
Raport
Komputerowa Analiza Szeregów Czasowych

Natalia Lach 262303, Alicja Myśliwiec 262275

Matematyka Stosowana
Wydział Matematyki Politechniki Wrocławskiej

Spis treści

1. Wstęp	2
2. Opis danych	2
3. Statystyki opisowe	2
4. Analiza graficzna rozkładów X i Y	3
5. Dobranie regresji liniowej dla danych	6
5.1. Wykres rozproszenia oraz prosta regresji	6
5.2. Wyznaczenie R^2	7
6. Przedziały ufności	8
6.1. Dla parametru β_1	9
6.2. Dla parametru β_0	9
6.3. Wyniki	9
7. Analiza residuów	9
7.1. Analiza średniej	10
7.2. Analiza wariancji	11
7.3. Analiza rozkładu	12
7.4. Analiza korelacji	14
8. Wnioski	15

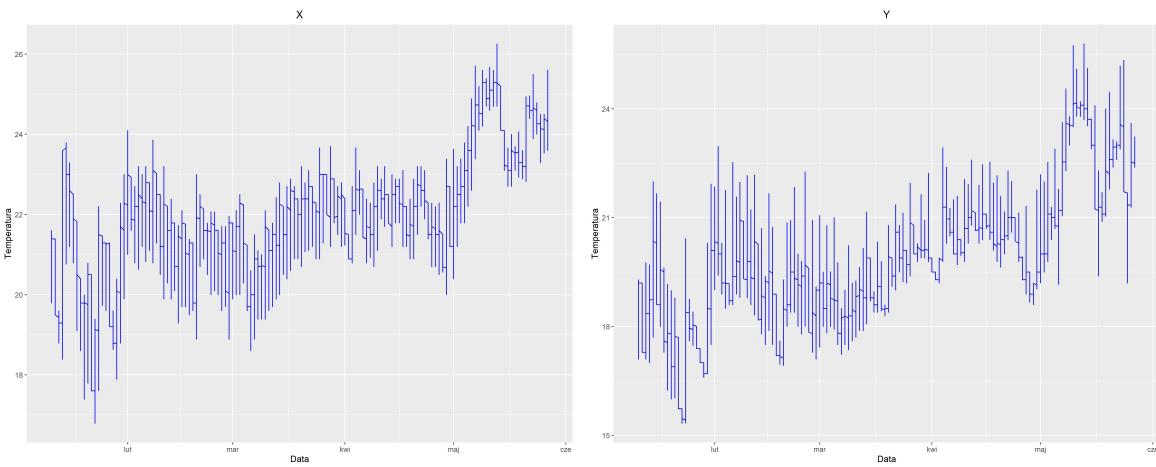
1. Wstęp

W niniejszej pracy przedstawiono analizę danych opisujących temperaturę zanotowaną w kuchni oraz łazience pewnego domostwa.¹

Analiza zostanie przeprowadzona przy pomocy regresji liniowej. Uzyskane wyniki zostaną porównane z teoretycznymi modelami. Wszelkie działania na wybranym zbiorze danych oraz wizualizacja przeprowadzono w języku R.

2. Opis danych

W wybranych zbiorach danych, X opisuje temperaturę w kuchni, Y zaś temperaturę w łazience odnotowaną w tym samym czasie. Pomiary były wykonywane podczas półrocznego okresu od stycznia do czerwca 2016 roku - każdy ze zbiorów zawiera 19735 wartości. Odczyt temperatur aktualizowano co 10 minut.



Rys. 1: Wykresy zależności temperatury od czasu odpowiednio w kuchni i łazience

Można odczytać z wykresów przedstawionych na rysunku 1, iż temperatura na przestrzeni miesiącą rosła. Biorąc pod uwagę odpowiadające im pory roku, można dojść do wniosku, że temperatura panująca na zewnątrz również mogła mieć wpływ na otrzymywane wartości. Zauważalne jest podobieństwo w charakterystyce obu wykresów.

3. Statystyki opisowe

W celu dokładniejszego opisania zbiorów, wykorzystane zostaną odpowiednie statystyki:

— średnia

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad (1)$$

— wariancja

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad (2)$$

— odchylenie standardowe

$$s = \sqrt{\sigma^2}, \quad (3)$$

¹ Dane pochodzą ze strony kaggle

<https://www.kaggle.com/datasets/loveall/appliances-energy-prediction!>

— kwantyle

$$Q_p = x_{[p(n-1)]+1}, \quad (4)$$

Szczególnym kwantylem jest mediana $x_{med} = Q_2 = \begin{cases} \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}), & \text{dla } n \text{ parzystych} \\ x_{(\frac{n+1}{2})}, & \text{dla } n \text{ nieparzystych} \end{cases}$.

Otrzymane wyniki zostaną przedstawione w tabeli.

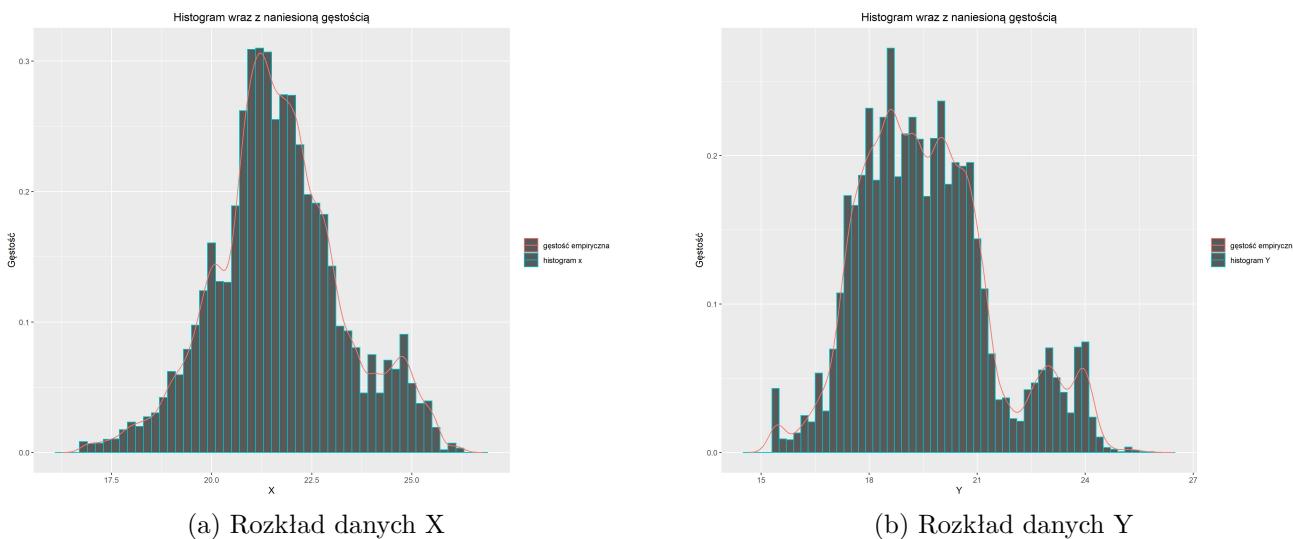
	X	Y
\bar{x}	21.687	19.592
σ^2	2.579	3.403
s	1.606	1.845
Q_1	20.76	18.2775
Q_2	21.6	19.39
Q_3	22.6	20.619

Tab. 1: Statystyki opisowe rozkładów X i Y

Z tabeli 1 wynika, że oba zbiory danych delikatnie różnią się od siebie. Średnia temperatura w kuchni jest ponad 2 stopnie wyższa niż w łazience, ale jej wariancja osiągnęła mniejszą wartość. Oznacza to nieco mniejsze wahania temperatur w tym pomieszczeniu. Rozstęp międzykwartylowy osiąga podobne wartości dla obu zbiorów danych.

