Lab 8 - zadanie domowe

Mateusz Kapusta

2022-04-26

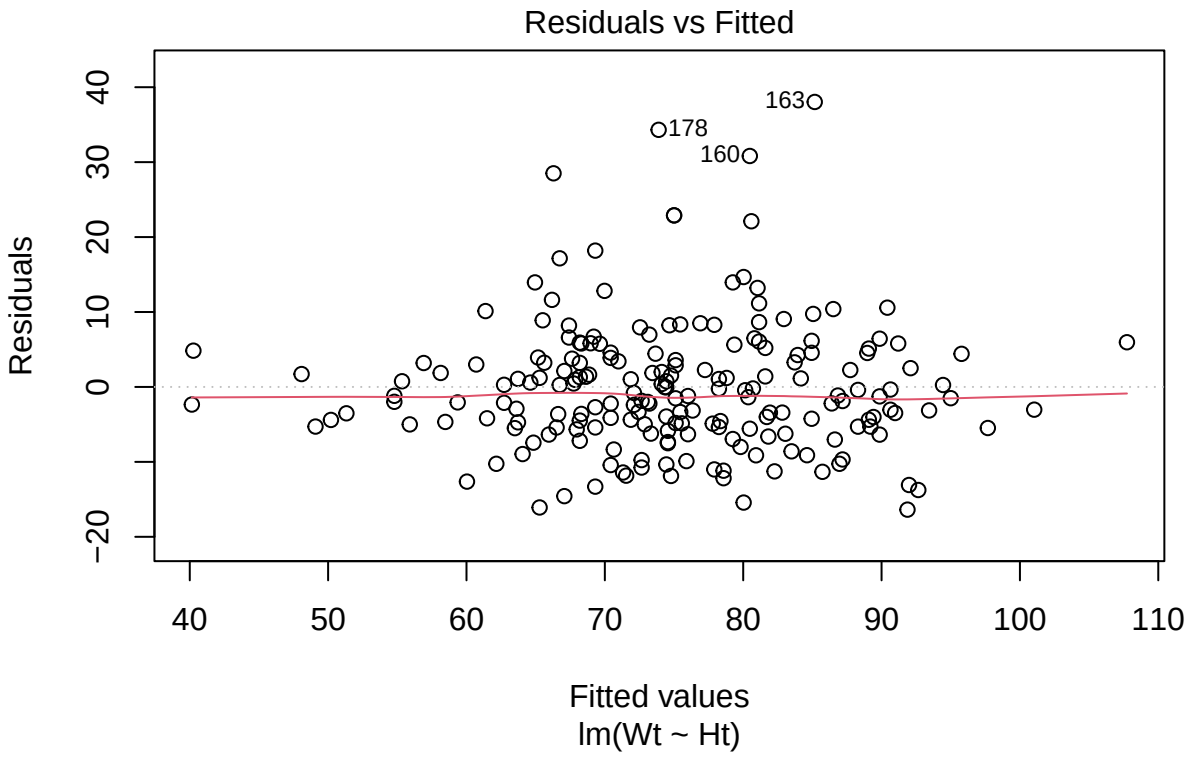
Zadanie nr 2

Ładujemy dane

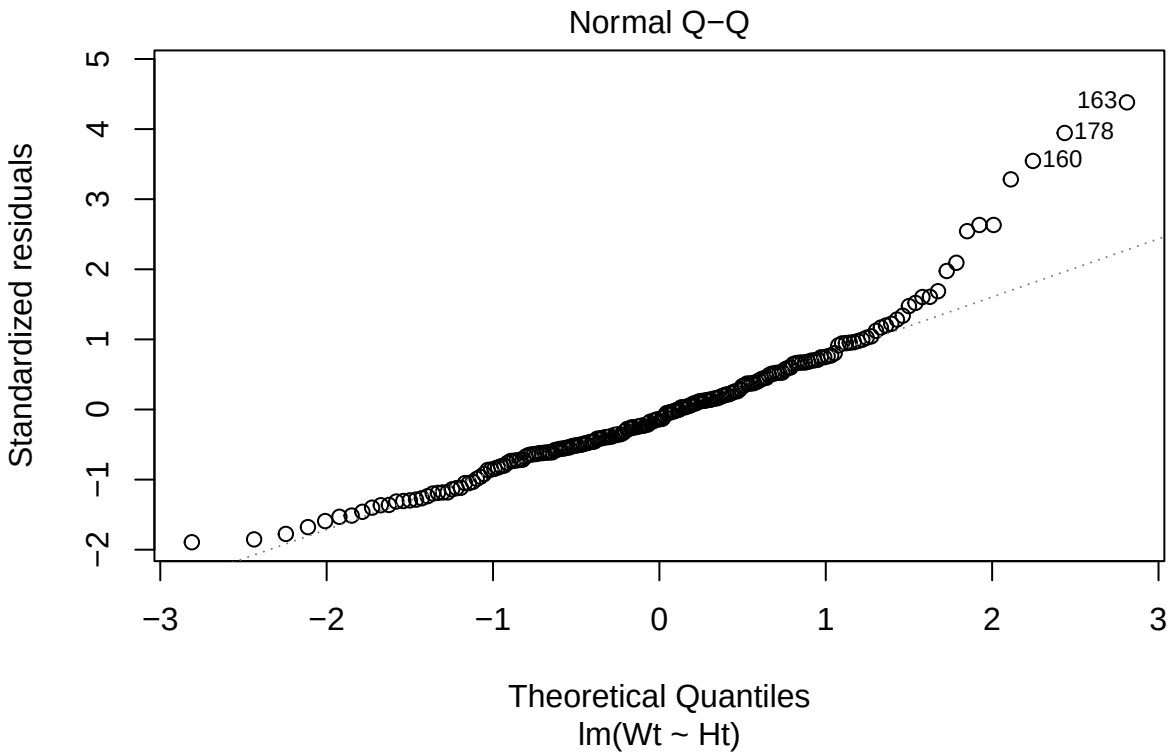
```
ais<-read.csv("ais.txt",sep="\t")
model<-lm(Wt~Ht,data=ais)
summary(model)
```

```
##
## Call:
## lm(formula = Wt ~ Ht, data = ais)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.372  -5.296  -1.197   4.378  38.030
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -126.18901   11.39656  -11.07   <2e-16 ***
## Ht             1.11712    0.06319   17.68   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.72 on 200 degrees of freedom
## Multiple R-squared:  0.6098, Adjusted R-squared:  0.6079
## F-statistic: 312.6 on 1 and 200 DF,  p-value: < 2.2e-16
```

```
plot(model, which=1)
```

