Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki

Metody statystyczne

Sprawozdanie z projektu

Autorzy	Marcin Mitręga Jakub Dusza Jakub Mieszczak Bartłomiej Gordon Mikołaj Gajos Paweł Kupczak
Grupa	3
Kierunek	Informatyka
Prowadząca	dr inż. Alina Momot

Spis treści

1 Treść projektu

W badaniu statystycznym zebrano dane dotyczące wzrostu i wagi dzieci z pewnej szkoły podstawowej w Polsce. Zarejestrowane dane przedstawiają 100 trójek: wzrost (wartość podana w cm), waga (wartość podana w kg), płeć (1 - chłopiec, 0 - dziewczynka).

Wzrost	Waga	Płeć
133	28	1
139	34	0
137	30	0
141	39	1
139	33	0
128	25	0
121	29	0
116	23	1
138	36	0
138	31	1
154	42	1
124	30	1
135	36	0
120	29	1
121	21	1
125	25	1

Rysunek 1: Fragment pliku z danymi

2 Zadania

2.1 Zadanie 1

Dokonać analizy wzrostu i wagi dzieci, na podstawie wyznaczonych wartości przeciętnych, kwartyli oraz odchylenia standardowego.

Kod programu:

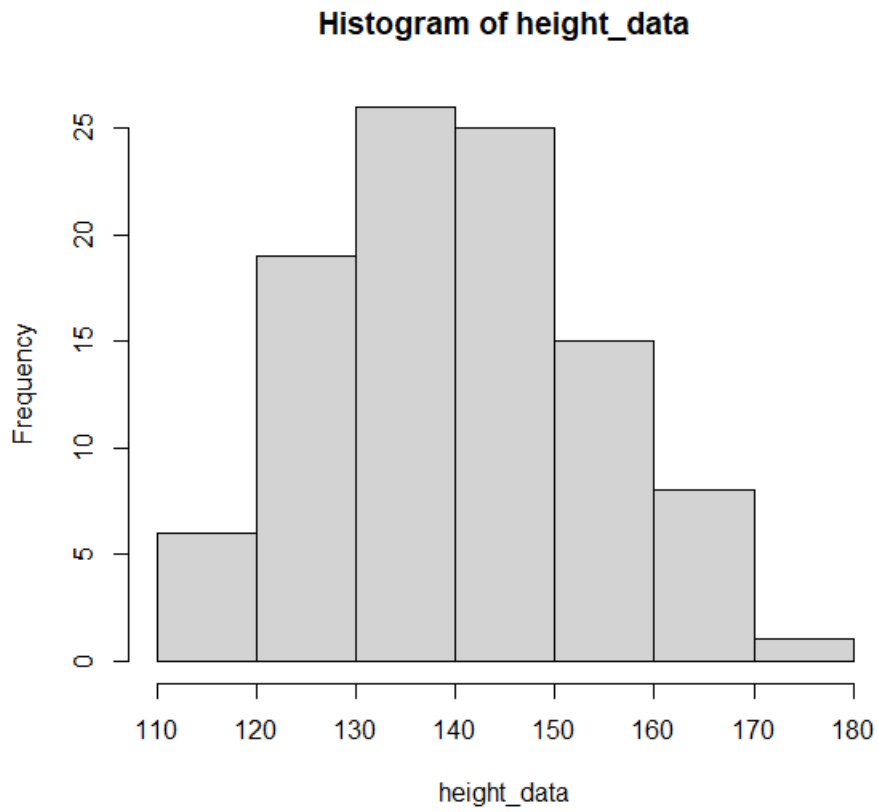
```
# WCZYTYWANIE DANYCH
data <- read.delim("dane.txt")

height_data <- data[ , c("wzrost")]
gender_data <- data[ , c("plec")]
weight_data <- data[ , c("waga")]


# ZADANIE 1
mean(height_data)
sd(height_data)
quantile(height_data)
hist(height_data)
hist(weight_data)
```

Na początku wczytano dane do zmiennej, po czym rozdzielono je na dane dotyczące wzrostu, płci oraz wagi. Następnie użyto wbudowanych funkcji języka R do wyliczenia średniego wzrostu, odchylenia standardowego wzrostu oraz kwantyli wzrostu. Potem użyto funkcji `hist()` do wyświetlenia histogramów wzrostu i wagi, które poddano analizie.

