Raport 1
Regresja Liniowa

Bartłomiej Gintowt

20.12.2022r.

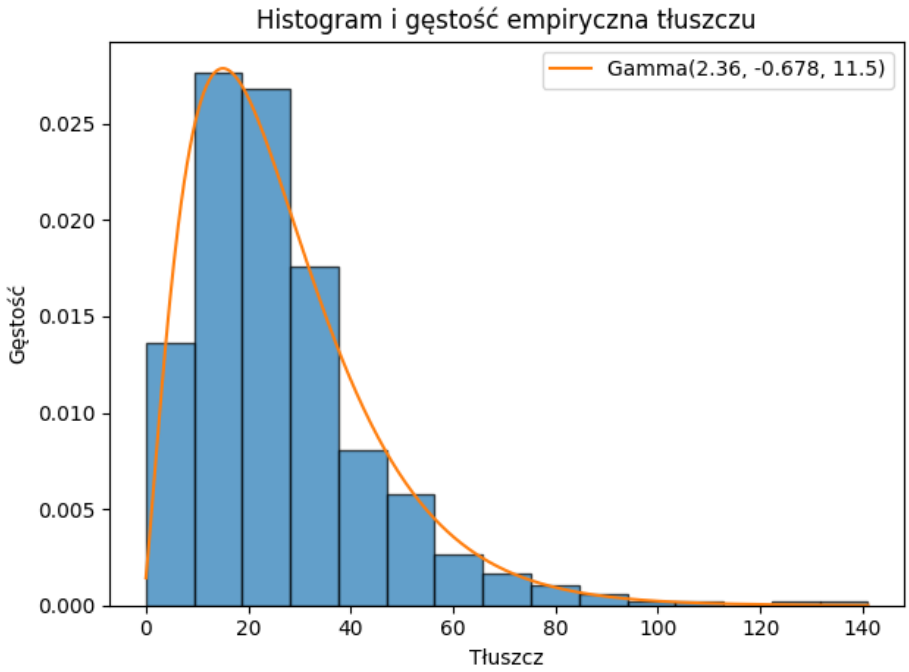
1 Wstęp

Dane wykorzystywane do wykonania raportu zostały pobrane ze strony: https://vincentarelbundock.github.io/Rdatasets/datasets.html [Item: fastfood]. W pliku możemy znaleźć informacje dotyczące wartości odżywczych dla pozycji w menu największych restauracji fastfoodowych w Stanach Zjednoczonych. Jako zmienną objaśniającą przyjmujemy dane z kolumny "calories", które opisują całkowitą wartość kalorii w daniu. Natomiast za zmienną objaśnianą wartości z kolumny "total_fat", podane w gramach, mówiące o całkowitej zawartości tłuszczu dla danej pozycji w menu. Długość naszej próbki danych wynosi 515. Celem sprawozdania będzie dopasowanie odpowiedniego modelu regresji liniowej do danych testowych oraz ocena czy model ten będzie dobrany całkowicie poprawnie.

2 Analiza jednowymiarowa zmiennej niezależnej oraz zmiennej zależnej

2.1 Zmienna zależna - Tłuszcz

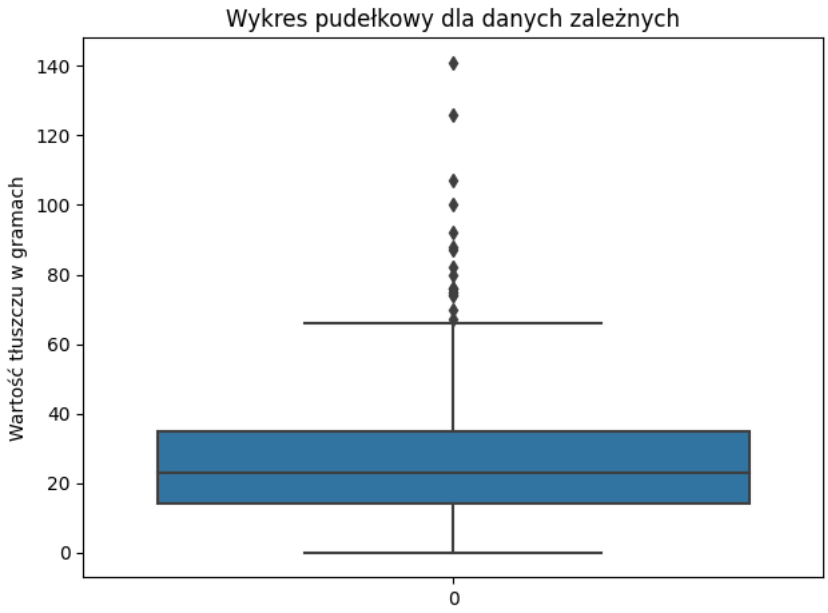
Utworzymy histogram danych z kolumny tłuszcze, a następnie za pomocą pakietu distfit dostępnym w języku Python dopasujemy do niego najlepszą możliwą gęstość wraz z odpowiednimi parametrami.