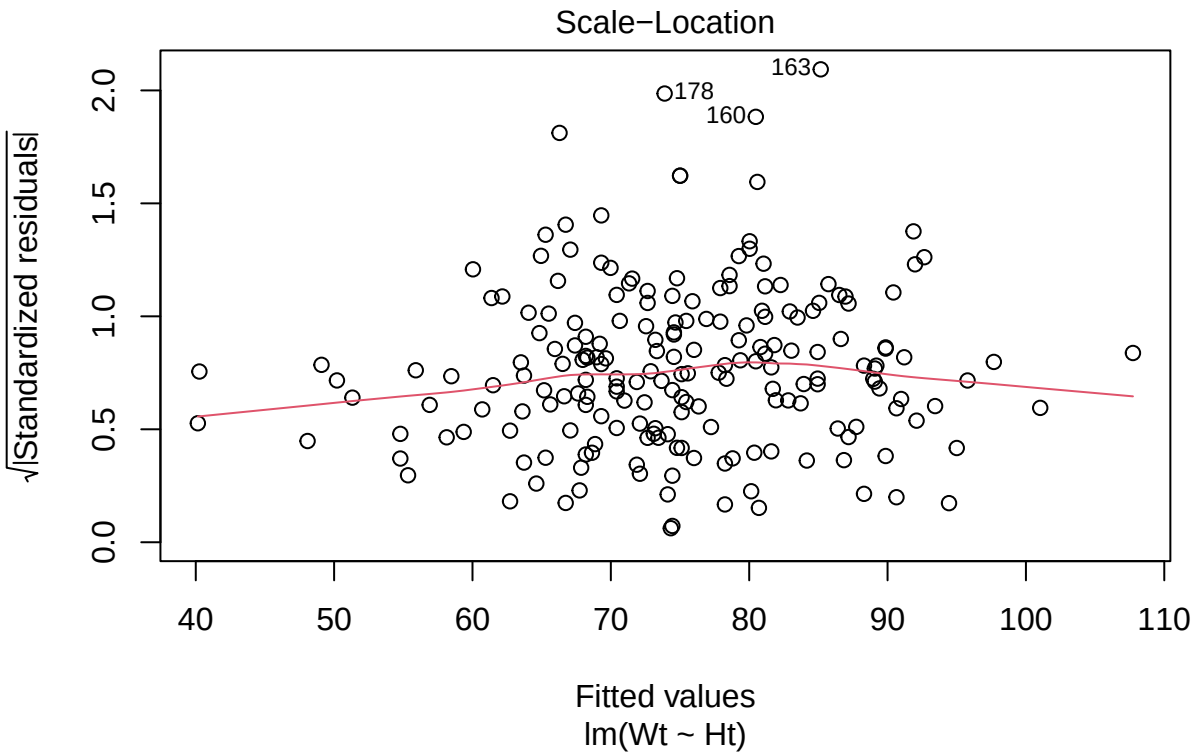



```
plot(model,which=2)
```



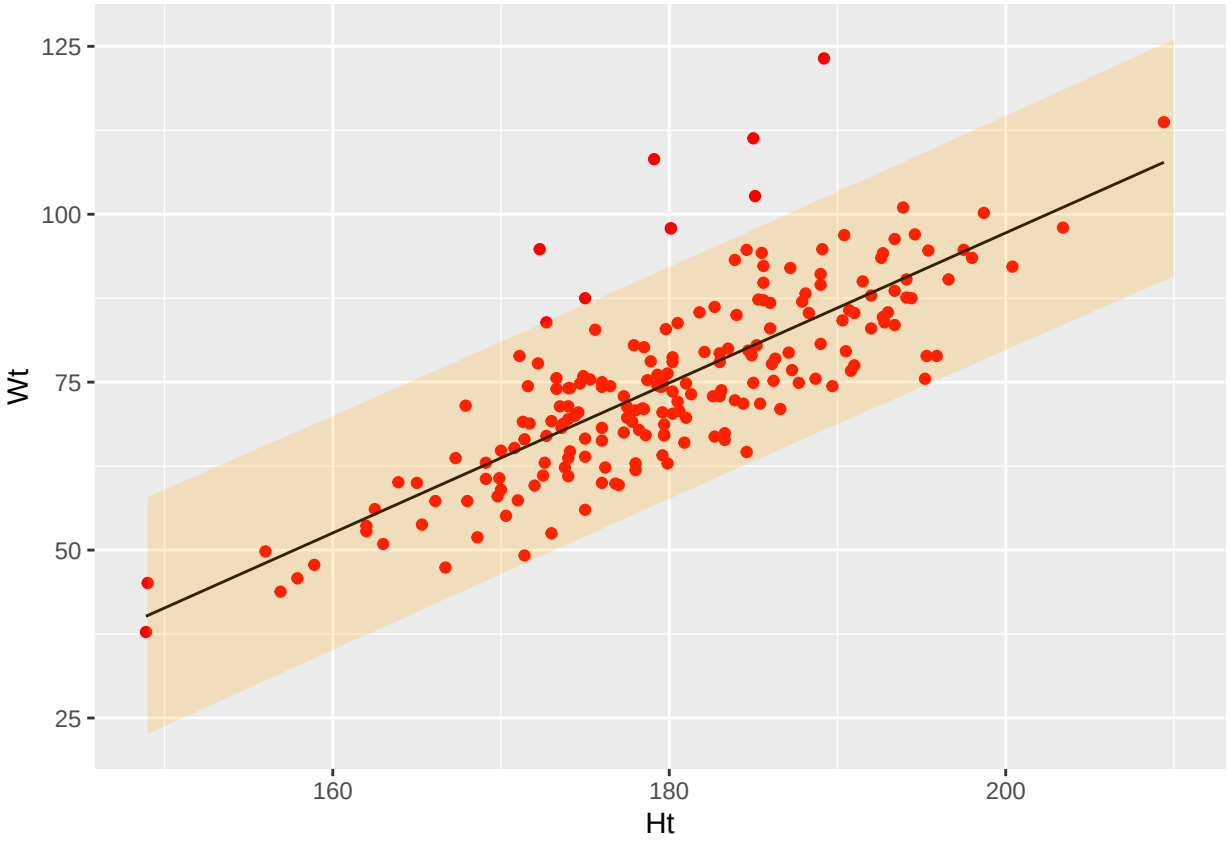
Wykres residuów w zależności od predykowanej wartości sugeruje, iż dane mają trend liniowy. Podsumowanie modelu skazuje, iż p-wartość dla naszych współczynników ma bardzo małą wartość co sugeruje, iż istnieją spore podstawy do stwierdzenia, iż współczynniki są statystycznie ważne. Widzimy, że residua z grubsza mają rozkład normalny za wyjątkiem dużych wartości kwantyli teoretycznych. Fakt ten sugeruje, że dla dużych wartości residuuów mamy więcej obserwacji niż się spodziewamy.

```
plot(model, which=3)
```



Widzimy też, że odchylenie standardowe residuów jest bardzo mało zależne od predykowanej masy jednakże spodziewać się należy zwiększenie przedziału ufności dla danych w środkowej części predykowanych wartości.