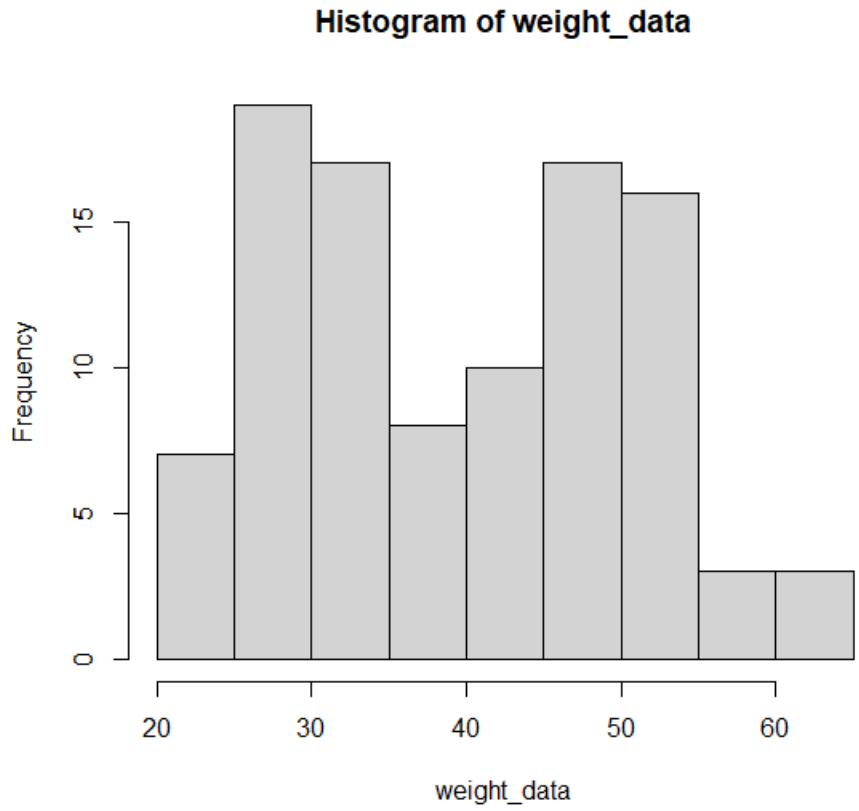
Średnia, odchylenie i kwantyle:

```
> mean(height_data)
[1] 140.6
> sd(height_data)
[1] 14.07914
> quantile(height_data)
    0%    25%    50%    75%   100%
113.00 130.50 139.50 149.25 174.00
```

Histogramy wzrostu i wagi:



Rysunek 2: Histogram wzrostu



Rysunek 3: Histogram wagi

Wnioski do zadania:
Z histogramu wzrostu można odczytać, że jego rozkład jest bliski normalnemu, natomiast z histogramu wagi nie można wyciągnąć tego samego wniosku - wynika to prawdopodobnie z tego, że dane na histogramie nie zostały rozdzielone na płci.
Ponadto, z analizy kwantyli i średniej można zauważyć, że średnia i mediana wzrostu są bardzo podobne, co wynika z normalności rozkładu.

2.2 Zadanie 2

Wyznaczyć modele regresji liniowej przedstawiające zależności wzrostu od wagi:

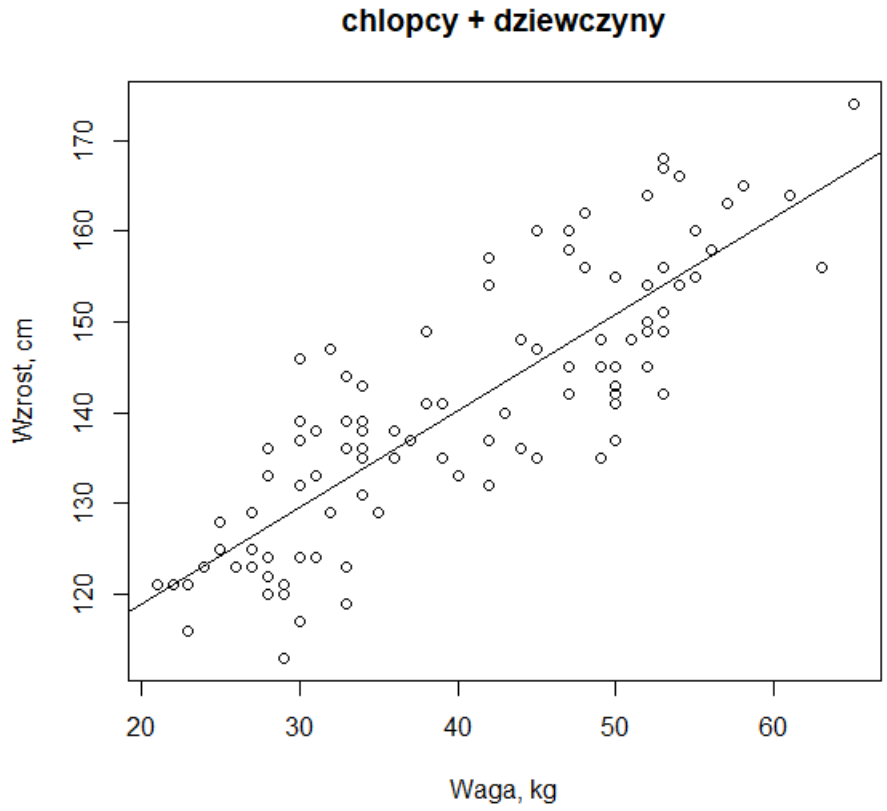
- dziewczynek,
- chłopców,
- razem (dziewczynek i chłopców).