Rysunek 1: Histogram wraz z dopasowaną gęstością do danych zależnych.

Otrzymujemy, iż najlepiej dobraną gęstością dla naszego histogramu jest gęstość rozkładu Gamma($a = 2.36122, loc = -0.679726, scale = 11.5491$). Dodatkowo zauważamy prawostronną skośność histogramu.

Przejrzymy również wykres pudełkowy naszych danych zależnych.



Rysunek 2: Wykres pudełkowy dla danych zależnych.

Zauważamy, że średnia wartości tłuszczu znajduje się w przedziale między 20 a 40. Wartości odstające natomiast zaczynają się powyżej wartości około 70 oraz poniżej 0.

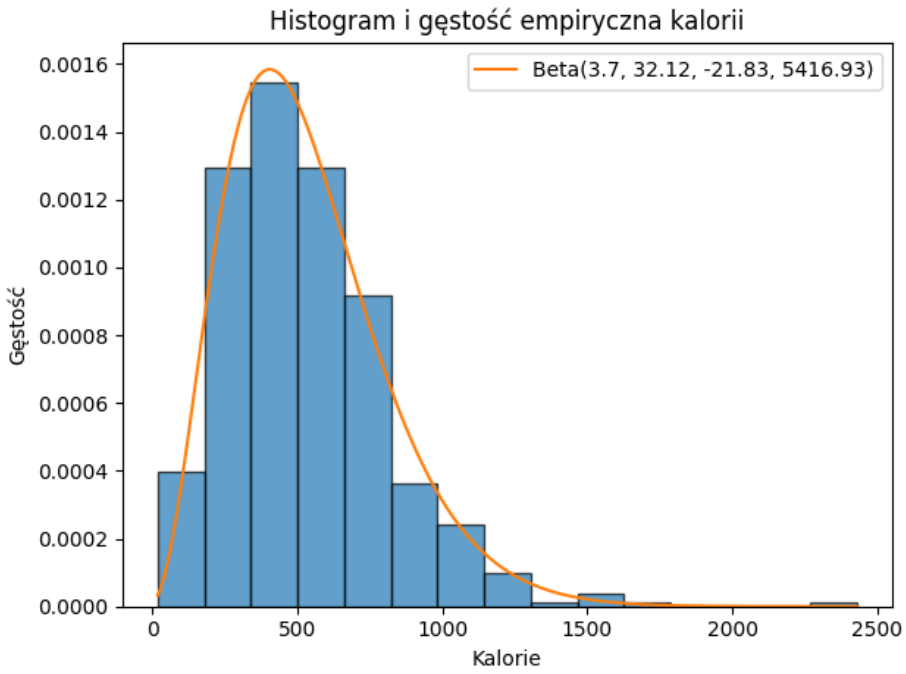
Zmierzymy teraz podstawowe miary dla naszego zbioru danych zależnych.

Tłuszcze	
Średnia	26.59
Mediana	23.0
$\text{Kwartyl}_{0.25}$	14.0
$\text{Kwartyl}_{0.75}$	35.0
Rozstęp z próby	141.0
Rozstęp międzykwartylowy	21.0
Wariancja z próby	338.339
Odchylenie standardowe z próby	18.39
Współczynnik zmienności	0.6918
Skośność	1.7869
Kurtoza	5.5419

Mediana naszych danych jest mniejsza niż ich średnia, zatem zgadza się, że histogram jest prawostronnie skośny. Kurtoza jest dodatnia, zatem intensywność wartości skrajnych jest większa niż dla rozkładu normalnego. Wartość współczynnika zmienności informuje nas o silnej zmienności w danych.