4. Analiza graficzna rozkładów X i Y

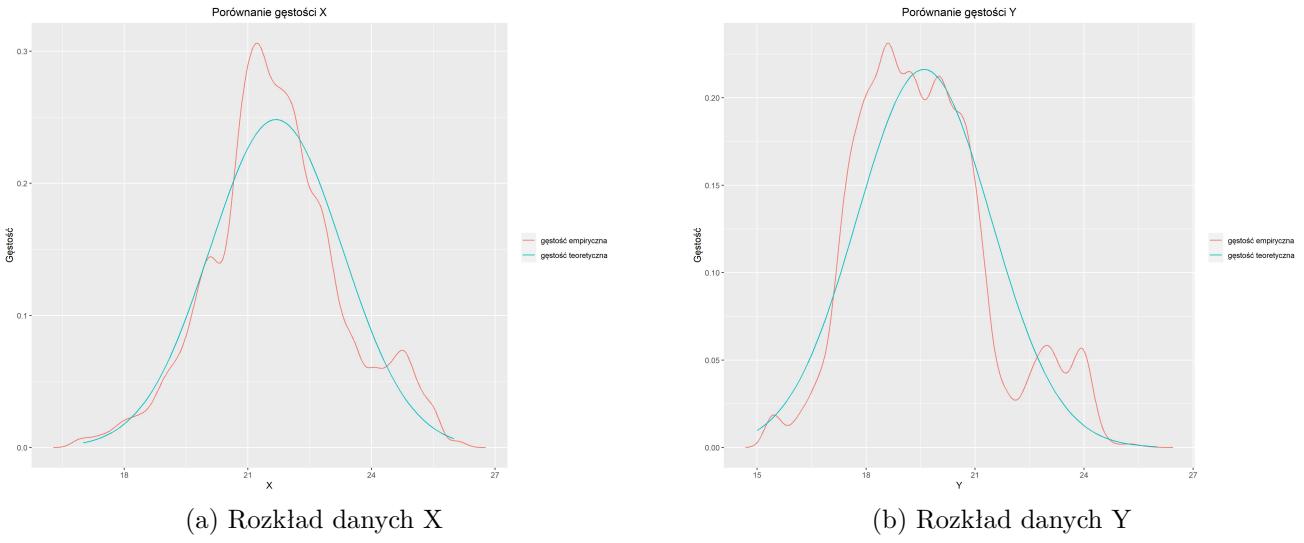
Poniżej przedstawiona zostanie analiza rozważanych rozkładów danych pod względem histogramów, gęstości oraz dystrybuant empirycznych. Korzystając z danych zawartych w tabeli 1, wykorzystamy obliczone średnią i wariancję do dopasowania teoretycznego rozkładu.



Rys. 2: Histogramy wraz z gęstością empiryczną

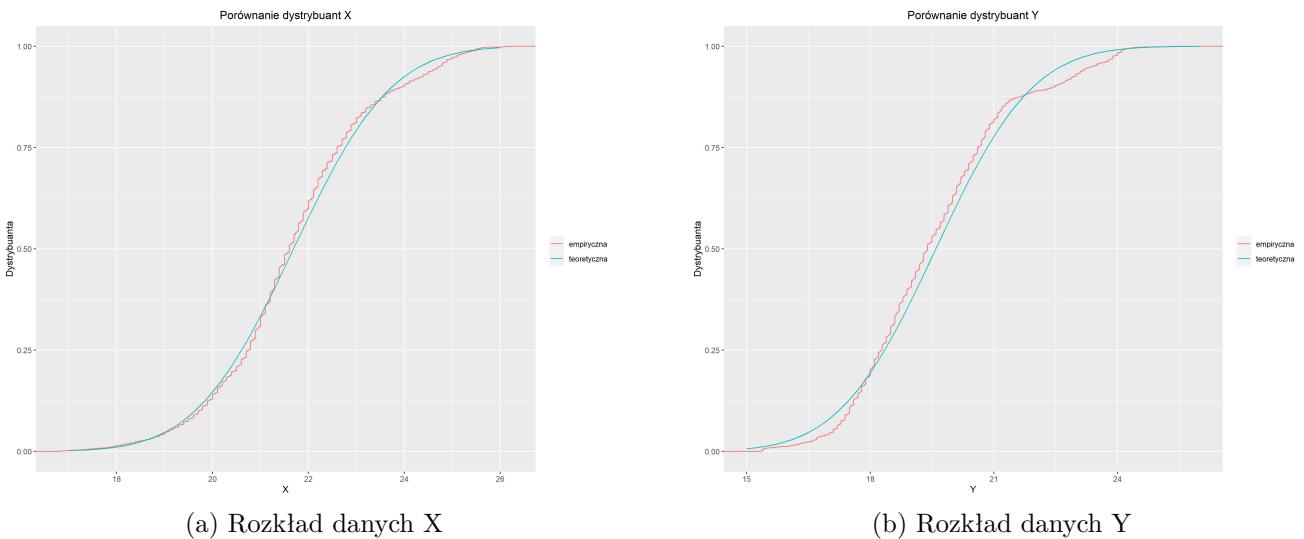
Przedstawione histogramy danych X i Y na rysunku 2 wraz z empirycznymi gęstościami, nie sugerują wprost żadnego konkretnego rozkładu. Na podstawie histogramu dla danych Y , można podejrzewać, iż jest to rozkład dwumodalny² (widoczne dwa wybrzuszenia). Sprawdzimy jednak kompatybilność rozkładów X i Y z rozkładem normalnym.

