposób bardzo dobitny. Możemy bez obaw przystąpić do dalszego modelowania danych przy użyciu szeregu X_t.

4.2 Parametry ARMA(p,q)

Dopasowując parametry klasy ARMA kierujemy się tzw. kryteriami informacyjnymi. Jak ich nazwa wskazuje, są one informacyjne, tzn ich wartości wyliczone dla konkretnych parametrów informują nas który model jest lepszy. Im niższa wartość, tym lepszy model. Kryteriów mamy do wyboru wiele, ja kierowałem się przede wszystkim kryteriami AIC oraz BIC wyliczanymi według wzorów

$$AIC = 2k - 2\ln\hat{L}$$

$$BIC = k \cdot \ln(n) - 2\ln\hat{L}$$

W powyższych wzorach n oznacza liczność danych, k to liczba estymowanych parametrów, a $\hat{L}$ jest maksimum osiąganym przez funkcję wiarogodności.

Najniższe wartości kryteriów ($AIC \approx 10.84534$, oraz $BIC \approx 9.897447$) otrzymałem dla modelu ARMA(1,0) co jest szokującą, ale bardzo zadowalającą nowiną. Już z wyglądu funkcji autokorelacji można było wnioskować o prostocie poszukiwanego modelu, ale szczerze nie spodziewałem się, że rzeczywiste dane dopasują się tak dobrze. W pakiecie R wyestymowałem współczynniki szeregu ARMA(1,0) dopasowanego do danych. Szereg jest rozwiązaniem równania

$$X_t - 0.3645 \cdot X_{t-1} = Z_t$$

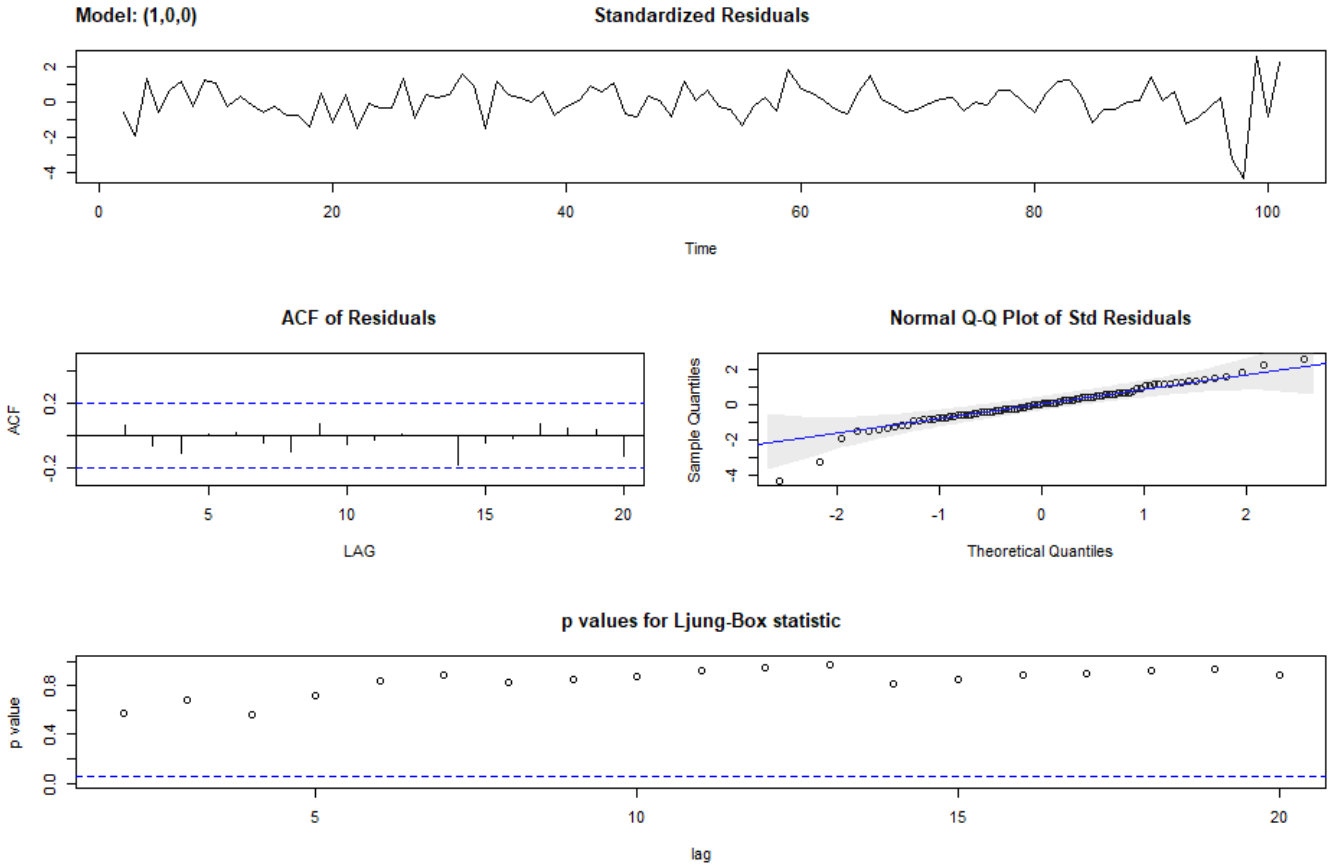
Inne modele, wyglądające równie dobrze w świetle kryteriów informacyjnych to między innymi ARMA(2,0) czy ARMA(1,1). Dopasowanie było na podobnym poziomie, lecz dodawanie kolejnych współczynników znikomie poprawiało dopasowanie, a komplikowało model. Udało się też dopasować kilka modeli szerszej klasy ARIMA(p,d,q), np ARIMA(1,1,1) tutaj model wpasowywał się równie dobrze, ale ta klasa wyrzucała gorsze rezidua. Z tych powodów zdecydowałem się na użycie modelu ARMA(1,0).