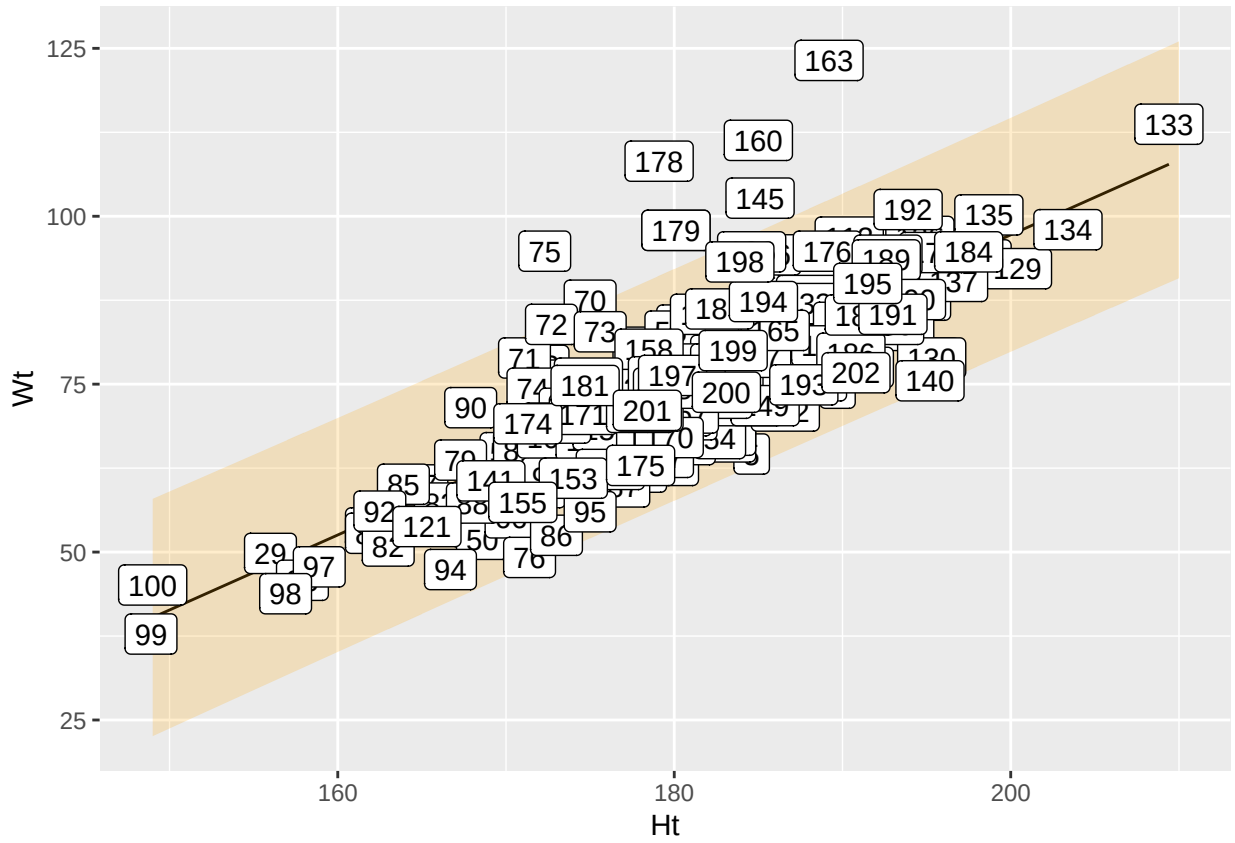
```
przedzialy <- predict(model, data.frame('Ht'=floor(min(ais$Ht)):floor(max(ais$Ht))+1),
                      interval='prediction')
przedzialy <- data.frame(przedzialy)
przedzialy$Ht <- floor(min(ais$Ht)):floor(max(ais$Ht))+1
predykcja <- data.frame('Ht'=ais$Ht, 'Wt' = model$coefficients[1]+model$coefficients[2]*ais$Ht)
ggplot()+geom_point(aes(x=Ht,y=Wt),data=ais,col="red")+geom_line(
  aes(x=Ht,y=Wt),data=predykcja)+geom_ribbon(
    aes(x=Ht, ymin=lwr, ymax=upr), data=przedzialy, alpha=0.2, fill='orange')
```



Podsumowując w powyższym modelu zarówno normalnośc residuuów, liniowość trendu, niezależność błędu jak i hemoskedastyczność są spełnione a otrzymane parametry charakteryzuje bardzo mała p-wartość. Wnioski jakie możemy wyciągnąć z analizy residuów to zwiększony przedział ufności dla środka danych.