² W późniejszej części zostanie wykonany test sprawdzający te własność



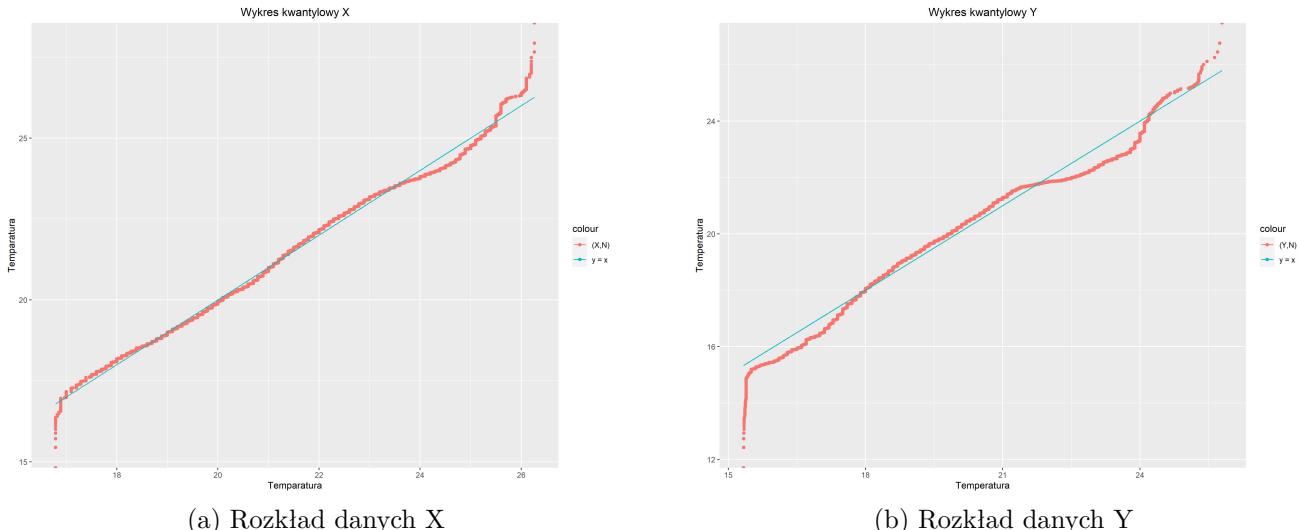
Rys. 3: Porównanie empirycznej gęstości z teoretyczną

Gęstości w zestawieniu z gęściami rozkładów odpowiednio $\mathcal{N}(21.687, 2.579)$ i $\mathcal{N}(19.592, 3.403)$, ukazane na rysunku 3, zdecydowanie się nie pokrywają. Predykowana dwumodalność jest jeszcze bardziej widoczna bez obecności histogramów.



Rys. 4: Porównanie empirycznej dystrybuanty z teoretyczną

W przypadku porównywania dystrybuant na rysunku 5, rozbieżności opisywanych rozkładów z teoretycznymi rozkładami normalnymi nie są na tyle duże, aby jedynie na ich podstawie wykluczyć ich zgodność. Wątpliwości mogłyby wystąpić przy danych z rozkładu Y. Mając jednak informacje na temat wyglądu zarówno dystrybuant jak i gęstości, jesteśmy w stanie wykluczyć normalność omawianych rozkładów.



Rys. 5: Wykresy kwantylowe

Także analiza wykresów kwantylowych wskazuje na to, że rozważane dane nie pochodzą z rozkładów normalnych o danych parametrach średniej i wariancji. Wygenerowane punkty nie znajdują się na prostej, lecz w mniejszym lub większym stopniu odchylają się od niej.

Wcześniej wspomniano jednak o przypuszczeniu, że owe zbiory danych przypominają wyglądem rozkłady multimodalne. Aby potwierdzić tę hipotezę wykorzystamy dostępny w języku R tak zwany *dip test Hartigana*³. Jest to test, którego hipotezę zerową stanowi unimodalność, a alternatywną multimodalność rozkładu. Jeśli w ramach przeprowadzonych obliczeń otrzymamy p -wartość < 0.05 , to będziemy mieli podstawy do odrzucenia hipotezy zerowej na rzecz hipotezy alternatywnej, to jest będziemy mogli stwierdzić, że rozkład jest przynajmniej dwumodalny. W ramach wykonywanego testu otrzymano następujące wyniki przedstawiony na rysunku 6.

```
Hartigans' dip test for unimodality / multimodality
data: X
D = 0.013023, p-value < 2.2e-16
alternative hypothesis: non-unimodal, i.e., at least bimodal
```

```
Hartigans' dip test for unimodality / multimodality
data: Y
D = 0.011822, p-value < 2.2e-16
alternative hypothesis: non-unimodal, i.e., at least bimodal
```

Rys. 6: Dip test X oraz Y

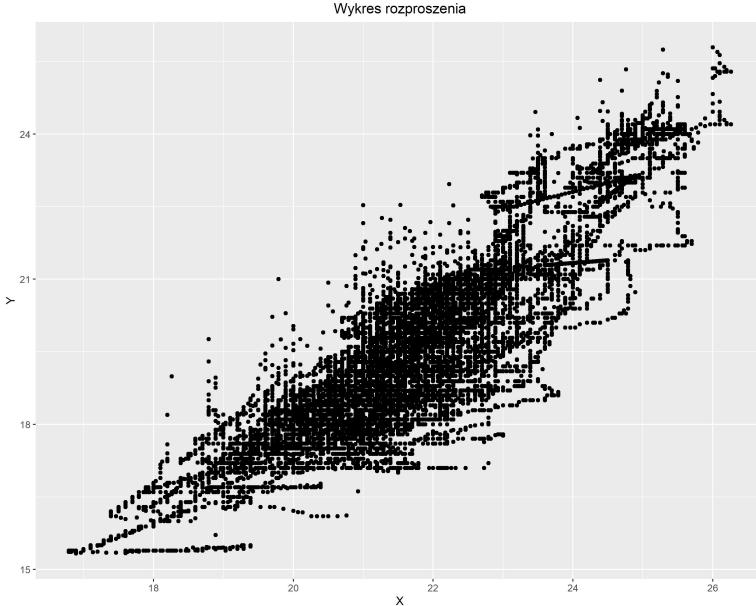
P -wartości wykonanych testów w obu przypadkach wyniosły mniej niż 0.05, co wiąże się z odrzuceniem hipotezy zerowej, czyli potwierdzeniem wysuniętych wcześniej przypuszczeń o multimodalności rozważanych rozkładów.

³ <https://cran.r-project.org/web/packages/diptest/diptest.pdf>

5. Dobranie regresji liniowej dla danych

5.1. Wykres rozproszenia oraz prosta regresji

Kolejnym krokiem jest znalezienie prostej regresji opisującej powiązanie danych ze zbiorów X oraz Y . W tym celu przedstawiony zostanie wykres rozproszenia obu tych zmiennych, by móc stwierdzić, że istnieje szansa na opisanie tej relacji za pomocą regresji liniowej.



Rys. 7: Wykres rozproszenia X i Y

Wykres na rysunku 7 wskazuje na pewną dodatnią zależność zmiennej objaśnianej oraz objaśniającej. Rozproszenie danych jest dość duże, jednak współczynnik korelacji próbkoowej, o wzorze

$$r = \frac{1}{n-1} \cdot \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}. \quad (5)$$

wyniósł aż 0.885. Można zatem użyć znanych nam metod i wyznaczyć szukaną prostą regresji dla danych X i Y . Zakładamy, że będzie ona postaci