Kod programu:

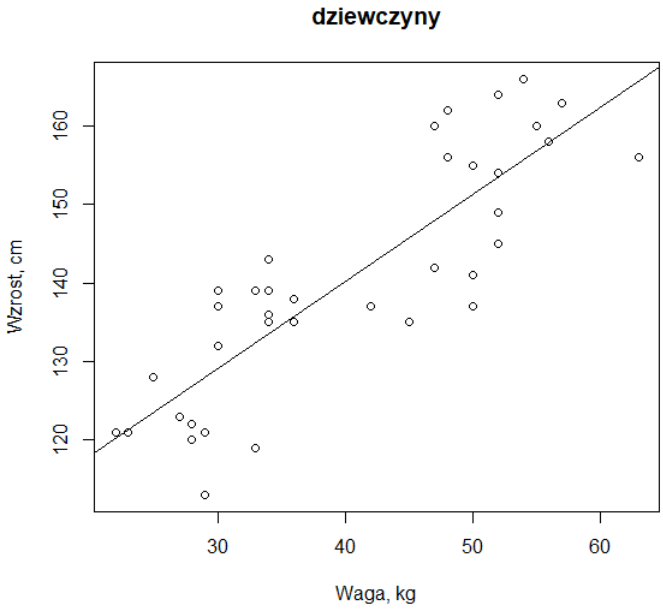
```
# ZADANIE 2
chlopcy  <- subset(data, plec==1)
dziewczyny <- subset(data, plec==0)

model_reglinp <- function(X) {
  return( lm(X$wzrost ~ X$waga) )
}

reglinp = function(X, name){
  fit <- model_reglinp(X)

  summary(fit)
  plot(X$waga, X$wzrost, main = name, xlab = "Waga, kg",
  ylab = "Wzrost, cm")
  abline(fit)

  return(fit)
}

boys_model     <- reglinp(chlopcy,    "chlopcy")
girls_model    <- reglinp(dziewczyny, "dziewczyny")
everyone_model <- reglinp(data,       "chlopcy + dziewczyny")
```

W kodzie źródłowym podzielono zestaw danych na osobne płcie. Zdefiniowaliśmy funkcję `model_reglinp()`, która zwraca model regresji liniowej dla danego zbioru danych. Następna funkcja - `reglinp()` - przedstawia model regresji liniowej dla zadanego zbioru danych oraz go zwraca. Na końcu użyto wcześniej wspomnianych funkcji do zdobycia modeli regresji liniowych dla każdej podgrupy wymienionej w treści zadania.

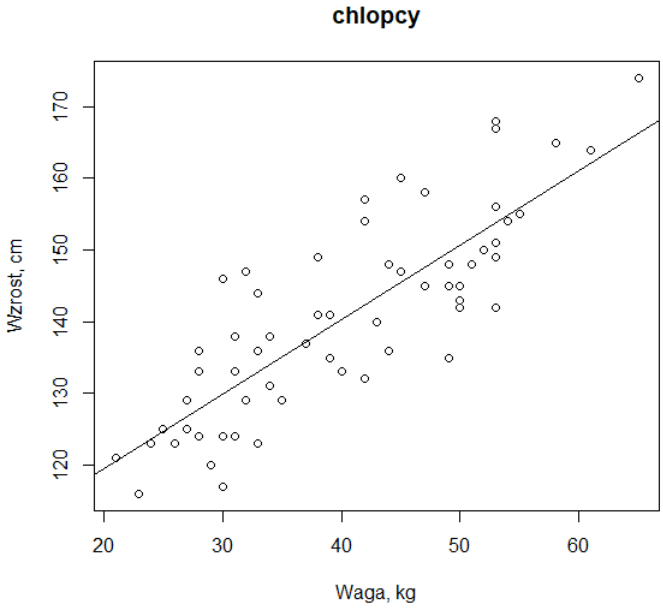
Modele regresji liniowych:



Rysunek 4: Model regresji liniowej dla dziewczynek + chłopców



Rysunek 5: Model regresji liniowej dla dziewczynek



Rysunek 6: Model regresji liniowej dla chłopców

Wnioski do zadania:
Na wykresach można zauważyć, że dla wszystkich podgrup nie występują błędy grube, co oznacza że zebrane pomiary mieszczą się w oczekiwanych przedziałach.

2.3 Zadanie 3

Wyznaczyć standardowe błędy szacunku dla estymatorów punktowych parametrów modelu i dokonać oceny dopasowania otrzymanych modeli.

Kod programu:

```
# Zadanie 3
slope_coefficients <- c(
        boys_model[["coefficients"]][["X$waga"]],
        girls_model[["coefficients"]][["X$waga"]],
        everyone_model[["coefficients"]][["X$waga"]]
)

slope_coefficients

const_term_coefficients <- c(
        boys_model[["coefficients"]][["(Intercept)"]],
        girls_model[["coefficients"]][["(Intercept)"]],
        everyone_model[["coefficients"]][["(Intercept)"]]
)

const_term_coefficients

std_err <- function(x) sd(x) / sqrt(length(x))

std_err(slope_coefficients)
std_err(const_term_coefficients)
```

Rozpoczęto od wyłuskania parametrów z modeli regresji liniowej. Następnie zdefiniowano funkcję do obliczania błędu standardowego, po czym wypisano błędy dla obu zbiorów parametrów modeli regresji liniowej.

Wyliczone błędy standardowe:

```
> std_err(slope_coefficients)
[1] 0.0200958
> std_err(const_term_coefficients)
[1] 0.8344543
```

Wnioski do zadania:
Z wyliczonych błędów standardowych można wywnioskować, że przyrost współczynnika kierunkowego dla wszystkich modeli był bardzo podobny, natomiast wartość wyrazu wolnego dużo bardziej różniła się na przestrzeni badanych modeli, co prawdopodobnie wynika z tego, że rozwój wzrostu i wagi różni się w zależności od płci.

2.4 Zadanie 4

Przedstawić graficznie dane dotyczące zależności wzrostu od wagi:

- dziewczynek,
- chłopców,
- razem (dziewczynek + chłopców).

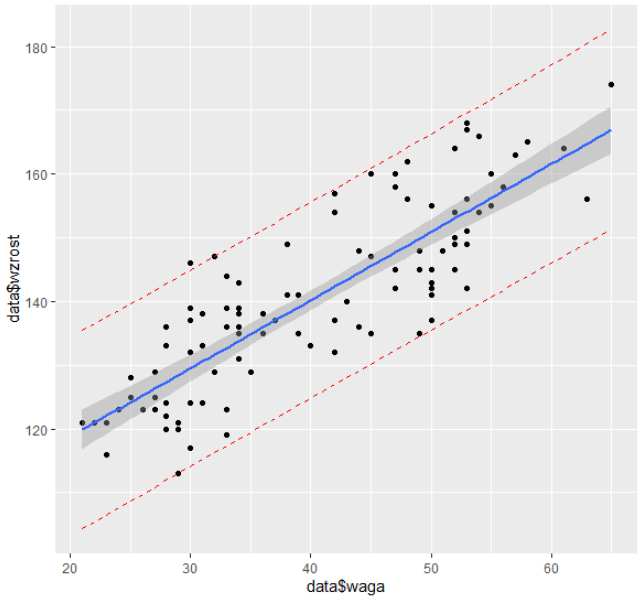
Na wykresach prezentujących zbiór par danych wejściowych dodać proste regresji oraz krzywe prezentujące pasma predykcji na poziomie ufności 0.95.

