Projekt z języka R

Radosław Kluczewski

26.01.2022

Zadanie 1

Do wykonania pierwszego zadania zostały użyte dane ze strony AstroStatistics.

a) W celu uzyskania podstawowych statystyk, które opisują dane zostało wykonane nastepujące polecenie:

```
summary(asteroids)
```

```
##    Asteroid              Dens            Err
##  Length:26          Min.   :0.800   Min.   :0.0300
##  Class :character   1st Qu.:1.343   1st Qu.:0.1350
##  Mode  :character   Median :2.060   Median :0.3000
##                     Mean   :2.182   Mean   :0.6073
##                     3rd Qu.:2.700   3rd Qu.:0.7500
##                     Max.   :4.900   Max.   :3.9000
```

Zostało również obliczone odchylenie standardowe σ, które zostanie wykorzystane w następującym wzorze na błąd standardowy mediany:

$$SE_{me} = \frac{1,253 \cdot \sigma}{\sqrt{N}}. \tag{1}$$

Poniżej zostały zaprezentowane wyniki po podstawieniu do powyższego wzoru, gdzie dla kolumny dens oraz err:

```
SE1
```

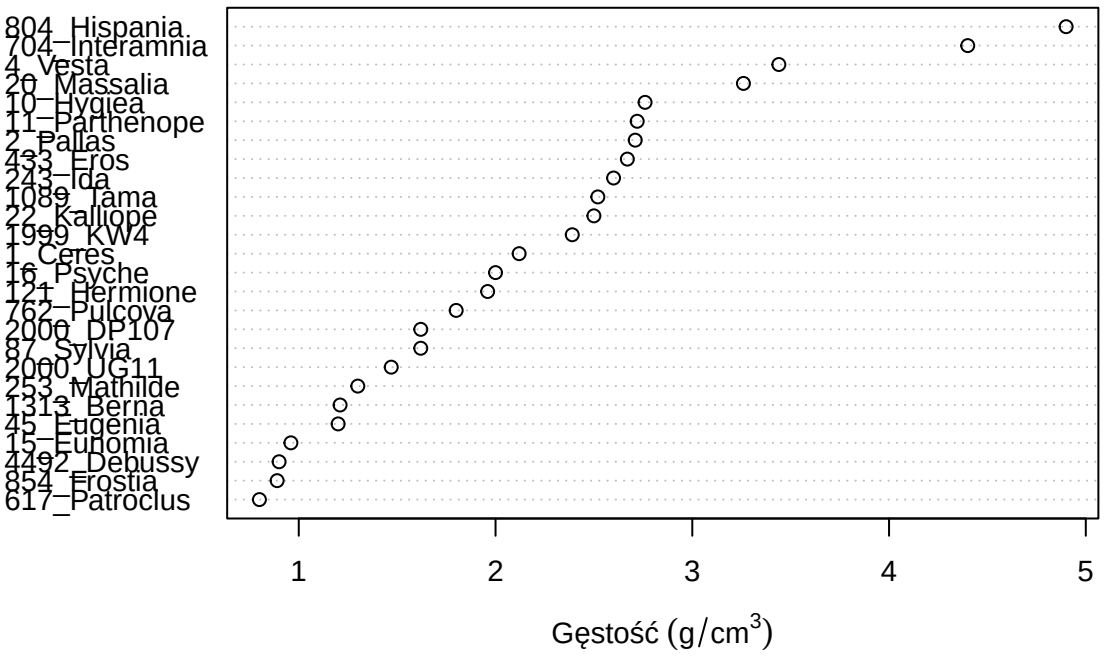
```
## [1] 0.2572554
```

```
SE2
```

```
## [1] 0.2010005
```

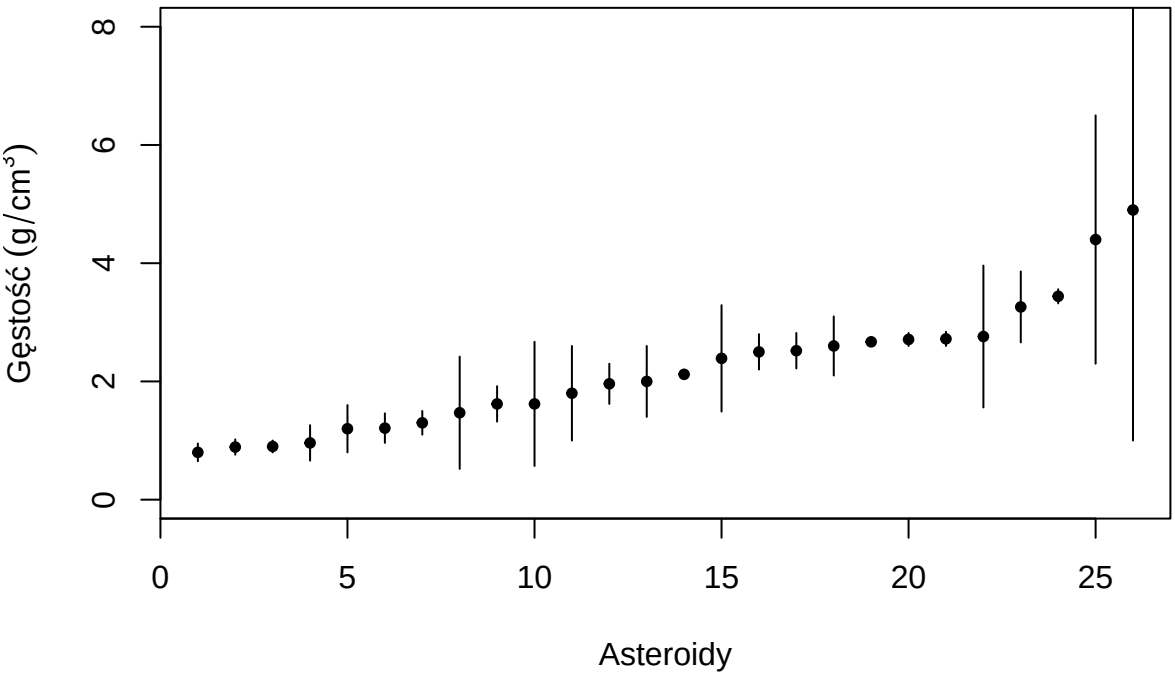
b) Na poniższych rysunkach pokazano rozkład danych, gdzie uwzględniono również błędy:

```
dotchart(dens, labels = names, cex = 0.9, xlab = expression(Gęstość ~ (g / cm ^ 3)))
```