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X. \quad (6)$$

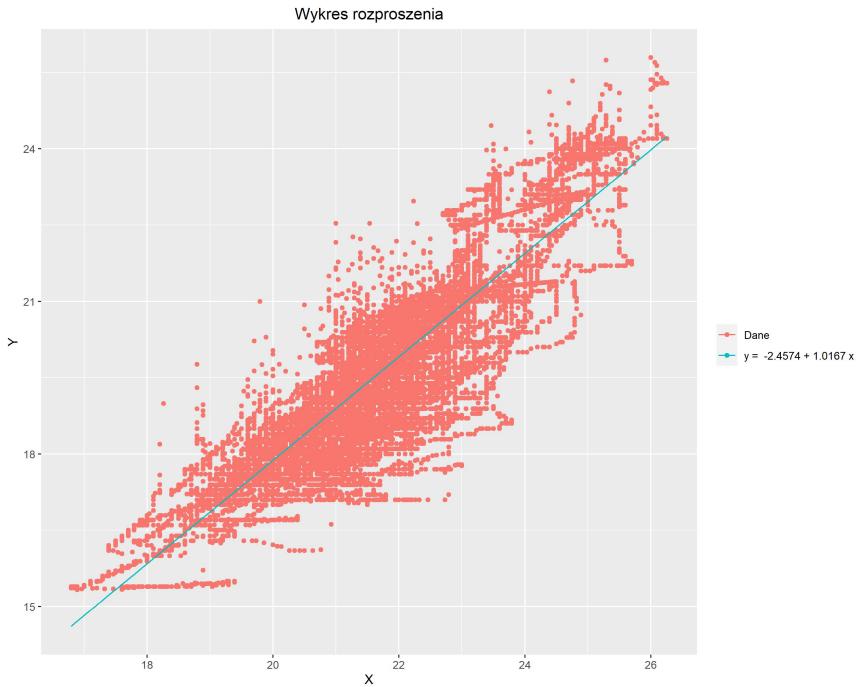
Korzystając z metody najmniejszych kwadratów, otrzymujemy wzory na estymatory współczynników β_0 oraz β_1 . Są one postaci

$$\begin{cases} \hat{\beta}_1 = r \frac{s_y}{s_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}, \quad (7)$$

gdzie r to współczynnik korelacji próbkoowej zadany wzorem 5. Otrzymano następujące wyniki.

$$\begin{cases} \hat{\beta}_1 = 1.0167 \\ \hat{\beta}_0 = -2.4574 \end{cases} \quad (8)$$

Stąd prosta regresji jest dana wzorem $y = -2.4574 + 1.0167x$.



Rys. 8: Wykres rozproszenia X i Y wraz z prostą regresji

Z powyższego rysunku 8 wynika, że znalezione wartości dość dobrze opisują szukaną zależność. Rozproszenie danych jest dość nieregularne, stąd lokalne odchylenia danych od prostej.

5.2. Wyznaczenie R^2

Aby ocenić jakość dopasowania modelu do danych, skorzystamy z takiej miary jak współczynnik determinacji R^2 . Określa ona w jaki sposób zmienna objaśniana opisana jest przez liniową funkcję zmiennej objaśniającej oraz przyjmuje wartości z przedziału $[0,1]$. Granice tego przedziału oznaczają odpowiednio brak liniowej zależności oraz idealne pokrycie się danych z prostą dopasowania. Można go opisać następującym wzorem.

$$R^2 = \frac{SSR}{SST} \quad (9)$$

Korzystając z faktu, iż $SST = SSR + SSE$, powyższy wzór możemy zapisać w postaci

$$R^2 = 1 - \frac{SSE}{SST} \quad (10)$$

. Użyte statystyki zostały opisane w tabeli 2.

Statystyka	Opis	Wzór	Wartość
SST	<i>sum of squares total</i>	$\sum_{i=1}^n (y_i - \bar{y})^2$	67147.599
SSR	<i>sum of squares regression</i>	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	52621.024
SSE	<i>sum of squares error</i>	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	14526.576

Tab. 2: Wybrane statystyki modelu regresji

Znając wzory na poszczególne statystyki, R^2 przedstawia się następująco.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (11)$$

gdzie

- y_i - i -ta wartość zmiennej objaśnianej,
- \bar{y} - średnia z wartości zmiennej objaśnianej y_i ,
- \hat{y}_i - wartość prostej regresji w punkcie x_i .

Na podstawie powyższych wzorów otrzymano wartość statystyki R^2 .

$$R^2 \approx 0.78366.$$

Jest to wartość bliższa 1, co oznacza całkiem poprawne dopasowanie modelu regresji do opisywanych danych, co zostało zauważone na wykresie rozproszenia z rysunku 8.

6. Przedziały ufności

Kolejnym krokiem analizy będzie skonstruowanie przedziałów ufności dla wcześniej znalezionych estymatorów współczynników β_0 oraz β_1 . W tym celu należy skonstruować taki przedział (A_1, A_2) , aby $P(A_1 < \beta < A_2) = 1 - \alpha$, gdzie α jest poziomem istotności. Skorzystamy z twierdzenia, które informuje nas o ich rozkładach. Mianowicie, jeśli $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ ⁴, to

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right), \quad (12)$$

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right). \quad (13)$$

Nie znamy jednak wartości wariancji, dlatego posłużymy się jej nieobciążonym estymatorem

$$S^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}. \quad (14)$$

Dla nieznanego parametru σ^2 statystyki

$$T_1 = \frac{\hat{\beta}_1 - \beta_1}{S} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sim t_{n-2}, \quad (15)$$

$$T_0 = \frac{\hat{\beta}_0 - \beta_0}{S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}, \quad (16)$$

pochodzą z rozkładu t -studenta z $n - 2$ stopniami swobody. Skonstruowanie więc przedziałów w obu przypadkach będzie polegało na wstawieniu owych parametrów do wzoru

$$P\left(-t_{n-2, 1-\frac{\alpha}{2}} < T < t_{n-2, 1-\frac{\alpha}{2}}\right) = 1 - \alpha, \quad (17)$$

gdzie $t_{n-2, 1-\frac{\alpha}{2}}$ jest kwantylem ze wspomnianego rozkładu.

⁴ Zmienna losowa opisana w sekcji nr 7 odpowiedzialna za szum w modelu regresji

6.1. Dla parametru β_1

Podstawmy parametr T_1 do wzoru 17.

$$P \left(-t_{n-2,1-\frac{\alpha}{2}} < \frac{\hat{\beta}_1 - \beta_1}{S} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} < t_{n-2,1-\frac{\alpha}{2}} \right) = 1 - \alpha \quad (18)$$

Zależy nam na znalezieniu przedziału dla β_1 , dlatego też należy przekształcić powyższe równanie.

$$P \left(\hat{\beta}_1 - S \frac{t_{n-2,1-\frac{\alpha}{2}}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} < \beta_1 < \hat{\beta}_1 + S \frac{t_{n-2,1-\frac{\alpha}{2}}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \right) = 1 - \alpha \quad (19)$$

6.2. Dla parametru β_0

Teraz podstawmy parametr T_0 do wzoru 17.

$$P \left(-t_{n-2,1-\frac{\alpha}{2}} < \frac{\hat{\beta}_0 - \beta_0}{S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} < t_{n-2,1-\frac{\alpha}{2}} \right) = 1 - \alpha \quad (20)$$

Dokonując analogicznych przekształceń jak w przypadku β_1 otrzymujemy

$$P \left(\hat{\beta}_0 - t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} < \beta_0 < \hat{\beta}_0 + t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right) = 1 - \alpha \quad (21)$$

6.3. Wyniki

Niech $\alpha = 0.05$.

Po podstawieniu wszystkich wartości liczbowych do skonstruowanych powyżej przedziałów, otrzymano następujące wyniki.

	A_1	A_2	$\hat{\beta}$
β_1	1.0093	1.0241	1.0167
β_0	-2.6195	-2.2953	-2.4574

Tab. 3: Skrajne wartości przedziałów ufności estymowanych parametrów

Porównując wartości z tabeli 3, widać, iż parametry β_1 oraz β_0 mieszczą się w wyznaczonych przedziałach ufności.