Kod programu:

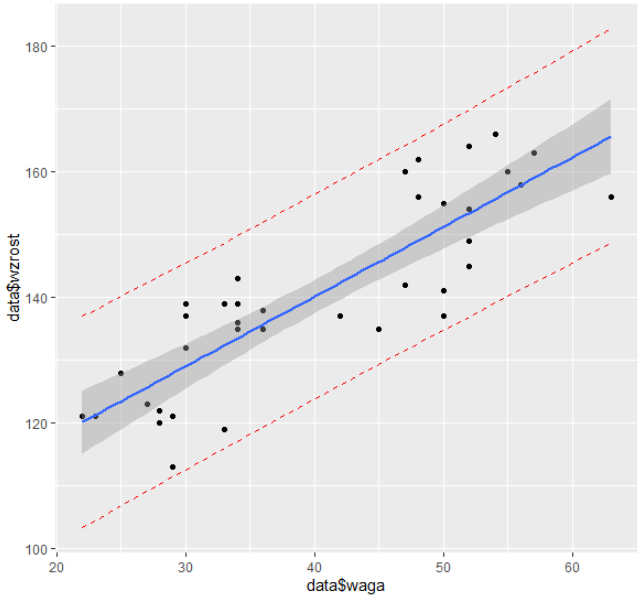
```
# Zadanie 4
prediction <- function(data) {
  model <- model_reglinp(data)

  pred.int <- predict(model, interval = "prediction")
  mydata   <- cbind(data, pred.int)

  p <- ggplot(mydata, aes(data$waga, data$wzrost)) +
       geom_point() +
       stat_smooth(method = lm)

  p + geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
      geom_line(aes(y = upr), color = "red", linetype = "dashed")
}

prediction(chlopcy)
prediction(dziewczyny)
prediction(data)
```

Najpierw zdefiniowano funkcję prezentującą wykres danych wraz z prostymi regresji oraz krzywych pasma predykcji, po czym ją wywołano dla podzbiorów podanych w treści zadania.

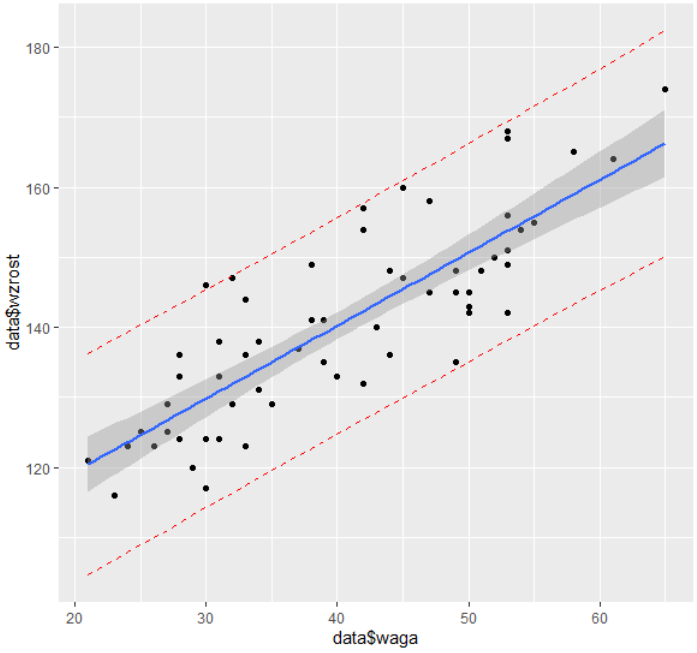
Wykresy z prostymi regresji i krzywymi pasma predykcji:



Rysunek 7: Wykres dla dziewczynek + chłopców



Rysunek 8: Wykres dla dziewczynek



Rysunek 9: Wykres dla chłopców

Wnioski do zadania:
Z wykresów można wyczytać, że dane mieszczą się w pasmach predykcji, co jest kolejnym potwierdzeniem, że wśród badanej próbki nie znajdują się błędy grube.

2.5 Zadanie 5

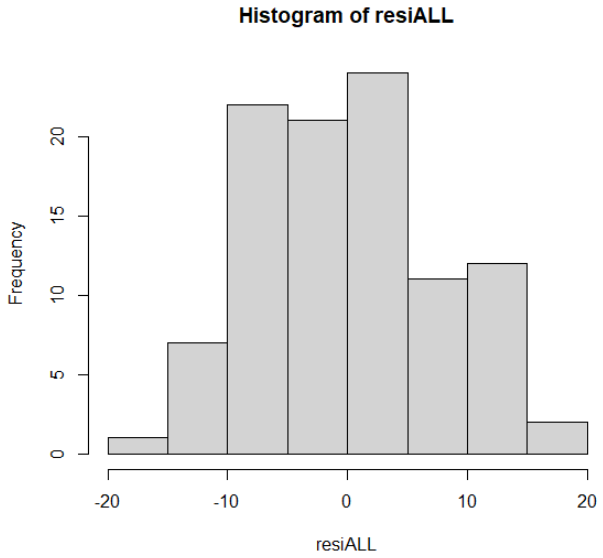
Opracować histogramy rezyduów. Sprawdzić, czy rezydua mają rozkład normalny.

