

# Komputerowa analiza szeregów czasowych



<b>Kierunek, nazwa wydziału</b> Matematyka stosowana, Wydział Matematyki	<b>Grupa ćwiczeniowa, termin zajęć</b> T00-79a Wtorek 15:15
<b>Imię, nazwisko, numer albumu</b> Małgorzata Kowalczyk 262295, Julia Mazur 262296	<b>Data oddania</b> 22.12.2022 r.
<b>Tytuł</b> Wykorzystanie poznanych metod dotyczących analizy zależności liniowej do wybranych danych rzeczywistych	<b>Prowadzący</b> Dr inż. Aleksandra Grzesiek

# Spis treści

<b>1 Wstęp</b>	<b>3</b>
1.1 Cel . . . . .	3
<b>2 Analiza jednowymiarowa zmiennych</b>	<b>4</b>
2.1 Analiza jednowymiarowa zmiennej objaśniającej . . . . .	4
2.1.1 Wizualizacja danych . . . . .	4
2.1.2 Podstawowe statystyki . . . . .	5
2.1.3 Porównanie rozkładów . . . . .	6
2.2 Analiza jednowymiarowa zmiennej objaśnianej . . . . .	8
2.2.1 Wizualizacja danych . . . . .	8
2.2.2 Podstawowe statystyki . . . . .	9
2.2.3 Porównanie rozkładów . . . . .	10
2.3 Wnioski . . . . .	11
<b>3 Analiza zależności liniowej pomiędzy zmienną objaśniającą a zmienną objaśnianą</b>	<b>13</b>
3.1 Analiza danych treningowych . . . . .	14
3.1.1 Wizualizacja danych . . . . .	14
3.1.2 Estymacja punktowa współczynników w klasycznym modelu regresji	14
3.1.3 Estymacja przedziałowa współczynników w klasycznym modelu regresji . . . . .	15
3.2 Ocena poziomu zależności . . . . .	17
3.3 Analiza danych testowych . . . . .	18
3.4 Wnioski . . . . .	20
<b>4 Analiza residuów</b>	<b>22</b>
4.1 Sprawdzenie założeń klasycznego modelu regresji . . . . .	22
4.2 Testy normalności rozkładu . . . . .	27
4.3 Wnioski . . . . .	28
<b>5 Podsumowanie</b>	<b>28</b>

# **1 Wstęp**

Raport dotyczy analizy zależności liniowej w danych rzeczywistych dotyczących wyników z testów, które zostały uzyskane przez uczniów liceum w Stanach Zjednoczonych. Dane te zostały pobrane ze strony <https://www.kaggle.com/>. Zawierają one 8 zmiennych, które opisują kolejno wynik z matematyki, czytania i pisania, informacje demograficzne o uczniu (wykształcenie rodziców, płeć, rasa) oraz przygotowanie do testu w postaci zrobienia kursu powtarzającego i zjedzenia posiłku. W pracy wykorzystano jedynie dwie z nich. Konkretnie jest to wynik testu matematycznego i testu czytania. Przygotowując dane do analizy, zweryfikowano także, czy nie występują w nich braki. Po sprawdzeniu otrzymano, że obie kolumny zawierają po 1000 obserwacji.

Wszystkie wyniki zaprezentowane w niniejszej pracy otrzymano przy pomocy języka R. Ponadto, w trakcie realizacji sprawozdania korzystano z wykładów dr hab. inż. Agnieszki Wyłomańskiej.

## **1.1 Cel**

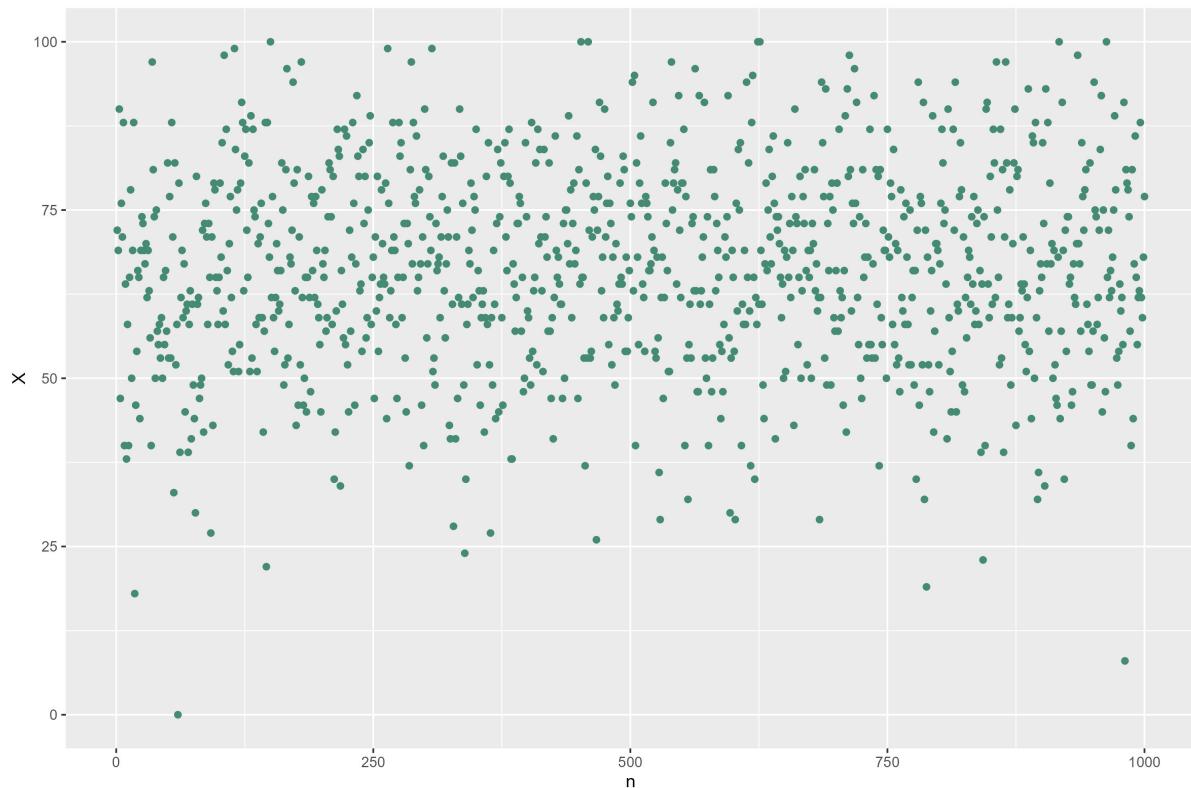
Celem niniejszej pracy jest zbadanie zależności między rezultatem uzyskanym przez uczniów z testu z matematyki a wynikiem testu z czytania. Dodatkowo sprawdzone zostanie zachowanie modelu regresji liniowej dla badanych danych. W tym celu dokonano analizy jednowymiarowej zmiennej zależnej i niezależnej. Zrealizowano to poprzez obliczenie podstawowych statystyk oraz wizualizacje danych. Zanalizowano także zależność pomiędzy zmiennymi oraz rozpatrzono residua badanego modelu.

## 2 Analiza jednowymiarowa zmiennych

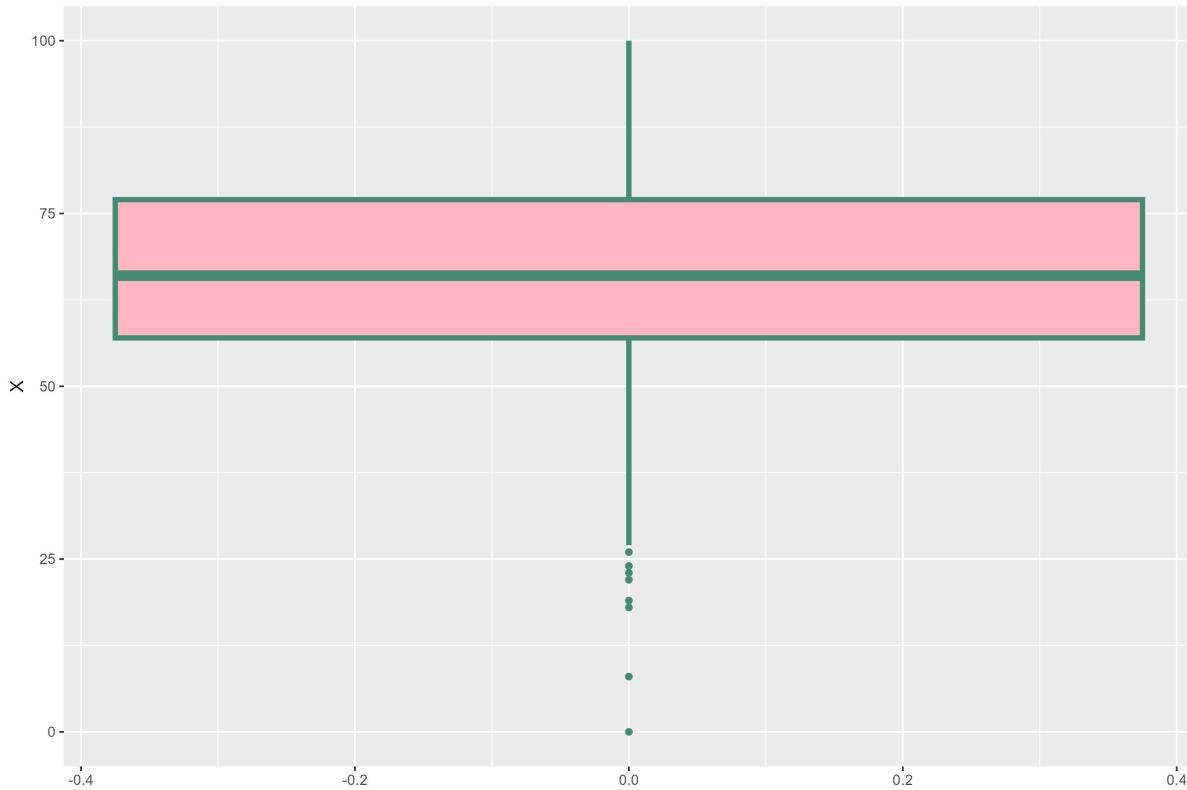
### 2.1 Analiza jednowymiarowa zmiennej objaśniającej

#### 2.1.1 Wizualizacja danych

Jako zmienną objaśniającą przyjęto wynik z matematyki (ozn.  $X$ ). Dzięki stworzeniu wykresu przedstawiającego rozkład sprawdzono, jakie wartości przyjmuje badana zmien- na. Zwizualizowano też ją na wykresie pudełkowym.



Wykres 2.1: Rozkład  $X$



**Wykres 2.2:** Boxplot dla  $X$

Na podstawie wykresu 2.1 stwierdzono, że wyniki otrzymane z testu z matematyki przyjmują wartości z przedziału  $[0, 100]$ . Zdecydowana większość uczniów uzyskała wynik powyżej 50 punktów, a zaledwie 7 osób uzyskało rezultat poniżej 25. Analogiczne wnioski można zauważać analizując wykres 2.2. Sugeruje on rozkład lewostronnie skośny, ponieważ obserwacje odstające, znajdują się pod pudełkiem. Oznacza to, że w grupie znalazło się kilka osób, którym test z matematyki poszedł znacznie słabiej.

### 2.1.2 Podstawowe statystyki

W tej sekcji wyliczono podstawowe statystyki dla  $X$ .

