

Politechnika Wrocławska
Wydział Matematyki

Skład grupy:	Agata Sobczak 268873 Katarzyna Kudelko 268762
Prowadzący laboratorium:	mgr inż. Wojciech Żuławiński
Prowadząca wykład:	dr hab. inż. Agnieszka Wyłomańska

Komputerowa Analiza Szeregów Czasowych

Raport 1.

Analiza danych z wykorzystaniem regresji liniowej

Spis treści

1	Wstęp	3
1.1	Wizualizacja danych	3
2	Podstawowe statystyki	4
2.1	Obliczanie statystyk	4
2.2	Wykresy pudełkowe X i Y	5
2.2.1	Wnioski	6
2.3	Porównanie rozkładów	7
2.3.1	Wnioski	7
3	Regresja liniowa	8
3.1	Dopasowanie regresji liniowej	8
3.2	Dodatkowe statystyki	9
4	Przedziały ufności β_0, β_1	10
4.1	Wnioski	11
5	Analiza residuów	12
5.1	Sprawdzenie $E(\epsilon_i) = 0$	12
5.2	Sprawdzenie czy $\text{Var}(\epsilon_i) = \sigma^2 < \infty$	13
5.3	Sprawdzenie $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$	13
5.4	Sprawdzenie czy ϵ_i są niezależne (nieskorelowane)	15
5.5	Wartości odstające	16
5.6	Wnioski	16
6	Predykcje	16
7	Wnioski	18

1 Wstęp

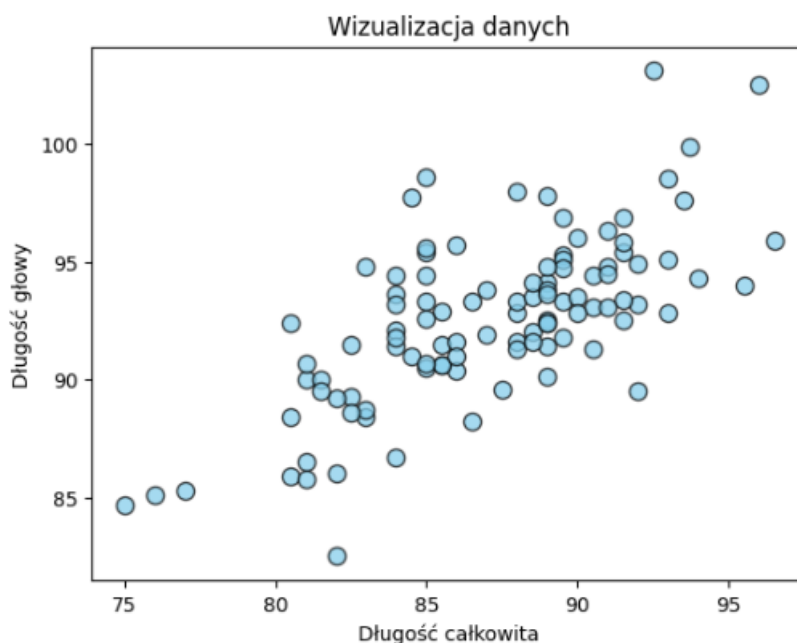
Raport przedstawia analizę danych dotyczących rozmiarów oposów, opartą na 104 obserwacjach dostępnych na platformie Kaggle: <https://www.kaggle.com/datasets/abrambeyer/openintro-possum/data>

Zbiór danych obejmuje 14 różnych zmiennych, które opisują różne aspekty oposów, takie jak ich długość oraz inne charakterystyki.

W ramach analizy skoncentrowano się na dwóch kluczowych zmiennych: "Długość całkowita" oraz "Długość głowy". Długość głowy została zmierzona w milimetrach, natomiast długość całkowita w centymetrach. Te dwie cechy zostały wybrane ze względu na ich potencjalną istotność i wpływ na zrozumienie ogólnego wymiaru i budowy oposów.

Przedstawienie wyników analizy tych zmiennych pozwoli lepiej zrozumieć charakterystyki oposów oraz zidentyfikować ewentualne wzorce czy zależności między nimi. Dzięki tym ustaleniom możliwe będzie lepsze poznanie populacji oposów oraz wnioskowanie na temat ich biologii czy środowiska życia.

1.1 Wizualizacja danych



Rysunek 1: Dane z dopasowaną prostą regresji

Do analizy wykorzystano "Długość całkowitą" jako zmienną X oraz "Długość głowy" jako Y.

2 Podstawowe statystyki

2.1 Obliczanie statystyk

	X - "Długość całkowita"	Y - "Długość głowy"
Mediana	88.00	92.80
Kwartył rzędu $\frac{1}{4}$	84.00	90.67
Kwartył rzędu $\frac{3}{4}$	90.00	94.72
Rozstęp międzykwartylowy	6.00	4.05
Rozstęp	21.50	20.60
Odchylenie standardowe	4.29	3.56
Wariancja	18.40	12.65
Kurtoza	-0.16	0.79
Współczynnik skośności	-0.28	-0.05
Współczynnik zmienności	4.92%	3.84%
Średnia arytmetyczna	87.09	92.60
Średnia geometryczna	86.98	92.53
Średnia harmoniczna	86.87	92.46
Minimum	75.00	82.50
Maksimum	96.50	103.10

Tabela 1: Podstawowe statystyki dla obydwu zmiennych

Uwaga 1 Wszystkie obliczone statystyki zostały podane z zaokrągleniem do dwóch miejsc po przecinku.

Uwaga 2 Podstawowe statystyki zostały policzone na podstawie następujących wzorów:

- **Mediana** - wartość dzieląca uporządkowany zbiór danych na dwie równe części, gdzie połowa wartości jest mniejsza lub równa medianie, a druga połowa jest większa lub równa medianie.
- **Kwartyle** - miary pozycyjne, podobne do mediany, dzielące uporządkowany zbiór danych na równe części, lecz w przeciwieństwie do mediany, dzielą zbiór na więcej niż dwie części.

$$F(x_p) = p$$

- **Rozstęp międzykwartylowy** - miara zmienności danych, wyznaczana jako różnica między trzecim a pierwszym kwartylem zbioru danych.

$$R_Q = Q_3 - Q_1$$

- **Rozstęp** - prosta miara zmienności danych, wyznaczana jako różnica między największą i najmniejszą wartością w zbiorze danych.

$$R = X_{max} - X_{min}$$

- **Wariancja** - miara zmienności danych, określająca, jak bardzo wartości w zbiorze różnią się od średniej arytmetycznej.