Rysunek przedstawiający gęstość poszczególnych asteroid. Jak można zauważyć największe zagęszczenie danych występuje w okolicach gęstości o wartościach od 1 do 3. Dla pozostałych danych zaobserwowano bardzo duże gęstości co może budzić podejrzenia. W celu sprawdzenia, czy dane o dużej gęstości są prawdziwe wyrysowano wykres z błędami, który został przedstawiony poniżej:

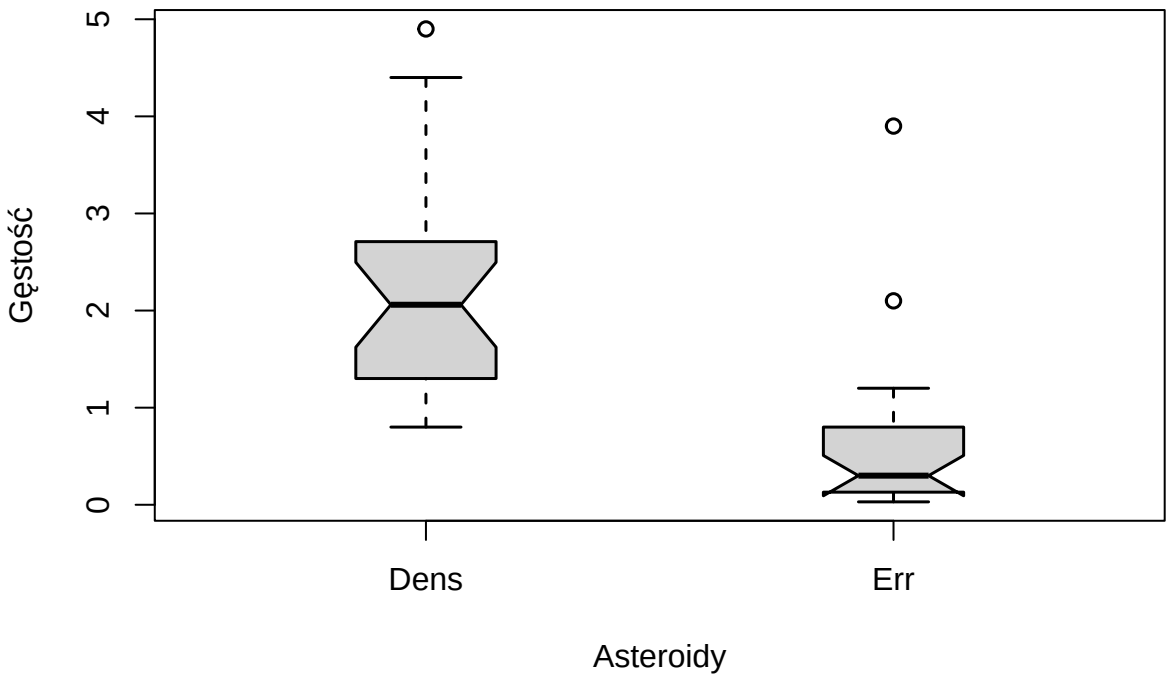
```
plot(dens, ylim = c(0, 8), xlab = "Asteroidy", ylab = expression(Gęstość ~ (g / cm ^ 3)), pch = 20)
num <- seq(1, length(dens))
segments(num, dens + err, num, dens - err)
```



Powyższy wykres pokazuje bardzo duże niepewności dla punktów o dużch gęstościach. Tak więc niekoniecznie punkty te mogą przyjmować tak duże wartości gęstości. Ostatnim wyrysowanym wykresem jest porównanie rozkładu gęstości asteroid do rozkładu błędów ich pomiarów. Wykres ten pozwala na sprawdzenie, czy dane

mają zbliżoną gęstość lub nie. Dodatkowo pozwala określić jak duże są błędy w prównaniu do pomiarów gęstości i czy te błędy są do siebie zbliżone lub czy są odstające.

```
boxplot(asteroids[, 2:3], varwidth = T, notch = T, xlab = "Asteroidy", ylab = "Gęstość", pars = list(bo
```



Jak możemy odczytać z rysunku dane mają zbliżoną do siebie gestość, a błędy są do siebie zbliżone.

c) W celu sprawdzenia, czy dane można opisać rozkładem normalnym zdecydowano się na dwa testy:

Test Kołmogorowa-Smirnowa:

```
lillie.test(dens)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dens
## D = 0.13644, p-value = 0.2435
```

```
lillie.test(err)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  err
## D = 0.24016, p-value = 0.0004775
```

Test ten jest jednym z najbardziej znanych testów zbiorczych na normalność, gdzie testuje on hipotezę zerową o tym, że rozkład naszej zmiennej jest zbliżony do normalnego. Jako statystykę testową przyjmujemy:

$$D_n = \sup_x |F_0(x) - S_n(x)|, \tag{2}$$

gdzie $s_n(x)$ to dystybuanta empiryczna, która jest usatlona na podstawie uporządkowania próbki nastepująco:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leqslant x}, \tag{3}$$

gdzie: X_i to wartość zmiennej X dla i-tej obserwacji. $I_{X_i \leqslant x}$ to funkcja charakterystyczna zbioru przyjmująca wartość jeden gdy $X_i \leqslant x$ i zero w przeciwnym wypadku. Duże wartości statystyki świadczą o dużej rozbieżności dystrybuanty F_0 i S_n, zatem konstruujemy prawostronny obszar krytyczny. Jeśli wartość D_n wpadnie w obszar krytyczny to hipotezę zerową odrzucamy na przyjętym poziomie istotności.