**Tabela 2.1:** Miary położenia  $X$

Średnia	Dominanta	Minimum	Kwartyl pierwszy	Kwartyl drugi	Kwartyl trzeci	Maksimum
66.09	65.00	0.00	57.00	66.00	77.00	100.00

**Tabela 2.2:** Miary rozproszenia  $X$

Rozstęp międzykwartylowy	Rozstęp	Wariancja	Odchylenie standardowe	Współczynnik zmienności
20.00	100.00	229.92	15.16	0.23

**Tabela 2.3:** Miary asymetrii i miary spłaszczenia  $X$

Współczynnik skośności	Kurtoza
-0.28	3.27

**Uwaga 1** Statystyki w tabelach 2.3 oraz 2.6 zostały obliczone na podstawie poniższych wzorów<sup>1</sup>:

- skośność:  $b_1 = \frac{m_3}{s^3}$ ,

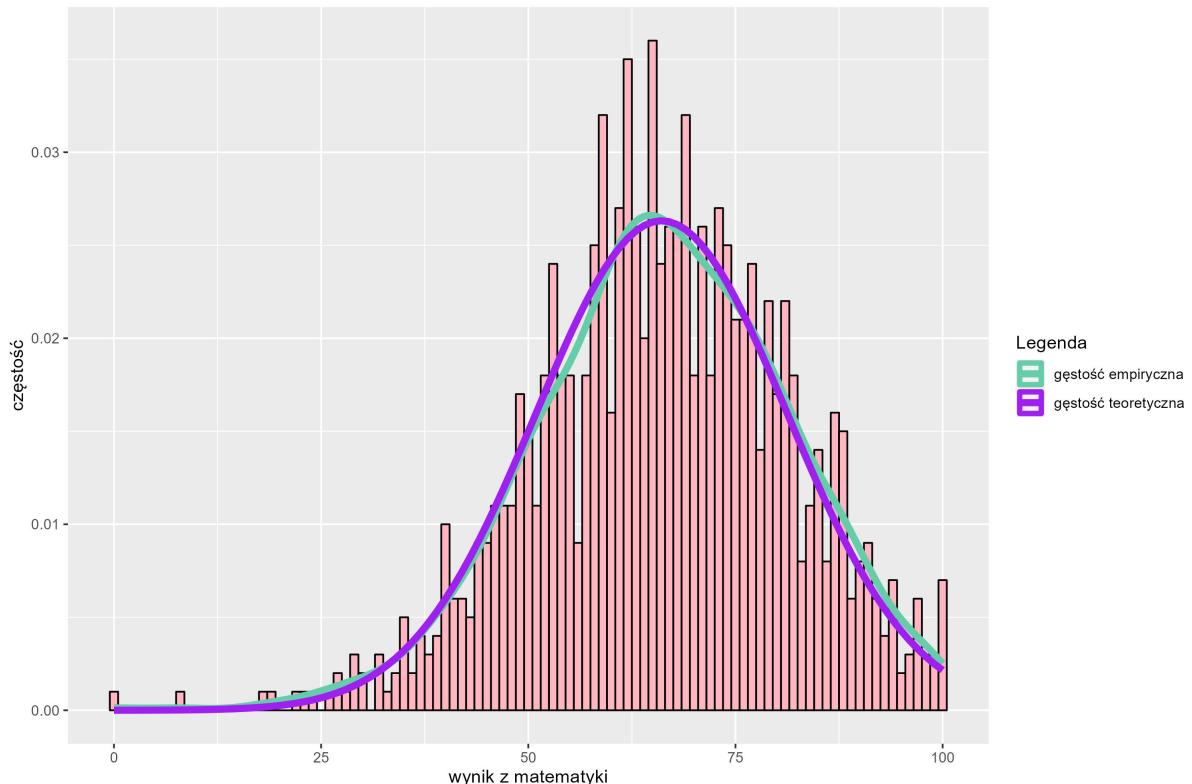
- kurtoza:  $b_2 = \frac{m_4}{s^4}$ ,

gdzie  $s$  to odchylenie standardowe, a  $m_k = \frac{1}{n} \sum_i^n (x_i - \mu)^k$  to próbkowy  $k$ -ty moment.

**Uwaga 2** Wszystkie obliczane statystyki zostały podane z zaokrągleniem do dwóch cyfr po przecinku.

### 2.1.3 Porównanie rozkładów

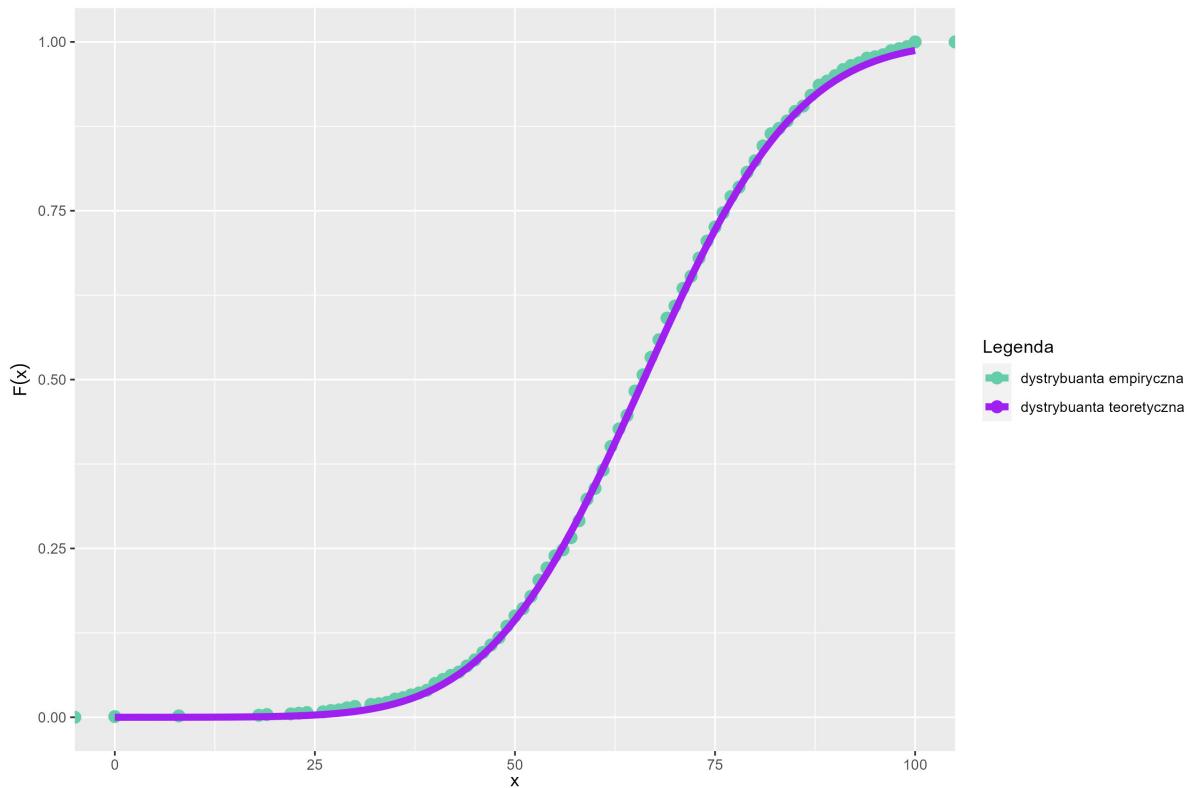
W celu porównania rozkładów zestawiono gęstości i dystrybuanty zmiennej  $X$  oraz rozkładu normalnego z odpowiednimi parametrami. W związku z tym, że średnia  $X$  wynosi 66.09 a odchylenie standardowe 15.16, ustalono teoretyczny rozkład jako  $\mathcal{N}(66.09, 15.16)$ .



**Wykres 2.3:** Porównanie gęstości empirycznej  $X$  i gęstości teoretycznej rozkładu  $\mathcal{N}(66.09, 15.16)$

---

<sup>1</sup>Wzory pochodzą z dokumentacji języka R.



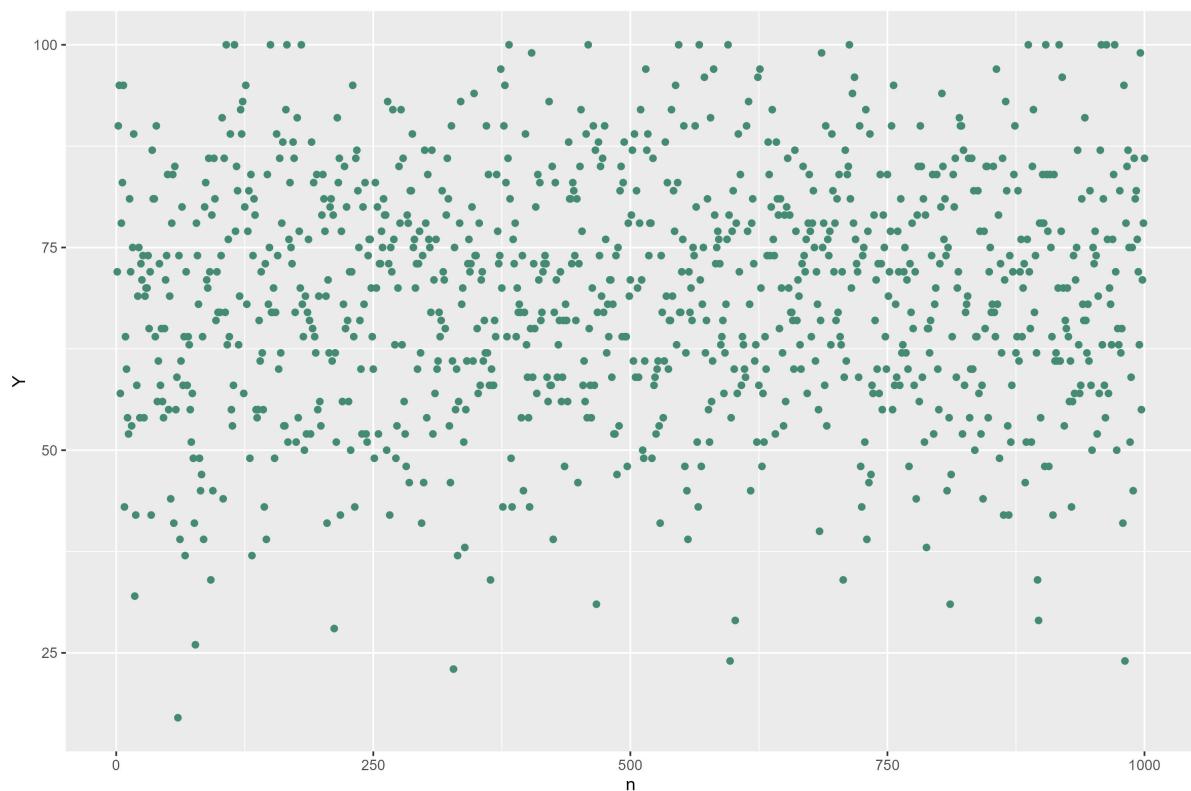
**Wykres 2.4:** Porównanie dystrybuanty empirycznej  $X$  i dystrybuanty teoretycznej rozkładu  $\mathcal{N}(66.09, 15.16)$

Nie ulega wątpliwości, że empiryczna gęstość i dystrybuanta przedstawione na wykresach 2.3 oraz 2.4 są bardzo zbliżone do ich odpowiedników teoretycznych.

