Iga Świtalska, Karolina Rakus

Model ARMA na podstawie danych rzeczywistych

7 lutego 2022

1. Opis danych

Dane, które będziemy analizować dotyczą pogody w Szegedzie (Węgry) w latach 2006 - 2016. Pomiary były notowane codziennie, co godzinę, przez 10 lat. Dane zawierają 96453 obserwacji i 12 niżej wymienionych kolumn:
— time - data i godzina pomiaru,
— summary - słowny opis pogody (np. częściowe zachmurzenie),
— precip type - rodzaj opadu (słownie),
— temperature - temperatura ($°C$),
— apparent temperature - temperatura pozorna ($°C$),
— humidity - wilgotność,
— wind speed - prędkość wiatru (km/h),
— wind bearing - kierunek, w którym porusza się obiekt (przeciwieństwo kierunku wiatru) (mierzony w stopniach),
— visibility - widoczność (km),
— pressure - ciśnienie (hPa),
— daily summary - słowne podsumowaie dnia.
Kolumnę „Loud Cover" zdecydowałyśmy się usunąć, ponieważ zawiera same wartości równe zero. Prawdopodobnie są to dane błędne lub brakujące. Do dalszej analizy wybrałyśmy temperaturę oraz czas. Ograniczyłyśmy także nasze dane do kwietnia 2016 roku. W celu ułatwienia dalszej analizy zmieniłyśmy nazwę pierwszej kolumny na „Date". Dodałyśmy również nową kolumę „Time", reprezentującą numer kolejnych pomiarów temperatury. Po tych modyfikacjach mamy 720 obserwacji i 3 kolumny. Pierwsze 6 wierszy zmodyfikowanych danych widocznych jest w tabeli 1.1.

	Date	Temperature	Time
87670	2016-04-01 00:00:00.000 +0200	11.81	1
87671	2016-04-01 01:00:00.000 +0200	11.81	2
87672	2016-04-01 02:00:00.000 +0200	11.44	3
87673	2016-04-01 03:00:00.000 +0200	10.72	4
87674	2016-04-01 04:00:00.000 +0200	10.23	5
87675	2016-04-01 05:00:00.000 +0200	11.00	6

Tabela 1.1. Pierwsze 6 obserwacji

2. Dekompozycja

Zanim przejdziemy do wizualizacji i dekompozycji danych, przytoczymy definicję autokorelacji (ACF) oraz częściowej autokorelacji ($PACF$). Obu tych funkcji będziemy używać w dalszej analizie.

Na początku zdefiniujemy funkcję autokowariancji jako kowariancję szeregu czasowego $\{X_t\}_{t\in\mathbb{Z}}$ i tego samego szeregu czasowego przesuniętego w czasie o h. Możemy ją zapisać wzorem:

$$\gamma(t,h) = Cov(X_t, X_{t+h}) = E[X_t\, X_{t+h}] - E[X_t]E[X_{t+h}].$$

Dla szeregów stacjonarnych w słabym sensie funkcja autokowariancji nie zależy od czasu, więc z możemy ją zapisać jako $\gamma(h)$. Funkcja autokorelacji jest dodatkowo unormowana przez funkcję autokowariancji dla $h = 0$. Wyraża się wzorem:

$$\delta(h) = \frac{\gamma(h)}{\gamma(0)}.$$

Funkcja częściowej autokorelacji jest współczynnikiem ϕ_{hh} w następującej reprezentacji:

$$X_t = \phi_0 + \phi_{h1}X_{t-1} + \ldots + \phi_{hh}X_{t-h}.$$

Czyli dla $h > 0$ jest ona współczynnikiem powyższej reprezentacji linowej, biorąc pod uwagę wszystkie X_{t-1}, X_{t-2},..., $X_{t-(h+1)}$.

Wizualizację naszego szeregu czasowego możemy zobaczyć na wykresie 2.1 temperatury od czasu. Na podstawie wykresu można stwierdzić, że dane nie są stacjonarne. Zauważalny jest wyraźny trend liniowy oraz sezonowość. Potwierdza to wykres autokorelacji 2.2 oraz częściowej autokorelacji 2.3. Dane wykazują zachowania okresowe, co znaczy, że autokorelacja jest zależna od czasu. Szereg czasowy jest stacjonarny w słabym sensie, jeśli ma funkcję średniej oraz autokowariancji niezależną od czasu. Nasze dane nie spełniają obu tych warunków.

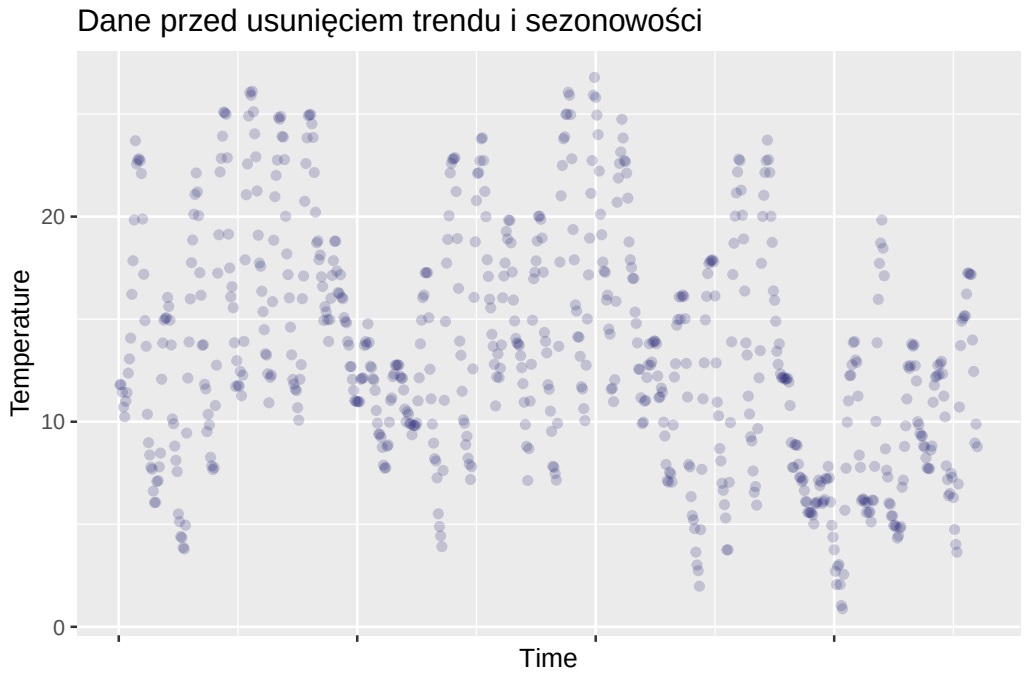
Zanim przejdziemy do dalszej analizy zajmiemy się usunięciem trendu liniowego i sezonowość.