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

- **Odchylenie standardowe** - pierwiastek kwadratowy z wariancji, określa, jak bardzo wartości w zbiorze różnią się od średniej arytmetycznej.

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

- **Współczynnik zmienności** - miara względnej zmienności, czyli stosunek odchylenia standardowego do średniej arytmetycznej w zbiorze danych.

$$\frac{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}}{\frac{1}{n} \sum_{i=1}^n x_i} \cdot 100\%$$

- **Średnia arytmetyczna** - średnia wartość w zbiorze danych, obliczana jako suma wszystkich wartości podzielona przez liczbę tych wartości.

$$\bar{x}_a = \frac{1}{n} \sum_{i=1}^n x_i$$

- **Średnia geometryczna** - średnia wartość, obliczana jako pierwiastek n-tego stopnia z iloczynu wszystkich wartości w zbiorze.

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

- **Średnia harmoniczna** - odwrotność średniej arytmetycznej odwrotności wartości w zbiorze.

$$\bar{x}_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$$

- **Kurtoza** - względna miara koncentracji i spłaszczenia rozkładu, określająca rozmieszczenie i koncentrację wartości w pobliżu średniej.

$$K = \frac{m_4}{s^4}$$

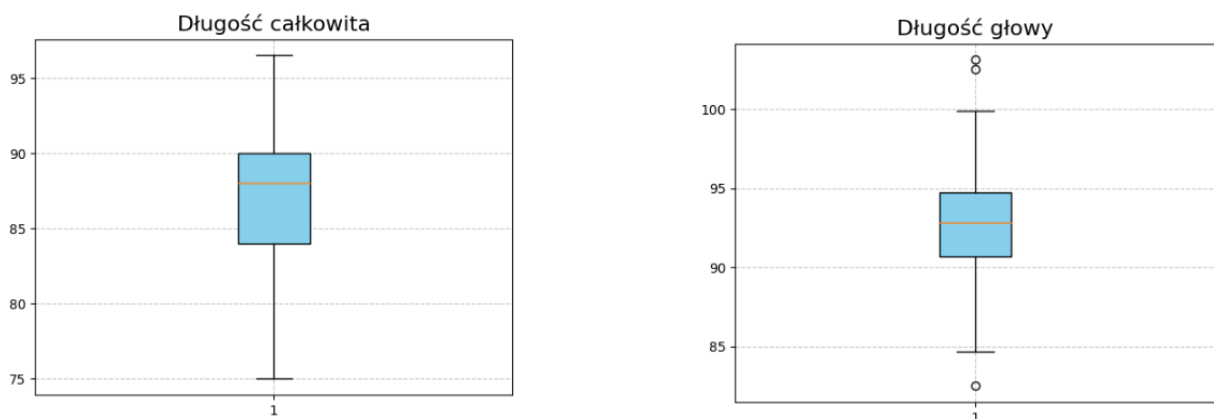
gdzie: m_4 - moment centralny rzędu czwartego, s^2 - odchylenie standardowe podniesione do czwartej potęgi.

- **Współczynnik skośności** - miara asymetrii rozkładu prawdopodobieństwa, wskazująca na to, w którą stronę rozkład jest asymetryczny i jak bardzo.

$$A_Q = \frac{Q_1 + Q_3 - 2m}{Q_3 - Q_1}$$

gdzie: m - mediana

2.2 Wykresy pudełkowe X i Y



Rysunek 2: Wykresy pudełkowe dla X i Y

2.2.1 Wnioski

Analizując *Tabelę 1* oraz *Rysunek 2* można stwierdzić, kurtoza wraz z współczynnikiem skośności dla danych X jest ujemna, co wskazuje na lekką płaskość ogona rozkładu w porównaniu do rozkładu normalnego i nieznaczną asymetrię, ze skierowaniem w lewo. Natomiast dla Y kurtoza jest dodatnia, a współczynnik skośności ujemny, co sugeruje nieco bardziej "spiczasty" rozkład z nieznaczną asymetrią w lewo.

Dla długości całkowitej, 25% oposów ma długość mniejszą niż 84.00 cm, a dla długości głowy, 25% ma długość mniejszą niż 90.67 mm. Dla długości całkowitej, 75% oposów ma długość mniejszą niż 90.00 cm, a dla długości głowy, 75% ma długość mniejszą niż 94.72 mm. Co więcej, najmniejszy opos ma całkowitą długość 75cm, a najdłuższy 96.50cm. Minimalna długość głowy to 82.50 mm, a maksymalna wynosi 103.10 mm.