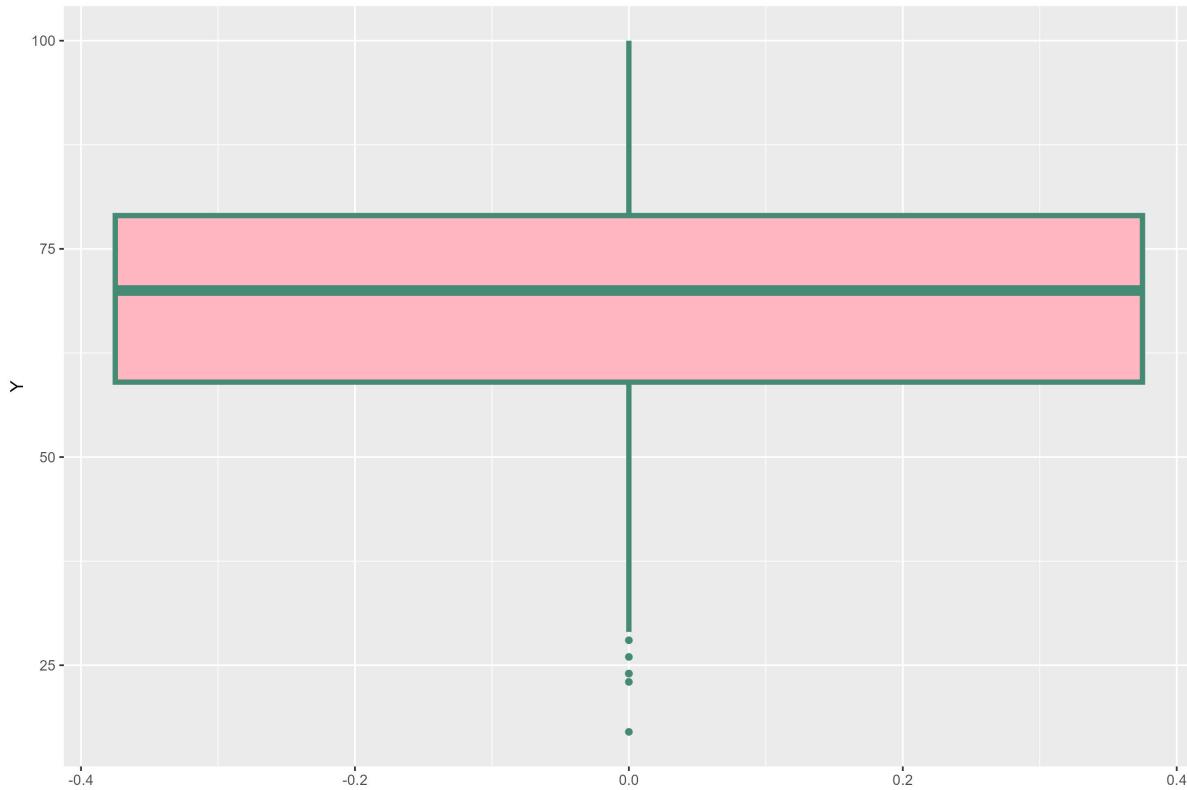
## 2.2 Analiza jednowymiarowa zmiennej objaśnianej

### 2.2.1 Wizualizacja danych

Jako zmienną objaśnianą przyjęto wynik z testu czytania (ozn.  $Y$ ). Zobrazowano rozkład badanej zmiennej oraz wykonano dla niej wykres pudełkowy.



Wykres 2.5: Rozkład  $Y$



**Wykres 2.6:** Boxplot dla  $Y$

Wykres 2.5 informuje, że wartość wyniku testu z czytania również mogła przyjąć wartości z przedziału  $[0, 100]$ . Bardzo dużo danych znajduje się powyżej 50, co może sugerować o rozkładzie lewostronnie skośnym. To przypuszczenie potwierdza wykres 2.6, gdzie pojawiło się kilka wartości odstających poniżej pudełka. Na tej podstawie można wnioskować, że w grupie są osoby, które uzyskały zdecydowanie gorszy wynik.

### 2.2.2 Podstawowe statystyki

W tej sekcji wyliczono podstawowe statystyki dla  $Y$ .

**Tabela 2.4:** Miary położenia  $Y$

Średnia	Dominanta	Minimum	Kwartyl pierwszy	Kwartyl drugi	Kwartyl trzeci	Maksimum
69.17	72.00	17.00	59.00	70.00	79.00	100.00

**Tabela 2.5:** Miary rozproszenia  $Y$

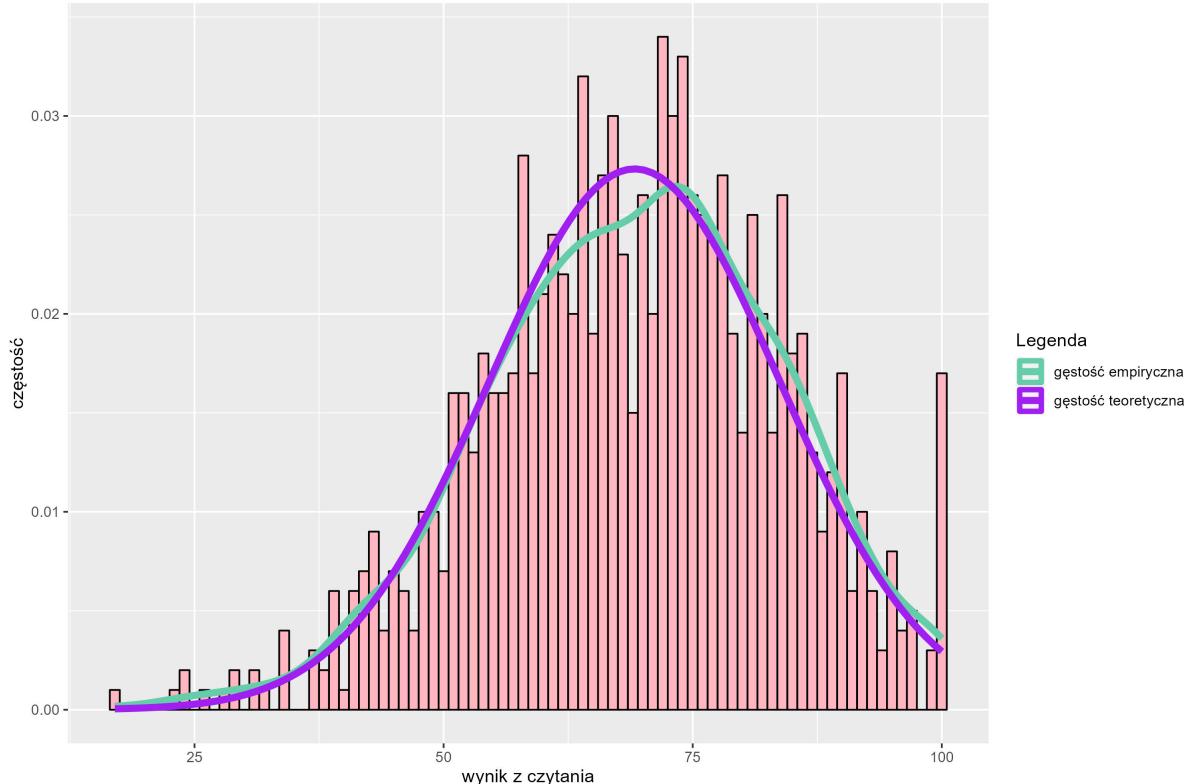
Rozstęp międzykwartylowy	Rozstęp	Wariancja	Odczylenie standardowe	Współczynnik zmienności
20.00	83.00	213.17	14.60	0.21

**Tabela 2.6:** Miary asymetrii i miary spłaszczenia  $Y$

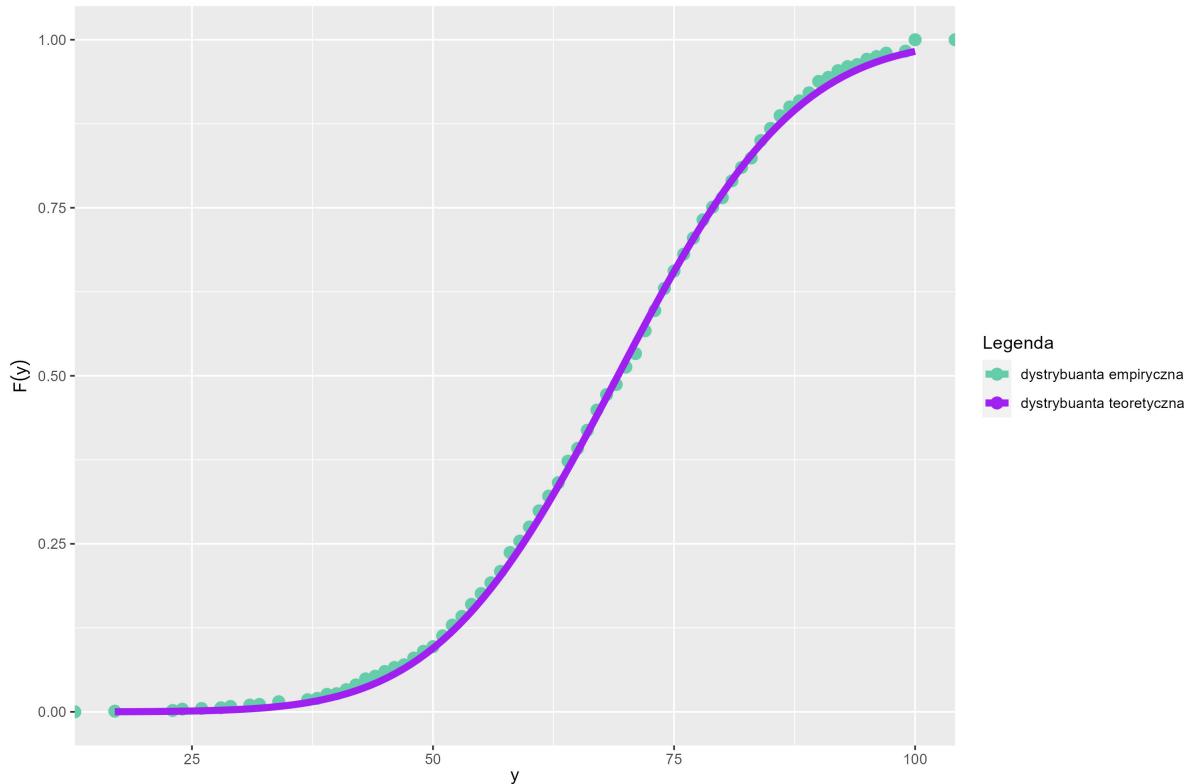
Współczynnik skośności	Kurtoza
-0.26	2.93

### 2.2.3 Porównanie rozkładów

Porównano gęstości i dystrybuanty  $Y$  oraz rozkładu normalnego z odpowiednimi parametrami. Jako że średnia  $X$  wynosi 69.17 a odchylenie standardowe 14.60, wybrano teoretyczny rozkład jako  $\mathcal{N}(69.17, 14.60)$ .



**Wykres 2.7:** Porównanie gęstości empirycznej  $Y$  i gęstości teoretycznej rozkładu  $\mathcal{N}(69.17, 14.60)$



**Wykres 2.8:** Porównanie dystrybuanty empirycznej  $Y$  i dystrybuanty teoretycznej rozkładu  $\mathcal{N}(69.17, 14.60)$

Podobnie jak w przypadku zmiennej  $X$ , empiryczne wartości gęstości i dystrybuanty zmiennej  $Y$  są bardzo zbliżone do ich odpowiedników teoretycznych.

