# Analiza danych rzeczywistych przy pomocy modelu ARMA

Natalia Klepacka, Joanna Kołaczek

9 lutego 2023

# Spis treści

**1 Wstęp** **2**

**2 Przygotowanie danych do analizy** **2**

**3 Modelowanie danych przy pomocy ARMA** **5**
3.1 Dobór rzędu modelu . . . . . 5
3.2 Estymacja parametrów modelu . . . . . 7

**4 Ocena dopasowania modelu** **7**

**5 Weryfikacja założeń dotyczących szumu** **7**
5.1 Średnia . . . . . 8
5.2 Wariancja . . . . . 8
5.3 Niezależność . . . . . 9

**6 Podsumowanie** **11**

**7 Źródła** **11**

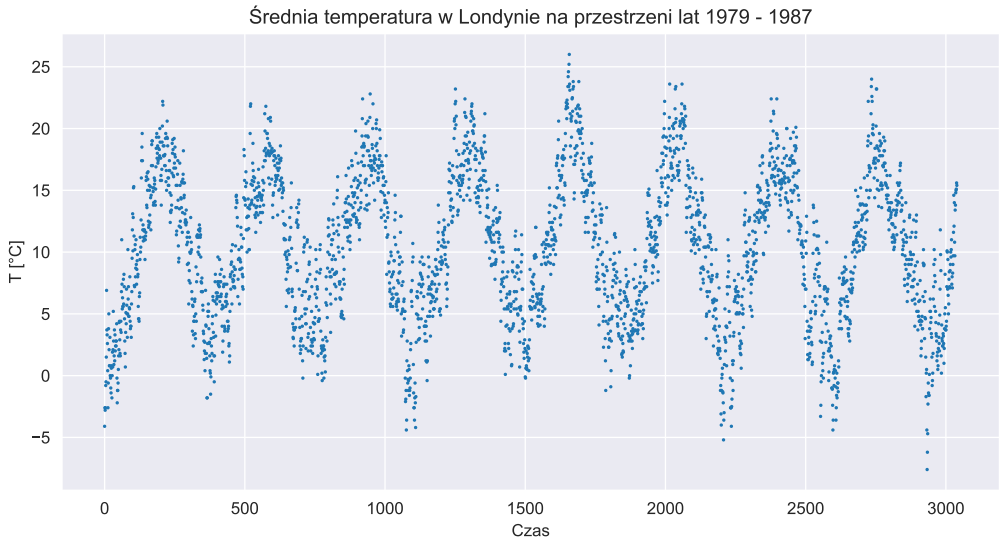
# 1 Wstęp

Niniejszy raport powstał na potrzeby realizacji laboratorium z Komputerowej Analizy Szeregów Czasowych, prowadzonych przez mgr Justynę Witulską, do wykładu prof. Agnieszki Wyłomańskiej. Będziemy analizować dane dotyczące średniej dziennej temperatury w Londynie, na przestrzeni lat 1979-1987. Dane pochodzą *z tej strony*. Są to informacje kolekcjonowane przez European Climate Assessment and Dataset - projekt zbierający dane o pogodzie w Europie. W raporcie przeprowadzimy dekompozycji Walda oraz przy pomocy kryteriów informacyjnych dobierzemy rzędy modelu ARMA, aby następnie wyestymować wartości parametrów tegoż modelu. Na koniec zweryfikujemy również, czy założenia dotyczące szumu są spełnione.

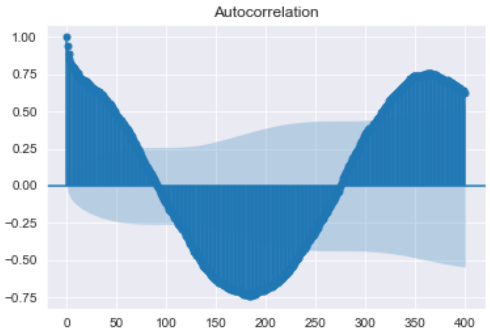
Życzymy Czytelnikowi miłej lektury.