7. Analiza residuów

W przeprowadzanej analizie danych głównym założeniem jest, że zmienną objaśnianą Y można wyrazić za pomocą pewnego teoretycznego modelu danego wzorem $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, gdzie

- Y_i - i -ta wartość zmiennej objaśnianej,
- X_i - i -ta wartość zmiennej objaśniającej,
- ϵ_i - zmienna losowa odpowiadająca za szum.

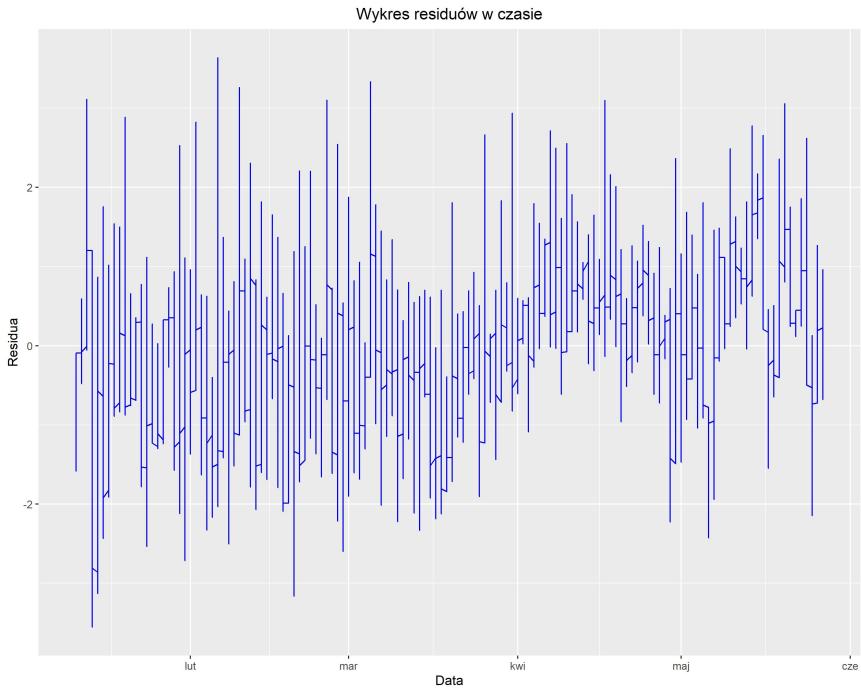
Residua $e_i = Y_i - \hat{Y}_i$ są w tym przypadku realizacjami zmiennej losowej ϵ_i . Korzystając z metody najmniejszych kwadratów podczas wyznaczania wzorów na współczynniki β_0 oraz β_1 posługiwieliśmy się kilkoma założeniami wobec rozkładu ϵ_i , które należy teraz sprawdzić, aby przetestować poprawność założonego przez nas modelu. Przedstawione zostały one poniżej.

1. $\forall_{i=1,2,\dots,n} \mathbb{E}\epsilon_i = 0$ - średnia każdej z ϵ_i jest stała i wynosi 0.
2. $\forall_{i=1,2,\dots,n} \text{Var}\epsilon_i = \sigma^2$ - wariancja każdej z ϵ_i jest stała i wynosi σ^2 .
3. $\{\epsilon_i\}_{i=1}^n$ jest ciągiem niezależnych zmiennych losowych

Ponadto skorzystano z dodatkowego założenia, definiując jednoznacznie rozkład z jakiego pochodzą zmienne losowe ϵ_i .

$$\text{— } \forall_{i=1,2,\dots,n} \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

W następnych sekcjach zostanie przedstawiona analiza związana z poszukiwaniem odpowiedzi na pytanie, czy powyższe założenia zostały spełnione. Wpierw jednak, przedstawiony zostanie wykres znalezionych residiów względem czasu.



Rys. 9: Residuum względem czasu

Już na tej podstawie można wysunąć wnioski co do rozkładu ϵ_i . Prezentowane wartości e_i na początku wydają się być oczekiwany przez nas szumem, jednak mniej więcej w połowie, zaczynają być widoczne pewne okresowe wzrosty i spadki. Skutkiem tego może być niespełnienie założenia o stałej średniej wynoszącej zero. Oszacowanie wielkości wariancji jest już trudniejsze, ale na pierwszy rzut oka nie widać rażących nieprawidłowości. O spełnieniu warunków na rozkład i nieskorelowaniu ϵ_i nie da się zadecydować na tym etapie analizy.

7.1. Analiza średniej

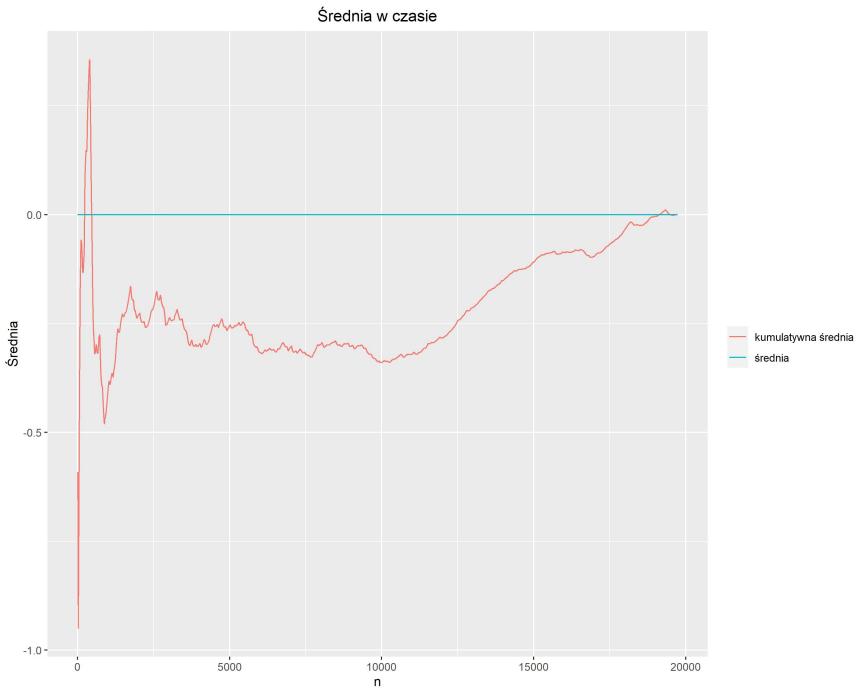
Wpierw przyjrzymy się średniej residiów. Korzystając ze wzoru 1, obliczono, że wynosi ona

$$\bar{e} = 3.5 \cdot 10^{-16}.$$

Jest to wynik bardzo bliski zeru, co może sugerować spełnienie wymaganego założenia. Sprawdzmy jednak jak wygląda prosta średnia ruchoma dla rozważanych danych. Wyznaczona ona zostanie jako średnia arytmetyczna z kolejnych kumulowanych wartości.

$$\bar{e}_k = \frac{1}{k} \sum_{i=1}^k e_i, \quad (22)$$

dla $k \leq n$, gdzie n ilość obserwacji. Generując wartości średniej dla wszystkich możliwych k , otrzymano następujący wykres ich wartości.



Rys. 10: Prosta średnia ruchoma residuum.