### 2.3 Wnioski

Na podstawie wyliczonych statystyk znajdujących się w tabelach 2.1, 2.2, 2.3 oraz 2.4, 2.5, 2.6 dla zmiennej objaśniającej i objaśnianej wyciągnięto następujące wnioski.

- Średni wynik testu z matematyki wynosi 66.09, natomiast dla testu z czytania przyjmuje on wartość 69.17. Mediana, która dzieli dane na dwa podzbiory (powyżej i poniżej jej wartości) o równej liczbie obserwacji wynosi dla zmiennej  $X$  66, a dla  $Y$  70. Oznacza to, że połowa wyników w badanym zbiorze danych jest taka bądź niższa. Te wyniki mogą sugerować, że uczniowie lepiej sobie radzą z przedmiotem humanistycznego.
- Kwartyl pierwszy informuje o tym, że 25% uczniów otrzymało wynik z testu matematycznego mniejszy bądź równy 57. W przypadku testu z czytania wartość ta wyniosła 59. Co więcej, dzięki kwartylowi wiadomo, że 75% uczniów otrzymało większe wyniki niż te wcześniej wspomniane. Natomiast przy pomocy kwartyla trzeciego wiadomo, że 75% uczniów otrzymało wynik mniejszy bądź równy 77 z testu z matematyki oraz 79 z testu czytania. 25% zdających uzyskało lepsze wyniki. Potwierdza to założenia, że uczniowie nieznacznie gorzej radzą sobie z przedmiotem ścisłym.
- Z testu sprawdzającego wiedzę matematyczną osiągnięto wynik minimalny równy 0, natomiast dla testu z czytania ta wartość wynosi 17. Wynik maksymalny w obu testach wynosi 100.

- Wariancje dla badanego zbioru danych wynoszą odpowiednio dla  $X$  i  $Y$  229.92 oraz 213.17, co świadczy o tym, że rozproszenie danych jest duże. Spodziewano się takiego wyniku, gdyż badano wynik testu z dużego zakresu wartości dla aż 1000 uczniów, z których każdy z nich ma różne podejście do nauki oraz predyspozycje do osiągania wysokich wyników. W szczególności nie każdy z nich ma takie same zainteresowanie w danej dziedzinie.
- Niskie współczynniki zmienności, bo wynoszące zaledwie 0.23 i 0.21 świadczą o małej zmienności danych. Wielu studentów napisało testy podobnie do siebie i są to dobre wyniki. Wartości odstających jest mało, czyli rzadko zdarzał się uczeń, który słabo poradził sobie na egzaminie. Potwierdzają to też boxploty na rysunkach 2.2 i 2.6.
- Dla badanych danych współczynnik skośności wynosi  $-0.28$  dla  $X$  oraz  $-0.26$  dla  $Y$ , co oznacza, że otrzymano niewielką lewostronną skośność. Pozwala to wysunąć wniosek, że w większości uczniowie uzyskiwali mimo wszystko względnie dobre wyniki.
- Kurtoza dla wyniku testu z matematyki osiągnęła wartość 3.27, co oznacza, że jest więcej skrajnych wartości odstających niż w rozkładzie normalnym. W związku z tym wiele wyników testów jest znacznie oddalonych od wartości średniej. Sugeruje to występowanie skrajnych wyników w grupie, która z naturalnych przyczyn jest zróżnicowana, ponieważ opisuje uczniów. Dla testu z czytania wynik ten wynosi 2.93, co oznacza, że rozkład jest bardziej spłaszczony. Prowadzi to do obserwacji, że wielu uczniów uzyskało podobne wyniki.
- Wiedząc, że kurtoza dla rozkładu normalnego wynosi 3, zaobserwowano, że dane mogą mieć w przybliżeniu rozkład normalny.
- Wyznaczone gęstości empiryczne i dystrybuanty empiryczne pokrywają się z teoretycznymi.

### 3 Analiza zależności liniowej pomiędzy zmienną objaśniającą a zmienną objaśnianą

W tej części raportu omówiono dopasowywanie modelu regresji liniowej do danych, których różne wskaźniki wskazują na zależność liniową.

Teoretyczny model regresji liniowej prezentuje się następująco:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

oraz przyjmuje założenia:

1.  $\mathbf{E}\varepsilon_i = 0, \forall i = 1, 2, \dots, n;$
2.  $Var(\varepsilon_i) = \sigma^2, \forall i = 1, 2, \dots, n;$
3.  $\{\varepsilon_i\}_{i=1}^n$  są niezależnymi zmiennymi losowymi.

**Uwaga 3** Powyższe założenia są prawdziwe w klasycznym modelu regresji. W tym raporcie rozważono jeszcze jedno założenie (o rozkładzie  $\varepsilon_i$ )

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \forall i = 1, 2, \dots, n.$$

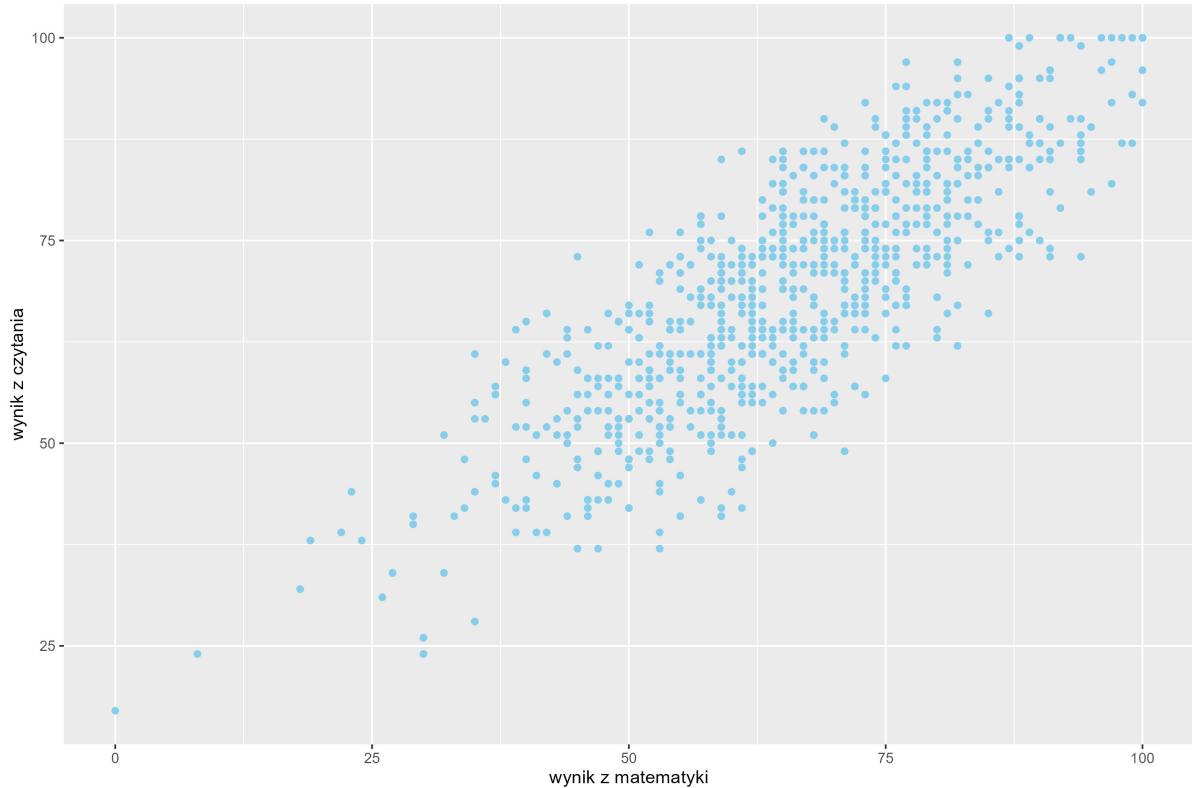
Spełnienie tego warunku nie jest jednak wymagane w klasycznym modelu regresji liniowej.

Ze zbioru par  $(x_i, y_i)$  dla  $i = 1, 2, \dots, 1000$  wybrano losowo 80% par jako dane treningowe, a pozostałe dane oznaczono jako dane testowe. Dalsza analiza została podzielona na dwie części. Do danych treningowych dopasowano model regresji liniowej, natomiast dla danych testowych sprawdzono, czy jest on dobrany poprawnie.

### 3.1 Analiza danych treningowych

#### 3.1.1 Wizualizacja danych

Zaczęto od zobrazowania zależności między zmiennymi treningowymi na wykresie rozproszenia.



**Wykres 3.1:** Rozproszenie danych treningowych

Ułożenie danych na wykresie 3.1 może wskazywać na dodatnią zależność liniową, zatem dopasowywanie modelu regresji liniowej ma sens.

#### 3.1.2 Estymacja punktowa współczynników w klasycznym modelu regresji

Korzystając z metody najmniejszych kwadratów, wyznaczono współczynniki prostej regresji<sup>2</sup> jako:

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

$$2. \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Po podstawieniu badanych danych otrzymano następujące wartości współczynników prostej regresji liniowej.

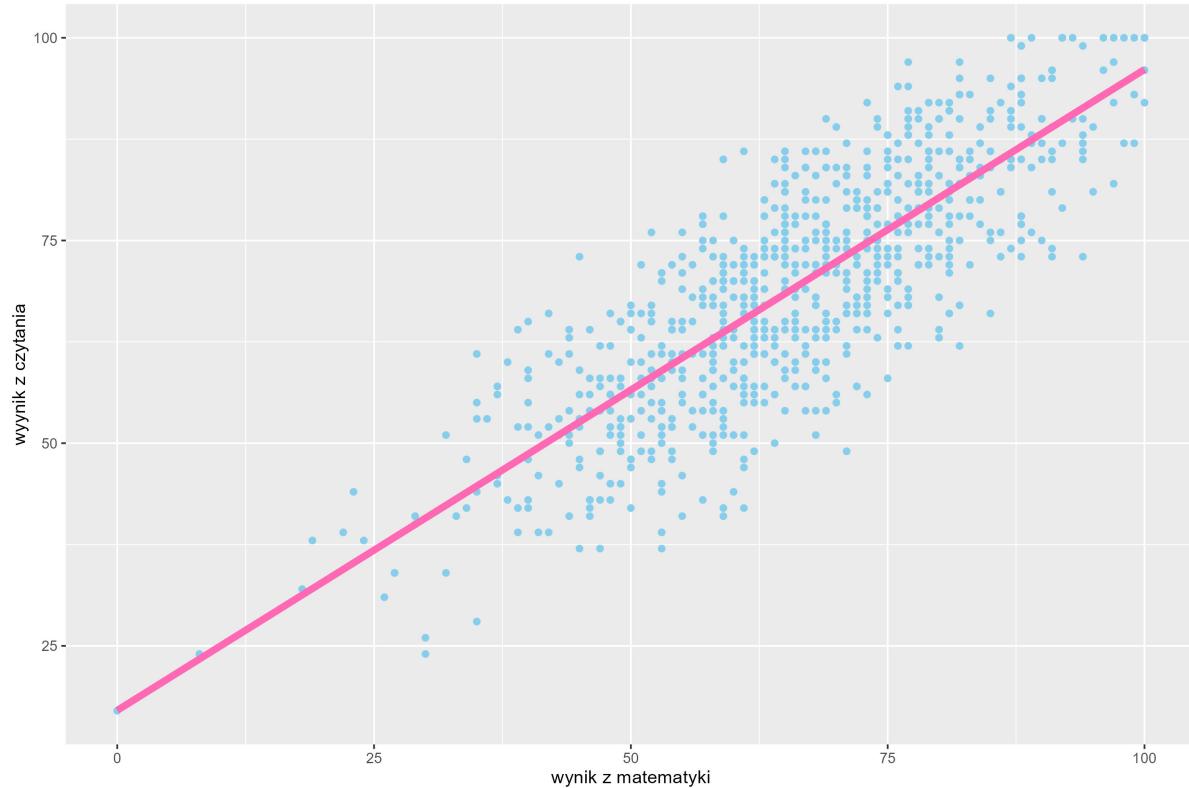
---

<sup>2</sup>Wykorzystano wzory wyprowadzone na wykładzie.