# 2 Przygotowanie danych do analizy

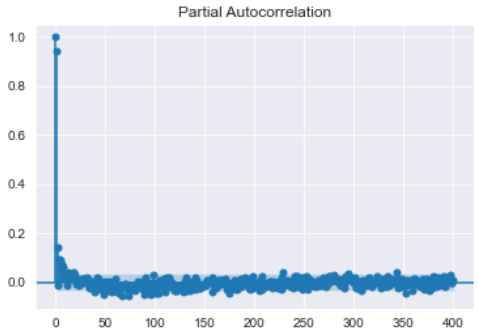
Na wykresie [1] przedstawiona jest zależność średniej dziennej temperatury w Londynie od czasu. Wyraźnie widoczna jest sezonowość, natomiast nie możemy być pewni co do obecności trendu. Wykresy autokorelacji [2] oraz częściowej autokorelacji [3] potwierdzają, że nie możemy mówić tu o szeregu stacjonarnym.




Rysunek 1: Wykres temperatury w Londynie

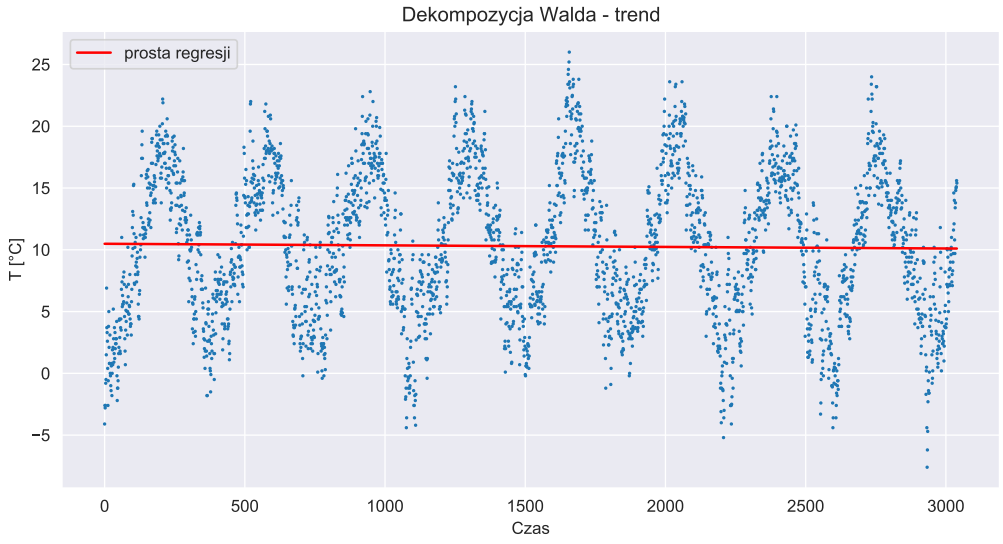



Rysunek 2: Autokorelacja




Rysunek 3: Częściowa autokorelacja

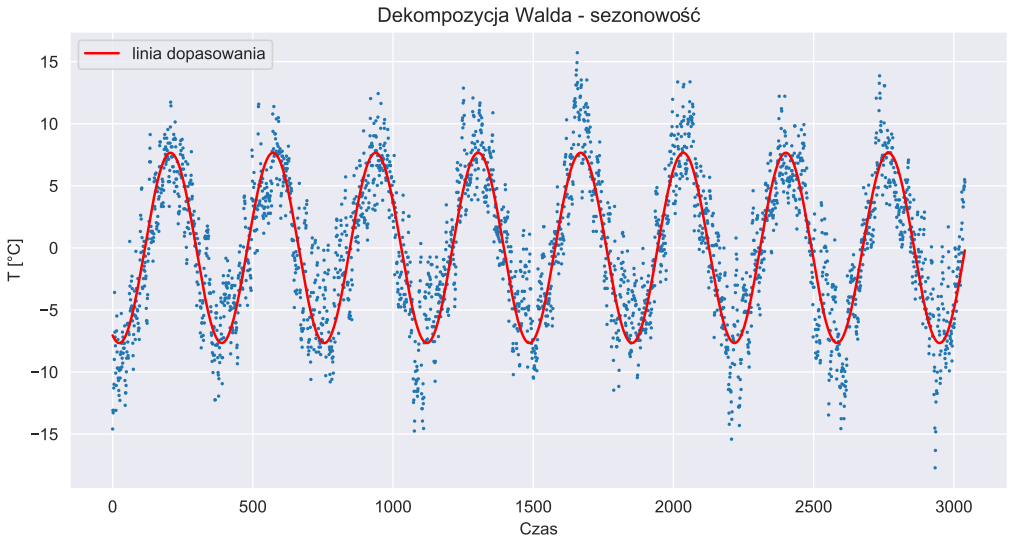
Aby usunąć możliwy trend, stosujemy dla naszych danych regresję liniową. Efekt widzimy na wykresie [4]. Obliczony współczynnik kierunkowy był bardzo bliski 0, co oznacza, że w danych nie występuje trend liniowy. Wyraz wolny wyniósł około 10, więc odejmujemy tę wartość od wartości oryginalnych, aby ich średnia była bliska 0.