Na początku dopasowałyśmy do danych następujący model regresji liniowej:

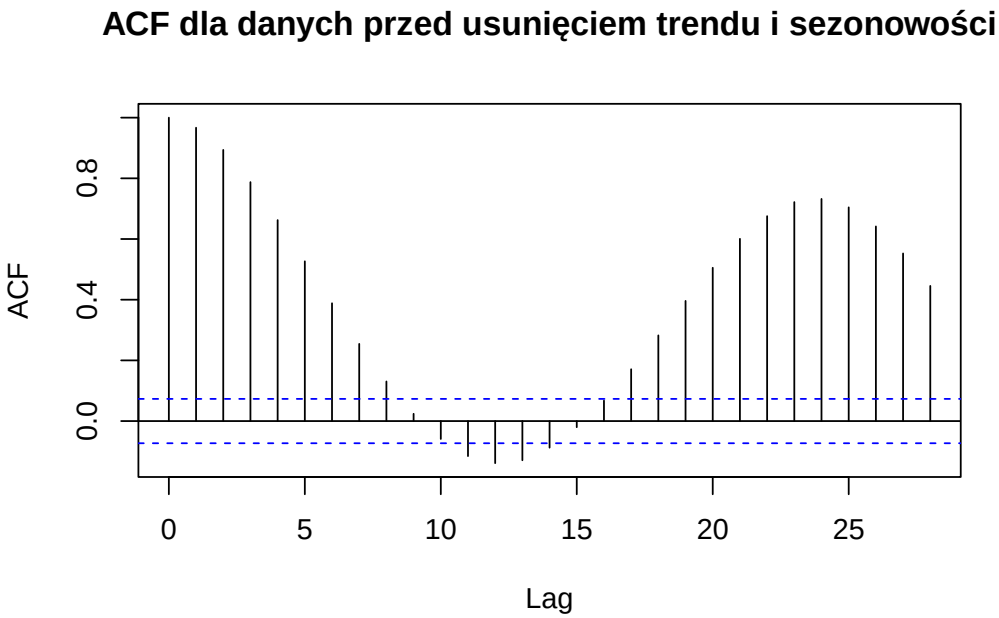
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

gdzie y_i - zmienna objaśniana, x_i - zmienna objaśniająca, ϵ_i - szum.

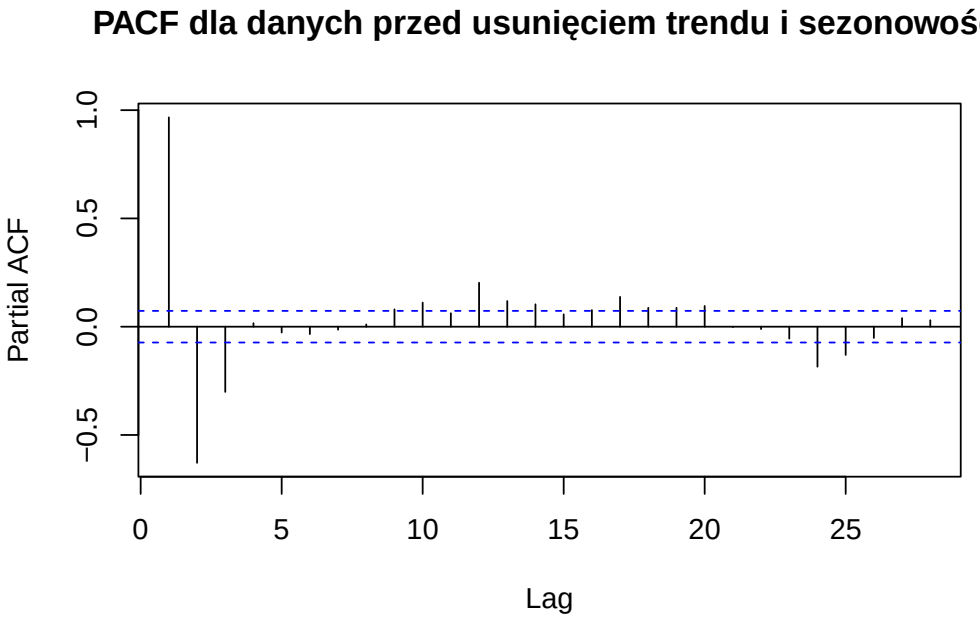
Współczynniki dobrane metodą najmniejszych kwadratów wynoszą $\beta_0 \approx 16,4797$ i $\beta_1 \approx -0,0087$. Dane z dobraną do nich prostą regresji zostały przedstawione na wykresie 2.4. Następnie usunęłyśmy trend liniowy odejmując funkcję liniową od danych.



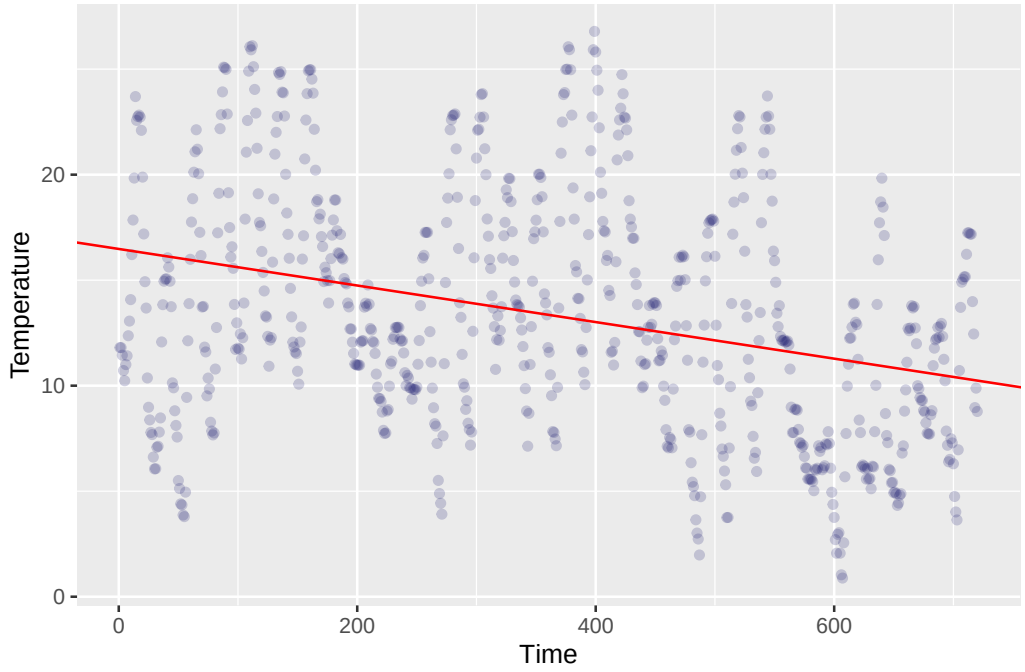
Rysunek 2.1. Wykres zależności temperatury od czasu przed usunięciem trendu i sezonowości.