2.2 Zmienna niezależna - Kalorie

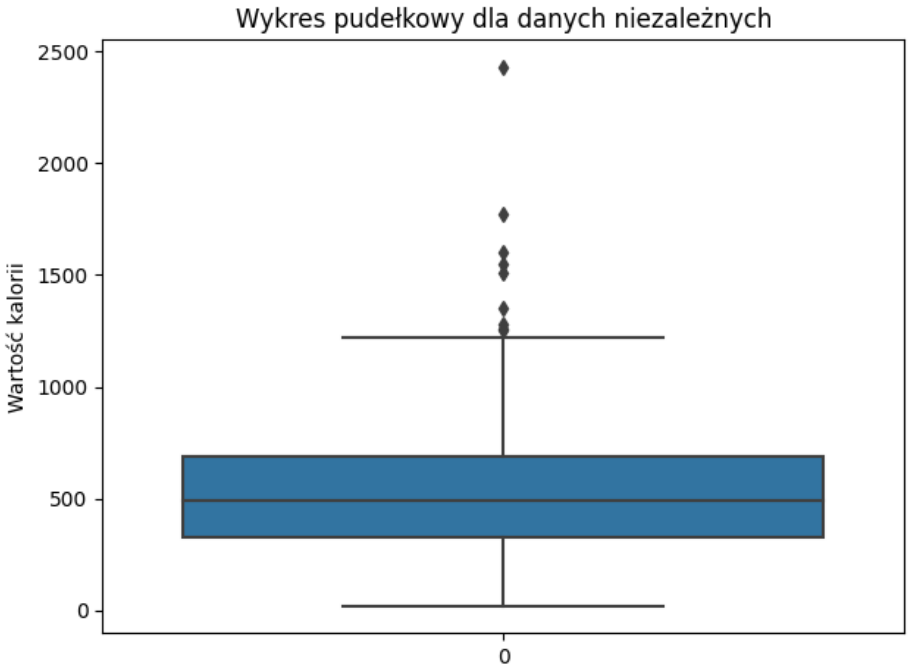
Podobnie jak dla tłuszczy zwizualizujemy sobie dane niezależne. Utworzymy histogram wartości i dopasujemy gęstość.



Rysunek 3: Histogram wraz z dopasowaną gęstością do danych niezależnych.

Otrzymujemy, iż najlepiej dobranym rozkładem będzie rozkład Beta($a = 3.66778, b = 32.1245, loc = -21.8256, scale = 5416.93$). Ponownie zauważyć możemy skośność prawostronną histogramu.

Przejrzymy również wykres pudełkowy naszych danych niezależnych.



Rysunek 4: Wykres pudełkowy dla danych niezależnych.

Zauważamy, że średnia wartości kalorii wynosi około 500. Wartości odstające natomiast zaczynają się powyżej wartości około 1300 oraz poniżej 0.

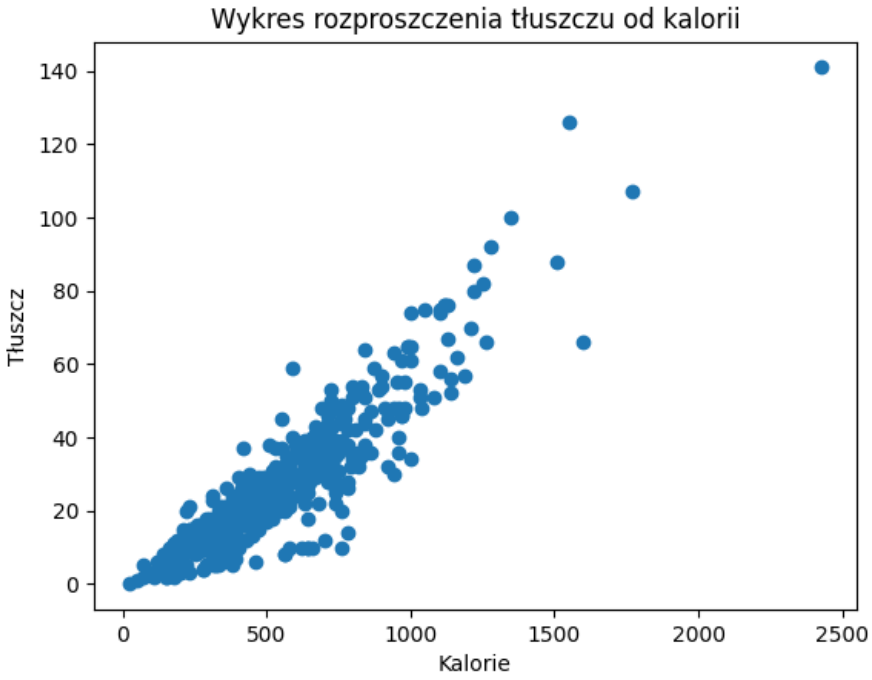
Zmierzymy teraz podstawowe miary dla naszego zbioru danych niezależnych.