Rysunek 4: Regresja liniowa

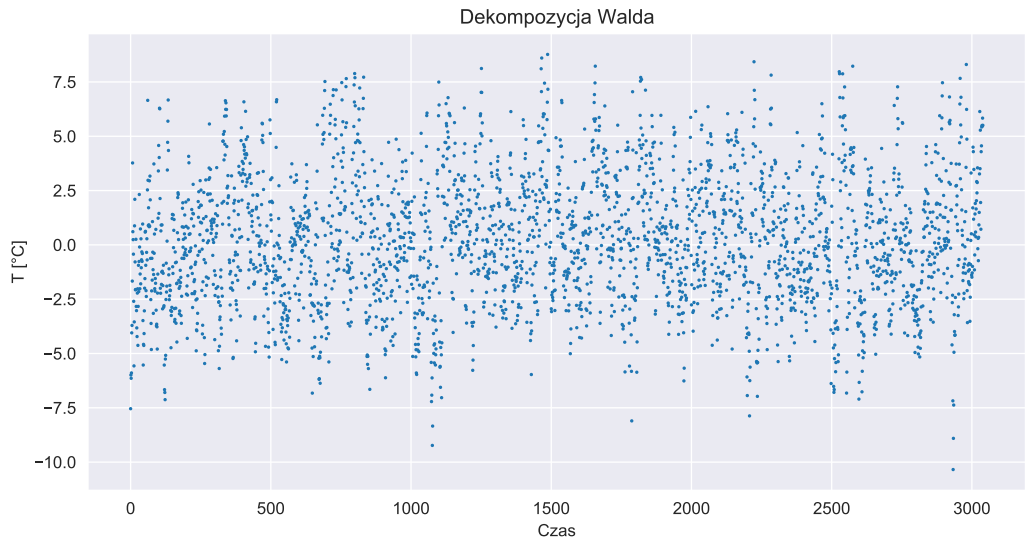
Kolejnym krokiem będzie próba usunięcia sezonowości. Na poprzednich wykresach mogliśmy zauważyć, że dane układają się w kształt funkcji sinusoidalnej. Załóżmy, że da się je opisać przy pomocy funkcji $f(x) = c \cdot \sin(d \cdot x + e)$. Używamy pakietu `scipy`, a konkretnie funkcji `optimize.curve_fit`,

aby dopasować odpowiednie współczynniki $c$, $d$, $e$. Efekt widzimy na wykresie 5.

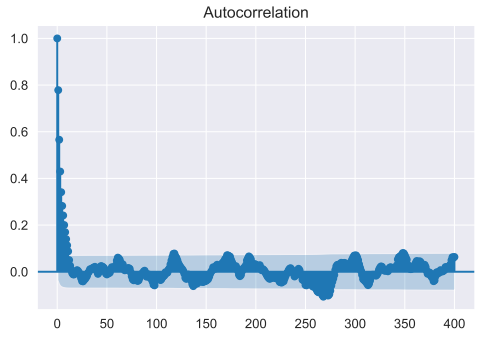



Rysunek 5: Krzywa sinusoidalna

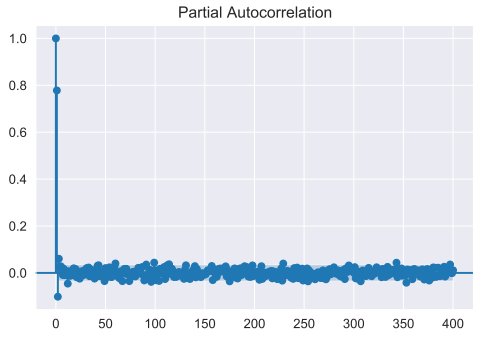
Aby dokończyć proces dekompozycji, wystarczy od naszych danych odjąć wartości otrzymanej funkcji w danym czasie [6]. Możemy teraz ponownie sprawdzić, jak prezentują się wykresy autokorelacji i częściowej autokorelacji. Od tego momentu będziemy zakładać, że szereg czasowy jest stacjonarny w słabym sensie.