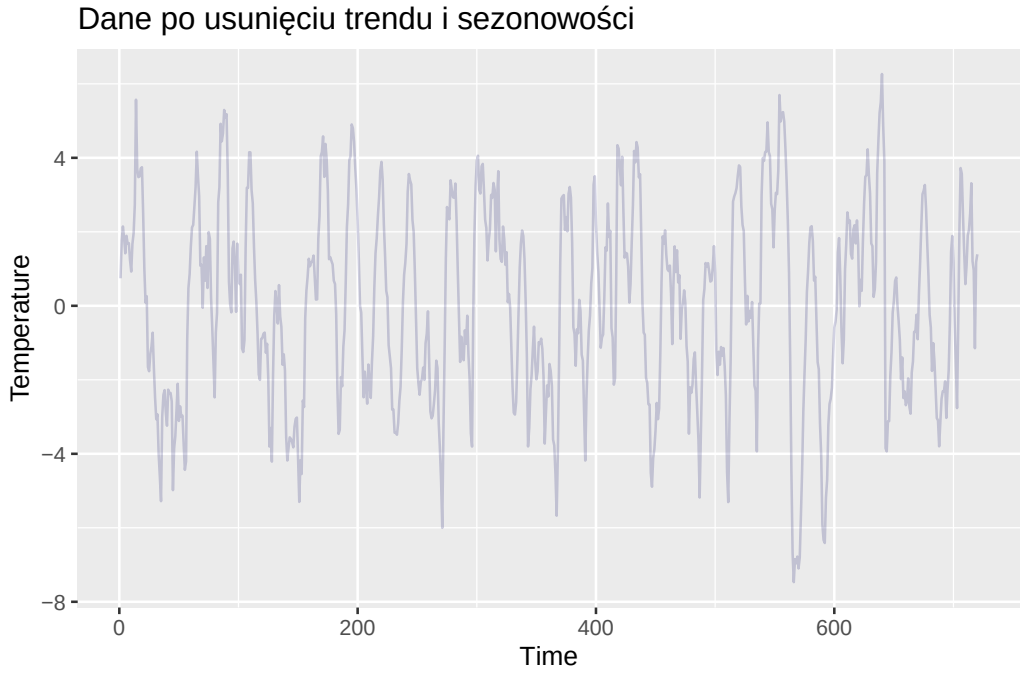
Rysunek 2.2. Wykres autokorelacji temperatury przed usunięciem trendu i sezonowości.



Rysunek 2.3. Wykres częściowej autokorelacji temperatury przed usunięciem trendu i sezonowości.



Rysunek 2.4. Wykres zależności temperatury od czasu z dopasowaną prostą regresji liniowej.



Rysunek 2.5. Wykres zależności temperatury od czasu po usunięciu trendu i sezonowości.

i	1	2	3	4	5
a_i	5.318	2.031	1.834	1.714	2.116
b_i	0.262	0.04791	0.03049	0.009047	0.02081
a_i	-2.804	1.034	3.679	-1.707	-0.8195

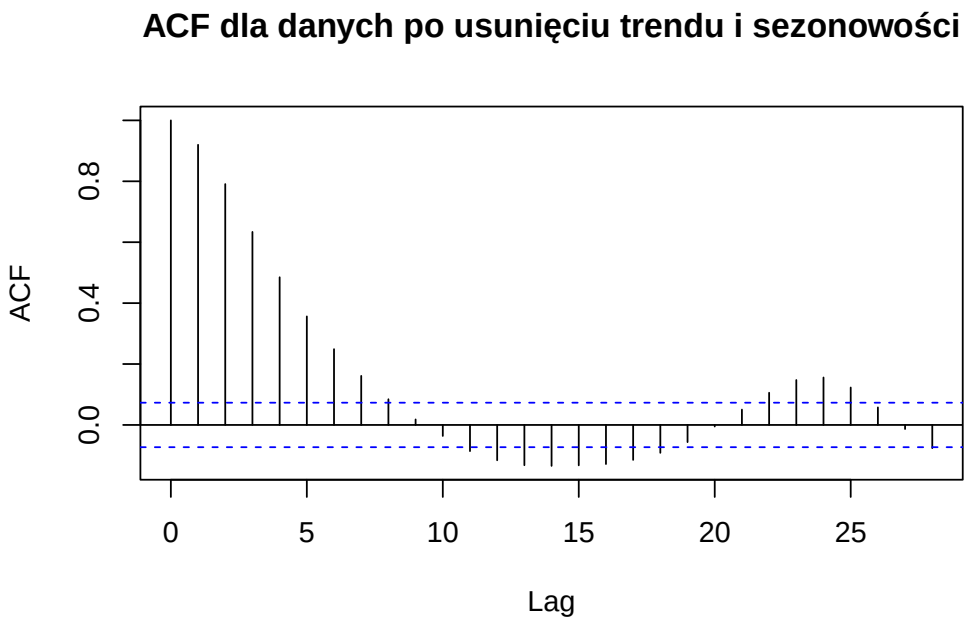
Tabela 2.1. Współczynniki modelu sumy sinusów.

W celu usunięcia sezonowości dopasowałyśmy do danych sumę pięciu sinusów. Model ten można zapisać wzorem:

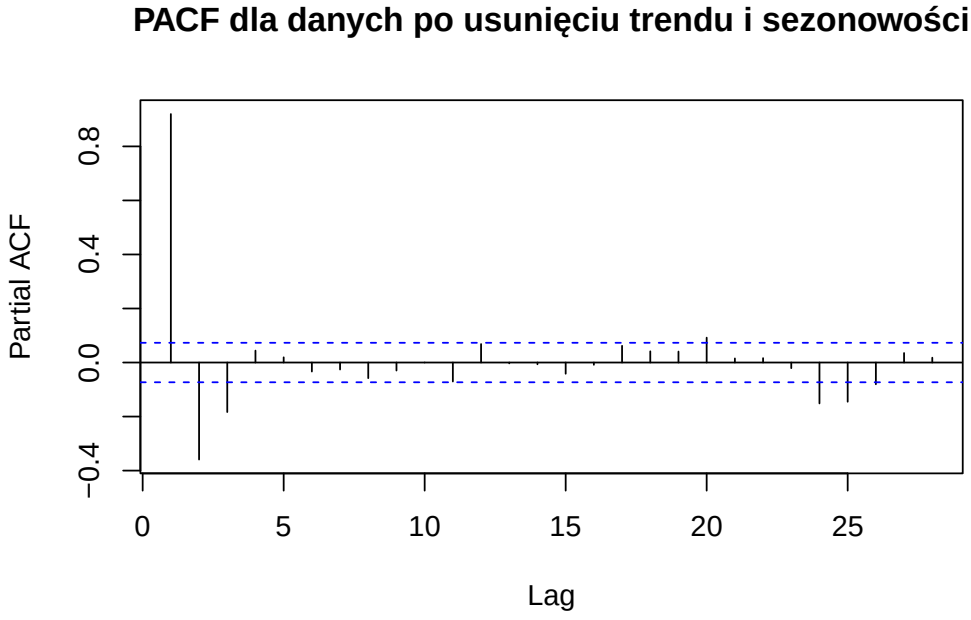
$$y_i = \sum_{i=1}^{5} a_i \, \sin(b_i \, x + c_i).$$

Wartości poszczególnych współczynników widoczne są w tabeli 2.1.

Po usunięciu trendu i sezonowości dane są dużo bardziej stacjonarne. Widoczne jest to na wykresie rozproszenia 2.5, na wykresie autokorelacji 2.6 oraz częściowej autokorelacji 2.7. Temperatura oscyluje wokół zera i ma stałą wariancję. Ponadto od pewnego momentu większość wartości autokorelacji oraz częściowej autokorelacji mieści się w przedziałach ufności. Niestety nie całą sezonowość udało się usunąć, jednak biorąc pod uwagę, że mamy do czynienia z danymi rzeczywistymi, jest to akceptowalne.



Rysunek 2.6. Wykres autokorelacji temperatury po usunięciu trendu i sezonowości.



Rysunek 2.7. Wykres częściowej autokorelacji temperatury po usunięciu trendu i sezonowości.

3. Dopasowanie modelu

Zanim przejdziemy do dopasowania modelu, przypomnimy definicję modelu ARMA. Szereg czasowy $\{X_t\}$ jest szeregiem ARMA(p,q), jeśli jest stacjonarny w słabym sensie (funkcja średniej i autokowariancji nie zależą od czasu) oraz spełnia następujące równanie:

$$X_t - \phi_1 X_{t-1} - \ldots - \phi_p X_{t-p} = Z_t - \theta_1 Z_{t-1} - \ldots - \theta_q X_{t-q},$$

gdzie $\{Z_t\} \tilde{} \; WN(0, \sigma^2)$ oraz wielomiany $\phi(z) = 1 - \phi_1 z - \ldots - \phi_p z^p$, $\theta(z) = 1 - \theta_1 z - \ldots - \theta_q z^q$ nie mają wspólnych pierwiastków.

Na początku znajdziemy optymalny rząd modelu, czyli p oraz q z powyższego wzoru, korzystając z kryteriów informacyjnych. Kryteria informacyjne AIC oraz BIC wyrażają się wzorami:

$$AIC = 2(p+q) - 2\ln L(z_1, \ldots, z_n),$$

$$BIC = \ln N \cdot (p+q) - 2\ln L(z_1, \ldots, z_n),$$

gdzie p - rząd części autoregresyjnej modelu, q - rząd części związanej ze średnią ruchomą, N - liczba obserwacji, L - funkcja największej wiarogodności.

Kryteria informacyjne są wskaźnikami dopasowania modelu. Można zauważyć, że maksymalizując funkcję największej wiarogodności, minimalizujemy kryteria informacyjne. Przyjmują one zatem najmniejszą wartość dla optymalnego rzędu modelu. Postanowiłyśmy dobrać model na podstawie kryterium Akaikego. Dla p oraz q z zakresu od 0 do 6 liczyłyśmy wartość AIC. Kryterium informacyjne przyjęło minimalną wartość dla $p = 4$ i $q = 6$. Nasz model ma więc postać:

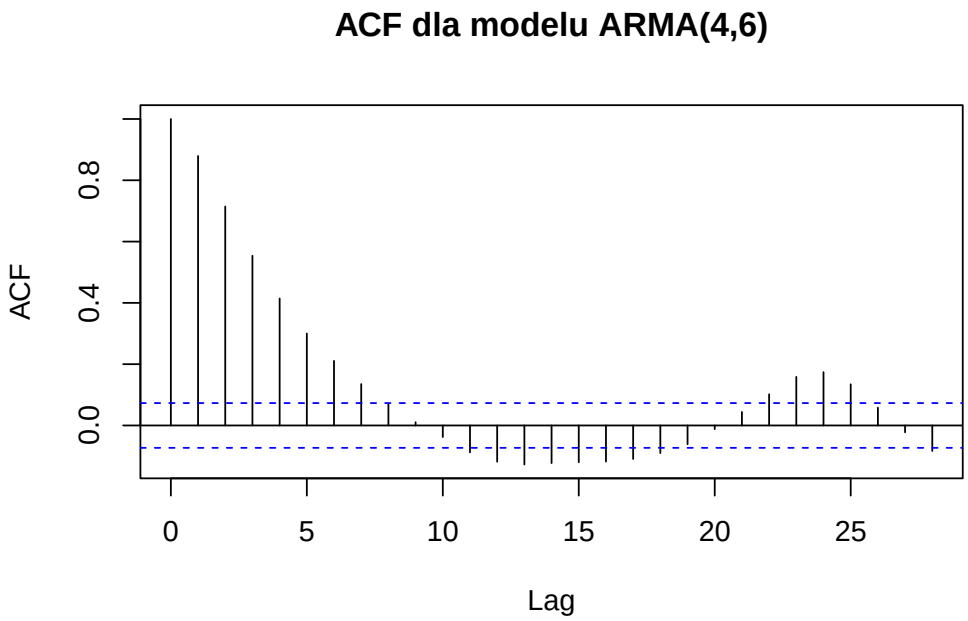
$$X_t - \phi_1 X_{t-1} - \ldots - \phi_4 X_{t-4} = Z_t - \theta_1 Z_{t-1} - \ldots - \theta_6 X_{t-6},$$

Następnie zajmiemy się estymacją parametrów ϕ_i oraz θ_i. Estymatory wyznaczone metodą największej wiarogodności przedstawione zostały w tabeli 3.1. Estymator wariancji białego szumu $\{Z_t\}$ wynosi 0.878.