Przyjmując poziom istotności na poziomie $\alpha = 0.05$ oraz porównując go z otrzymaną wartością p-value możemy stwierdzić, czy hipotezę można przyjąć lub odrzucić. Dla gęstości hipotezę możemy uznać za poprawną, ponieważ wartość ta jest znacznie większa od przyjętej wartości α, natomiast dla niepewności hipotezę odrzucamy, ponieważ wartość p-value jest znacznie mniejsza od α.

Test Shapiro-Wilka:

```
shapiro.test(dens)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dens
## W = 0.93021, p-value = 0.07841
```

```
shapiro.test(err)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  err
## W = 0.64214, p-value = 9.4e-07
```

Test ten pozwala sprawdzić, czy próba pochodzi z populacji o rozkładzie normalnym. Statystyką testową jest:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{4}$$

gdzie: x_i jest i-tą najmniejszą liczbą w próbce, $\overline{x}$ to średnia z próbki, natomiast a_i wyrażaja się przez wzór: $(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{C}$. C z wcześniejszego wzoru to norma wektora, która jest opisana wzorem: $C = ||m^T V^{-1}|| = \sqrt{m^T V^{-1} V^{-1} m}$, gdzie $m = (m_1, \ldots, m_n)^T$ zawiera obserwacje dla wartości niemalejących.

Hipotezą zerową to pochodznie próby z populacji o rozkładzie normalnym. Dla wcześciej przyjętego poziomu istotności możemy stwierdzić, że hipoteze dla gęstości asteriod możemy przyjąść za słuszną, natomiast hipotezę dla niepewności należy odrzucić z powodu znacznie mniejszej wartości p-value.

d) Na podstawie testów Dixon'a i Grubb'a zostało sprawdzone, czy w danych znajdują się punkty, które są outliersami. Outliersy to punkty, dla których pomiar może być wynikiem błedu grubego. Pierwszym testem jest test Dixona, którego wyniki zostały zaprezentowane poniżej:

```
dixon.test(dens)
```

```
##
##  Dixon test for outliers
##
## data:  dens
## Q = 0.365, p-value = 0.1709
## alternative hypothesis: highest value 4.9 is an outlier
```

Test Dixona to test służących do sprawdzenia, czy próbka zawiera dane powstałe w wyniku popełnienia błędu grubego. Statystyka testowa to: $Q = \frac{\text{gap}}{\text{range}}$, gdzie gap to moduł z różnicy pomiędzy podejrzanym pomiarem, a wartością najbliższego pomiaru. Range jest różnicą pomiędzy największą wartości z próbki, a najmniejszą wartościa z próbki.

Kolejnym testem na wykrycie błędu grubego obarczonego błędem jest test Grubbs'a, który polega na zdefiniowaniu hipotezy H_0 : o braku odchyleń w zbiorze danych oraz H_a : czyli, czy istnieje ryzyko odchylenia w zbiorze danych. Statystyka testowa jest określona jako:

$$G = \frac{max|X_i - \overline{X}|}{\sigma}, \tag{5}$$

gdzie $\overline{X}$ - średnia, σ - odchylenie standardowe. Statystykę Grubbsa uznaje się największe odchylenie od średniej w zbiorze o rozkładzie normalnym. Wyniki testu prezentują się następująco:

```
grubbs.test(dens)
```

```
##
##  Grubbs test for one outlier
##
## data:  dens
## G = 2.5967, U = 0.7195, p-value = 0.07011
## alternative hypothesis: highest value 4.9 is an outlier
```

Analizujac wyniki testów można stwierdzić, że żaden punkt z nie możemy uznać za błąd gruby. Szczególnie punkty, które mają największą gęstość.

Zadanie 2

a) Poniżej zaprezentowano statystyki dla dwóch zestawów danych, gdzie pierwszym zestawem danych są obserwacje gromad kulistych w Drodze Mlecznej:

```
summary(galaxies1)
```

```
##     M31_GC                K
##  Length:360         Min.   :10.75
##  Class :character   1st Qu.:13.85
##  Mode  :character   Median :14.54
##                     Mean   :14.46
##                     3rd Qu.:15.33
##                     Max.   :18.05
```

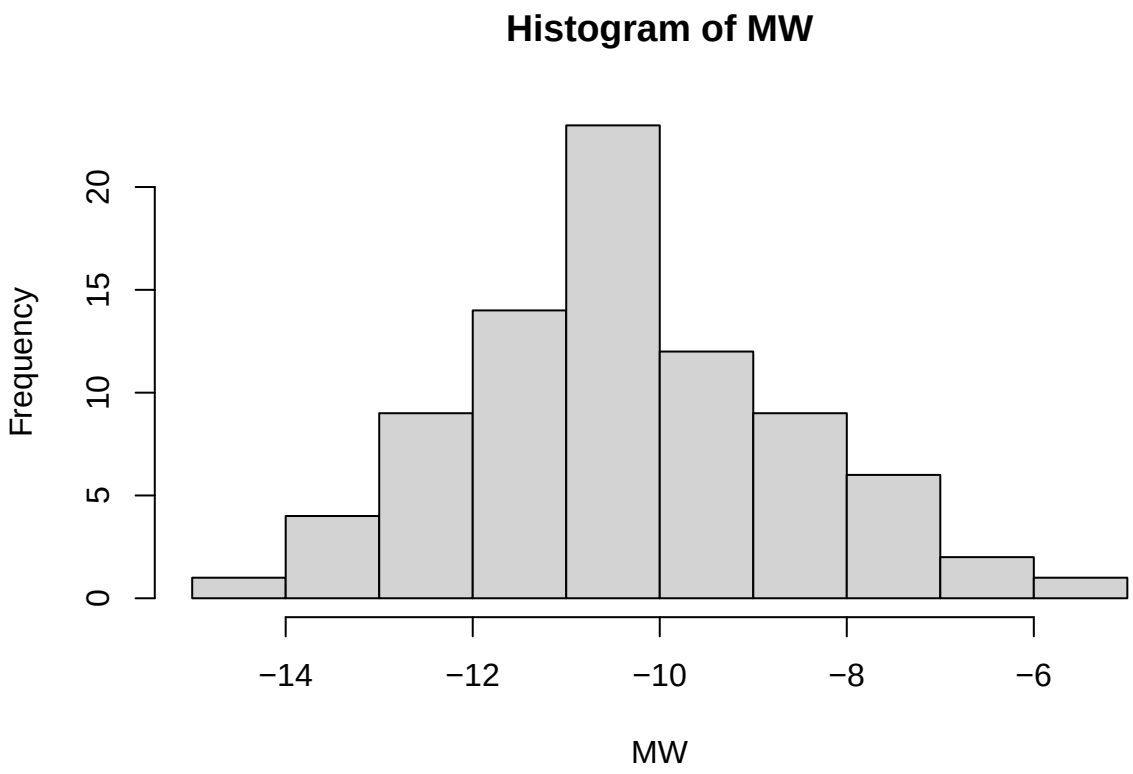
natomiast drugim są obserwacje gromad kulistych w M31:

```
summary(galaxies2)
```