Kalorie	
Średnia	532.3495
Mediana	490.0
$\text{Kwartyl}_{0.25}$	330.0
$\text{Kwartyl}_{0.75}$	690.0
Rozstęp z próby	2410.0
Rozstęp międzykwartylowy	360.0
Wariancja z próby	79438.169
Odchylenie standardowe z próby	281.848
Wspólczynnik zmienności	0.5294
Skośność	1.4088
Kurtoza	4.6603

Średnia jest większa od mediany, zatem zgadza się, że histogram jest prawostronnie skośny. Kurtoza jest większa od zera, zatem intensywność wartości skrajnych jest większa niż dla rozkładu normalnego. Współczynnik zmienności informuje nas o że dane nie są jednorodne.

3 Analiza zależności liniowej pomiędzy zmienną objaśniającą a zmienną objaśnianą

Zaczniemy od sprawdzenia wykresu rozproszenia tłuszczy w zależności od kalorii.



Rysunek 5: Wykres rozproszenia tłuszczy w zależności od kalorii.

Zauważamy, że dane na wykresie układają w sposób liniowy. Będziemy dzięki temu mogli dostosować do nich odpowiedni model regresji liniowej wyrażony wzorem $Y = B_0 + B_1 * x_i + \varepsilon_i$.

Wzory na teoretyczne estymatory współczynników wyznaczonych przy pomocy metody najmniejszych kwadratów:

$$\hat{B_1} = B_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x}) * \varepsilon_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\hat{B_0} = \overline{Y} - \hat{B_1} * \overline{x}$$

Jako, że bazujemy na danych rzeczywistych posłużymy się estymowanym modelem regresji $\hat{Y} = \hat{B_0} + \hat{B_1} * x_i$.

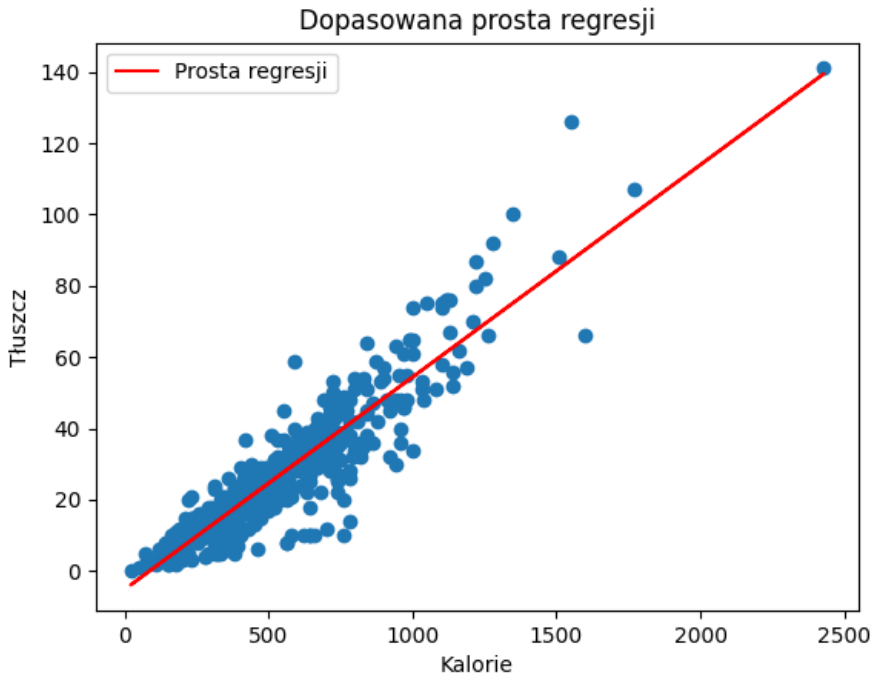
Współczynniki dla takiego modelu wyrażone są wzorem:

$$\hat{B_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$
$$\hat{B_0} = \overline{Y} - \hat{B_1} * \overline{x}$$

Dla naszego zbioru danych otrzymujemy następujące wartości estymowanych współczynników:

$$\hat{B_1} = 0.059510306626776636$$
$$\hat{B_0} = -5.08999158213026$$

Nałożymy wyestymowaną prostą regresji na wykres rozproszenia.