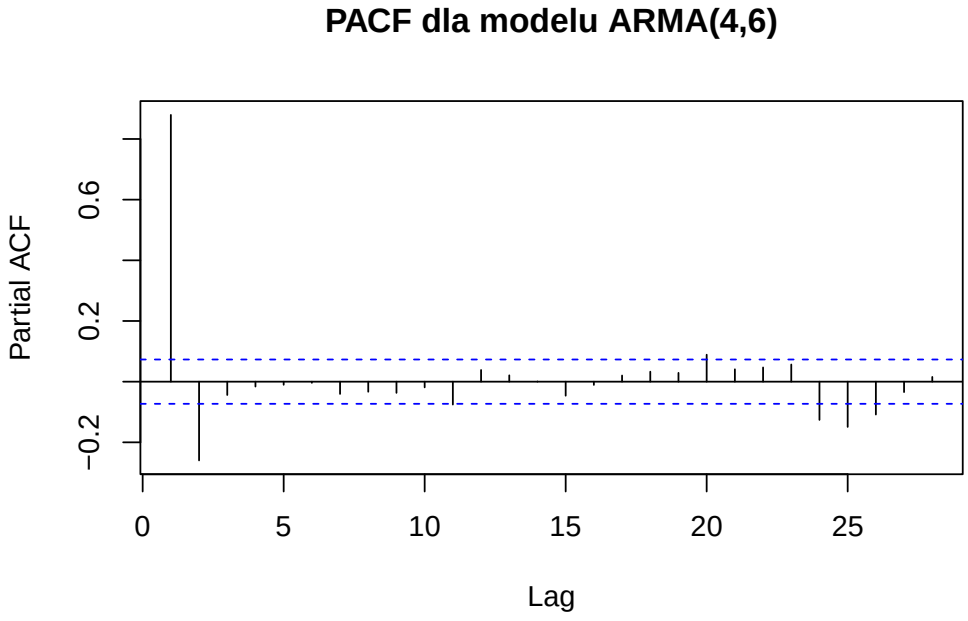
Widzimy, że funkcja autokorelacji 3.1 i częściowej autokorelacji 3.2 modelu teoretycznego dopasowanego do danych zachowuje się podobnie do wykresów 2.6 i 2.6 dla danych, co świadczy o tym, że model został dobrze dopasowany.

ϕ_1	ϕ_2	ϕ_3	ϕ_4	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
-0.437	0.543	0.643	-0.321	1.645	1.317	0.482	0.345	0.089	0.022

Tabela 3.1. Współczynniki modelu arma.



Rysunek 3.1. Wykres autokorelacji dla modelu teoretycznego.



Rysunek 3.2. Wykres częściowej autokorelacji dla modelu teoretycznego.

4. Weryfikacja poprawności modelu

4.1. Analiza residuów

Założenia o residuach $\{\epsilon_i\}_{i=1,2,...,n}$:

— $\forall_{i=1,...,n}\ \epsilon_i \backsim N(0, \sigma^2)$,
— $E\epsilon = 0$,
— $\epsilon_1, ..., \epsilon_n \rightarrow$ niezależne,
— $Var\epsilon_i = \sigma^2 \quad \forall_{i=1,...,n}$ - wariancja jest stała.

4.1.1. Normalność

W celu sprawdzenia czy residua mają rozkład normalny przyjrzyjmy się najpierw wykresom. Po spojrzeniu na histogram 4.1 i wykres dystrybuanty 4.2 zauważamy obecność skrajnych wartości, poza nimi obserwujemy podobieństwo do rozkładu normalnego. Natomiast wykres kwantylowy 4.3 wskazuje na brak normalności rozkładu residuów, widoczne sa różnice w ogonach.

Następnie policzmy kurtozę - 6.4626514. Dla rozkładu normalnego powinna ona być równa 3. W naszym przypadku jest jednak większa, zatem residua mają rozkład leptokurtyczny. Oznacza to, że intensywność wartości skrajnych jest większa niż dla rozkładu normalnego (ogony są cięższe).

Powyższe metody nie potwierdzaja w pełni zgodności rozkładu residuów z rozkładem normalnym. W celu dalszej analizy wykonamy test Kołomogorwa-Smirnowa, Andersona-Darlinga oraz Jarque-Bera.

	test	p.value
1	Kołogorwa-Smirnowa	0.07
2	Andersona-Darlinga	0.00
3	Jarque-Bera	0.00

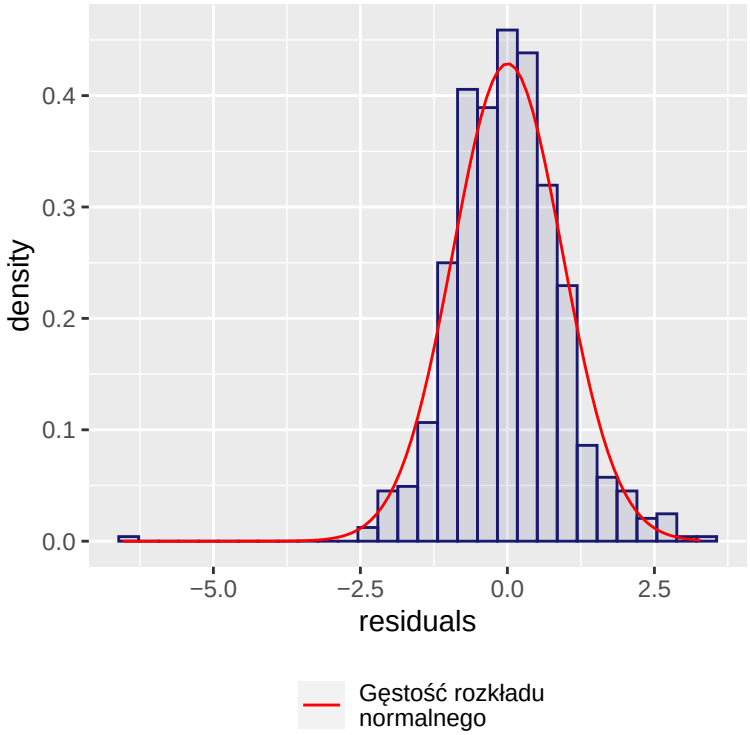
Tabela 4.1. Testy sprawdzające normalność residuów

Z tabeli 4.1 możemy odczytać, że dla testu Kołomogorowa-Smirnowa p-wartość wynosi 0.07 a zatem nie mamy podstaw do odrzucenia hipotezy zerowej - czyli normlaności rozkładu. Dla pozostałych testów p wartość jest bardzo bliska 0 (są one tak małe, że zostały zaokrąglone przez pakiet statystyczny do zera). Oznacza to, że odrzucamy hipotezę zerową o normalności residuów na rzecz hipotezy alternatywnej.