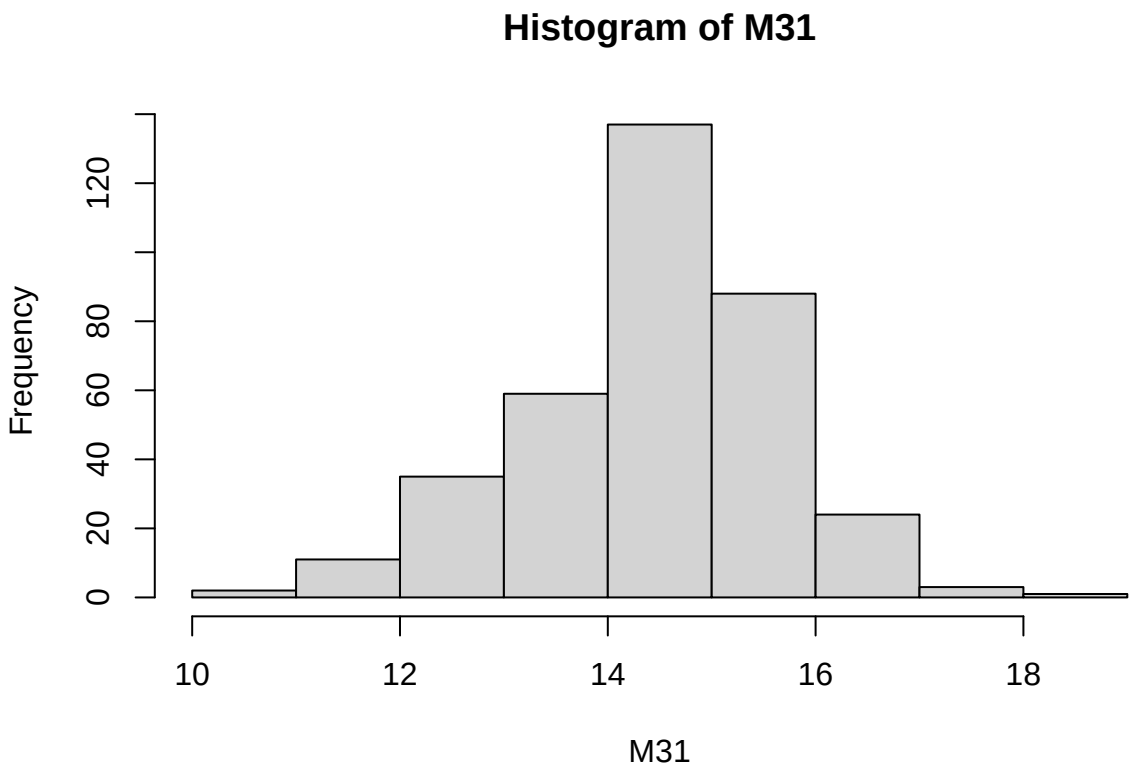
```
##     MWG_GC                K
##  Length:81          Min.   :-14.205
##  Class :character   1st Qu.:-11.478
##  Mode  :character   Median :-10.557
##                     Mean   :-10.324
##                     3rd Qu.: -9.199
##                     Max.   : -5.140
```

b) Poniżej zaprezentowano rozkłady danych, dla których zdecydowano się wyrysować histogramy w celu przedstawienia graficznego częstotliwości występowania jasności w danym przedziale.

```
H1 <- hist(MW)
```




```
H2 <- hist(M31)
```



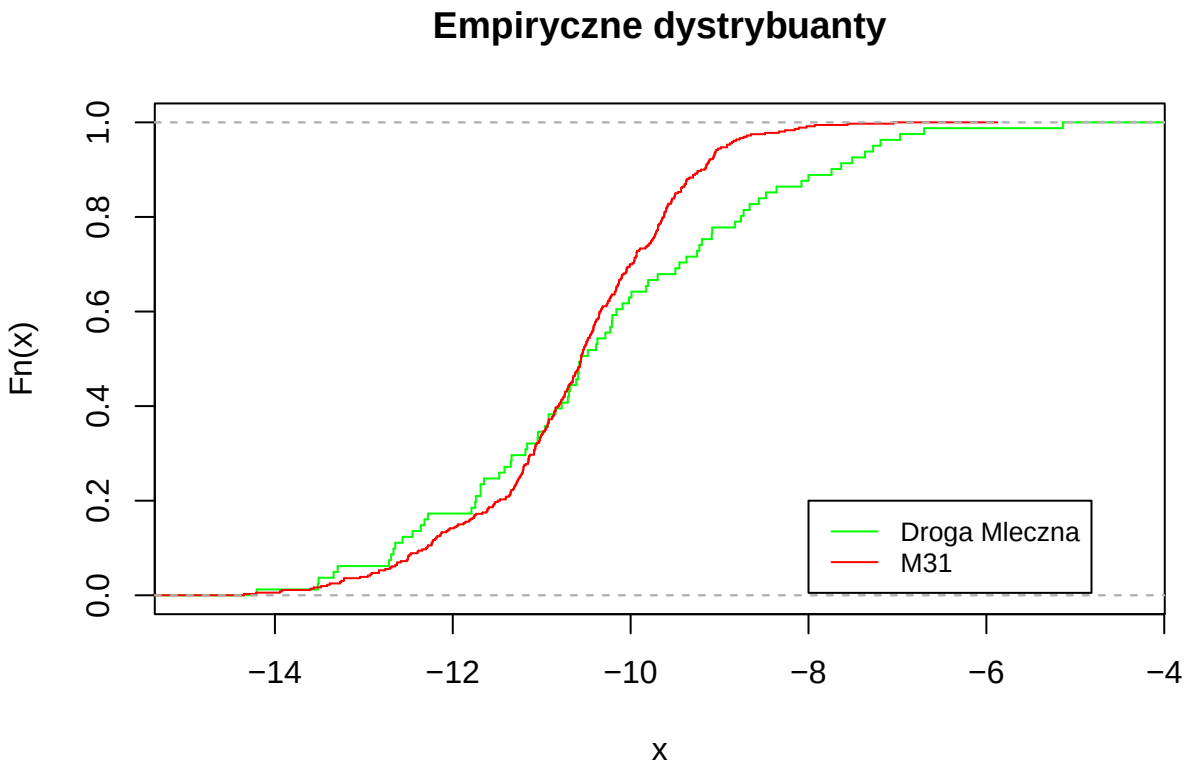
Z wykresów można zauważyć, iż wartość jasności jest na obu histogramach całkowicie różna. Chcąc przedstawić rozkłady na jednym wykresie należy uwzględnić poprawkę odejmując różnicę median Drogi Mlecznej oraz M31. W tym celu obliczono średnie, które następnie odjęto od siebie:

```
module <- median(MW)-median(M31)
module
```

```
## [1] -25.0965
```