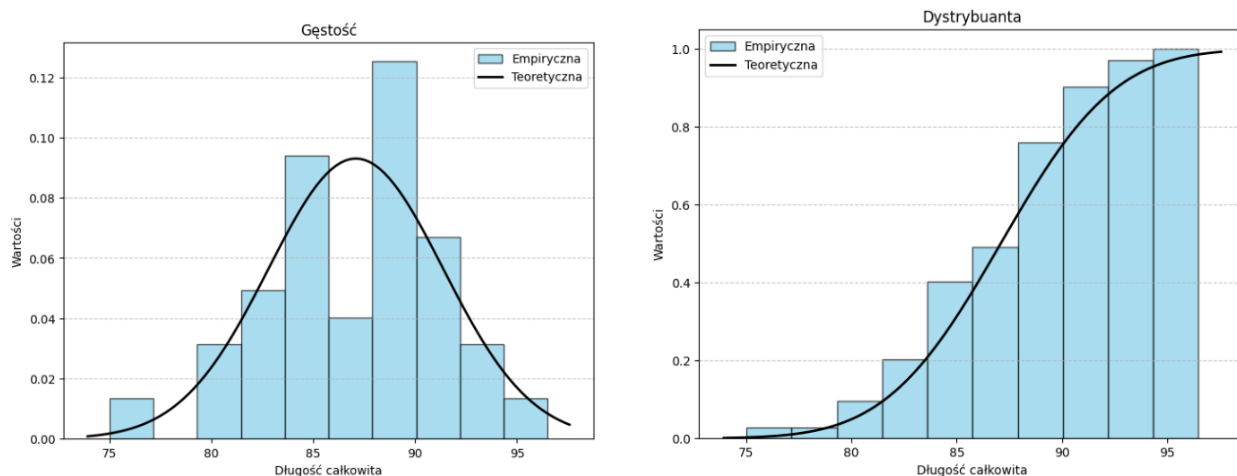
Współczynnik zmienności 4.92%, sugeruje, że dane dotyczące długości całkowitej oposów wskazują umiarkowaną zmienność w stosunku do ich średniej. Te długości są rozproszone wokół średniej w miarę tej zmienności. Dla długości głowy oposów współczynnik zmienności wynosi 3.84% co sugeruje niższą zmienność w porównaniu do długości całkowitej, a co za tym idzie są mniej rozproszone wokół średniej arytmetycznej.

Odchylenie standardowe odpowiednio dla X i Y, wynoszące 4.29, 3.56, mówi o tym że dane są średnio rozproszone wokół średniej arytmetycznej, a wariancje 18.40, 12.65 mówi o tym, że dane dotyczące Y mają mniejszą zmienność danych niż dla X.

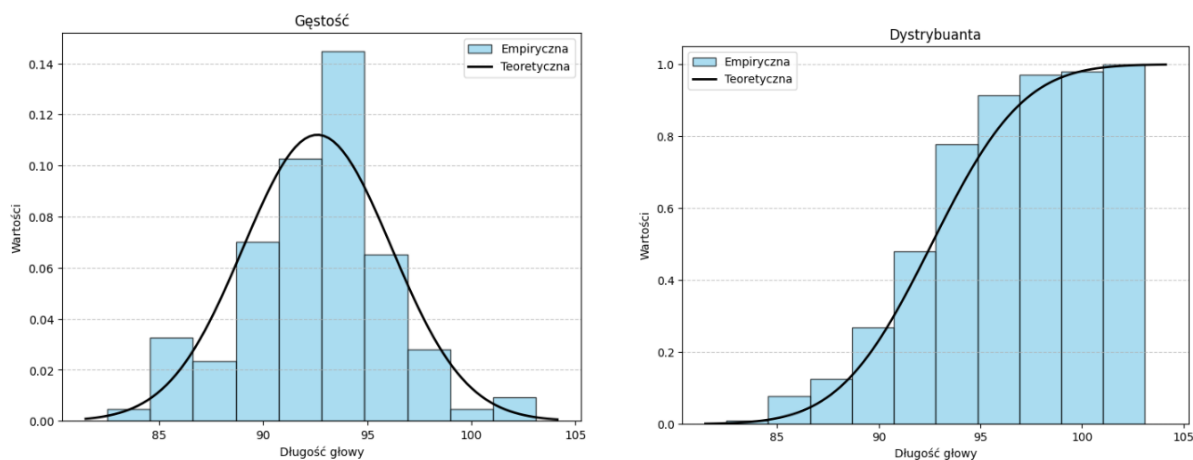
2.3 Porównanie rozkładów

W celu porównania rozkładów, zestawiono gęstości i dystrybuanty zmiennej X i rozkładu normalnego z odpowiednimi parametrami. Jako, iż średnia i odchylenie X wynoszą odpowiednio 87.09 i 4.29, wybrano teoretyczny rozkład jako $\mathcal{N}(87.09, 4.29)$.

Natomiast w przypadku zmiennej Y średnia i odchylenie X wynoszą odpowiednio 92.60 i 3.56, także został wybrany teoretyczny rozkład normalny $\mathcal{N}(92.60, 3.56)$.



Rysunek 3: Porównanie empirycznych i teoretycznych gęstości, i dystrybuant dla X



Rysunek 4: Porównanie empirycznych i teoretycznych gęstości, i dystrybuant dla Y

2.3.1 Wnioski

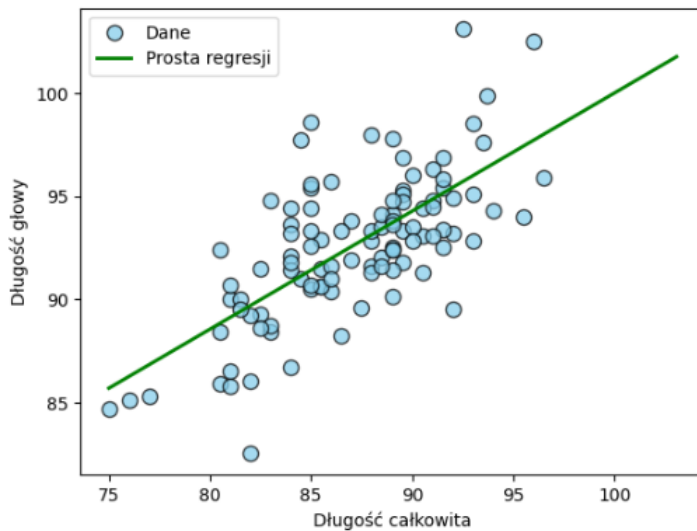
Wyniki analizy wykresach zarówno dla X jak i dla Y rozkładu empirycznego ukazują, że badane dane nie wykazują cech charakterystycznych dla idealnego rozkładu normalnego. Zamiast tego, rozkłady zdają się posiadać pewne odstępstwa od symetrycznej formy, typowej dla rozkładu normalnego.