Rysunek 6: Wykres rozproszenia oraz dopasowana prosta regresji.

Dopasowana krzywa regresji oddaje stosunkowo dobrze faktyczny przebieg naszych danych.

Korzystając z estymacji przedziałowej możemy wyznaczyć przedziały ufności, do których powinny należeć nasze teoretyczne współczynniki. Wykorzystamy do tego poprzednio wyznaczone wartości estymatorów prostej regresji.
Przedziały ufności dla teoretycznych współczynników przy nieznanej sigmie wyrażają się wzorami:

$$(\hat{B_0} - t_{(1-\frac{\alpha}{2},n-2)} \cdot s\sqrt{\frac{1}{n} + \frac{(\overline{x_n})^2}{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}} \leqslant B_0 \leqslant \hat{B_0} + t_{(1-\frac{\alpha}{2},n-2)} \cdot s\sqrt{\frac{1}{n} + \frac{(\overline{x_n})^2}{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}})$$

$$(\hat{B_1} - t_{(1-\frac{\alpha}{2},n-2)} \cdot s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}} \leqslant B_1 \leqslant \hat{B_1} + t_{(1-\frac{\alpha}{2},n-2)} \cdot s\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x_n})^2}}),$$

gdzie $t_{(1-\frac{\alpha}{2},n-2)}$ oznaczają kwantyle dla rozkładu t-studenta rzędu $1-\frac{\alpha}{2}$ o $n-2$ stopniach swobody oraz o nieobciążonym estymatorze odchylenia standardowego

$$s = \sqrt{\frac{1}{n-2} \cdot \sum_{i=1}^{n}(Y_i - \hat{Y})^2}.$$

Podstawiając wartości danych do wyznaczonych wzorów otrzymujemy przedziały ufności równe:

$$(-6.489682740150616 \leqslant B_0 \leqslant -3.690300424109904)$$

$$(0.057186616962654174 \leqslant B_1 \leqslant 0.0618339962908991).$$

Teraz skupimy się ocenie poziomu zależności zmiennych. Posłużymy się do tego współczynnikiem korelacji Pearsona ($r_{pearsona}$), całkowitą sumą kwadratów (SST), całkowitą sumą błędów (SSE), całkowitą sumą kwadratów regresji (SSR) oraz współczynnikiem determinacji (R^2). Dane są one następującymi wzorami:

$$r_{pearsona} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})}}$$

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$$

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$$

$$R^2 = r^2_{pearson} = \frac{SSR}{SST}$$

Korzystając z powyższych wzorów otrzymujemy następujące wartości dla danych rzeczywistych:

Ocena poziomu zależności	
$r_{pearsona}$	0.9119
SST	174244.5515
SSE	29360.4169
SSR	144884.1346
R^2	0.8315

Wartość współczynnika korelacji Pearsona mówi nam o bardzo silnej zależności liniowej danych rzeczywistych. Skoro jego wartość jest dodatnia to wraz ze wzrostem wartości niezależnych rosną wartości zależne (funkcja liniowa jest rosnąca). Współczynnik R^2 jest wskaźnikiem jakości dopasowania danych do modelu regresji. Z jego wartości wynika, że model jest dobrze dobrany.

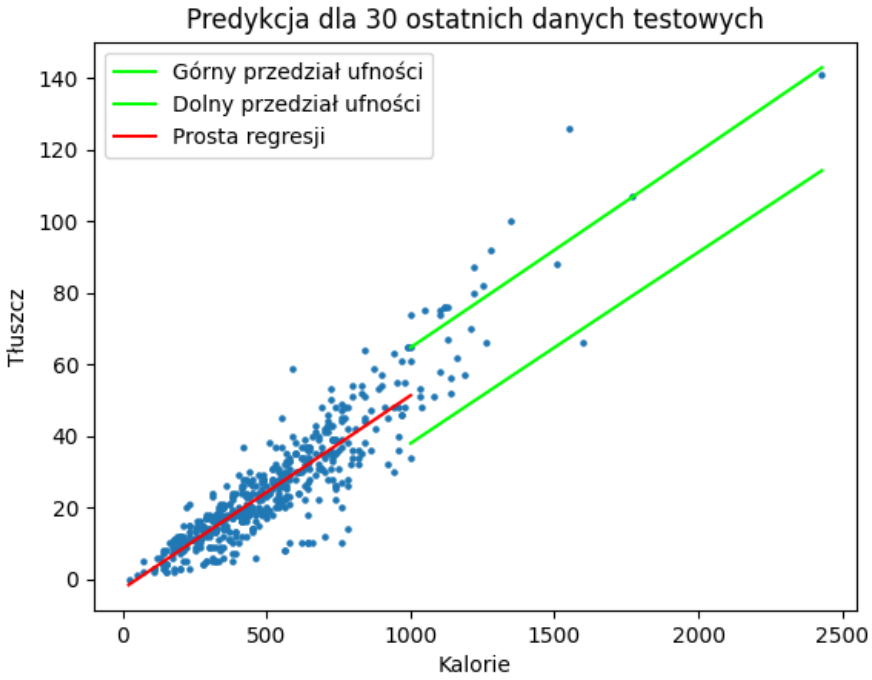
Teraz zajmiemy się wyznaczeniem predykcji dla naszego modelu. Będziemy ucinać 30 wartości z naszych danych aby następnie wyznaczyć dla nich przedziały ufności. Po usunięciu danych przedziały ufności dla prognozowanych wartości będą wyrażać się wzorami:

$$(\hat{B_1}x_{n-30+i}+\hat{B_0}-t_{(1-\frac{\alpha}{2},n-32)}\cdot s\sqrt{1+\frac{1}{n-30}+\frac{(x_{n-30+i}-\overline{x_{n-30}})^2}{\sum_{i=1}^{n}(x_i-\overline{x_n})^2}} \leqslant Y_{n-30+i} \leqslant$$

$$\hat{B_1}x_{n-30+i}+\hat{B_0}+t_{(1-\frac{\alpha}{2},n-32)}\cdot s\sqrt{1+\frac{1}{n-30}+\frac{(x_{n-30+i}-\overline{x_{n-30}})^2}{\sum_{i=1}^{n}(x_i-\overline{x_n})^2}}),$$

gdzie $\hat{B_1}$ i $\hat{B_0}$ to estymatory wyliczone metodą najmniejszych kwadratów dla n-30 pierwszych danych.

Korzystając z powyższych wzorów jesteśmy w stanie utworzyć wykres wartości przewidywanych.



Rysunek 7: Wykres wraz z przedziałami ufności predykcji.

Zauważamy, że część usuniętych wartości wpada do wyznaczonego przedziału ufności, są jednak takie, które dość wyraźnie do niego nie należą. Oznacza to, że model może być nie najlepiej dopasowany. Wynikać to może chociażby ze zmiany zachowania dla wartości kalorii większych od 1000, co może świadczyć o chociażby delikatnej zmianie ich zależności względem reszty danych. W mocnym przybliżeniu możemy jednak założyć, iż model się sprawdza.

4 Analiza residuów

W tej części skupimy się na analizie residuów (błędów) naszego modelu regresji liniowej. Sprawdzimy czy spełniają one następujące warunki:

$$E\varepsilon_i = 0 \ \forall_{i=1,2,..,n}$$

$$Var\varepsilon_i = \sigma^2 \ \forall_{i=1,2,..,n}$$

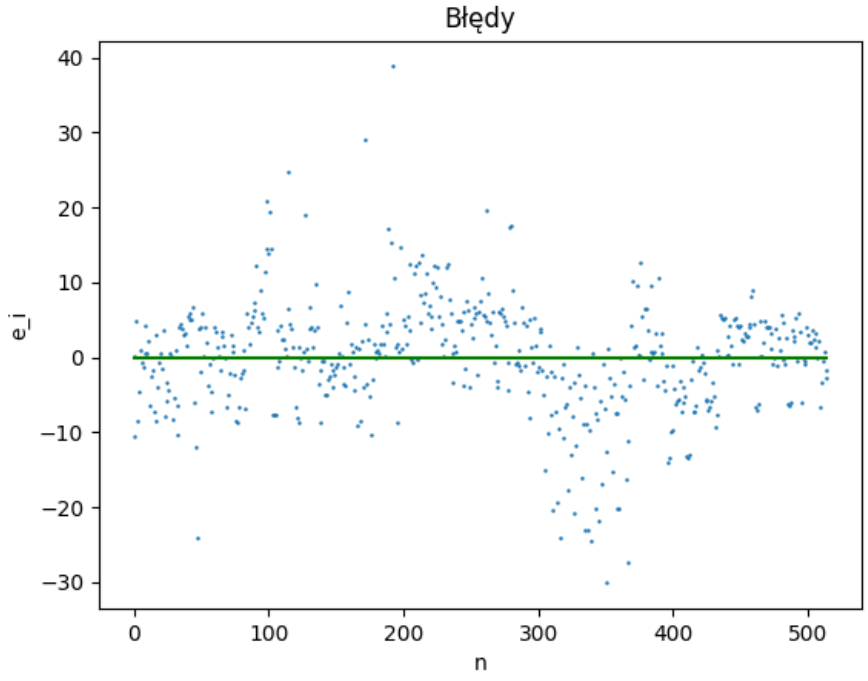
ε_i to niezależne zmienne losowe $\forall_{i=1,2,..,n}$.