**Tabela 3.1:** Przybliżone wartości współczynników prostej regresji

$\hat{\beta}_1$	$\hat{\beta}_0$
0.7904	17.0455

Na podstawie otrzymanych współczynników dopasowano prostą do danych.



**Wykres 3.2:** Rozproszenie danych treningowych z dopasowaną prostą regresji

### 3.1.3 Estymacja przedziałowa współczynników w klasycznym modelu regresji

W celu estymacji przedziałowej wyznaczono przedziały ufności dla  $\beta_0$  i  $\beta_1$  na poziomie istotności  $\alpha = 0.02$ .

**Uwaga 4** Jako że  $\sigma^2$  tego modelu jest nieznana, wykorzystano estymator  $\sigma^2$  o następującym wzorze

$$S^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 2}, \quad \text{dla } n > 2.$$

Powyższy estymator jest nieobciążony.

Wiadomo<sup>3</sup>, że:

- $\mathcal{B}_0 = \frac{\hat{\beta}_0 - \beta_0}{S \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2},$
- $\mathcal{B}_1 = \frac{\hat{\beta}_1 - \beta_1}{S \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2},$

gdzie t oznacza rozkład t-studenta z  $n - 2$  stopniami swobody.

Dodatkowo z symetrii rozkładu t-studenta wiadomo, że jeśli  $A$  pochodzi z tego rozkładu to

$$P\left(-t_{n-2,1-\frac{\alpha}{2}} \leq A \leq t_{n-2,1-\frac{\alpha}{2}}\right) = 1 - \alpha,$$

dla  $t_{n-2,1-\frac{\alpha}{2}}$  będącego kwantylem rozkładu t-studenta z  $n - 2$  stopniami swobody, rzędu  $1 - \frac{\alpha}{2}$ .

Po podstawieniu  $\mathcal{B}_0$  (odpowiednio  $\mathcal{B}_1$ ) za  $A$  uzyskano

$$P\left(-t_{n-2,1-\frac{\alpha}{2}} \leq \mathcal{B}_0 \leq t_{n-2,1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Wykonując proste przekształcenia, otrzymano przedziały ufności, które prezentują się następująco:

- $\beta_0 : \left[ \hat{\beta}_0 - t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{\beta}_0 + t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right],$
- $\beta_1 : \left[ \hat{\beta}_1 - t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{\beta}_1 + t_{n-2,1-\frac{\alpha}{2}} S \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right],$

gdzie  $\bar{x}$  to średnia próbowa.

Z danych otrzymano

- $\beta_0 : [13.96712, 20.12397],$
- $\beta_1 : [0.74501, 0.83586].$

**Uwaga 5** Dodatkowo obliczono wartość  $S^2 = 71.65695$ .

---

<sup>3</sup>Informacja z wykładu.

### 3.2 Ocena poziomu zależności

Skorzystano z następujących wzorów na sumy kwadratów<sup>4</sup>:

- całkowita suma kwadratów:  $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ ,
- suma kwadratów błędów:  $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ ,
- suma kwadratów odchyleń regresyjnych:  $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ .

Wyznaczono również współczynnik determinacji jako

$$R^2 = \frac{SSR}{SST}.$$

Współczynnik ten określa, w jaki sposób zmienna objaśniana jest opisana przez liniową funkcję zmiennej objaśniającej<sup>5</sup>.

**Tabela 3.2:** Miary zależności

r	SST	SSR	SSE	$R^2$
0.8206	175081.28	117899.03	57182.24	0.6734

Zauważono, że dla współczynników wyznaczonych metodą najmniejszych kwadratów faktycznie  $r^2 = 0.8206^2 = 0.6734 = R^2$ . Dodatkowo na podstawie tabeli 3.2 wyznaczono wariancję  $y$  oraz średni błąd kwadratowy<sup>6</sup> jako

- $\frac{SST}{n-1} = 219.12551 = var(y)$ ,
- $\frac{SSE}{n-2} = 71.65695$ .

**Uwaga 6** Jako, że  $S^2 = \frac{SSE}{n-2}$  jest nieobciążonym estymatorem  $\sigma^2$  dla współczynników wyznaczonych metodą najmniejszych kwadratów otrzymana wartość pokrywa się z wartością  $S^2$  otrzymaną w uwadze 5.

---

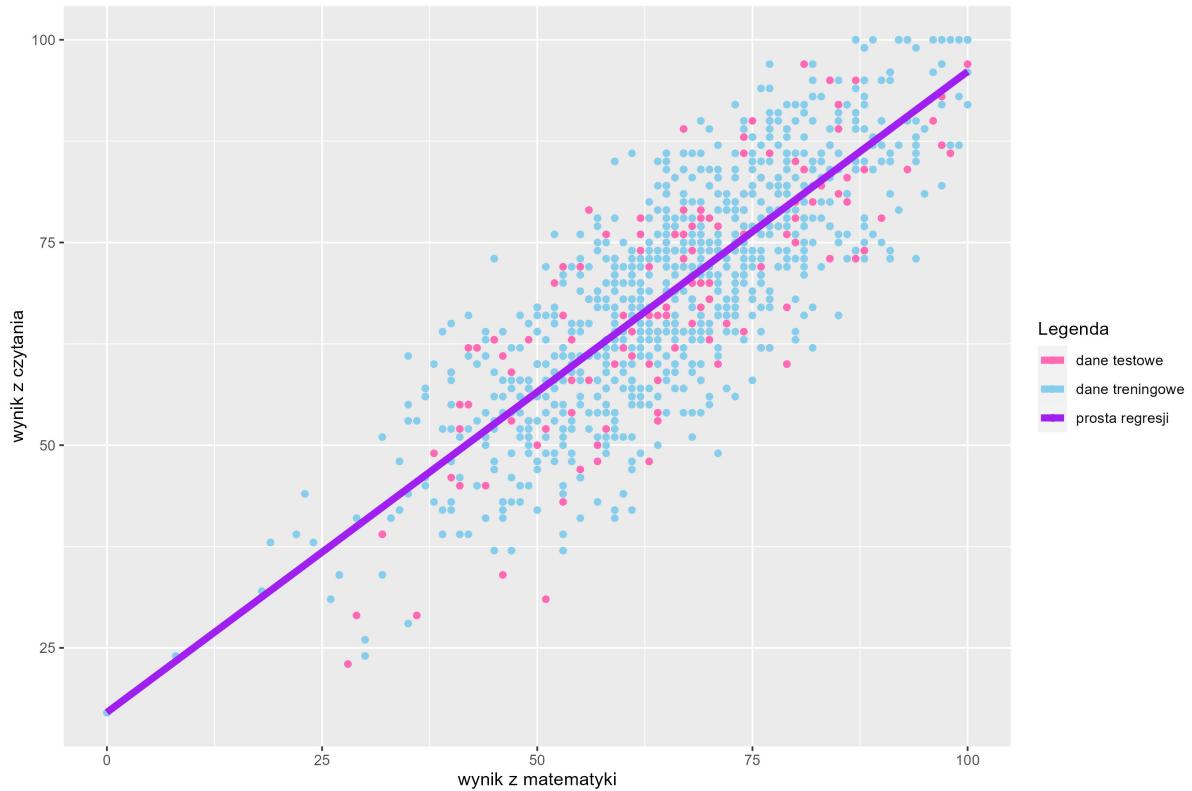
<sup>4</sup>Odpowiednio "total sum of squares", "sum of square errors", "regression sum of squares".

<sup>5</sup>Definicja z wykładu.

<sup>6</sup>Dotyczy to danych treningowych.

### 3.3 Analiza danych testowych

W tej części sprawdzono poprawność dopasowanego modelu regresji.



**Wykres 3.3:** Rozproszenie wszystkich danych z dopasowaną prostą regresji

Wyznaczono przedział ufności dla danych predykowanych ( $\hat{Y}_0$ ), przy nieznanej  $\sigma^2$ , na poziomie istotności  $\alpha = 0.02$ . W tym celu zdefiniowano estymator punktowy na podstawie współczynników wyznaczonych w tabeli 3.1

$$\hat{Y}_0 = \hat{\beta}_1 x_0 + \hat{\beta}_1.$$

Wiadomo<sup>7</sup>, że:

$$\tilde{Y}_0 = \frac{Y_0 - \hat{Y}_0}{S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \sim t_{n-2}.$$

**Uwaga 7** Warto zauważyć, że  $\bar{x}$  to średnia próbкова  $n$  danych, czyli niezawierająca  $x_0$ .

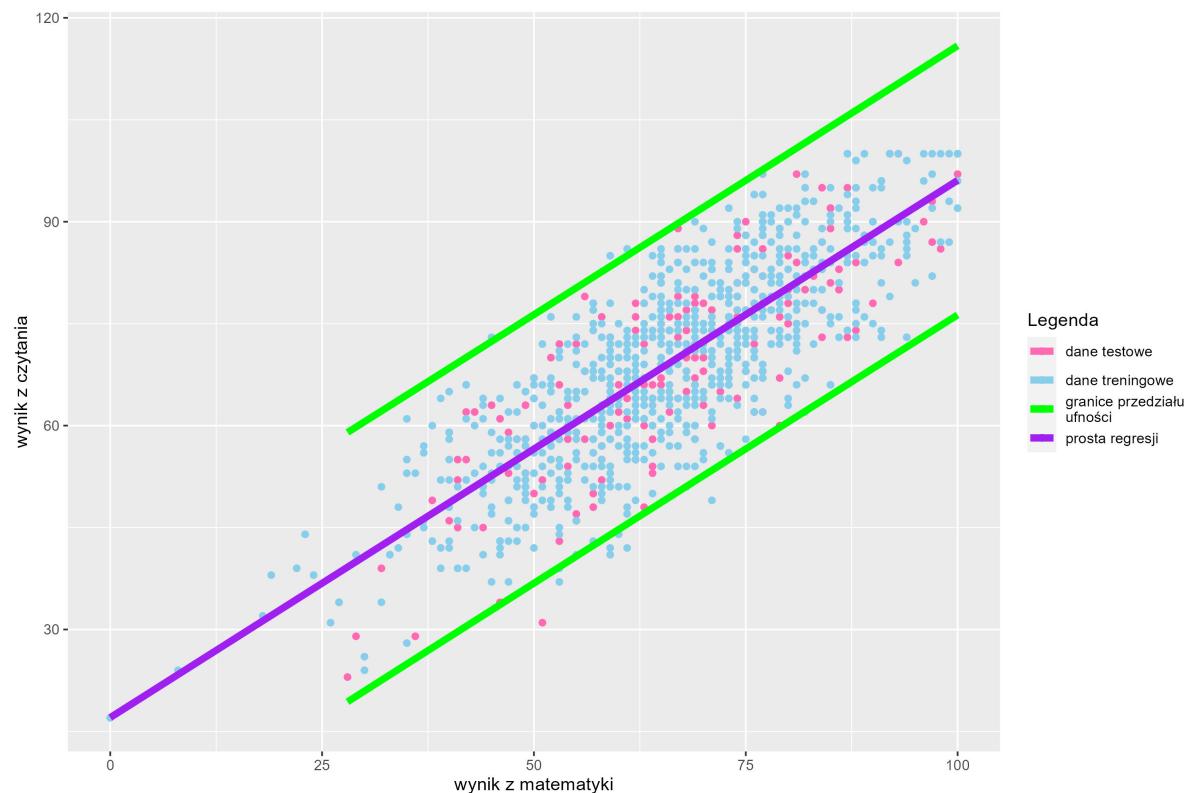
W sposób analogiczny jak dla  $\hat{\beta}_0$  i  $\hat{\beta}_1$  wyznaczono przedział ufności.