Analizując wykres na rysunku 11 można zauważyc pewne nieprawidłowości. Przez dużą część czasu średnia oscyluje w okolicy wartości -0.2 , aby od mniej więcej połowy zacząć rosnąć i osiągnąć ostateczną wartość bliską zeru. Takie zachowanie przeczy założeniu o stałej wartości oczekiwanej zmiennej losowej ϵ_i , przez co należy uznać ten warunek za niespełniony.

Otrzymane wyniki można jednak uznać za realne. Analizując ponownie wykres rozproszenia z rysunku 8 można zauważyc, że w mniej więcej połowie rozważanych wartości, występuje duże zagęszczenie punktów znajdujących się poniżej wyznaczonej prostej regresji. Jednak wraz z rosnącymi wartościami X i Y , obserwuje się więcej punktów danych znajdujących się ponad prostą. Tłumaczy to nieregularne zachowanie się rozważanej ruchomej średniej.

7.2. Analiza wariancji

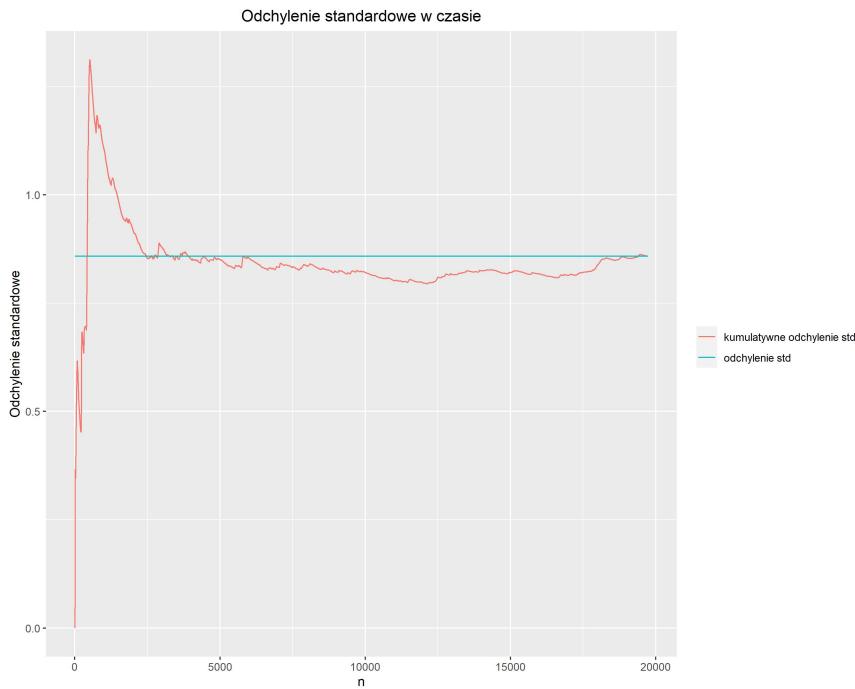
Analiza wariancji przeprowadzona zostanie w podobny sposób do badania średniej z poprzedniej sekcji. Wpierw wyliczona zostanie jej wariancja próbкова ze wzoru 2. Wyniosła ona

$$\sigma^2 = 0.736.$$

Ponownie zbadana zostanie prosta wariancja ruchoma, polegająca na wyliczaniu jej wartości na podstawie kolejnych ilości obserwacji k .

$$\sigma_k^2 = \frac{1}{k-1} \sum_{i=1}^k (e_i - \bar{e})^2 \quad (23)$$

Wszelkie oznaczenia pozostają takie same, jak w poprzednio podanych wzorach. Na tej podstawie wygenerowano wykres w zależności od $k \leq n$, ale wartości odchylenia standardowego, wynoszącego w ogólnosci $\sigma = 0.858$. Zatem na wykresie pojawią się pierwiastki wartości ze wzoru 23. Otrzymane wyniki przedstawiono poniżej.

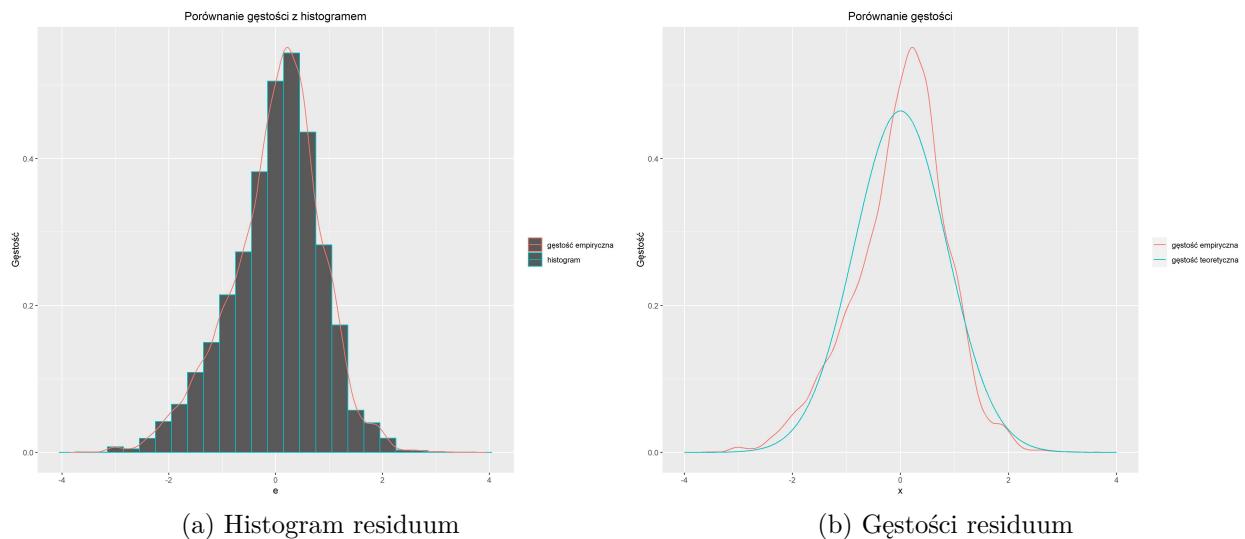


Rys. 11: Proste ruchome odchylenie standardowe residuum.

W przeciwieństwie do analizowanego wcześniej przypadku średniej, odchylenie standardowe (a co za tym idzie także i wariancja) wydaje się być stałe. Dla większości k oscyluje wokół swojej ogólnej wartości σ , ostatecznie do niej zbiegając, gdy $k = n$. Można zatem stwierdzić, że wariancja rzeczywiście jest stała w czasie i wynosi $\sigma^2 = 0.736$.