Dodatkowo, histogramy ukazują asymetrię i nieregularności w kształcie, co jest charakterystyczne dla danych, które nie podlegają idealnemu rozkładowi normalnemu. Odstępstwa od oczekiwanej symetrii i nietypowe kształty rozkładu mogą wskazywać na obecność innych czynników wpływających na zbiór danych.

3 Regresja liniowa

Dalszą częścią analizy danych jest dopasowanie regresji liniowej do przedstawionych danych oraz wyliczenie kolejnych statystyk.

3.1 Dopasowanie regresji liniowej



Rysunek 5: Dane z dopasowaną prostą regresji

Na wykresie można zauważyć liniową zależność między danymi. Priorytetem jest dostosowanie linii w taki sposób, aby dopasowanie do punktów było jak najlepsze. W celu wyliczenia współczynników prostej regresji użyto metody najmniejszych kwadratów.

$\hat{\beta}_0$	$\hat{\beta}_1$
0.57	42.70

Tabela 2: Wartości współczynników prostej regresji

Uwaga 3 Współczynniki prostej regresji zostały podane do dwóch miejsc po przecinku i zostały wyliczone na podstawie następujących wzorów:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

3.2 Dodatkowe statystyki

W celu określenia zależności liniowej między zmiennymi policzone zostały dodatkowe statystyki w postaci korelacji Pearsona i współczynnika determinacji.

R^2	ρ
0.48	0.69

Tabela 3: Współczynnik determinacji i korelacja Pearsona

Współczynnik korelacji Pearsona wynosi 0.69, a więc można uznać, że jest silna, dodatnia zależność między zmiennymi, co pasuje do rozkładu punktów na *Rysunku 5*. Współczynnik R^2 wskazuje jak dobre jest dopasowanie modelu. W tym przypadku wynosi on 0.48 co mówi o tym, że prawie połowa punktów leży na wyznaczonej prostej. Na *Rysunku 5* dostrzec można potencjalne wartości odstające.

Uwaga 4 Statystyki zostały zaokrąglone do dwóch miejsc po przecinku i wyliczone na podstawie następujących wzorów:

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}},$$
$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

4 Przedziały ufności β_0, β_1

Korzystając ze wcześniej wyliczonych współczynników, można wyliczyć przedziały ufności dla β_0, β_1 . Skorzystano z metody dla nieznannej wartości σ . W tym celu wyliczono wariancję z próbki S^2 . Aby uzyskać odchylenie standardowe potrzebne do obliczeń, wzięto pierwiastek z tej liczby.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Uwaga 6 Jako α przyjęto $\alpha = 0.05$, zatem współczynniki znajdują się w przedziałach z prawdopodobieństwem 95%.

Następnie wygenerowano przedziały według następujących wzorów:

Przedziały ufności dla β_0 :

$$\left[\hat{\beta}_0 - t_{n-2, 1-\alpha/2} \cdot S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{\beta}_0 + t_{n-2, 1-\alpha/2} \cdot S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right],$$

Przedziały ufności dla β_1 :

$$\left[\hat{\beta}_1 - t_{n-2, 1-\alpha/2} \cdot S \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{\beta}_1 + t_{n-2, 1-\alpha/2} \cdot S \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right],$$

gdzie:

- \bar{x} to średnia próbkowa,
- $t_{n-2, 1-\alpha/2}$ oznacza rozkład t -studenta z $n - 2$ stopniami swobody, rzędu $1 - \alpha/2$.

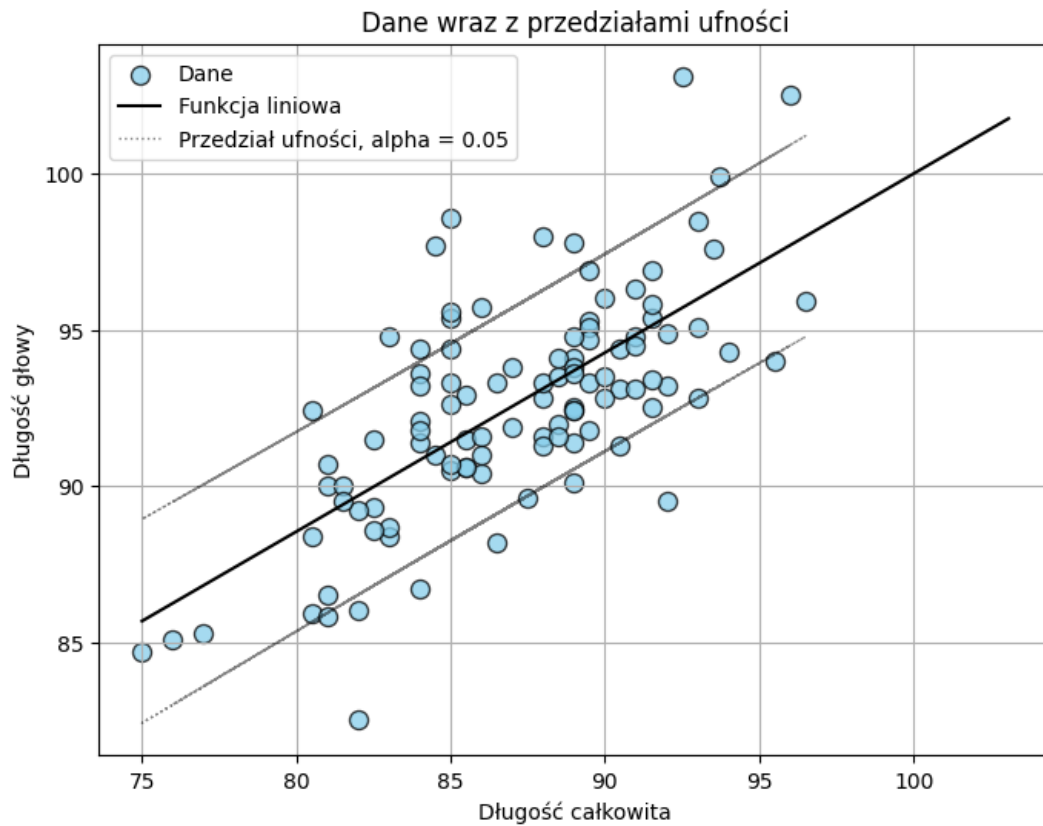
Przedziały prezentują się następująco:

β_0	β_1
[32.55612953160007, 52.863456766327495]	[0.45645787842959795, 0.6893446902467627]

Tabela 4: Obliczone przedziały ufności

Dodatkowo obliczono wartość $S^2 = 7.010622411312688$.