$$\left[ \hat{Y}_0 - t_{n-2,1-\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \hat{Y}_0 + t_{n-2,1-\frac{\alpha}{2}} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right]$$

---

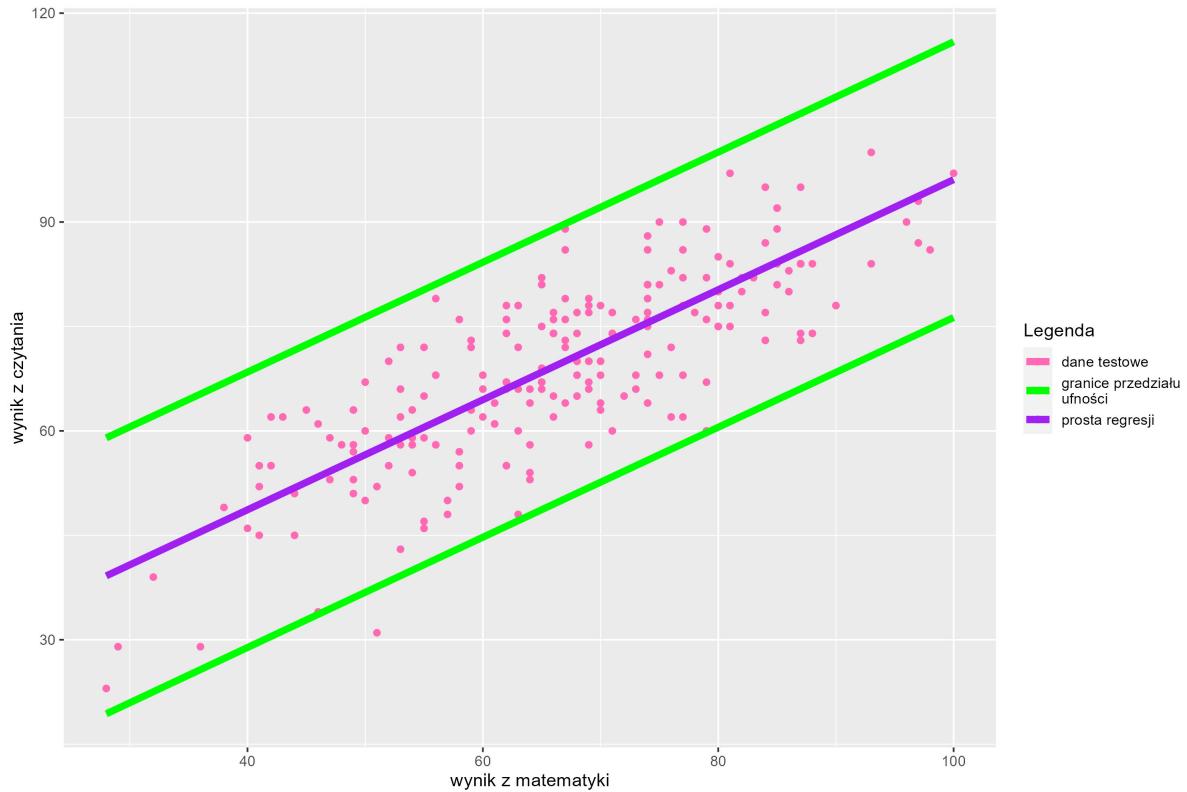
<sup>7</sup>Informacja z wykładu.

Na wykresie 3.4 zaznaczono przedziały ufności dla danych testowych.



**Wykres 3.4:** Rozproszenie danych z prostą regresji i przedziałami ufności dla danych testowych

Dla zwiększenia czytelności wykresu 3.4 przedstawiono prostą regresji i przedziały ufności na wykresie rozproszenia danych testowych 3.5.



**Wykres 3.5:** Rozproszenie danych testowych z prostą regresji i przedziałami ufności

Prawie wszystkie dane znajdują się w przedziale predykcji, zatem na tej podstawie można stwierdzić, że dopasowanie jest sensowne. Aby dokładniej określić jakość dopasowania modelu do danych testowych, obliczono wartość współczynnika determinacji  $R^2 = 0.67766$ . Wartość ta jest porównywalna z wynikiem dla danych treningowych przedstawionym w tabeli 3.2.

### 3.4 Wnioski

Na podstawie analizy danych testowych i treningowych wyciągnięto następujące wnioski:

- Prosta regresji ze współczynnikami  $(\hat{\beta}_0, \hat{\beta}_1)$  wyznaczonymi metodą najmniejszych kwadratów to  $y = 0.7904x + 17.0455$ .
- Na podstawie wykresu 3.2 stwierdzono poprawność dopasowanej prostej<sup>8</sup>.
- Wyestymowane współczynniki  $\hat{\beta}_0, \hat{\beta}_1$  zawierają się przedziałach ufności odpowiednio dla  $\beta_0, \beta_1$ , co świadczy o dobrym ich dopasowaniu.
- Wartość współczynnika korelacji Pearsona potwierdza wnioski wyciągnięte z wykresu rozproszenia danych treningowych 3.1, mówiące o silnej dodatniej korelacji zmiennych.

<sup>8</sup>Jedynie poglądową poprawność, a nie jakość dopasowania, ponieważ do dalszych wniosków potrzebne są estymacje współczynników.

- Na podstawie współczynnika determinacji stwierdzono, że jakość dopasowania jest jedynie zadowalająca, co można wytlumaczyć tym, że wyniki z matematyki i czytania nie zależą jedynie od siebie, ale również m.in. od predyspozycji badanych, czasu poświęconego na naukę, czy nawet poziomu ich skupienia w danym dniu. Co więcej, w zestawie danych znajdują się też inne zmienne, jak zjedzony posiłek, czy wykształcenie rodziców, których nie uwzględniono. Wartość  $R^2$  sugeruje dalszą analizę jakości dopasowania przyjętego modelu regresji liniowej oraz ewentualnego rozpatrzenia innych modeli.
- Wysoka wartość całkowitej sumy kwadratów (SST) może świadczyć o dużym rozproszeniu danych i tłumaczyć słabsze dopasowanie modelu do danych<sup>9</sup>.
- Większość danych testowych zawiera się w wyznaczonych przedziałach predykcji, co świadczy o dobrym dopasowaniu współczynników modelu.
- Jako że wartości współczynników determinacji zarówno dla danych treningowych, jak i testowych są w okolicach wartości 0.67, można stwierdzić, że dopasowanie jest sensowne. Podział na dane treningowe i testowe dokonano w sposób losowy, co oznacza, że procent danych wyjaśnionych przez model w każdej z tych grup powinien być podobny.

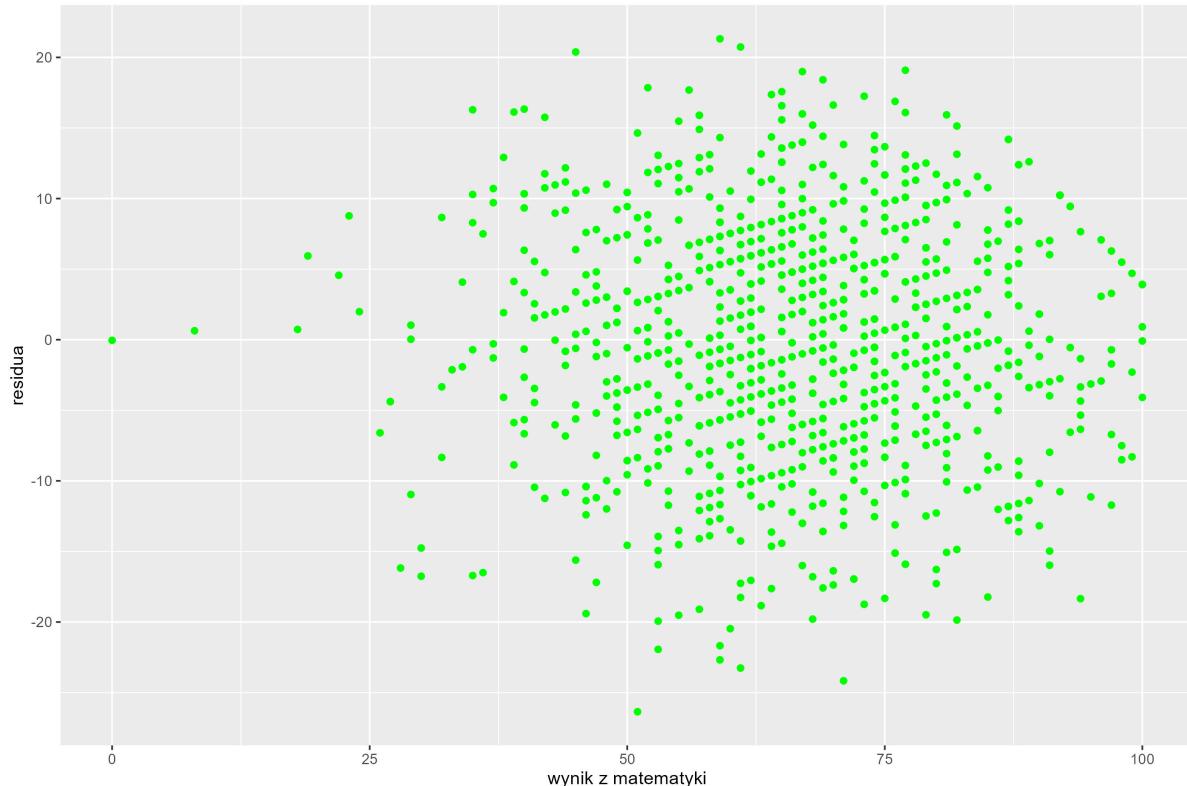
---

<sup>9</sup>Rozproszone dane będą w większej odległości od dopasowanej prostej.

## 4 Analiza residiów

### 4.1 Sprawdzenie założeń klasycznego modelu regresji

Analiza residiów w teoretycznym modelu regresji liniowej polega na sprawdzeniu, czy spełnione są założenia teoretycznego modelu regresji liniowej<sup>10</sup>. Pierwszym krokiem jest przedstawienia błędów na wykresie 4.1.