Rysunek 6: Dane po dekompozycji




Rysunek 7: Autokorelacja




Rysunek 8: Częściowa autokorelacja

# 3 Modelowanie danych przy pomocy ARMA

## 3.1 Dobór rzędu modelu

W celu dobrania rzędu modelu wyliczyłyśmy wartości kryteriów informacyjnych

- Kryterium Informacyjne Akaikego (AIC),
- Bayesowskie Kryterium Informacyjne (BIC),
- Kryterium Informacyjne Hannana-Quinna (HQIC)

dla wartości $p$ i $q$ z przedziału [0,9]. Otrzymałyśmy winiki widoczne w tabelach [1][2][3]. Najlepiej dopasowane pary wg poszczególnych kryteriów to

- $p = 5$, $q = 6$ dla AIC,
- $p = 1$, $q = 1$ dla BIC,
- $p = 3$, $q = 0$ dla HQIC.

Jako że nie da się określić modelu jednoznacznie najlepiej dopasowanego, do dalszej analizy zdecydowałyśmy się przyjąć najmniej skomplikowany model, czyli ARMA(1, 1).

| p | q | AIC | BIC | HQIC |
|---|---|---|---|---|
| 5 | 6 | 10933.960861 | 11010.577298 | 10961.678129 |
| 7 | 8 | 10934.090625 | 11034.281350 | 10970.336283 |
| 9 | 7 | 10934.984168 | 11041.068465 | 10973.361923 |
| 7 | 9 | 10935.015702 | 11041.100000 | 10973.393458 |
| 4 | 4 | 10935.166047 | 10994.101768 | 10956.487023 |

Tabela 1: Kryteria informacyjne wg AIC

| p | q | AIC | BIC | HQIC |
|---|---|---|---|---|
| 1 | 1 | 10945.997929 | 10969.572218 | 10954.526320 |
| 3 | 0 | 10940.398608 | 10969.866469 | 10951.059096 |
| 1 | 2 | 10941.710972 | 10971.178832 | 10952.371459 |
| 2 | 1 | 10944.132071 | 10973.599931 | 10954.792559 |
| 2 | 2 | 10938.737268 | 10974.098700 | 10951.529853 |

Tabela 2: Kryteria informacyjne wg BIC

| p | q | AIC | BIC | HQIC |
|---|---|---|---|---|
| 3 | 0 | 10940.398608 | 10969.866469 | 10951.059096 |
| 2 | 2 | 10938.737268 | 10974.098700 | 10951.529853 |
| 3 | 1 | 10938.935705 | 10974.297137 | 10951.728290 |
| 1 | 2 | 10941.710972 | 10971.178832 | 10952.371459 |
| 1 | 3 | 10940.538468 | 10975.899900 | 10953.331053 |

Tabela 3: Kryteria informacyjne wg HQIC

Jako że nie da się określić modelu jednoznacznie najlepiej dopasowanego, do dalszej analizy zdecydowałyśmy się przyjąć najmniej skomplikowany model, czyli ARMA(1, 1).

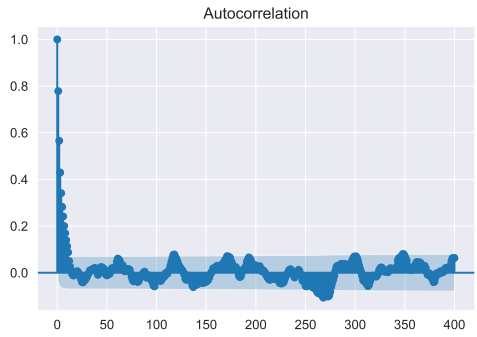
## 3.2 Estymacja parametrów modelu

Do estymacji parametrów wykorzystamy metodę największej wiarygodności zaimplementowaną w pythonowym pakiecie `statsmodels`. Metoda ta zakłada, że zmienne składające się na szum mają rozkład normalny. Wartości współczynników można zobaczyć w tabeli [4]