Następnie, używając tych danych, obliczono przedziały dla całej funkcji i zestawiono te dane z wykresem regresji liniowej.



Rysunek 6: Dane z dopasowaną prostą regresji

4.1 Wnioski

Większość danych zawiera się w zakresie predykcji, co sugeruje, że dopasowanie modelu jest sensowne. Niemniej jednak, spodziewane prawdopodobieństwo miało być równe 0.95, gdzie obliczenia empiryczne wskazują na prawdopodobieństwo bliskie 0.83. Taka niezgodność może wynikać z niewystarczającej ilości prób Monte Carlo, lub małej próbki danych. Wracając jednak do wyliczonego wcześniej współczynnika determinacji $R^2 = 0.48$ widzimy, że dopasowanie modelu nie jest obiecujące, co potwierdza zakres predykcji. Niedopasowanie modelu można również spróbować wytłumaczyć tym, że długość głowy i ciała nie zależą jedynie od siebie, ale również np. od genetycznych uwarunkowań oposów.

5 Analiza residuów

Analiza residuów jest potrzebna do procesu oceny jakości modelu statystycznego poprzez badanie jego błędów (residuów). Do sprawdzenia czy dane spełniają założenia modelu regresji, czyli:

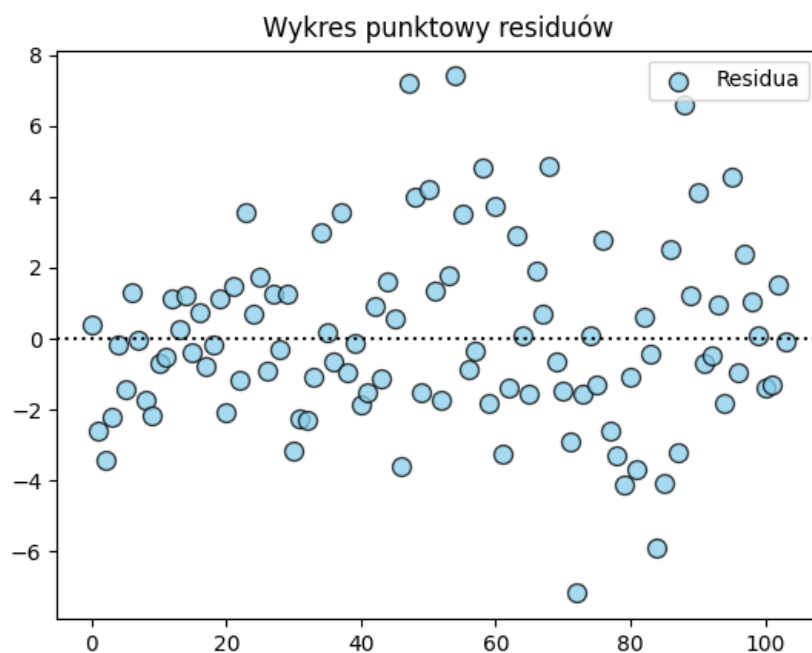
1. $E(\epsilon_i) = 0$,
2. $\text{Var}(\epsilon_i) = \sigma^2 < \infty$,
3. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$,
4. ϵ_i są niezależne (nieskorelowane).

przeprowadzimy analizę residuów $e_i = y_i - \hat{y}_i$, czyli realizacji zmiennych ϵ_i .

5.1 Sprawdzenie $E(\epsilon_i) = 0$

Ponieważ estymatory $\hat{\beta}_1, \hat{\beta}_0$ zostały wyznaczone metodą najmniejszych kwadratów, wiadomo, że pierwsze założenie na pewno jest spełnione.

Sporządzono również wykres reprezentujący rozłożenie reszt.



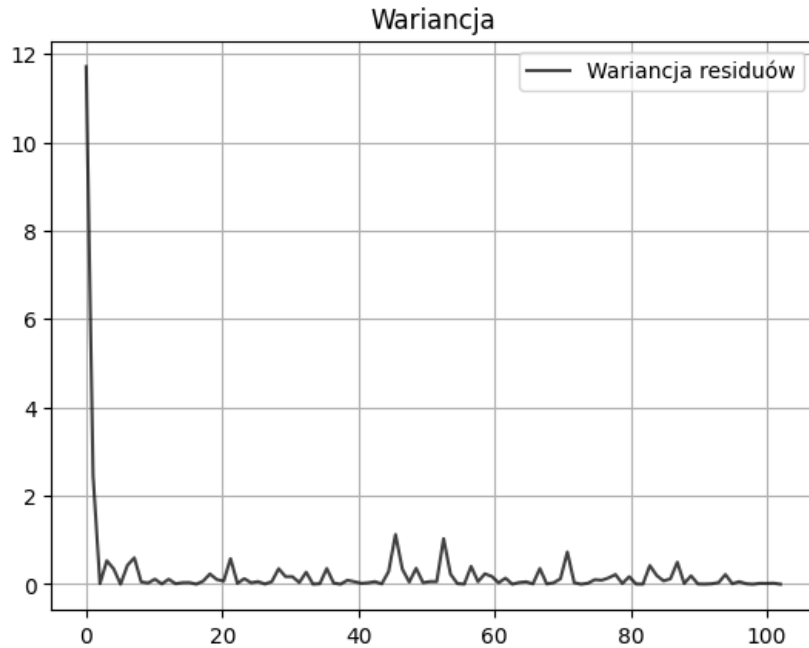
Rysunek 7: Rozłożenie reszt

Zauważalny jest fakt, że im bliżej poziomej linii 0, tym znajduje się tam więcej punktów. Reszty są ułożone losowo wokół 0, zatem można przyjąć poprawność pierwszego warunku.