Efektem będzie otrzymanie podobnego poziomu jasności, który jest obserwowalny w naszej Galaktyce. Wyniki porównania wykresów zostały przedstawione poniżej:

```
plot(ecdf(MW),cex.points=0,verticals=T, main="Empiryczne dystrybuanty", col="green")
plot(ecdf(M31+module),cex.points=0,verticals=T,add=T,main="Empiryczne dystrybuanty", col="red")
legend(-8,0.2,legend = c("Droga Mleczna","M31"),col = c("green","red"),lty = 1,cex=0.8)
```



c) Wynik wartości **module** porównano z wynikiem testu rang Wilcoxon'a, który został obliczony poniżej:

```
wilcox.test(MW,M31+module)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  MW and M31 + module
## W = 15669, p-value = 0.2936
## alternative hypothesis: true location shift is not equal to 0
```

Jak można zauważyć wartości wyznaczone za pomocą median oraz testu są zbliżone. Tak więc moża przyjąść, iż różnica median jako przesunięcie jest zadowalająco wystarczające.

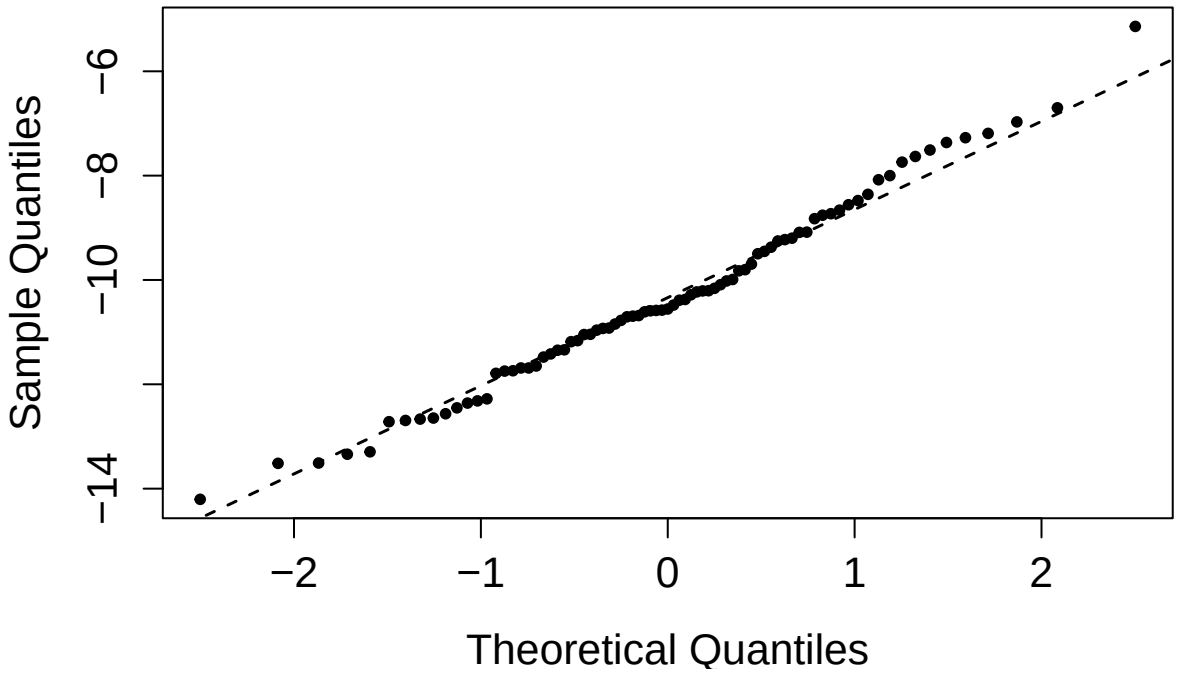
d) Q-Q – wykres kwantyl-kwantyl służy między innymi do porównania, czy dystrybuancja danych jest zgodna z modelem. Dwuwymiarowy wykres zmiennych przedstawia punkty, które powstają na wykresie w wyniku dwóch odpowiadających sobie wartości kwantyli zmiennych losowych ich rozkładu. warto zauważyć, że jeśli wszystkie punkty odpowadają lini prostej to zmienne losowe na wykresie są opisane tym samym rozkładem. Procedura utworzenia wykresu przedstawia się następująco:

- porządkowanie rosnąco residuł, które są kantylami empirycznymi,

- obliczenie rzędów kwantyli,
- obliczenie kwantyli teoretycznych,
- sortowanie residuł rosnąco dla drugiego zestawu danych,
- zestawienie kwantyli n wykresie.