4.3 Analiza reziduów

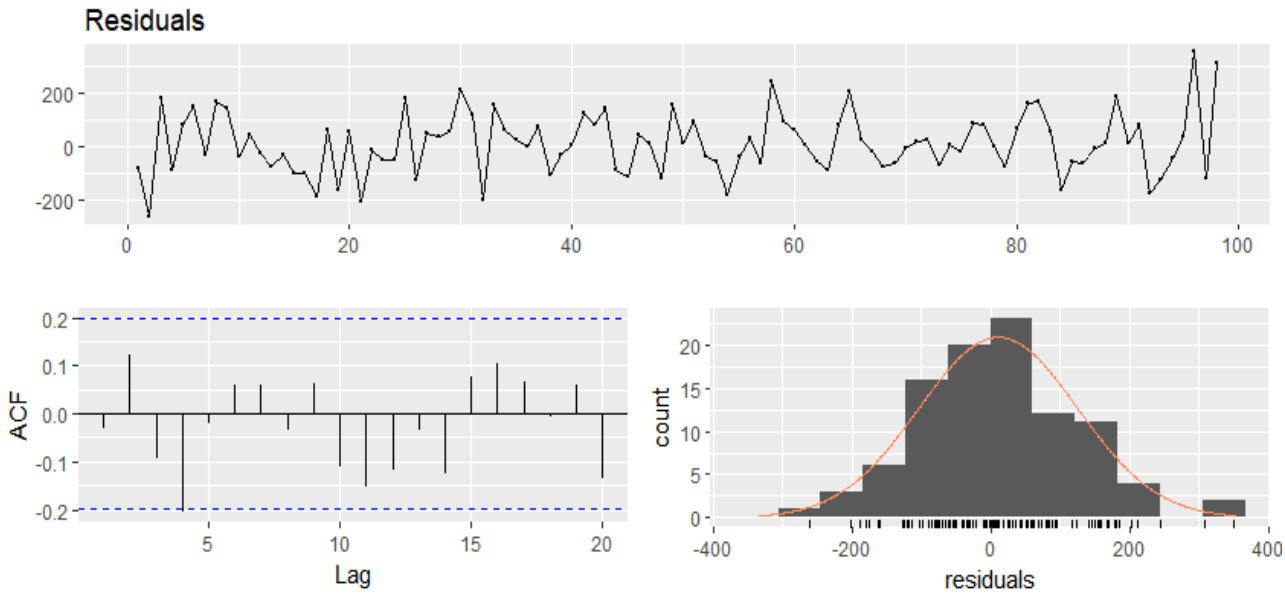
Bardzo dobrze wypadła analiza dopasowania przy pomocy funkcji *sarima*(), którą możemy obejrzeć na rysunku (5). Zacznijmy analizę reziduów od wykresów wyrzuconych przez tę funkcję. Górny panel przedstawia wykres ustandaryzowanych reziduów. Rezidua oscylują ładnie wokół zera, widać też obiecująco niezmienną wariancję. Na środkowym panelu, po lewej widzimy próbkową autokorelację reziduów. Brak znaczących wartości informuje nas, że kolejne obserwacje są kompletnie nieskorelowane. Po prawej natomiast mamy wykres kwantylowy kwantyli rozkładu normalnego, przeciw kwantylom próbkowym. Punkty układają się tu wzdłuż linii prostej, pięknie się do niej dopasowując. Widzimy kilka punktów które nie trzymają się dokładnie prostej, ale są one na samych skrajach wykresu, gdzie możliwe są niedokładności ze względu na małe ilości obserwacji. Dolny panel ilustruje p-wartości testu Ljunga-Boxa przeprowadzonego dla kolejnych wartości *lagow*. Test sprawdza hipotezę zerową zakładającą niezależność między reziduami odległymi o odpowiedni *lag*. U nas, p-wartości testu dla każdego *lagu* są wysokie, znacznie większe od 0.05, zatem test nie zgłasza wątpliwości co do niezależności między kolejnymi reziduami.