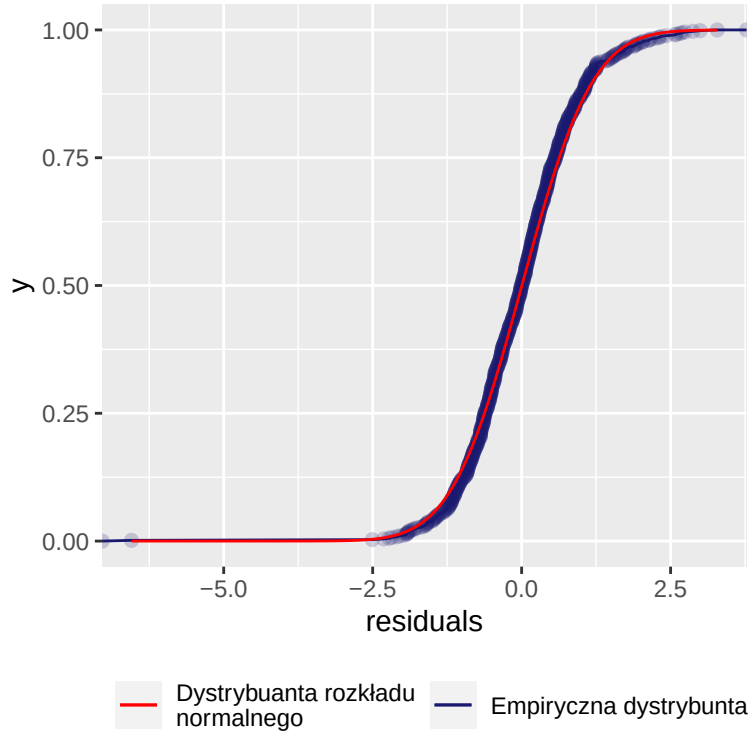
Możemy założyć, że reszty nie mają rozkładu normalnego.Mimo wszystko przeprowadzimy analizę pozostałych założeń.

4.1.2. Średnia

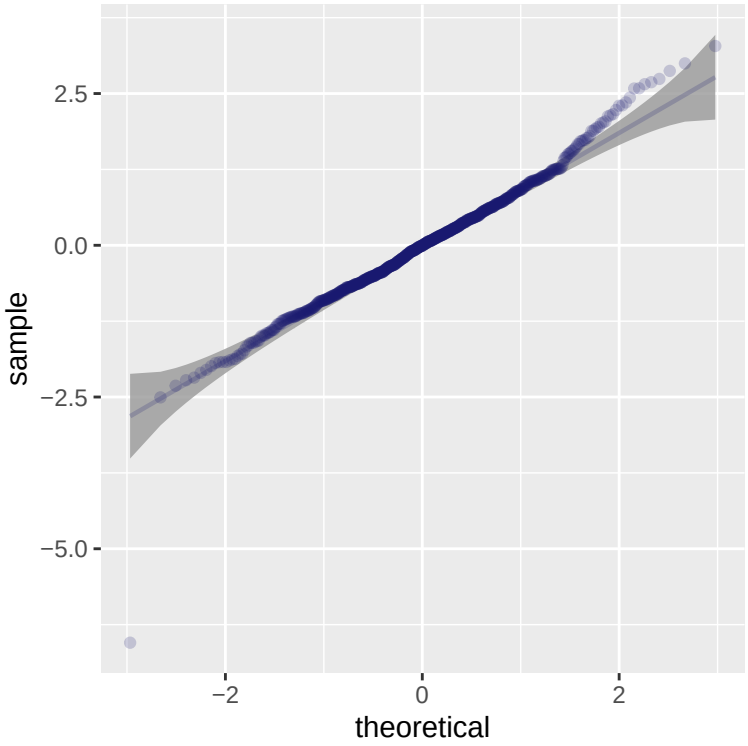
Aby sprawdzić czy $E\epsilon = 0$ narysujmy najpierw wykres rozproszenia.



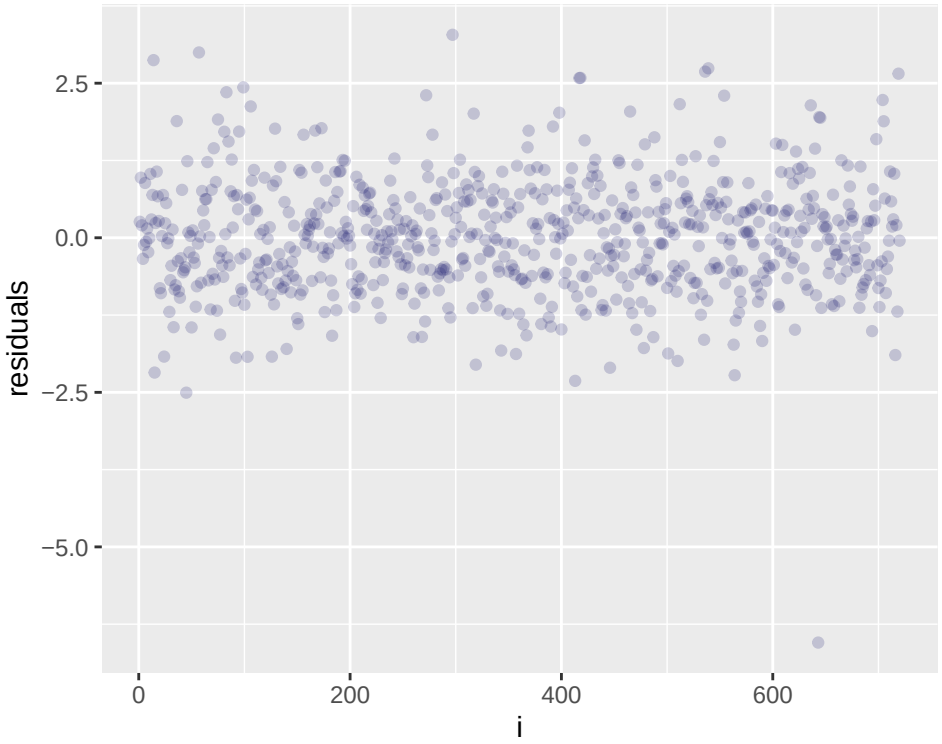
Rysunek 4.1. Histogram residuów oraz gęstość rozkładu normalnego