5.2 Sprawdzenie czy $\text{Var}(\epsilon_i) = \sigma^2 < \infty$

Aby sprawdzić ten warunek, obliczono wariancje częściowe S_k^2 residuów dla $k = 2, \dots, n$ i przedstawiono je na wykresie. Wariancje obliczono według następującego wzoru:

$$S_k^2 = \frac{1}{k-1} \sum_{i=1}^k e_i^2, \quad k = 2, \dots, n.$$

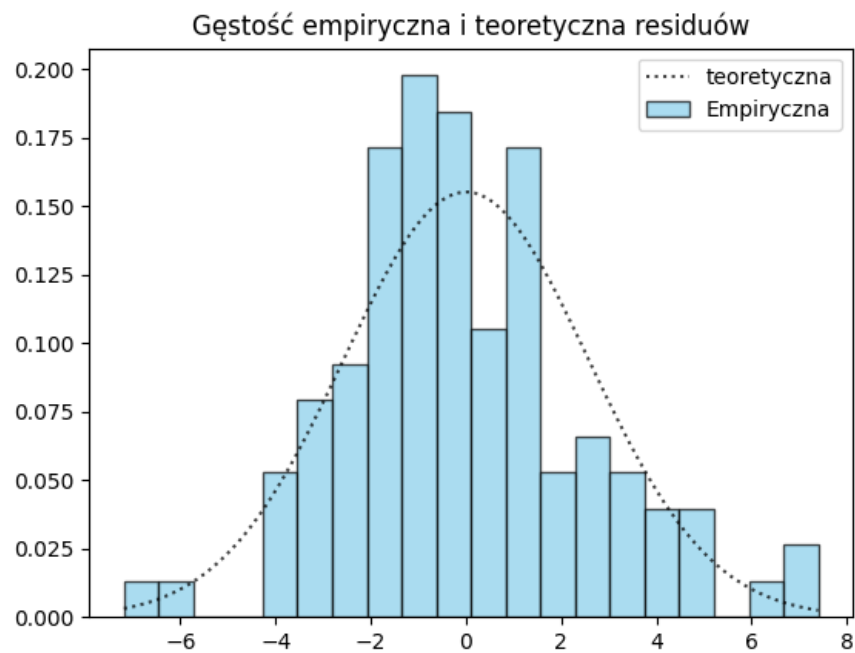


Rysunek 8: Wariancja częściowa S_k^2 residuów dla $k = 2, \dots, n$

Pomimo początkowych skoków wartości wariancji, wraz ze wzrostem k , wykres stopniowo zbliża się do 0. Na tej podstawie można przypuszczać, że wariancja jest skończona.

5.3 Sprawdzenie $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

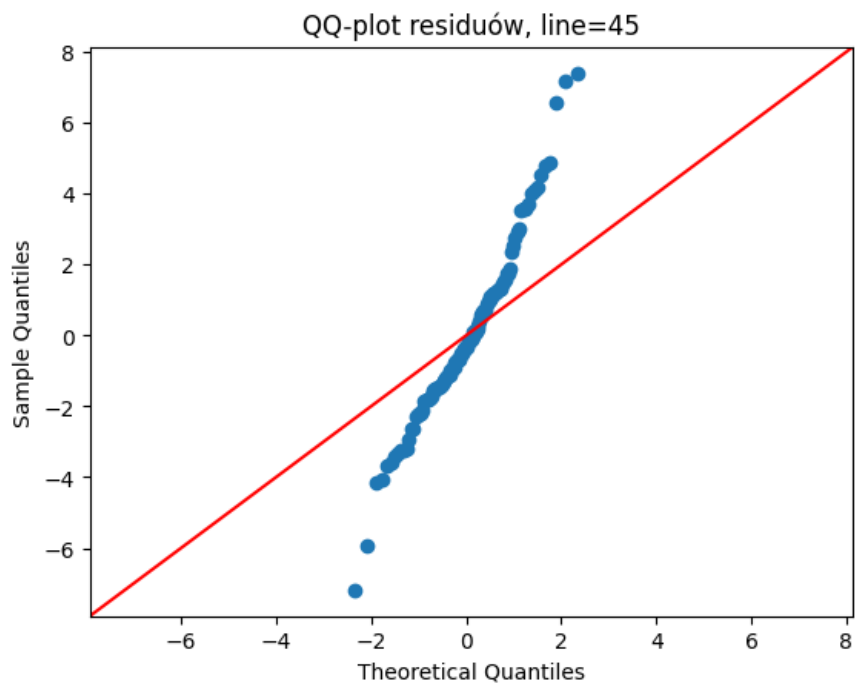
Aby zweryfikować to założenie, wygenerowano wykres gęstości empirycznej z próbki residuów i teoretycznej rozkładu normalnego.



Rysunek 9: Wykres gęstości empirycznej i teoretycznej residuów

Jak widać na powyższym rysunku, obie gęstości nie do końca się pokrywają. Daje to solidne podstawy do zwątpienia w normalność rozkładu reszt.

Aby sprawdzić normalność także na inne sposoby, wygenerowano również wykres kwantylowy danych. Wybrano wykres z nachyleniem linii kwantylowej do osi rzędu 45° .



Rysunek 10: Wykres kwantylowy residuów

Mimo iż nie spodziewamy się tutaj bardzo dokładnego pokrywania ze względu na niewielką liczbę danych, widoczne są spore odchylenia od teoretycznych wartości. Wykres kwantylowy zupełnie nie pokrywa się z empirycznymi wartościami. Podczas wykorzystywania testów normalności takich jak test Shapiro-Wilka, otrzymujemy odrzucenie hipotezy zerowej H_0 - dane pochodzą z rozkładu normalnego. Należy zatem odrzucić założenie o normalności residuów.