Ostatnim elementem analizy reziduów jest formalne sprawdzenie ich rozkładu. Do tego celu użyłem testów: KS, Lilleforse oraz Shapiro-Wilka.

W tym miejscu zostałem srogo pokarany. Założenie normalności w iście spektakularny sposób



Rysunek 5: Wykres reziduów (górny panel), funkcja autokorelacji reziduów i wykres kwantylowy reziduów (panel środkowy, odpowiednio) oraz p-wartości testu Ljunga-Boxa dla kolejnych lagów (panel dolny)



Rysunek 6: Wykres reziduów (górny panel), funkcja autokorelacji reziduów (lewy, dolny panel) oraz histogram reziduów wraz z dopasowanym rozkładem normalnym (prawy dolny panel)

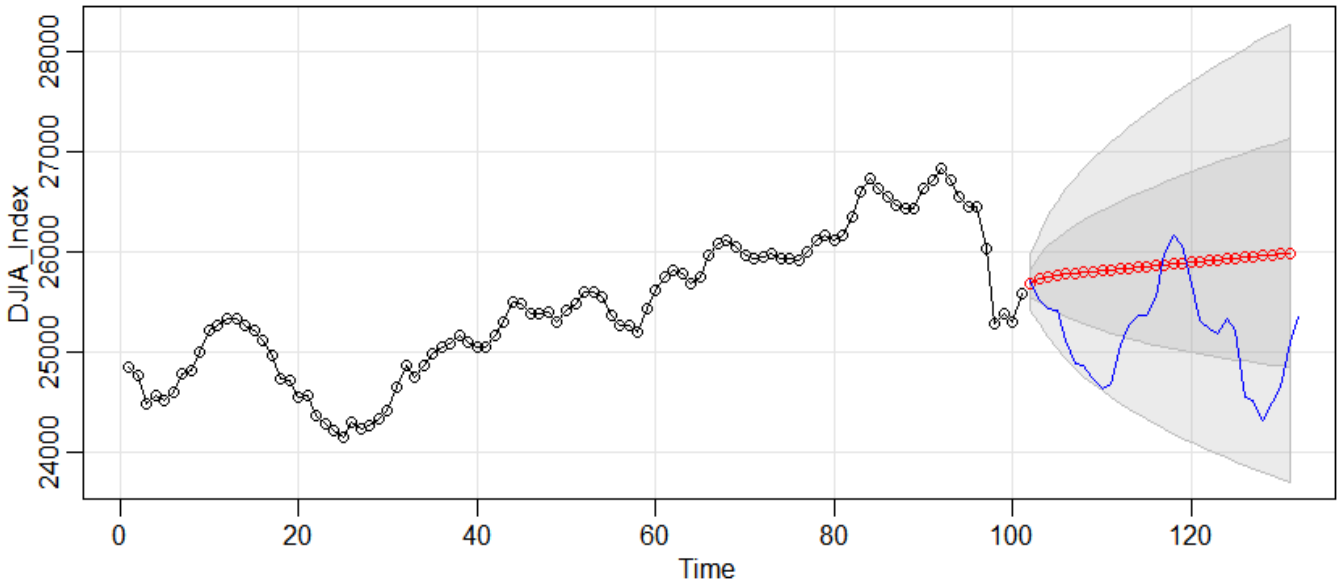
rozbił test Shapiro-Wilka, p-wartość była rzędu 10^{-4}. Winowajcą, okazały się dwie najbardziej ekstremalne obserwacje które oszczędziłem na samym początku. Po ich usunięciu wszystkie trzy testy nie zgłaszają wątpliwości co do normalności rozkładu i to z bardzo wysokimi p-wartościami ($pv_{KS} \approx 0.9914, pv_{LF} \approx 0.817, pv_{SW} \approx 0.732$).