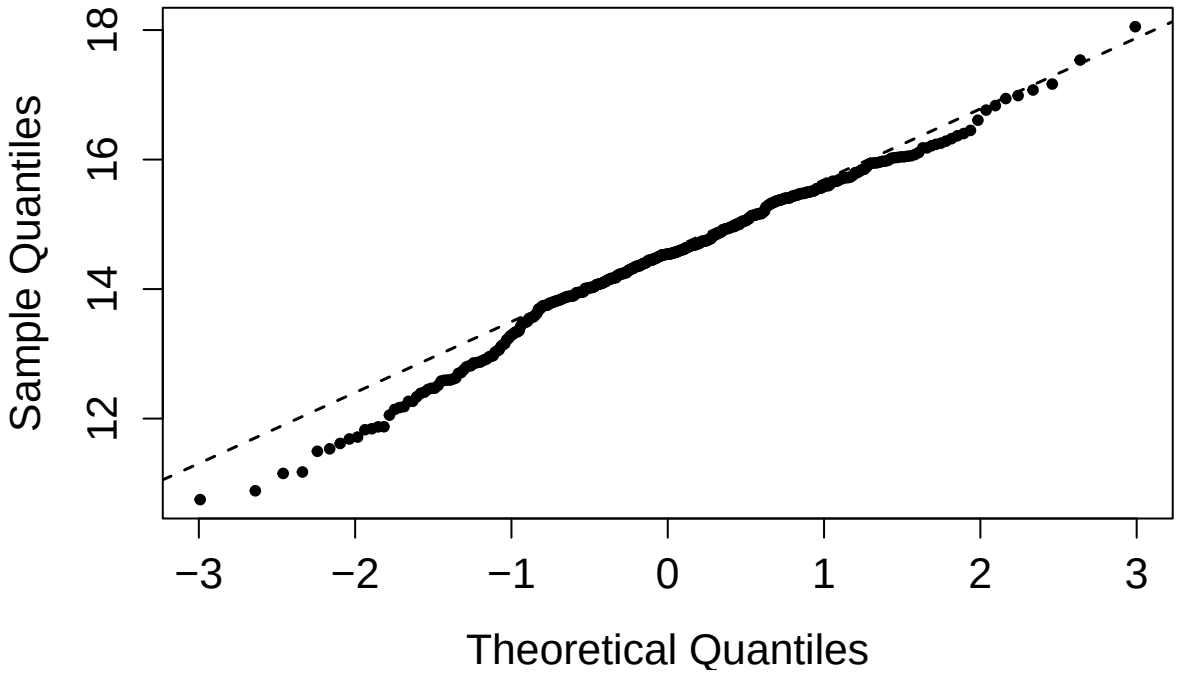
Poniżej zostały zaprezentowane wykresy kwantyl kwantyl:

```
qqnorm(MW, pch=20, cex.axis=1.3, cex.lab=1.3, main="")
qqline(MW, lty=2, lwd=1.5)
```



Oś OY powyższego wykresu przedstawia empiryczną dystrybucję jasności gromad w Drodze Mlecznej, natomiast oś OX teoretyczną dystrybucję rozkładu normalnego. Jak można zauważyć kwantyle empiryczne zgadzają się z modelem teoretycznym.

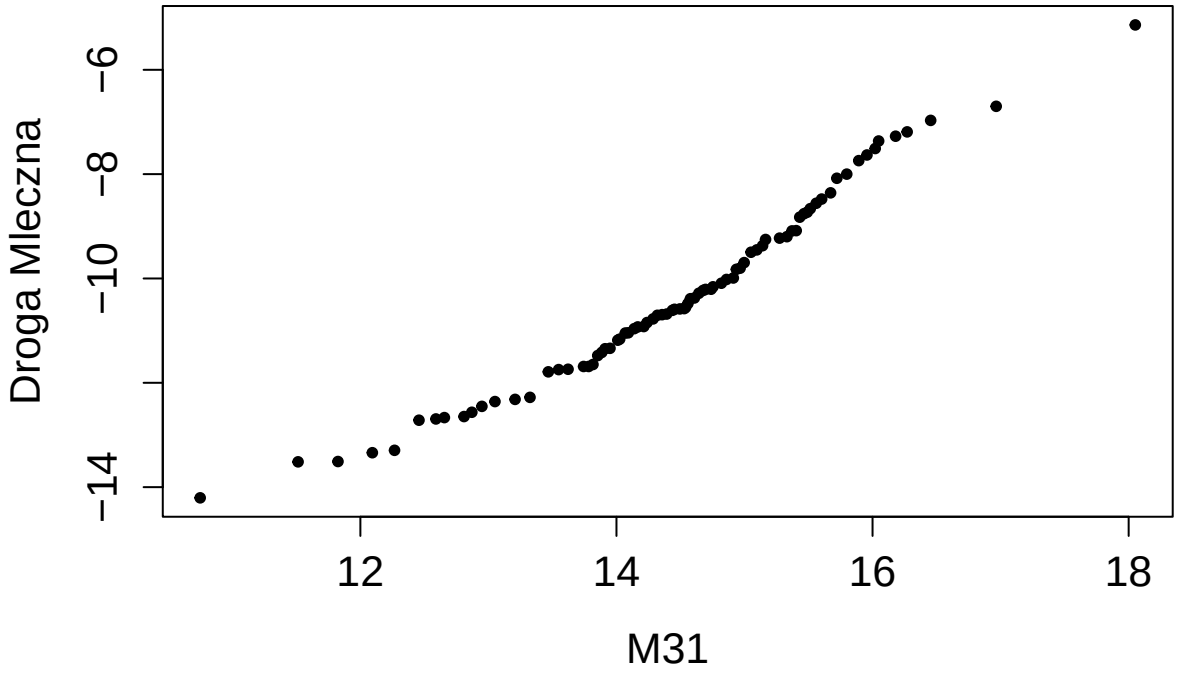
```
qqnorm(M31, pch=20, cex.axis=1.3, cex.lab=1.3, main="")
qqline(M31, lty=2, lwd=1.5)
```



Analogicznie jak poprzednio oś OY powyższego wykresu przedstawia empiryczną dystrybucję jasności gromad w M31, natomiast oś OX teoretyczną dystrybucję rozkładu normalnego. Można również zauważyć, iż model teoretyczny nie pokrywa się z niższymi wartościami.