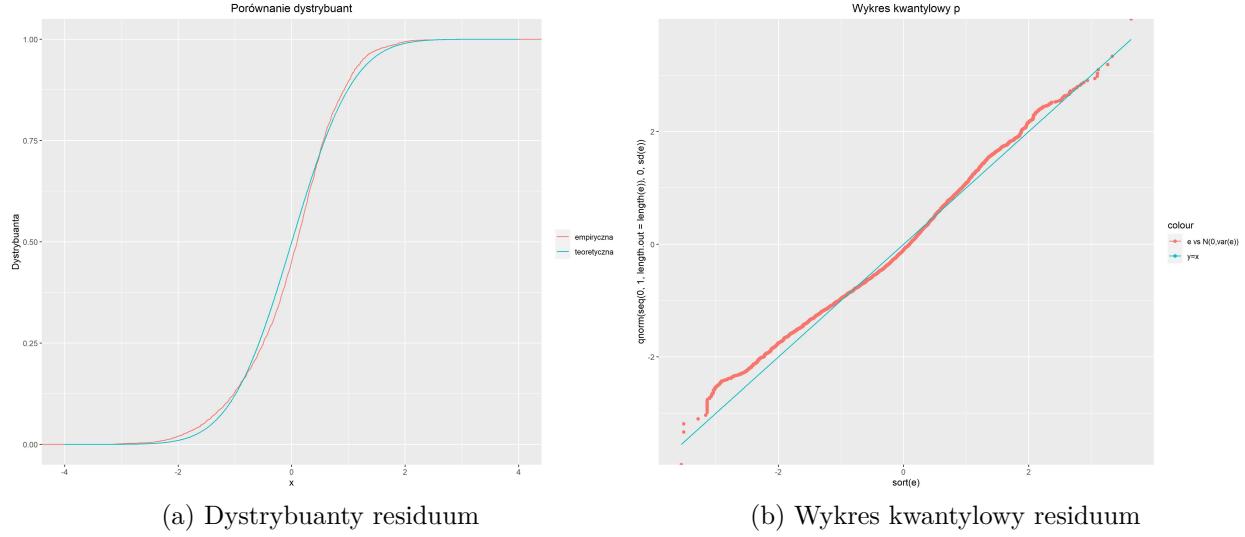
7.3. Analiza rozkładu

Teraz sprawdzone zostanie dodatkowe założenie o rozkładzie residuów, to jest czy $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. W tym celu przedstawione zostanie histogram wraz z wyestymowaną gęstością empiryczną oraz podobnie jak w graficznej analizie X i Y - dystrybuanty, gęstości oraz wykres kwartylowy porównane z teoretycznym rozkładem normalnym o danych parametrach. Zakładamy, że średnia wynosi 0 oraz wariancja równa jest $\sigma^2 = 0.736$.



Rys. 12: Histogram oraz gęstości residuów

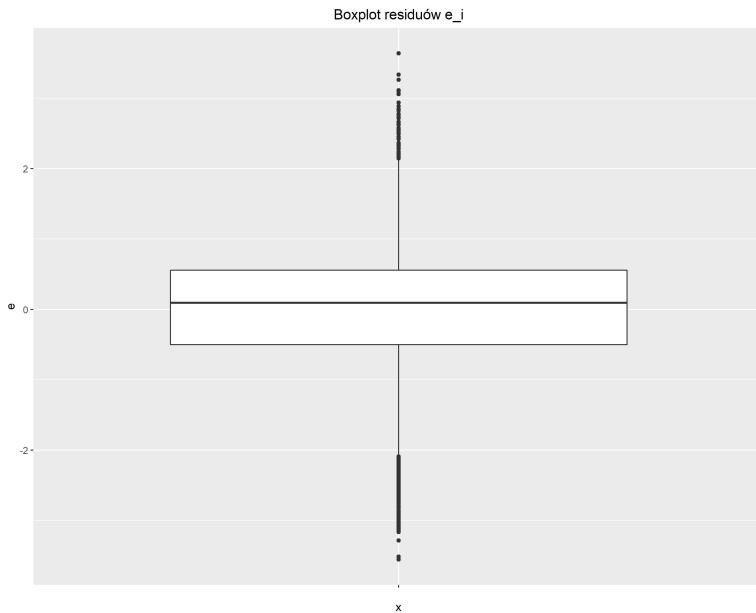
Na pierwszy rzut oka, histogram e_i jest zbliżony kształtem do wyglądu funkcji gęstości rozkładu normalnego. Jest on jednak delikatnie niesymetryczny. Mediana residiów wynosi 0.095, a więc jest większa od średniej. Świadczy to o lewoskośności rozkładu, która jest widoczna także na rysunku 12. Własność ta nie jest zgodna z charakterystyką rozkładu normalnego, który jest symetryczny. Stąd też rozbieżności widoczne na porównaniu gęstości empirycznej residiów z gęstością teoretyczną rozkładu normalnego. Wykres empiryczny osiąga wyższe wartości i jest bardziej wysmukły.



Rys. 13: Wykres kwantylowy oraz dystrybuanty residiów

Dystrybuanty na rysunku 13 także są do siebie dość zbliżone, jednak nie na tyle, by móc stwierdzić idealne dopasowanie. Występują lokalne odchylenia ich wartości, które zaburzają ich zgodność. Wykres kwantylowy nie przypomina prostej linii, szczególnie podczas porównywania ze sobą skrajnych kwantyli obu rozkładów.

Na koniec przedstawiony zostanie wykres pudełkowy e_i .



Rys. 14: Wykres pudełkowy residiów.

Wynika z niego, że mediana znajduje się delikatnie ponad zerem (co zostało sprawdzone wcześniej). Ponadto w zbiorze danych znajduje się wiele wartości odstających. Mogą one znacznie wpływać

na badane residua i poprzednio sprawdzanie założenia. Skutki wystąpienia tylu wartości odstających można było zaobserwować przykładowo na wykresie kwantylowym z rysunku 13, który dla wartości skrajnych najbardziej oddalał się od prostej.

Wszystkie powyższe wykresy wskazują na to, że rozważane założenie o normalności rozkładu residiów nie jest spełnione. W ramach upewnienia się, przeprowadźmy test Kołmogorowa-Smirnowa. Jego hipotezą zerową jest normalność badanych danych, a alternatywną to, że pochodzą one z pewnego innego rozkładu. Jeśli otrzymana p -wartość będzie większa od 0.05, nie będzie podstaw do odrzucenia hipotezy zerowej. Test przeprowadzono na ustandaryzowanych danych, to jest $A_i = \frac{e_i - \bar{e}}{\sigma}$. Otrzymano następujący wynik.

```
Asymptotic one-sample Kolmogorov-Smirnov test
```