Wykres nowych reziduów, ich funkcja autokorelacji oraz histogram wraz z dopasowaną gęstością można zobaczyć na rysunku (6) Być może wartoby wrócić teraz do samego początku konstruowania modelu i zastąpić te obserwacje sąsiednimi. Nie będę teraz tego robił, zakładam że model po takiej podmianie nadal będzie ARMA(1,0) z dobrym dopasowaniem.

4.4 Predykcja

Mając ustalony model ARMA(1,0), oraz po wykonaniu pomyślnej analizy reziduów możemy przystąpić do problemu predykcji. W tym podejściu dokonujemy predykcji reziduów, na podstawie której przechodzimy do predykcji szeregu ARMA, a na koniec odwracamy transformacje które wykonaliśmy na danych tak, aby dostać wyjściowy szereg i predykcje dla niego.

Dokonałem predykcji danych na 30 okresów do przodu, pod założeniem modelu ARMA(1,0). Predykcję widzimy na wykresie (7). Na rysunku, kolorem czerwonym naniesiono przewidywaną wartość średnią ($E[X_{t+h} - \hat{X}_{t+h}|X_t = x_t] = \phi^h x_t$) procesu, a szare pola to możliwe od niej od-



Rysunek 7: Predykcja wartości indeksu DJIA. Niebieskim kolorem wartość rzeczywista

chyły. Kolorem ciemnoszarym zakolorowano odchył od przewidywanej wartości średniej o jeden błąd standardowy, a kolorem jasnoszarym odchył o dwa błędy standardowe. Błąd standardowy jest tu rozumiany jako dodatni pierwiastek

$$Var[X_{t+h} - \hat{X}_{t+h}] = \sigma^2 \frac{1-\phi^{2h}}{1-\phi^2}$$

Dokonana w ten sposób predykcja wydaje się być trafna, rzeczywiste dane wpasowują się do niej.

5 Wnioski

Dane zostały zróżnicowane o jeden lag, a następnie dopasowany został model AR(1). Model wpasował sie dobrze, z najniższymi wartościami kryteriów informacyjnych ($AIC \approx 10.84534$, oraz $BIC \approx 9.897447$). Rezidua tak dobranego modelu przechodzą najpopularniejsze testy normalności wyśmienicie ($pv_{KS} \approx 0.9914, pv_{LF} \approx 0.817, pv_{SW} \approx 0.732$), co więcej- można z czystym sumieniem wnioskować o braku korelacji między kolejnymi reziduami, ponieważ żadna wartość ich funkcji autokorelacji nie przyjmuje wartości znaczących.

Mimo tych pomyślnych wniosków, nie mogę oprzeć się wrażeniu, że niewłaściwe było przyjęcie przeze mnie do analizy szczególnie wybranego okresu. Nie występowały w nim spore wahania indeksu, więc model AR(1) nie będzie się nadawał do modelowania wartości indeksu na innym

interwale czasowym. Widzieliśmy jaki wpływ miały zaledwie dwie obserwacje, które kompletnie rozwaliły założenie normalności reziduów. Co dopiero gdyby indeks wahał się bardziej.
Jednak na analizowanym okresie czasu, oraz gdy dane wykazują podobną wariancję, model AR(1) powinien być bardzo dobrym predyktorem.