W przypadku, gdy któryś z warunków nie jest spełniony nasze model regresji jest błędny i należy go odrzucić. Dodatkowo sprawdzimy również normalność residuów.

W celu wyznaczenia ich wartości skorzystamy ze wzoru

$$e_i = Y_i - \hat{Y}_i.$$

Sprawdzimy teraz wykres rozproszenia naszych błędów.

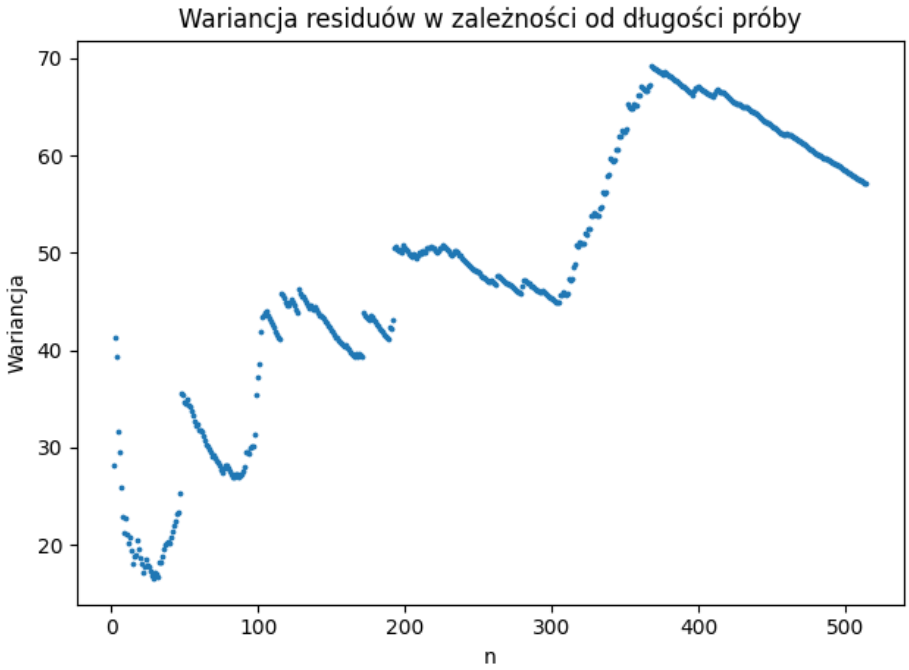


Rysunek 8: Wykres rozproszenia residuów.

Na jego podstawie jesteśmy w stanie "na oko" stwierdzić, że średnia błędów oscyluje w okolicach zera, a wariancja wyraźnie nie ma wartości stałej. Dla błędów w przedziałach chociażby $0, 100$ oraz $300, 400$ zaobserwować możemy zupełnie inne wartości wariancji. Przeprowadzimy jednak dokładniejsze testy dla średniej oraz wariancji w celu oceny poprawności dopasowanego modelu.

Wartość średnia wartości wszystkich residuów wynosi $2.4282625532772355 \cdot 10^{-15}$ co oznacza, że faktycznie średnia jest bardzo bliska wartości 0. Można zatem założyć, że pierwszy warunek jest spełniony.