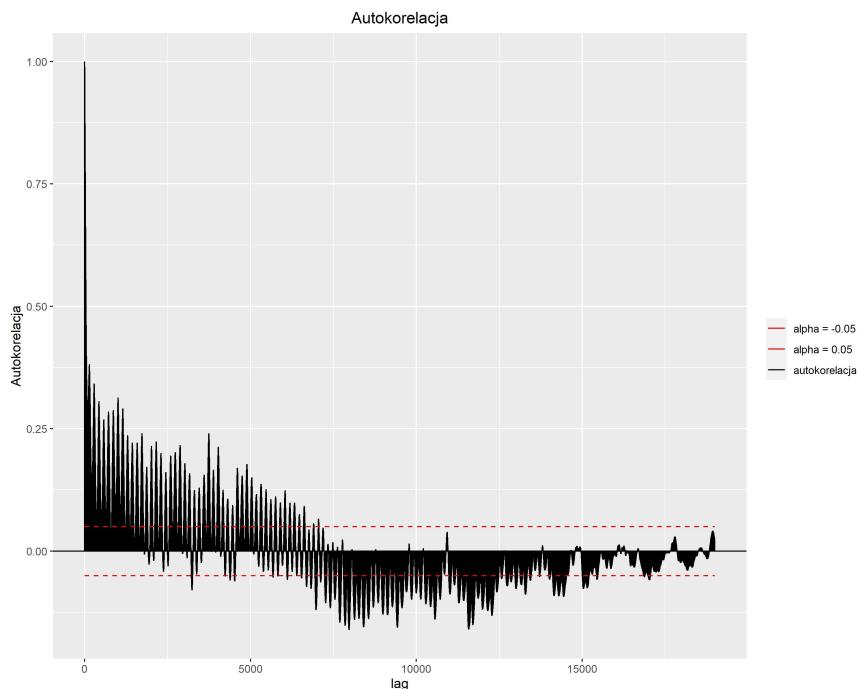
```
data: e_n
D = 0.053442, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Rys. 15: Wynik testu Kołmogorowa-Smirnowa.

Według opisu przeprowadzonego testu, p -wartość jest bliska零, co wiąże się z odrzuceniem hipotezy zerowej na rzecz hipotezy alternatywnej. Oznacza to, że podane dane nie pochodzą z rozkładu normalnego, co tylko potwierdziło wcześniej wysunięte wnioski.

7.4. Analiza korelacji

Ostatnim krokiem w analizie residiów jest sprawdzenie założenia o ich niezależności. W tym celu można posłużyć się funkcją autokorelacji. Mierzy ona zależność pomiędzy wartościami aktualnymi i tymi przesuniętymi w czasie o tak zwany *lag*. Na wykładzie podany był jej wzór. Jednakże, w języku R dostępna jest funkcja *acf*,⁵, poświęcona temu właśnie zagadnieniu, dzięki której otrzymano poniższy wykres autokorelacji.



Rys. 16: Autokorelacja residiów.

⁵ <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/acf>

Na wykresie 16 przerywaną linią zaznaczono poziom ± 0.05 . Są to wartości referencyjne, które mają za zadanie pokazać siłę korelacji. Gdyby dane były nieskorelowane, ich wartości oscylowałyby wokół zera, osiągając minimalne wartości (przykładowo poniżej 0.05). Jeśli zależność byłaby silna, autokorelacja byłaby bliska ± 1 . W rozważanym przypadku wartości są przedziałami bliskie około ± 0.2 , bądź nawet zbliżają się do zera. Jednakże pozostały one w wyznaczonym referencyjnym przedziale dla niewielkiej ilości *lag*. Ponadto zauważać można pewną sezonowość w otrzymanych wynikach. Wysokie dodatnie wartości zmieniają się w ujemne, by później z powrotem wrócić w pobliże zera. Można zatem wysunąć wniosek, że otrzymane residua są ze sobą w pewien sezonowy sposób skorelowane.

Także tym razem język R oferuje pomoc w zweryfikowaniu postawionej tezy. Jedna z dostępnych funkcji z modułu *car* przeprowadza tak zwany test Durbina - Watsona⁶. Bada on obecność autokorelacji dla $lag = 1$. W omawianym modelu regresji jest w stanie sprawdzić czy autokorelacja residuów występuje i jak silna jest. Hipotezą zerową jest w tym przypadku brak autokorelacji residuów, a alternatywną ich obecność. Gdy *p*-wartość jest mniejsza od 0.05 odrzucamy hipotezę zerową na rzecz alternatywnej. Statystyka D-W omawianego testu może osiągać wartości pomiędzy 0 a 4, gdzie wartości bliskie zeru wskazują na silną dodatnią zależność, te bliskie 4 na silną ujemną zależność, a wartość bliska 2 oznacza brak zależności. Przetestujmy zatem rozważane residua.

```
lag Autocorrelation D-W Statistic p-value
 1      0.9885986   0.02277309     0
Alternative hypothesis: rho != 0
```

Rys. 17: Wynik testu Durbina- Watsona.

Otrzymano wartość statystyki D-W około 0.0227 oraz *p*-wartość równą 0. Świadczy to o bardzo silnej dodatniej zależności wśród residuów. Potwierdza tym samym przypuszczenia wysunięte podczas analizy wykresu autokorelacji. Podsumowując, także założenie o niezależności residuów nie zostało spełnione.

8. Wnioski

Główym celem pracy było wyprowadzenie oraz analiza modelu regresji liniowej na podstawie znalezionych danych. Najpierw przeprowadzona została dokładana analiza rozważanych zmiennych *X* i *Y*, opisujących temperatury w dwóch różnych pomieszczeniach w pewnym mieszkaniu na przestrzeni czasu. Zauważone zostały pewne zależności, typu wzrost średnich wartości wraz z upływem czasu oraz podobieństwo trajektorii obu tych zbiorów. Następnie stworzony został model regresji liniowej, opisujący liniową zależność między zmiennymi, którego jakość dopasowania okazała się być bliska 0.8, co jest całkiem dobrym wynikiem. Skonstruowane zostały także przedziały ufności.

Wałą częścią pracy była analiza residuów modelu. Sprawdzone zostały wszystkie założenia dotyczące szumu modelu, włączając z tym dodatkowym, mówiącym o rozkładzie. Niestety, analiza wykazała, że większość z wymienionych warunków nie była spełniona. W szczególności te, dotyczące niezależności residuów oraz stałej wartości oczekiwanej. Obie te nieprawidłowości mogą wynikać z nie do końca poprawnego doboru danych do analizy. Oba te zbiory dotyczą temperatury w pewnym pomieszczeniu zależnej od czasu, a więc i oba są poddatne na efekty związane ze zmianą, przykładowo, pory roku. Odczyty temperatur zaczynają się w styczniu a kończą końcem maja. Jest to okres, w którym występują duże wahania pogodowe, a co za tym idzie, także odpowiednie reakcje domowników na daną aurę. Taka zależność mogła wpływać na funkcję autokorelacji oraz założenie o niezależności residuów. Ponadto dobór pomieszczeń jest tutaj ważny. Kuchnia i łazienka są pomieszczeniami, na

⁶ <https://rdrr.io/cran/car/man/durbinWatsonTest.html>

których temperaturę wpływają różne urządzenia oraz czynności w nich wykonywane. Podczas gotowania kuchnia się nagrzewa, lecz niekoniecznie musi wpływać to na temperaturę w łazience. O innej porze za to domownicy biorą prysznic, suszą włosy, ale nie podnosi to automatycznie temperatury w innych pokojach. Stąd też pewne wartości odstające widoczne przykładowo na wykresie rozproszenia obu zbiorów danych.

W związku z niespełnieniem podanych przez model założeń, analiza przedziałów ufności nie została przeprowadzona w pełni poprawnie. Ważnym założeniem było tam między innymi to o rozkładzie residiów, a dokładniej, że $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Analiza wykazała jednak, że residua nie pochodzą z rozkładu normalnego. W takim wypadku przedziały ufności powinny zostać skonstruowane w inny sposób.

Mimo wszystko, znaleziona przez nas prosta regresji dość dobrze opisuje omawiane przez nas dane. Istnieje między nimi widoczna liniowa zależność, a ów prosta pozwala w mniej lub bardziej sprawny sposób opisać ich relację. Można także spróbować predykcji przyszłych wartości, jednak wymagałoby to głębszej analizy wpływu pory roku na otrzymane dane i odszukania pewnych sezonowych zależności pomiędzy nimi.