**Wykres 4.1:** Rozkład błędów

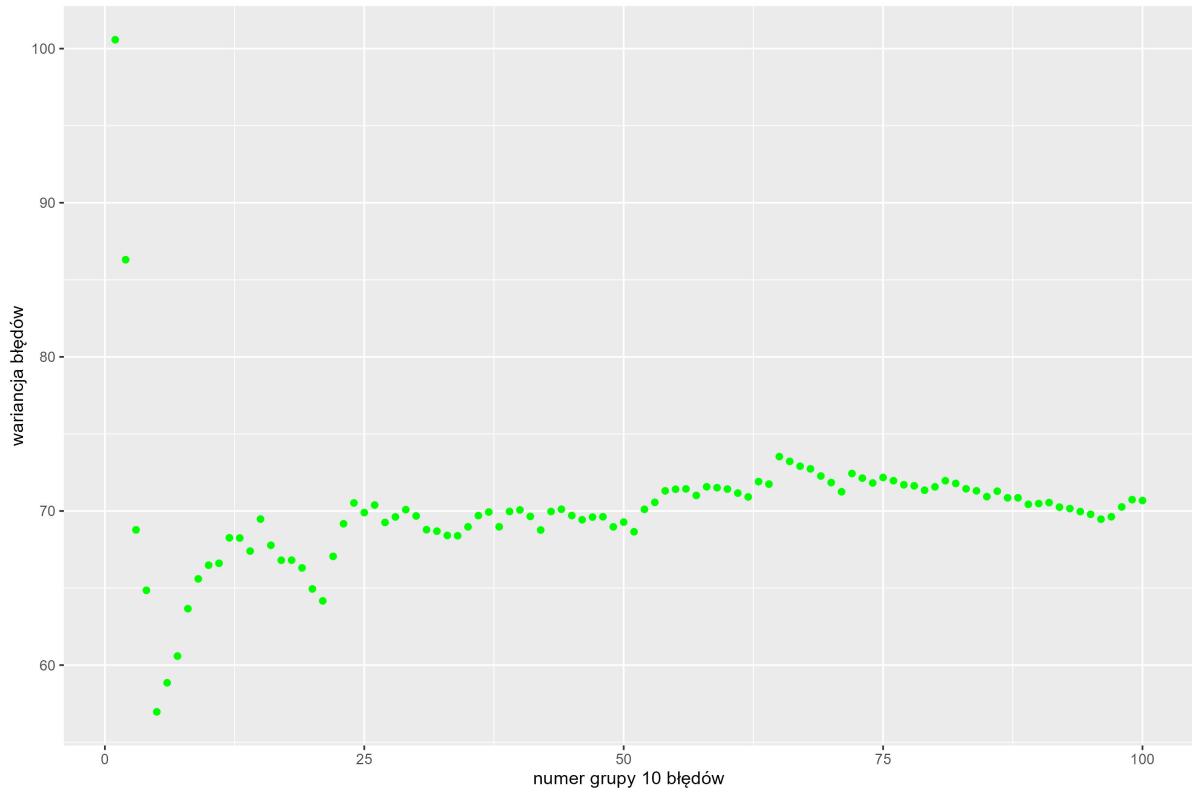
Z samego wykresu nie można stwierdzić, czy dane spełniają założenia modelu regresji, ponieważ żadne z założeń nie zostało widocznie złamane. Wyliczono zatem podstawowe wartości i umieszczone je w tabeli 4.1.

**Tabela 4.1:** Średnia i wariancja błędów

Średnia	Wariancja
-0.11542	70.68029

Na tej podstawie zauważono, że średnia jest blisko 0, co zgadza się z pierwszym założeniem modelu. Dla lepszego zobrazowania wariancji błędów pogrupowano zbiór wartości resztkowe po dziesięć, a następnie obliczono wariancję każdej z grup. Otrzymane wyniki zobrazowano na wykresie 4.2.

<sup>10</sup>Założenia te przedstawiono na początku sekcji 3.



**Wykres 4.2:** Wartości wariancji dla kolejnych grup błędów

Następnie sprawdzono niezależność wartości resztkowych. W tym celu zdefiniowano funkcję empirycznej autokowariancji.

$$\hat{\gamma}(h) = \sum_{i=1}^{n-|h|} (e_{i+|h|} - \bar{e})(e_i - \bar{e}),$$

gdzie  $h \in \mathbf{Z}$  to opóźnienie (lag) i  $-n < h < n$ .

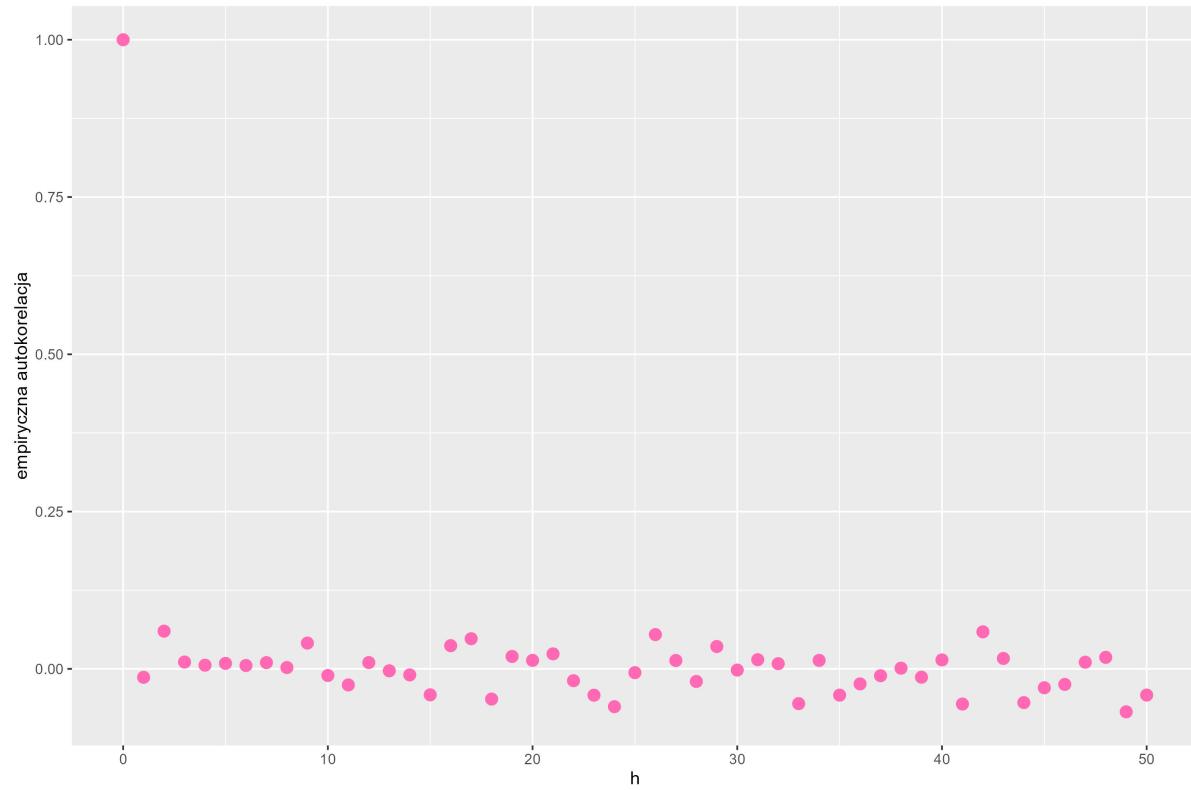
Na jej podstawie utworzono funkcję empirycznej autokorelacji jako

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

**Tabela 4.2:** Wartości funkcji empirycznej autokorelacji dla przykładowych wartości  $h$

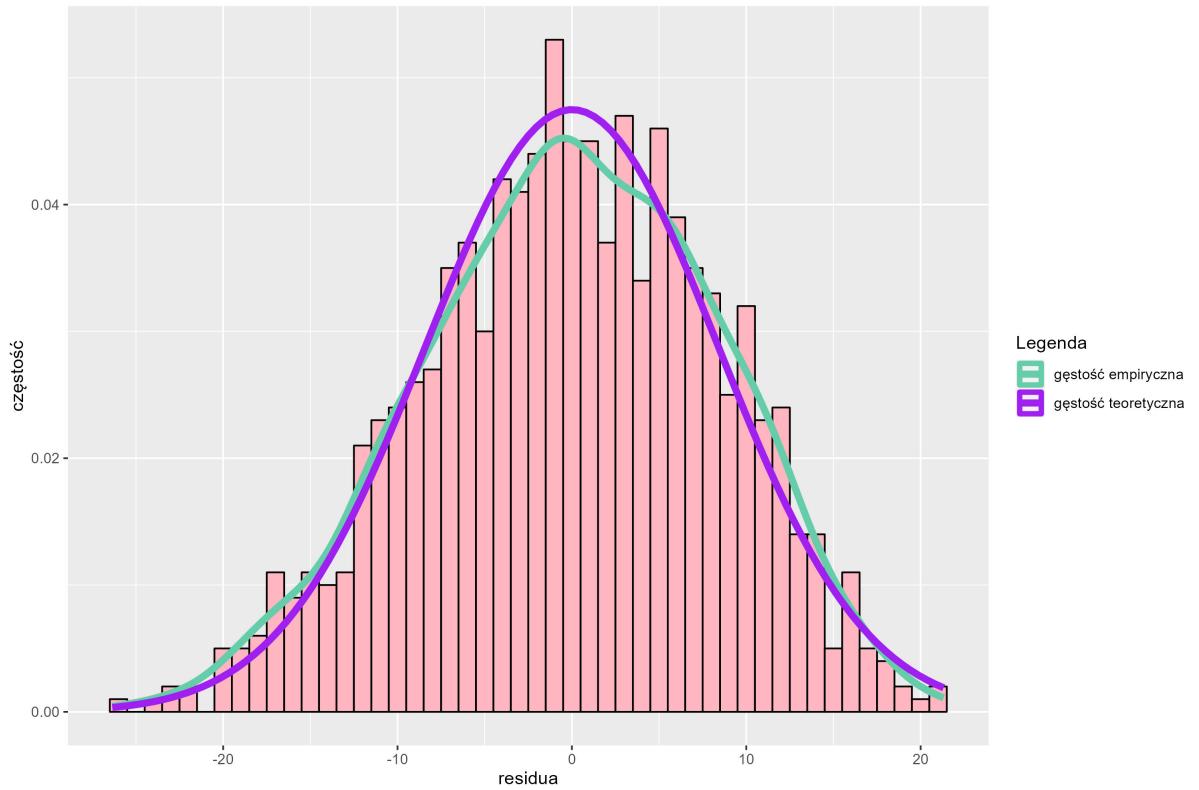
$h$	0	1	2	3	4
$\hat{\rho}(h)$	1	-0.0134	0.0600	0.0109	0.0058

Dla  $h = 0$  oczywiście otrzymano 1, jednak dla pozostałych  $h$  wartości empirycznej autokorelacji są bliskie 0. Dla dokładniejszego przedstawienia zachowania tej funkcji zamieszczono wykres 4.3 przedstawiający jej wartości dla  $h = 0, 1, 2, \dots, 50$ .