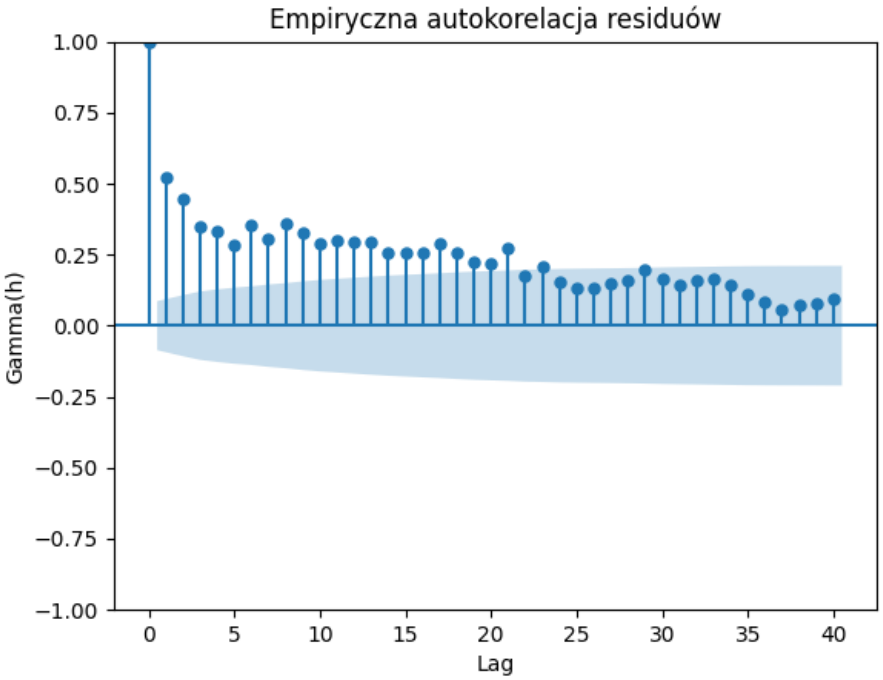
Sprawdzimy teraz wartość wariancji błędów w zależności od długości próby.



Rysunek 9: Wariancja residuów w zależności od długości próby.

Z wykresu wynika, iż wariancja nie jest stała, zmienia się drastycznie w zależności od długości próby. Na podstawie tego możemy odrzucić prawdziwość drugiego założenia i uznać dobrany model regresji za niepoprawny.

W celu sprawdzenia warunku niezależności residuów utworzymy wykres empirycznej funkcji autokorelacji residuów ($\hat{\gamma}(h)$).



Rysunek 10: Empiryczna autokorelacja residuów.

Z wykresu możemy zauważyć, że wartości funkcji autokorelacji przez długi czas nie wpadają do przedziału, dla którego residua byłyby niezależne. Skorzystaliśmy również z testu Durbina-Watsona i otrzymaliśmy wartość 0.95365, która mówi nam o istnieniu dodatniej autokorelacji. Oznacza to, że występuje korelacja błędów, a model można uznać za niepoprawny.

Na koniec sprawdzimy jeszcze czy residua są z rozkładu normalnego. Pierwszym czynnikiem przeczącym tej tezie jest zmienność wariancji, która uniemożliwia, że błędy są rozkładu $N(0, \sigma^2)$. Wykorzystać możemy również test statystyczny Shapiro-Wilka. Dla rozpatrywanych residuów zwraca nam on p-wartość równą $4.1939706915683495 \cdot 10^{-12}$. Rozkład testowanych zmiennych jest normalny w przypadku p-wartości większych niż 0.05, zatem zauważyć możemy fałszywość hipotezy jakoby residua były rozkładu normalnego.

5 Podsumowanie

Przeprowadzając analizę danych tłuszczy w zależności od kalorii wyznaczyliśmy ich podstawowe statystyki oraz miary. Byliśmy w stanie dopasować dla wartości zmiennych zależnych i niezależnych odpowiednie histogramy wraz z wykresami gęstości. Następnie przeprowadziliśmy analizę zależności pomiędzy zmiennymi. W tym celu posłużyliśmy się prostą regresji ze współczynnikami wyznaczonymi metodą najmniejszych kwadratów, wyznaczyliśmy przedziały ufności dla teoretycznych współczynników oraz wyznaczyliśmy przedziały ufności dla prognozowanych wartości, które okazały się w przybliżeniu poprawne. Wyliczyliśmy wartości liczbowe statystyk umożliwiających nam ocenę zależności liniowej zmiennych, jak chociażby współczynnik korelacji Pearsona czy determinacji. Okazały się one bardzo wysokie dzięki czemu stwierdzić możemy liniowość rozpatrywanych przez nas danych. Dokonaliśmy również oceny poprawności residuów oraz sprawdziliśmy czy są one z rozkład normalnego. Okazało się, że warunek na stałość wariancji nie został spełniony, jak i również warunek na niezależność błędów. Co oznacza, że dobrany model regresji liniowej w rzeczywistości źle działa dla danych testowych. Błędy również nie są z rozkładu normalnego o czym mogliśmy się przekonać wykonując test Shapiro-Wilka czy spoglądając na poprzednio wyznaczoną zmienność wariancji.