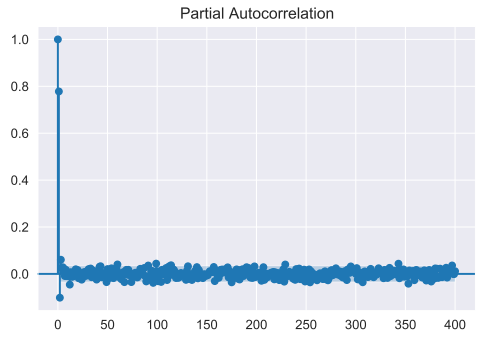
| | coef | std err | z | P>\|z\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.2230 | 0.150 | 1.486 | 0.137 | -0.071 | 0.517 |
| ar.L1 | 0.7205 | 0.017 | 42.821 | 0.000 | 0.687 | 0.753 |
| ma.L1 | 0.1546 | 0.023 | 6.580 | 0.000 | 0.109 | 0.201 |
| sigma2 | 3.4663 | 0.096 | 35.996 | 0.000 | 3.278 | 3.655 |

Tabela 4: Współczynniki modelu

# 4 Ocena dopasowania modelu




Rysunek 9: Autokorelacja




Rysunek 10: Częściowa autokorelacja

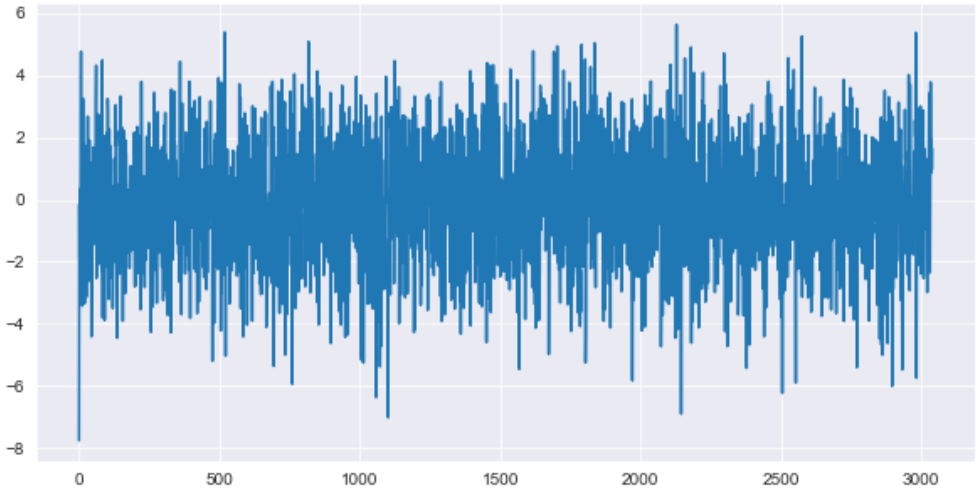
Od pewnego momentu wszystkie wartości funkcji autokorelacji i częściowej autokorelacji mieszczą się w przedziałach ufności, co sugeruje, że model jest dobrze dopasowany do danych.

# 5 Weryfikacja założeń dotyczących szumu

Szum $Z_t$ w modelu ARMA powinien spełniać poniższe założenia:

1. Średnia $Z_t$ jest bliska 0
2. Wariancja $Z_t$ jest taka sama dla każdego $t$
3. $Z_t$ są niezależne

4. $Z_t$ mają rozkład normalny




Rysunek 11: Residua modelu

## 5.1 Średnia

Na podstawie wykresu[11] możemy stwierdzić, że średnia prawdopodobnie jest równa 0. Dodatkowym sprawdzeniem może być wykonanie testu t. Test ten sprawdza hipotezę zerową "Średnia z próby jest równa $\mu$" przeciwko hipotezie alternatywnej "Średnia z próby jest różna od $\mu$". Jak widać poniżej[1], wynik testu potwierdza wnioski wyciągnięte z wykresu — nie ma podstaw do odrzucenia hipotezy zerowej.

```
1    sp.stats.ttest_1samp(resid, popmean=0)
2
3    #output
4    Ttest_1sampResult(statistic=0.04706184306961072, pvalue=0.9624674462497231)
```

Kod 1: Test t

## 5.2 Wariancja

Na wykresie[11] nie widać znaczących zmian w wariancji zależnych od czasu, jednak dla pewności wykonamy test Levene'a jednorodności wariancji. Test ten sprawdza hipotezę zerową "Wszystkie próbki pochodzą z populacji o tej samej wariancji" przeciwko hipotezie alternatywnej "Przynajmniej dwie próbki pochodzą z populacji o różnych wariancjach". W tym celu dzieli próbkę na mniejsze podgrupy i porównuje ze sobą ich mediany.