5.4 Sprawdzenie czy ϵ_i są niezależne (nieskorelowane)

Sprawdzono, czy rozważane dane są niezależnymi zmiennymi losowymi. Do tego celu wykorzystano dwie funkcje: Empiryczną autokowariancję:

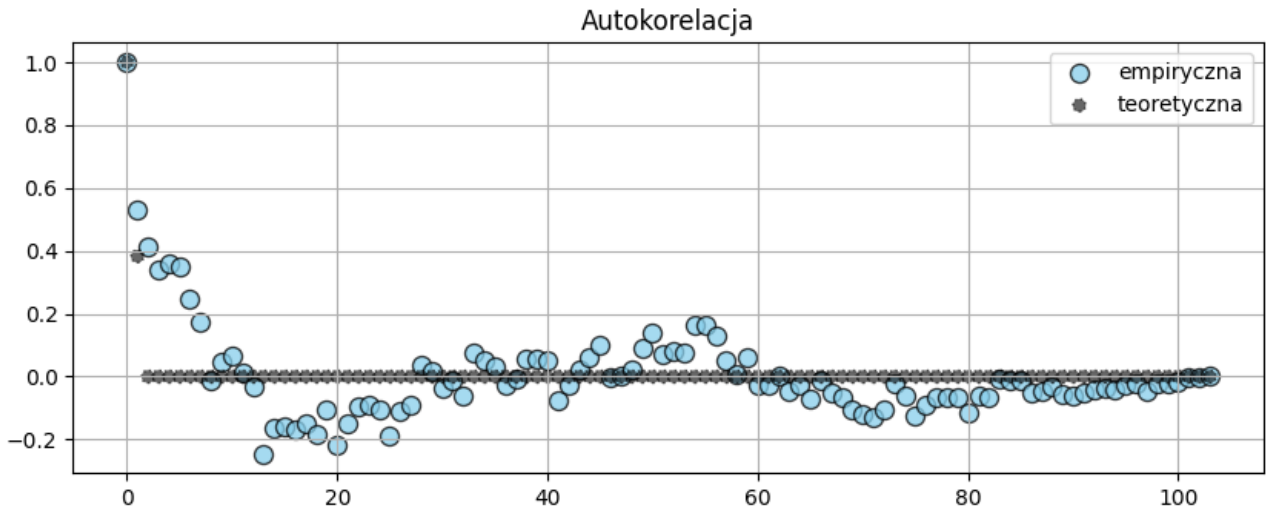
$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} (e_{i+|h|} - \bar{e})(e_i - \bar{e}), \quad 0 \leq h < n, \quad h \in \mathbb{Z},$$

Empiryczną autokorelację:

$$\hat{g}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)},$$

gdzie: h to opóźnienie.

Na ich podstawie utworzono funkcję empirycznej autokorelacji i przedstawiono te dane na wykresie.

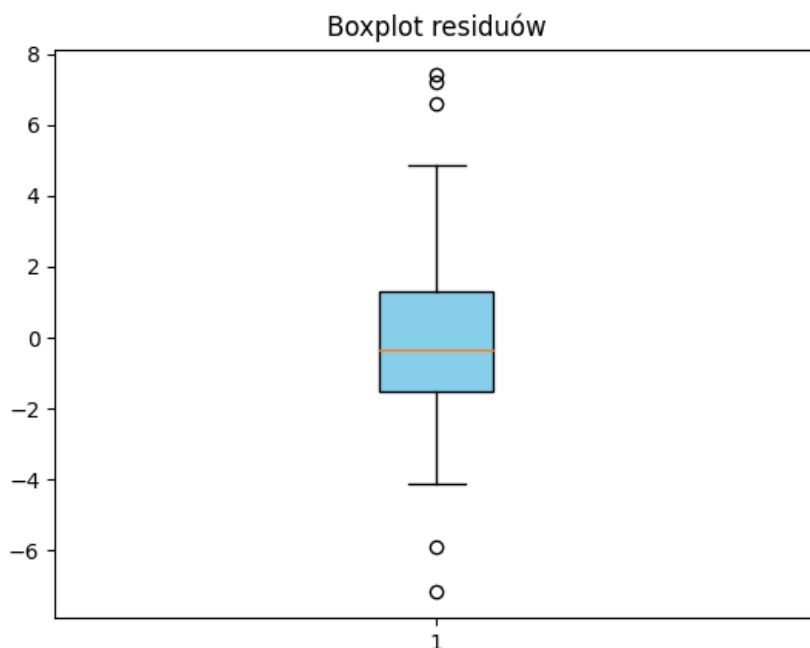


Rysunek 11: Autokorelacja residuów

Dla $h = 0$ oczywiście otrzymano 1, jednak dla pozostałych h wartości empirycznej autokorelacji są bliskie 0. Tak wyglądająca funkcja autokorelacji empirycznej dowodzi, że badane reszty są zmiennymi nieskorelowanymi.

5.5 Wartości odstające

Wartościom odstającym przypatrzone się bliżej za pomocą wykresu pudełkowego residuów.



Rysunek 12: Wykres pudełkowy residuów

Zauważono kilka odstających wartości, wykraczających poza zwykły zakres boxplotu. Mogą one wskazywać na przypadki, w których model słabo radzi sobie z ich wyjaśnieniem lub ich wartości są znacząco różne od reszty danych. Mimo że jest ich niewiele, mogą one mieć duży wpływ na trafność modelu, ponieważ przy niewielkiej liczbie danych dokładność się zmniejsza (większy wpływ na wynik mają wartości odstające).

5.6 Wnioski

Powyższa analiza residuów pokazała, że choć reszty są nieskorelowane, o średniej bliskiej 0 i o ograniczonej wariancji, niestety nie pochodzą z rozkładu normalnego, choć są do niego lekko zbliżone.

6 Predykcje

Proces predykcji oraz oceny jej efektywności przeprowadzany jest w następujący sposób:

1. Podzielono dane na zbiór treningowy i testowy.
2. Na zbiorze treningowym obliczane są współczynniki $\hat{\beta}_0$ i $\hat{\beta}_1$ dla prostej regresji.
3. Na podstawie zbioru treningowego budowana jest prosta regresji, a następnie obliczane są błędy predykcji oraz analizowany jest ich rozkład.
4. Na zbiorze testowym generowane są wartości predykowane, uwzględniając błędy znanego rozkładu.
5. Tworzone są przedziały ufności dla zbioru testowego na określonym poziomie istotności $1-\alpha$ oraz obliczane, jaka część predykowanych wartości oraz rzeczywistych danych testowych znajduje się w nich.