Porównując dystrybuancję dwóch zestawów danych otrzymujemy nastepujący wykres:

```
qqplot(M31,MW, pch=20, cex.axis=1.3, cex.lab=1.3, main="", xlab = "M31", ylab = "Droga Mleczna")
```



W powyższym przypadku na osiach są empiryczne dystrybucje jasności gromad w Drodze Mlecznej i w M31. Na powyższym rysunku możemy zaobserwować nieliniowy trend, który prowadzi do wniosku, iż rozkłady jasności obu galaktyk są różne.

e) W celu potwierdzenia powyższego wniosku należy przeprowadzić nastepujący test:

```
ks.test(MW, M31+module)
```

```
## Warning in ks.test(MW, M31 + module): p-value will be approximate in the
## presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  MW and M31 + module
## D = 0.18333, p-value = 0.02348
## alternative hypothesis: two-sided
```

Statystyka testu została opisana w poprzednim zadaniu. Dla poziomu istotności $\alpha = 0.05$ hioptezę należałoby odrzucić, ale ponieważ poziom istotności możemy przyjąść również na poziomie $\alpha = 0.01$ nie mamy pewności co do poprawności odrzucenia hipotezy.

Zadanie 3

a) Estymator zgodny, który jest nieobciążony oraz najefektywniejszy wartości przeciętnej to wartość średnia. Poniżej została przedstwiona wartość obliczonej średniej:

```
Estym <- mean(Height, na.rm = T)
Estym
```

```
## [1] 172.3809
```

b) Dla nieznanego odchylenia standargowego oraz nieznanej wartości przeciętnej stosujemy statystykę studenta. Szukam wartość błędu przy przedziale ufności 95%, gdzie wykonuję następujące obliczenia:

```
#rozkład studenta
s2 <- sd(Height, na.rm = T)
n <- length(Height)
t <- qt(0.975, n-1)
t
```

```
## [1] 1.970067
```

```
#przedział ufnosci
Q1=qt(0.05/2, 237-1)*s2/sqrt(n)+ Estym
Q1
```

```
## [1] 171.1207
```

```
Q2=qt(1-0.05/2, 237-1)*s2/sqrt(n)+ Estym
Q2
```

```
## [1] 173.641
```

Ostatecznie przedziałem ufności jest wartość:

$$[171.1207; 173.641],$$

natomiast wartość średnia wraz z błędem pochodzącym z tego przedziału to:

$$172.38 \pm 1.34[cm].$$

c) W celu sprawdzenia hipotezy zerowej, czyli czy studenci palą niezależnie od ilości ćwiczeń zastosowano test chi-kwadrat. Statystyka opisująca ten test to:

$$\chi^2 = \sum_{i=1}^{r} \sum_{k=1}^{s} \frac{(n_{ik} - n \cdot p_{ik})^2}{n \cdot p_{ik}}, \tag{6}$$

gdzie: n_{ik} - liczności odpowiadające danej parze kategorii, r=3 -liczba kategorii częstotliwości palenia, s=4 - liczba kategorii częstości ćwiczeń, n - suma n_{ik} oraz p_{ik} które jest obliczane ze wzoru:

$$p_{ik} = p_{i.}p_{.k} = \frac{n_{i.}}{n} \cdot \frac{n_{.k}}{k}. \tag{7}$$

W sumie tatystyka ma 6 stopni swobody, natomiast obszar krytyczny jest prawostronny. Otrzymano nastepujący wynik testu chi-kwadrat:

```
chisq.test(Exer, Smoke)
```

```
## Warning in chisq.test(Exer, Smoke): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  Exer and Smoke
## X-squared = 5.4885, df = 6, p-value = 0.4828
```

Z wyniku testu można stwiedzić, iż dla poziomu istotności $\alpha = 0.05$ nie mamy podstaw do odrzucenia hipotezy zerowej.