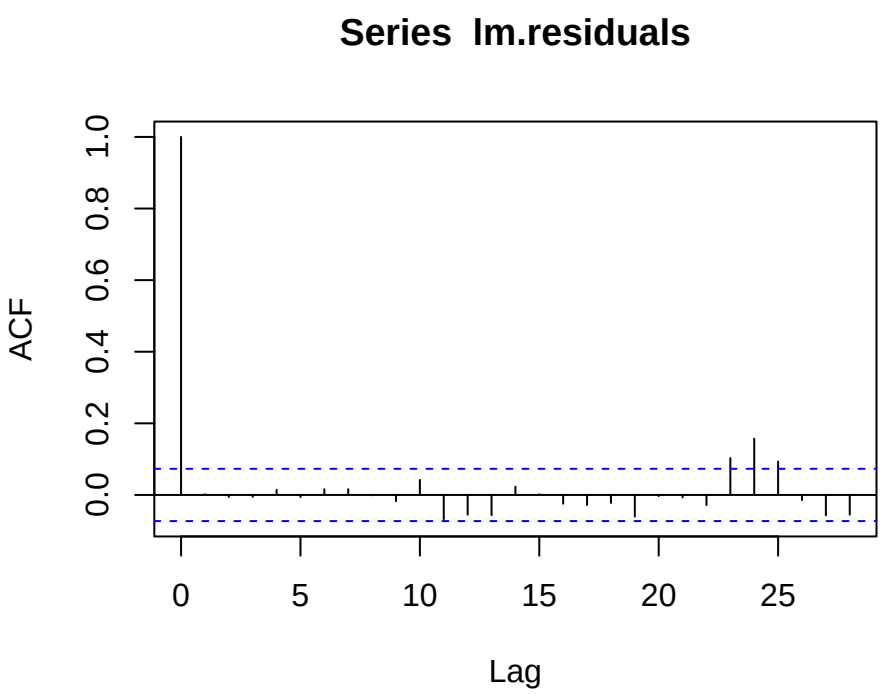
Rysunek 4.2. Dystrybuanta rozkładu normalnego oraz empiryczna residuów



Rysunek 4.3. Wykres kwanytlowy residuów



Rysunek 4.4. Wykres rozproszenia residuów



Rysunek 4.5. Funkcja autokorelacji residuów

Z wykresu 4.4 widzimy że ϵ_i oscyluja blisko 0. Średnia próbkowa wynosi 0.004355, więc jest bliska zeru. Wykonajmy jednak jeszcze test dla średniej (t.test). P wartość dla t testu, wynosi: 0.9000826. Nie mamy podstaw do odrzucenia hipotezy zerowej, więc możemy założyć, że średnia jest
równa 0.

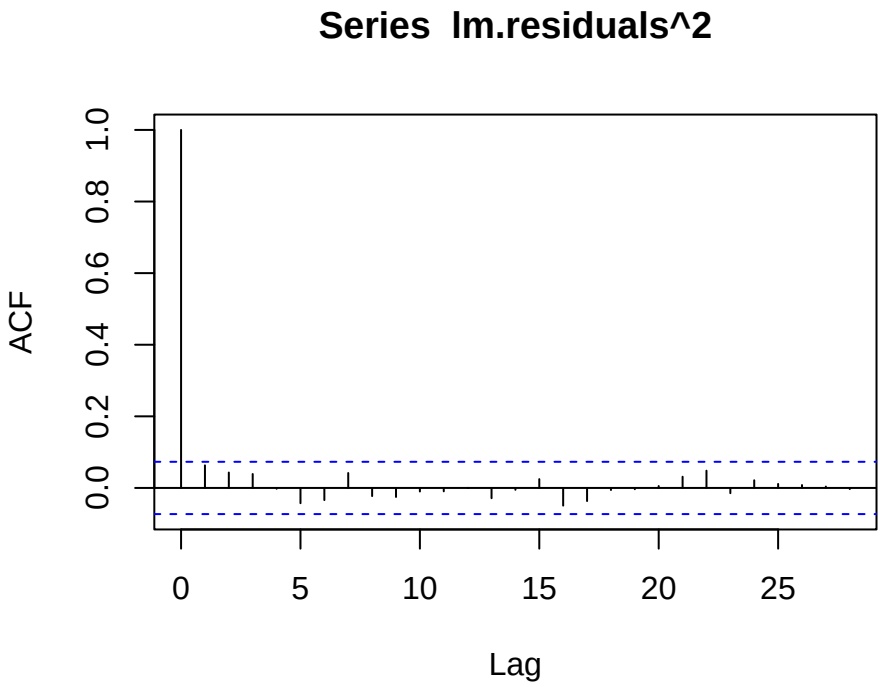
4.1.3. Niezależność reszt

Sprawdzimy teraz, czy residua są niezależne. Zaczniemy od narysowania wykresu autokorelacji. Jak widać na wykresie 4.5, punkty oscylują wokół zera, jendak nie wszystkie mieszczą się w przedziale ufności. Możemy także zauważyć pewną sezonowość.

Następnie przeprowadzimy test Ljunga-Boxa, który również pozwala ocenić czy występuje autokorelacja wśród reszt. P wartość w naszym przypadku wynosi 0.9565985. Oznacza to, że przy poziomie istotności 0.05 nie ma podstaw do odrzucenia hipotezy zerowej o braku korelacji między residuami. Na podstawie testu możemy, więc założyć, że są one nieskorelowane.

4.1.4. Stałość wariancji

Na koniec sprawdzimy założenie o stałości wariancji. Wykres 4.6 przedstawia autokorelację reszt podniesionych, do kwadratu. Jeśli residua podniesione do kwadratu są nieskorelowane, oznacza to, że są one homoskedastyczne, czyli mają stałą wariancję. Potwierdza to arch test, dla którego p wartość wynosi: 0.6596531. Jest ona większa od ustalonego poziomu istotności $\alpha = 0.05$.