Analizowane dane zostały podzielone na zbiór treningowy i testowy. Zbiór testowy składa się z 50 obserwacji o największej długości całkowitej ciała oposa.

Współczynniki regresji liniowej wyliczone dla zbioru treningowego wynoszą:

$\hat{\beta}_0$	$\hat{\beta}_1$
54.39283695048063	0.43985772052711536

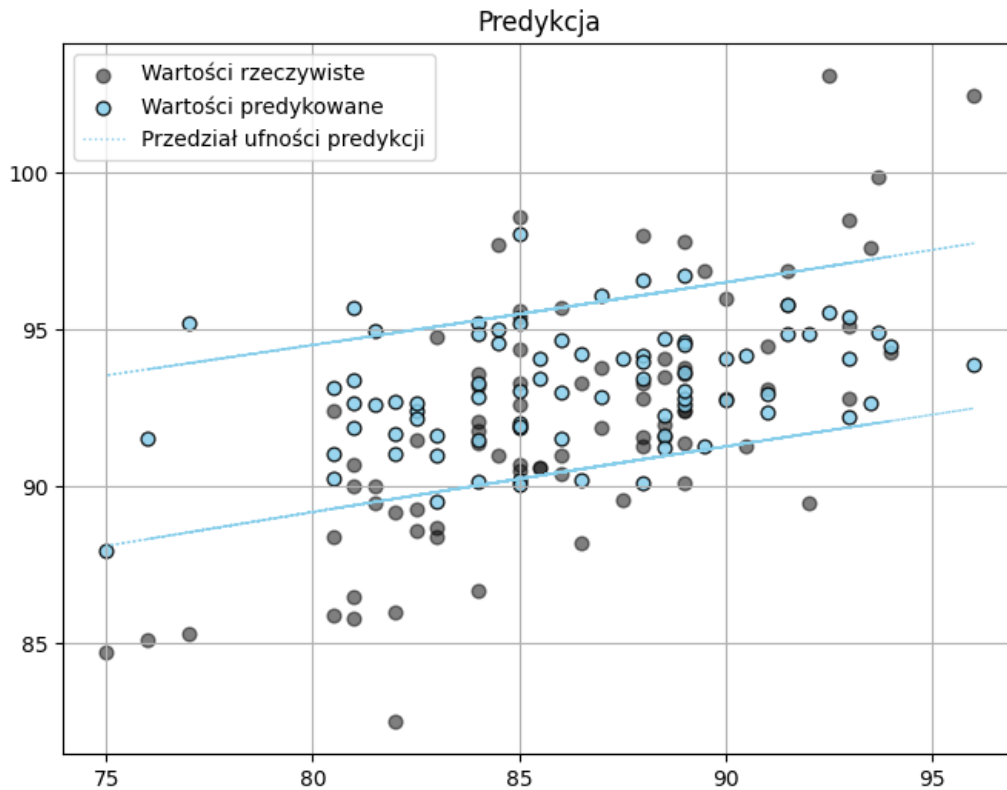
Tabela 5: Obliczone współczynniki dla predykcji

Przedziały ufności zostały wyznaczone za pomocą:

$$\left[\hat{y}(x_0) - t_{1-\alpha/2, n-2} \cdot S_{ut} \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 \sum_i (x_i - \bar{x})^2}, \hat{y}(x_0) + t_{1-\alpha/2, n-2} \cdot S_{ut} \sqrt{1 + \frac{1}{n} + (x_0 - \bar{x})^2 \sum_i (x_i - \bar{x})^2} \right]$$

W obliczeniach wykorzystano $\alpha = 0.05$ oraz nieobciążony estymator odchylenia standardowego:

$$S = \sqrt{\frac{1}{n-2} \sum_i (Y_i - \hat{Y}_i)^2}.$$



Rysunek 13: Wykres danych przewidywanych, ich przedziałów ufności oraz wartości rzeczywistych

Różnice między wartościami rzeczywistymi a przewidywaniami, czyli residua, są zauważalne. Duże residua sugerują, że model może nie w pełni objaśniać zmienność danych. Przedział ufności jest szeroki, co sugeruje, że model może być niepewny co do swoich prognoz.

7 Wnioski

Wnioski z przeprowadzonej analizy dotyczącej długości ciała oraz głowy oposów pozwalają na kilka istotnych spostrzeżeń. Pomimo pewnych trudności w dopasowaniu modelu, analiza sugeruje istnienie liniowej zależności między badanymi zmiennymi. Model regresji liniowej nie jest idealny, co potwierdzają zarówno statystyki jak i analiza residuów. Wartość współczynnika determinacji wskazuje, że model nie jest w stanie wyjaśnić części zmienności danych. Jednakże, istniejące wartości mogą wskazywać na obecność dodatkowych czynników wpływających na długość ciała oposów, które nie zostały uwzględnione w analizie.

Analiza residuów pokazała, że reszty nie pochodzą z rozkładu normalnego, sugerując, że model regresji może nie być najoptymalniejszym narzędziem do prognozowania długości ciała i głowy oposów. Trzeba również wspomnieć, że predykcje modelu na zbiorze testowym, mimo uwzględnienia przedziałów ufności, nie zawsze pokrywają się z rzeczywistymi danymi, co może wynikać z trudności w generalizacji modelu na nowe obserwacje,

W związku z powyższym, sugeruje się kontynuację badań, które uwzględnią dodatkowe zmienne, takie jak genetyczne uwarunkowania czy dieta, w celu pełniejszego zrozumienia czynników wpływających na długość ciała i głowy oposów. Taki podejście pozwoli na stworzenie bardziej dopasowanego modelu, lepiej odzwierciedlającego złożoność analizowanego zjawiska.