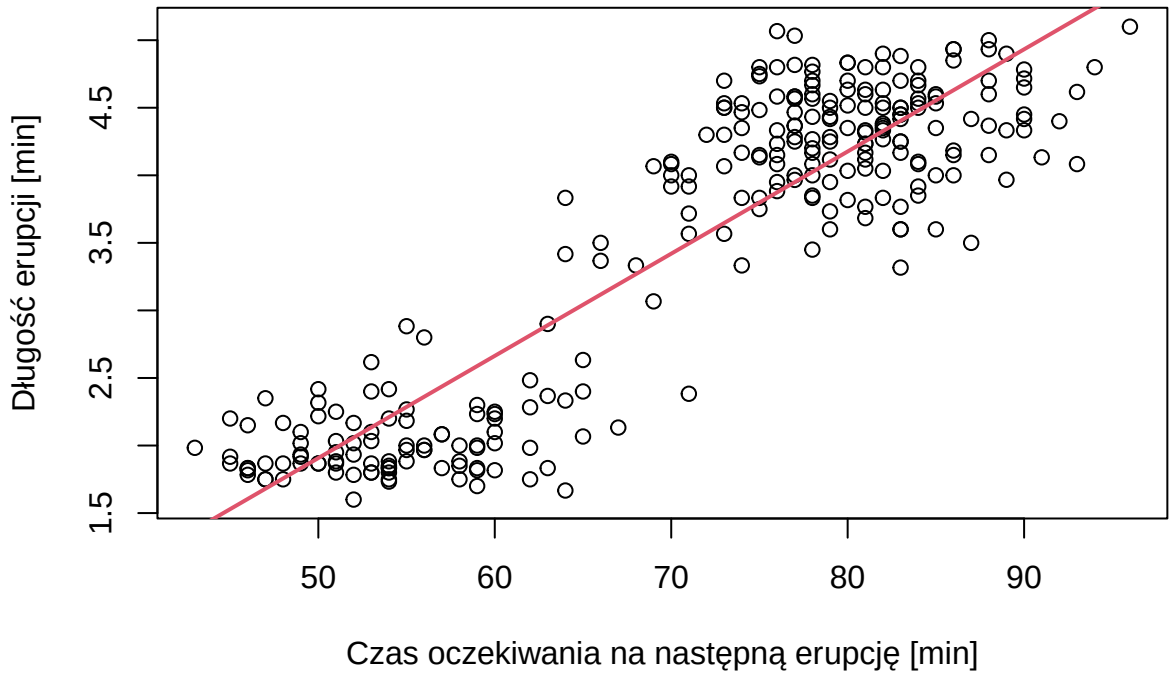
Zadanie 4

a) W celu dopasowania regresji liniowej zostały wykonane następujące polecenia:

```
x <- faithful$eruptions
y <- faithful$waiting
regline <- lm(x~y)
summary(regline)
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.29917 -0.37689  0.03508  0.34909  1.19329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.874016   0.160143  -11.70   <2e-16 ***
## y            0.075628   0.002219   34.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4965 on 270 degrees of freedom
## Multiple R-squared:  0.8115, Adjusted R-squared:  0.8108
## F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

```
plot(x~y, ylab = "Długość erupcji [min]", xlab = "Czas oczekiwania na następną erupcję [min]")
abline(regline,lwd=2,col=2)
```



Powyżej zostały przedstawione statystyki oraz wykres zależności długości erupcji od czasu oczekiwania na kolejną erupcję wraz z dopasowaną prostą za pomocą regresji liniowej. Wartościami parametrów dopasowania dla prostej:

$$y = a \cdot x + b, \tag{8}$$

są między innymi:

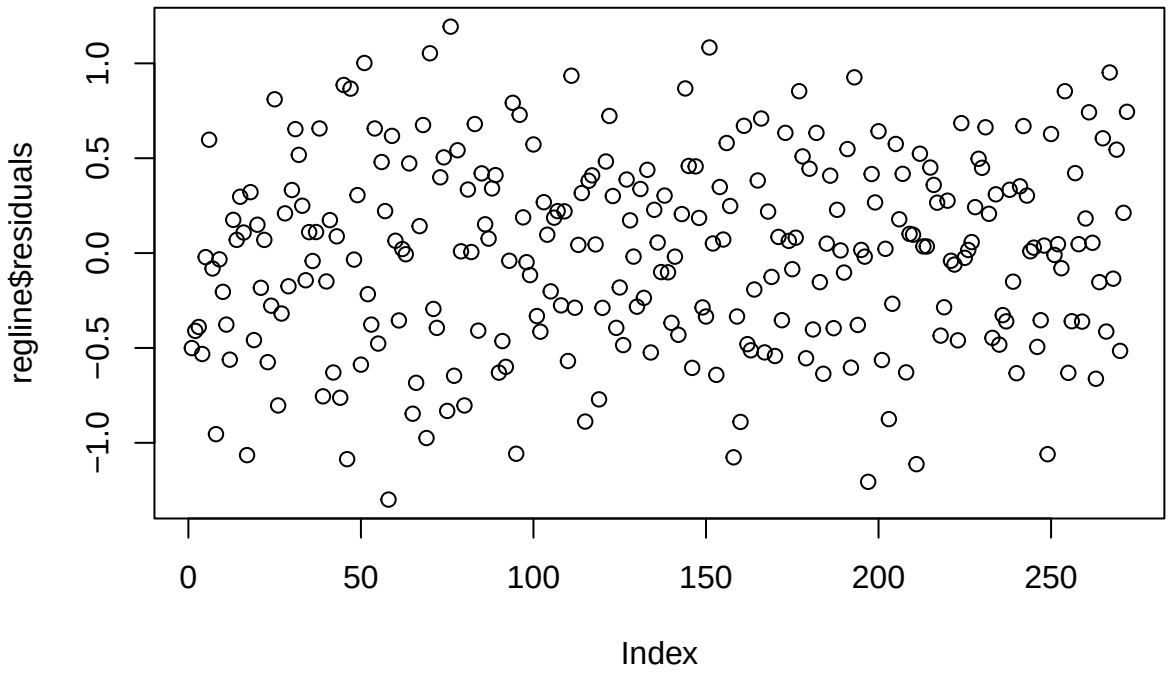
- $a = 0.0756 \pm 0.0023$,
- $b = -1.87 \pm 0.17$.

Do oszacowania czasu trwania następnej erupcji wykorzystano powyższy wzór wraz z współczynnikami dopasowania, gdzie $x = 80$ min. Po podstawieniu otrzymano następujący wynik:

$$4.18 \pm 0.36.$$

b) Poniżej zostały wyrysowany wykres residuł dopasowanej regresji:

```
plot(regline$residuals)
```



Jak moża zauażyć zastosowany model regresji liniowej jest poprawny, ponieważ dane są rozproszone oraz nie wykazują żadnego trendu.

c) Kryteria **Akaike** oraz **Bayesowskie** stosuje się w celu porównania zastosowanych modeli do danych. Wartości uzyskane z tych kryteriów są estymatorami błędu oraz zgodności próbek z modelem. Ich wartość może wskazać, który z modeli jest lepszym wyborem, gdzie niższa wartość oznacza, iż model dla poszczególnych danych jest lepszy. Poniżej zaprezentowano wartości tych kryteriów dla regresji liniowej.

```
AIC(regline)
```

```
## [1] 395.0159
```

```
BIC(regline)
```

```
## [1] 405.8333
```

d) Korzystając z 95% poziomu ufności średniego czasu trwania erupcji dla czasu 80 min zostały wyznaczone za pomocą funkcji predict wartości przewidywane y. Poniżej przedstawiono wyniki działania funkcji:

```
regline1 <- lm(eruptions~waiting,data=faithful)
time80=predict(regline1,data.frame(waiting=c(80)),interval = "confidence")
print(time80)
```

```
##       fit      lwr      upr
## 1 4.17622 4.104848 4.247592
```

Ostatecznie odczytanym z powyższych wartości jest przedział ufności: $(4.1; 4.25)$.