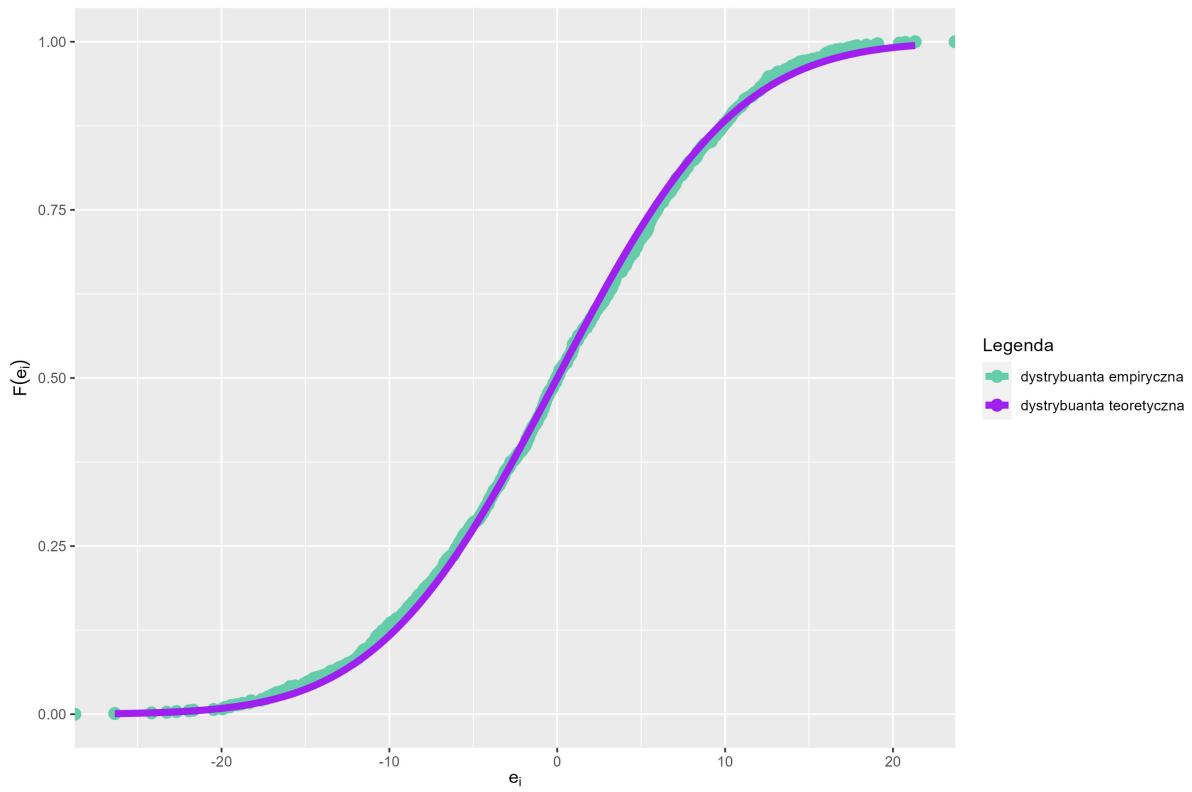
**Wykres 4.3:** Wartości funkcji autokorelacji w zależności od  $h$

Ostatecznie sprawdzono, czy wartości resztkowe faktycznie mają rozkład normalny. Z założień modelu regresji wiadomo, że  $\mu = 0$ . Z wyliczeń, które otrzymano wcześniej, przyjęto, że  $\sigma \approx \sqrt{S^2} \approx 8.4$ . Na wykresie 4.4 porównano gęstość empiryczną i teoretyczną.

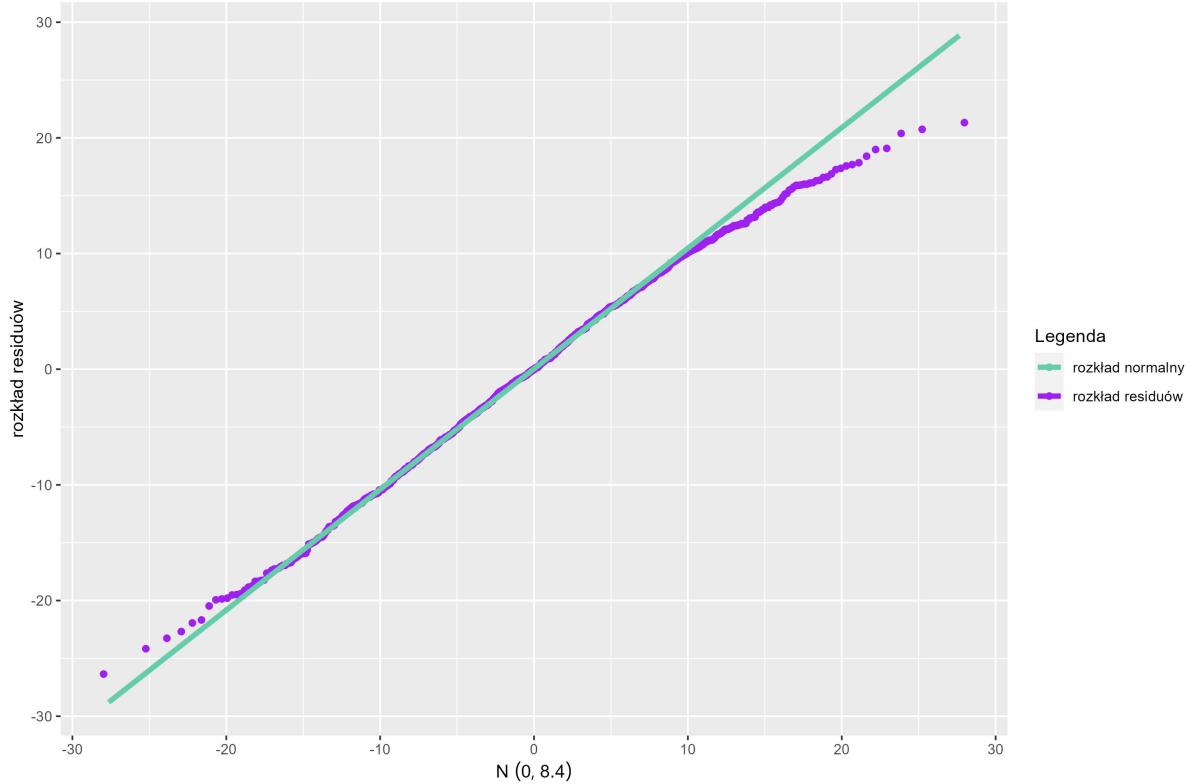


**Wykres 4.4:** Porównanie gęstości empirycznej  $e_i$  i gęstości teoretycznej rozkładu  $\mathcal{N}(0, 8.4)$

Zauważono, że gęstości się niemal pokrywają, co może wskazywać na to, że dopasowanie jest sensowne. Sprawdzono również dystrybuanty i zachowanie wartości resztkowych na wykresie kwantylowym.



**Wykres 4.5:** Porównanie dystrybuanty empirycznej  $e_i$  i dystrybuanty teoretycznej rozkładu  $\mathcal{N}(0, 8.4)$



**Wykres 4.6:** Porównanie kwantylów  $e_i$  i teoretycznych kwantylów rozkładu  $\mathcal{N}(0, 8.4)$

Podobnie porównując dystrybuanty i wykresy kwantylowe zauważono, że badane dane mają rozkład bliski teoretycznemu. Jednak z powodu niektórych odchyleń na wykresach 4.4, 4.5, 4.6 zdecydowano się na wykonanie testów normalności rozkładu. W szczególności z wykresu 4.6 zauważono, że punkty po lewej stronie wykresu znajdują się powyżej linii, a po prawej poniżej linii, co świadczy o tym, że rozkład charakteryzuje większa obecność wartości odległych od średniej, niż jest w rozkładzie normalnym, czyli o ujemnej kurtozie.

## 4.2 Testy normalności rozkładu

Aby sprawdzić, czy  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ ,  $\forall i = 1, 2, \dots, n$ . wykonano test badający normalność rozkładu. Przyjęto następujące hipotezy.

$$H_0 : \text{rozkład błędów jest rozkładem normalnym},$$

$$H_1 : \text{rozkład błędów jest różny od rozkładu normalnego}.$$

Postawienie takiej hipotezy zerowej dokonano na podstawie graficznej reprezentacji danych na wykresach 4.4, 4.5, 4.6, które mogą sugerować rozkład normalny. Następnie wyznaczoną na podstawie statystyki testowej p-wartość porównano z poziomem istotności  $\alpha = 0.02$ . Jeżeli p-wartość  $\leq \alpha$  to hipoteza zerowa jest odrzucana, na rzecz hipotezy alternatywnej, oznaczającej, że rozkład błędów nie jest normalny. Gdy p-wartość  $> \alpha$  to nie ma podstaw do odrzucenia hipotezy zerowej. Nie oznacza to jednak przyjęcia hipotezy o normalności rozkładu.

Do sprawdzenia, czy dane pochodzą z rozkładu normalnego, służą między innymi testy Kołmogorowa-Smirnova, Shapiro-Wilka oraz Jarque-Bera.

**Tabela 4.3:** Testy normalności rozkładu

Test	Statystyka	P-wartość
Kołmogorowa-Smirnova	0.020961	0.7718
Shapiro-Wilka	0.9956	0.005716
Jarque-Bera	9.7454	0.007653

Na podstawie tabeli 4.3 oraz przy przyjęciu poziomu istnotności  $\alpha = 0.02$  stwierdzono następujące fakty.

1. W przypadku testu Kołmogorowa-Smirnova, ponieważ p-wartość jest większa niż poziom istotności, nie ma podstaw do odrzucenia hipotezy, że błędy w modelu są z rozkładu normalnego.
2. W teście Shapiro-Wilka oraz teście Jarque-Bera otrzymano p-wartości mniejsze od 0.02, zatem odrzucono hipotezę zerową.

### 4.3 Wnioski

Na podstawie analizy wartości resztkowych sformułowano następujące wnioski:

- Średnia wartość błędów z pełnego zbioru danych<sup>11</sup> jest bliska 0, zatem pierwsze z założeń jest zachowane.
- Na wykresie 4.2 zauważono, że wariancja jest stała, ponieważ wartości wariancji dla przyjętych grup utrzymują się na w przybliżeniu stałym poziomie.
- Wartości odstające na wykresie 4.2 mogą zostać wyjaśnione małą licznoscią grup. Przy niewielkiej liczbie danych dokładność się zmniejsza, ponieważ większy wpływ na wynik mają wartości odstające (anomalie). Niemniej jednak wykres dobrze sprawuje swoją rolę, jako przybliżony obraz stałej wariancji.
- Na podstawie tabeli 4.2 oraz wykresu 4.3 stwierdzono, że zmienne są niezależne, ponieważ wartości funkcji empirycznej autokorelacji dla  $h > 0$  są bliskie 0.
- Na podstawie powyższych wniosków przyjęto, że model regresji liniowej został poprawnie dopasowany.
- Na podstawie wykresów 4.4, 4.5, 4.6 zauważono, że rozkład residiów jest bliski rozkładowi normalnemu z parametrami  $\mu = 0, \sigma = 8.4$ . Po przeprowadzeniu testu Kołmogorowa-Smirnowa otrzymano, że nie ma podstaw do odrzucenia hipotezy zerowej. Analizując testy, które są bardziej odporne na dane wrażliwe, otrzymano, że dane nie pochodzą z rozkładu normalnego. Rozbieżność w tych testach może jedynie sugerować zbliżenie wartości analizowanych residiów do rozkładu normalnego.

## 5 Podsumowanie

Wykorzystując poznane metody dotyczące analizy zależności liniowej do danych opisujących wyniki uczniów z testu z matematyki oraz z czytania, uzyskano dopasowanie prostej regresji liniowej. Nie jest ono całkowicie dokładne. Wynika to z faktu, że przedstawiony w niniejszej pracy problem nie wyczerpuje w całości swoich możliwości. Pracę można byłoby rozszerzyć poprzez wykorzystanie modelu bardziej złożonego niż ten, który został przedstawiony. Uwzględniałby on więcej zmiennych, które znajdowały się w głównym zbiorze danych. Warto też zauważyć, że na osiągane rezultaty wpływają także takie zmienne jak czas poświęcony na naukę, czy nawet to, jakie samopoczucie miał uczniów w dniu egzaminu. To wszystko powoduje, że znalezienie dokładnego rozwiązania jest utrudnione.

---

<sup>11</sup>Zarówno treningowych, jak i testowych.