Zadanie nr 3

```
ggplot()+geom_line(aes(x=Ht,y=Wt),data=predykcja)+geom_ribbon(
  aes(x=Ht, ymin=lwr, ymax=upr), data=przedzialy, alpha=0.2, fill='orange')+geom_text(
    aes(x=Ht,y=Wt,label=rownames(ais)),data=ais)+geom_label(
    aes(x=Ht,y=Wt,label=rownames(ais)),data=ais)
```



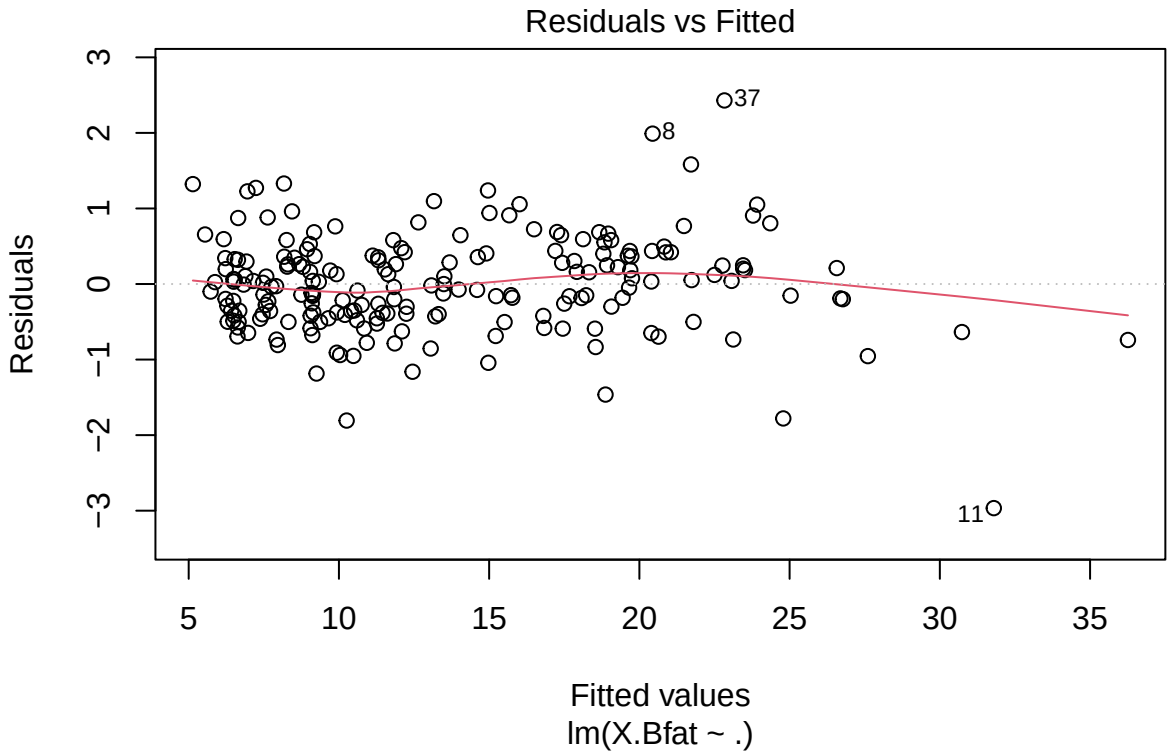
Zadanie nr 4

```
modelp<-lm(X.Bfat~.,data=ais)
summary(modelp)
```

```
##
## Call:
## lm(formula = X.Bfat ~ ., data = ais)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.96590 -0.40031 -0.00434  0.36446  2.42951
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0364151  7.7681757   0.133   0.8940
## Sexmale      -1.1884394  0.2577289  -4.611 7.55e-06 ***
## SportField   -0.6421617  0.2822778  -2.275   0.0241 *
## SportGym     -0.9114721  0.5092958  -1.790   0.0752 .
## SportNetball  0.2770035  0.2311573   1.198   0.2324
## SportRow      0.1269433  0.1936222   0.656   0.5129
## SportSwim    -0.4892245  0.2302937  -2.124   0.0350 *
## SportT400m   -1.0176295  0.2374992  -4.285 2.97e-05 ***
## SportTennis  -0.3665629  0.2872657  -1.276   0.2036
## SportTSprnt  -0.7145078  0.2918772  -2.448   0.0153 *
```