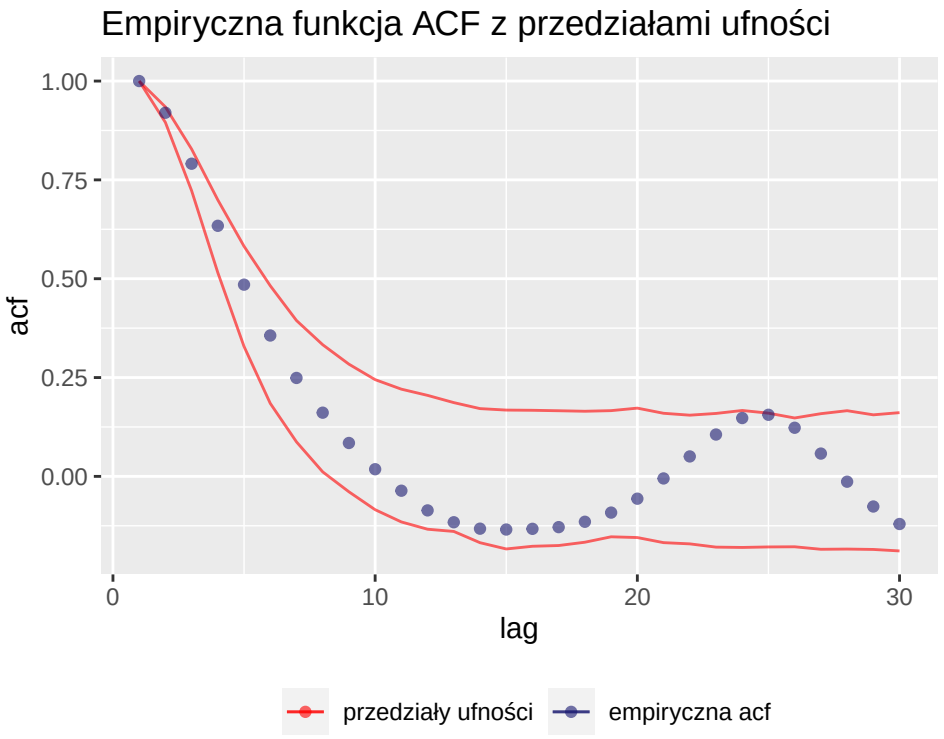


Rysunek 4.6. Funkcja autokorelacji residuów podniesionych do kwadratu

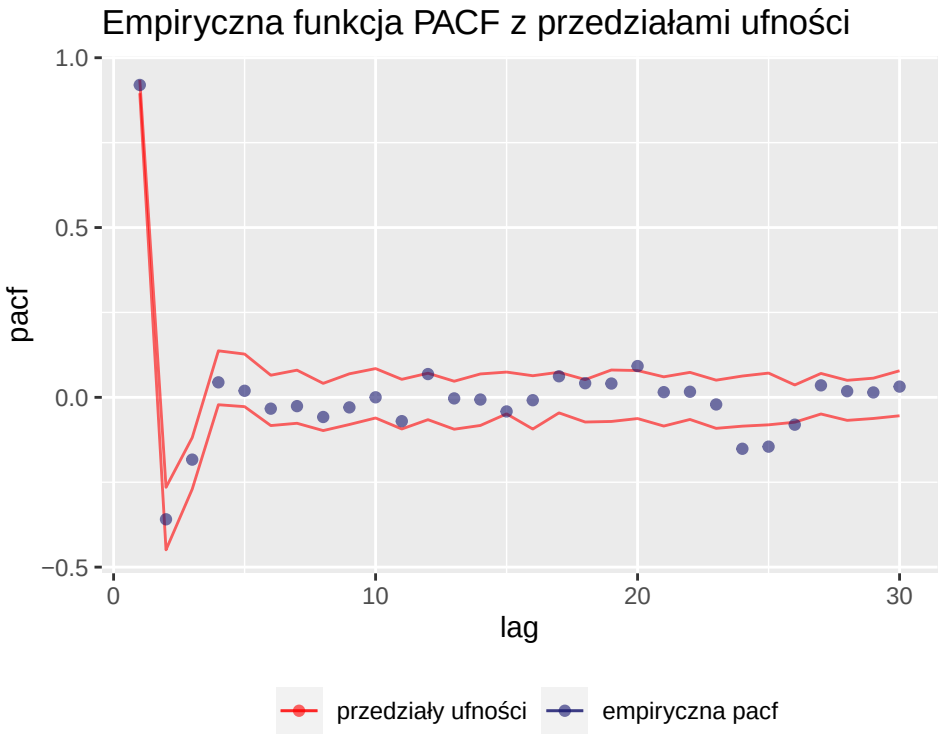
Możemy więc założyć, że hipoteza zerowa jest prawdziwa i residua są homoskedastyczne.

4.2. Przedziały ufności dla funkcji autokwariancji i autokorelacji

W celu dalszej weryfikacji naszego modelu porównamy empiryczne funkcje autokorelacji oraz częściowej autokorelacji z przedziałami ufności wyznaczonymi symulacyjnie. Jak widzimy na 4.7 oraz 4.8, zarówno ACF jak i PACF całkiem dobrze wpasowują się w przedziały ufności. Jednak zauważamy również pewną sezonowość, która momentami sprawia że funkcje znajdują się poza przedziałami ufności.



Rysunek 4.7. Empiryczna funkcja autokorelacji z przedziałami ufności wyznaczonymi symulacyjnie



Rysunek 4.8. Empiryczna funkcja częściowej autokorelacji z przedziałami ufności wyznaczonymi symulacyjnie

5. Podsumowanie

W celu analizy szeregu czasowego wybrałyśmy dane pogodowe dotyczące temperatury odnotowanej w Szegedzie w kwietniu 2016 roku. Pierwszym krokiem do doboru modelu ARMA jest dekompozycja danych. W tym celu dopasowałyśmy model regresji liniowej, aby pozbyć się trendu liniowego. Po otrzymaniu współczynników $\beta_0 = 16,4797$ oraz $\beta_1 = -0.0087$ odjęłyśmy funkcję liniową od danych. W celu usunięcia sezonowości dopasowałyśmy sumę pięciu sinusów. Powyższe kroki pozwoliły na usunięcie trendu liniowego, jednak w danych została pewna sezonowość. Kolejnym krokiem, który można w przyszłość zastosować to dopasowanie sumy większej liczby sinusów lub zastosowanie metody różnicowania. Efekt dekompozycji to bardziej stacjonarne dane, różnicę możemy zobaczyć porównując funkcję autokorelacji (2.2, 2.6) oraz częściowej autokorelacji (2.3, 2.7) odpowiednio dla danych surowych oraz po usunięciu części sezonowości i trendu liniowego. Na podstawie kryterium Akaikiego (AIC) dobrałyśmy parametry $p = 4$ oraz $q = 6$. Poprawność dobranych współczynników ϕ_i oraz θ_i wnioskujemy po przyjrzeniu się wykresom funkcji autokorelcji oraz częściowej autokorelacji dla danych (2.6, 2.7) oraz wysymulowanej trajketori modelu (3.1, 3.2). Są one niemal identyczne. W celu weryfikacji poprawności modelu sprawdziłyśmy założenia o residuach. Niestety założenie o normalności nie zostało spełnone. Kolejnym sposób na weryfikacje modelu to wykresy funkcji autokorelacji i częściowej autokorelacji z przedziałami ufności. Patrząc na wykresy 4.7 oraz 4.8 widzimy, że nasz model został dosyć dobrze dopasowany. Nieznaczne wyjście poza przedziały ufności może być spowodowane nie do końca usuniętą sezonowością. W dalszej analizie możemy również spróbować dobrać parametry p oraz q na podstawie innych kryteriów informacyjnych.