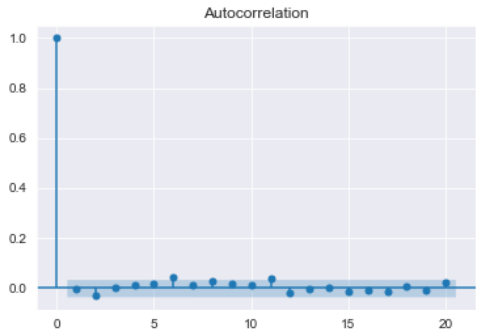
```
stat, p_value = scipy.stats.levene(random.sample(model.resid,500),random.sample(
model.resid,500))
if p_value > 0.05:
print("Wariancja jest raczej stala.")
else:
print("Wariancja raczej nie jest stala.")

#output
Wariancja jest raczej stala.
```

Kod 2: Test Levene'a

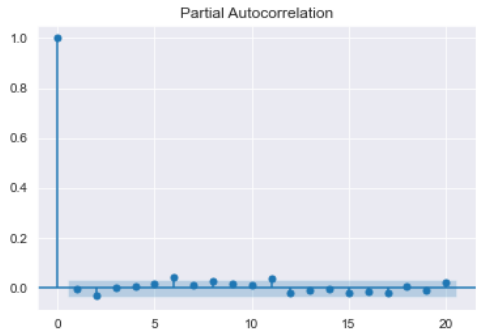
Zgodnie z wynikiem testu[2] nie ma podstaw do odrzucenia hipotezy zerowej, zatem możemy przyjąć, że wariancja residuów prawdopodobnie jest stała.

## 5.3 Niezależność

Z wykresów funkcji autokorelacji[12] i częściowej autokorelacji[13] można łatwo wywnioskować, że residua są realizacjami niezależnych zmiennych losowych.




Rysunek 12: Autokorelacja

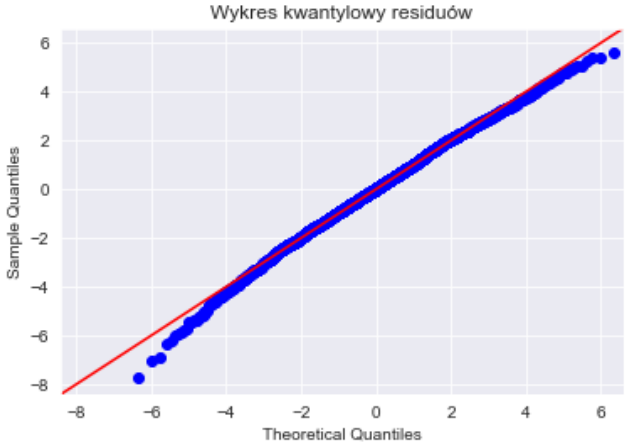



Rysunek 13: Częściowa autokorelacja

Dystrybuanta rozkładu residuów niemal idealnie pokrywa się z dystrybuantą rozkładu normalnego[14]. Podobnie histogram wartości[15] dobrze przybliża teoretyczną gęstość. Wykres kwantylowy[16] pokrywa się odrobinę gorzej, ale nadal wystarczająco dobrze, żebyśmy mogli uznać, że dane pochodzą z rozkładu normalnego.


Rysunek 14: Dystrybuanta residuów


Rysunek 15: Gęstość residuów




Rysunek 16: Wykres kwantylowy

# 6 Podsumowanie

Na podstawie przeprowadzonej analizy możemy stwierdzić, że między poziomem szczęścia a PKB per capita dla danego kraju występuje dosyć silna zależność liniowa. Niestety badanie residuów wykazało, że nie mają one rozkładu normalnego. Oznacza to, że co prawda punktowe estymacje zostały przez nas wykonane poprawnie, jednak wszystkie estymacje przedziałowe oparte były o założenie o normalności rozkładu residuów, zatem nie mają zastosowania w tym modelu. Niestety nie znając rozkładu $\epsilon$, nie możemy wyznaczyć faktycznych przedziałów ufności.

# 7 Źródła


- Wykłady
- Dokumentacja pakietów `statsmodels` i `scipy`
- `https://www.kaggle.com/datasets/emmanuelfwerr/london-weather-data`
- `https://en.wikipedia.org/wiki/Student%27s_t-test`
- `https://en.wikipedia.org/wiki/Levene%27s_test`