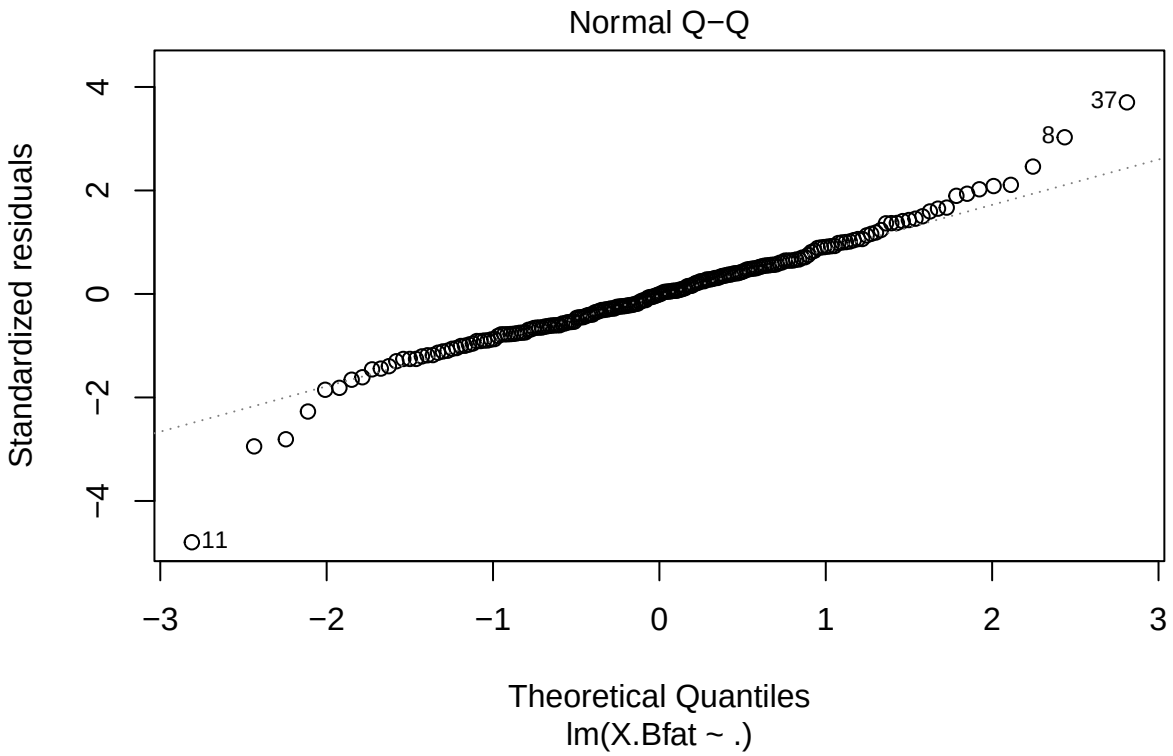
```
## SportWPolo   -0.2415956  0.2459495  -0.982   0.3273
## RCC          -0.0843242  0.3179322  -0.265   0.7911
## WCC          -0.0263891  0.0305244  -0.865   0.3884
## Hc            0.0570532  0.0557547   1.023   0.3075
## Hg           -0.0843821  0.1282557  -0.658   0.5114
## Ferr          0.0003584  0.0012469   0.287   0.7741
## BMI           0.2211776  0.1751068   1.263   0.2082
## SSF           0.0366942  0.0077008   4.765 3.86e-06 ***
## LBM          -0.9175196  0.0549361 -16.702  < 2e-16 ***
## Ht            0.0520338  0.0429649   1.211   0.2274
## Wt            0.7366858  0.0712304  10.342  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6824 on 181 degrees of freedom
## Multiple R-squared:  0.9891, Adjusted R-squared:  0.9878
## F-statistic: 817.8 on 20 and 181 DF,  p-value: < 2.2e-16
```

```
plot(modelp, which=1)
```