Kod programu:

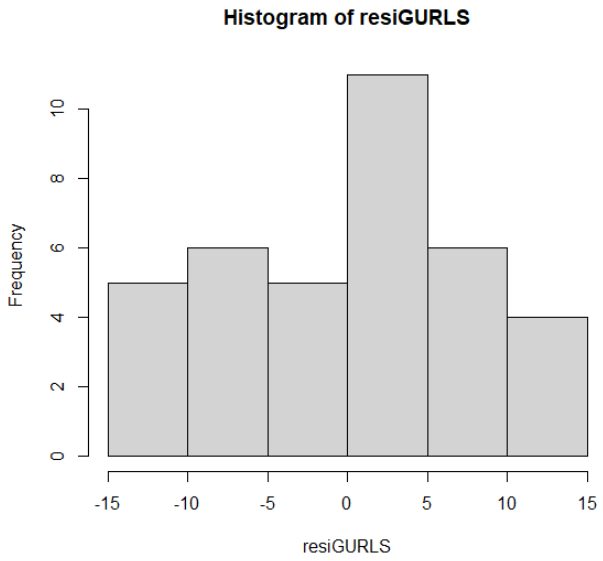
```
# Zadanie 5
resiDUDES <- residuals(boys_model)
resiGURLS <- residuals(girls_model)
resiALL   <- residuals(everyone_model)

hist(resiDUDES)
hist(resiGURLS)
hist(resiALL)

shapiro.test(resiDUDES)
shapiro.test(resiGURLS)
shapiro.test(resiALL)
```

Na początku użyto wbudowanej funkcji `residuals()` w celu wyliczenia rezyduów dla podgrupy chłopców, dziewczynek, oraz dla obu na raz, po czym wyświetlono ich histogramy. Na sam koniec użyto funkcji `shapiro.test()`, aby sprawdzić, czy konkretne rezydua mają rozkład normalny.

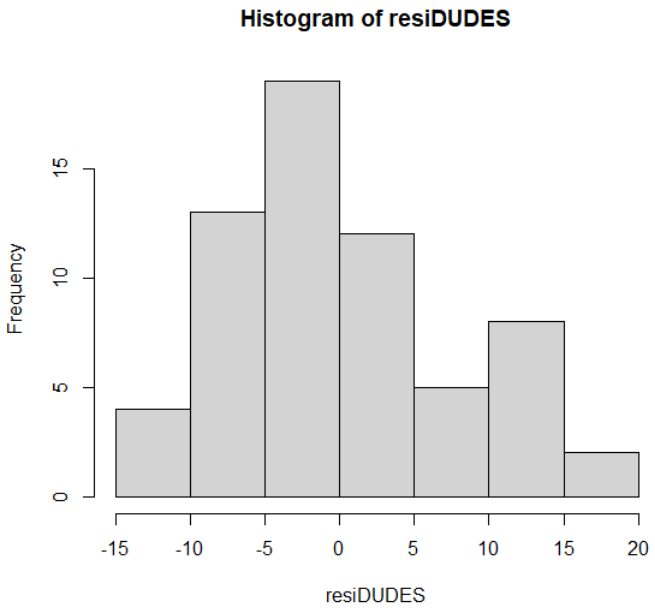
Histogramy rezyduów:



Rysunek 10: Rezydua dziewczynek + chłopców



Rysunek 11: Rezydua dziewczynek



Rysunek 12: Rezydua chłopców

Wyniki testów Shapiro-Wilk:

```
Shapiro-Wilk normality test

data:  resiDUDES
W = 0.9609, p-value = 0.04325

> shapiro.test(resiGURLS)

    Shapiro-Wilk normality test

data:  resiGURLS
W = 0.96607, p-value = 0.3122

> shapiro.test(resiALL)

    Shapiro-Wilk normality test

data:  resiALL
W = 0.98198, p-value = 0.1892
```

Wnioski do zadania:
Z wyliczonych wartości `p-value` można określić, czy dany zestaw rezyduów jest normalny (`p-value > 0.05` - rozkład jest normalny). Z uzyskanych wyników można wywnioskować, że:

- rezydua chłopców nie mają rozkładu normalnego,
- rezydua dziewczynek mają rozkład normalny,
- rezydua dziewczynek + chłopców mają rozkład normalny.