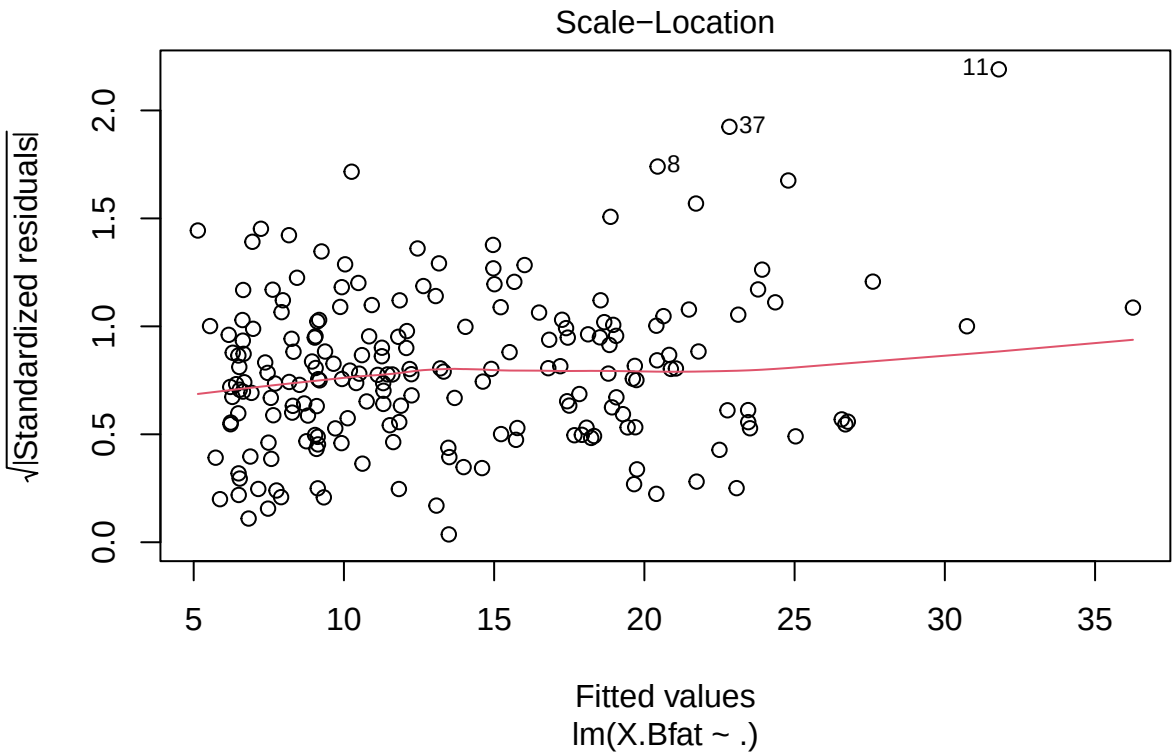
Pierwszy wykres przedstawiający residuua w zależności od fitowanej wartości pozwala nam stwierdzić, że trend z bardzo dużym przbyliżeniem jest liniowy.

```
plot(modelp,which=2)
```



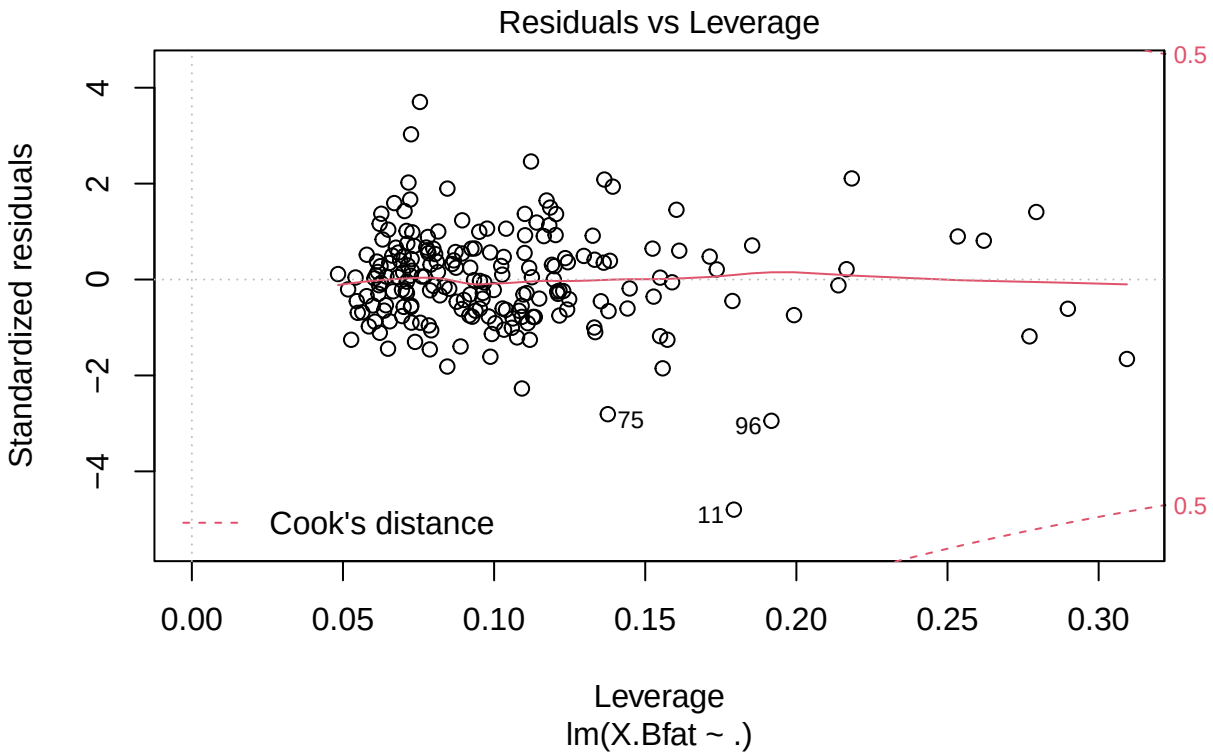
Wykres kwantylowy pozwla nam potwierdzić normalność błędów w bardzo dużym zakresie.

```
plot(modelp,which=3)
```



Wykres zależności pierwiasta z standaryzowanych residuuów w zależności od predykowanej wartości mówi nam, że błędy jakie popełniamy są hemoskedastyczne.

```
plot(modelp,which=5)
```



Ostatni wykres pozwala nam ustalić pomiary o dużej dźwigni. Na wszystkich wykresach wyraźnie widać że jedną z odstających obserwacji jest pomiar o numerze 11. Obserwacje 8,37 oraz 75 oraz 96 także pojawiają się na wykresach jako obserwacje odstajace i wykluczenie ich z danych jest warte rozważenia. Przejdzmy teraz do analizy wyników regresji liniowej. Niskie wartości p wartości sugerują, iż statystyczne ważnymi zmiennymi są

1. Płeć
2. Dyscyplina sportu jeżeli jest to
    - Pływanie
    - Bieg na 400 m
    - Sprint
    - Field?
3. waga
4. chuda masa ciała
5. suma grubości skóry

Zauważmym że wszystkie współczynniki dla rodzaju sportu które są statystycznie ważne mają ujemne wartości. Oznacza to, że sportowcy z tych dyscyplin mają średnio mniej prcentu tłuszczu w porównaniu do kolegów z innych dyscyplin o takich samych danych. Widzimy też, że dla mężczyzn współczynnik także będzie mniejszy co spowodowane jest najprawdopodobniej faktami biologicznymi (według znalezionych danych medycznych kobiety mają średni poziom tłuszczu większy o 10% większy niż u facetów).

Zadanie nr 6

```
tren<-ais[seq(2,nrow(ais),2),]
test<-ais[seq(2,nrow(ais),2)-1,]
modelk<-lm(X.Bfat~.,data=tren)
trenerr<-sum(sapply(1:nrow(tren),function (x)
  (predict(modelk,tren[,-10],interval="prediction")[x]-tren[x,10])^2))/(nrow(tren))
testerr<-sum(sapply(1:nrow(test),function (x)
  (predict(modelk,test[,-10],interval="prediction")[x]-test[x,10])^2))/(nrow(test))
r1<-summary(modelk)$adj.r.squared
```

Testy średniokwadaratowe danych treningowych oraz testowych to odpowiednio 0.3190558 oraz 0.6405177 natomiast współczynnik dopasowania R^2 to 0.9886545. Teraz dla okrojonego modelu

```
model2<-lm(X.Bfat~Sport+Sex+SSF+LBM+Wt,data=tren)
trenerr2<-sum(sapply(1:nrow(tren),function (x)
  (predict(model2,tren[,-10],interval="prediction")[x]-tren[x,10])^2))/(nrow(tren))
testerr2<-sum(sapply(1:nrow(test),function (x)
  (predict(model2,test[,-10],interval="prediction")[x]-test[x,10])^2))/(nrow(test))
r2<-summary(modelk)$adj.r.squared
```

Analogiczne błedy dla okrojonego modelu to 0.352069, 0.5817049 (błędy średniokwadratowe) oraz 0.9886545 (współczynnik R^2). Widzimy więc, że błęd treningowy jest większy a testowy mniejszy co zwiazane jest ze zjawiskiem overfittingu, im większa jest liczba parametrów które wykorzystujemy tym mniejszy jest bład treningowy ale testowy większy (oczywiscie w pewnych granicach tak jak u nas).