Raport 3

Aleksander Milach

2 December 2018

Zadanie 1

```
tc=qt(.975,10)
Fc=qf(.95,1,10)
tc^2-Fc
```

```
## [1] 8.881784e-16
```

Różnica między kwadratem tc, a fc jest bardzo bliska zeru. Zatem rzeczywiście $t_c^2 = F_c$.

Zadanie 2

A)dfM=dfE+1=1+20+1=22 obserwacje

B)Wartość estymatora σ wynosi $\sqrt{MSE} = \sqrt{\frac{SSE}{dfE}} = \sqrt{\frac{400}{20}} = 2\sqrt{5}$

C)H: $\beta_1 = 0$ K:$\beta_1 \neq 0$ Wartość statystyki testowej wynosi $F = \frac{\frac{SSM}{dfM}}{\frac{SSE}{dfE}} = \frac{\frac{100}{1}}{\frac{400}{20}} = 5$, przy H ta statystyka ma rozkład F-Snedecora z (1,20) stopniami swobody, wartość krytyczna qf(.95,1,20)=4.35 $F_c < F$, odrzucamy H na rzecz K.

D)$R^2 = \frac{SSM}{SSM+SSE} = \frac{100}{100+400} = 0.2$

E)Wartość próbkowego współczynnika korelacji między zmienną wyjaśniającą a wyjaśnianą wynosi $r = \sqrt{R^2} = 0.447$.

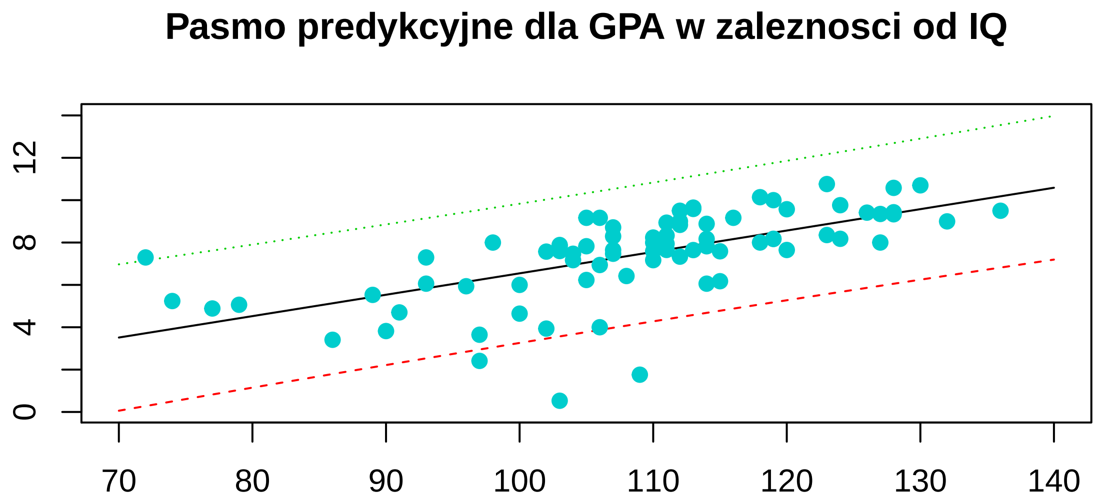
Zadanie 3

```
t=read.table(url("http://math.uni.wroc.pl/~mbogdan/Modele_Liniowe/Dane/table1_6.TXT"))
attach(t)
z3ab=function(V2,V3) {
v=c(summary(lm(V2~V3))$r.squared,
summary(lm(V2~V3))$coefficients[,1],
summary(lm(V2~V3))$coefficients[2,3:4],
summary(lm(V2~V3))$sigma^2)
names(v)=c('r2','b0','b1','t','p-wartosc','wariancja')
v}
z3ab(V2,V3)
```

```
##           r2            b0            b1             t     p-wartosc
##  4.016146e-01 -3.557056e+00  1.010217e-01  7.142020e+00  4.737341e-10
##     wariancja
##  2.672475e+00
```

```
P100l=predict(lm(V2~V3),data.frame(V3=100),interval="prediction",level=.9)[2]
P100p=predict(lm(V2~V3),data.frame(V3=100),interval="prediction",level=.9)[3]
```

```r
matplot(70:140,predict(lm(V2~V3),data.frame(V3=70:140),interval="prediction",level=.95),type="l",xlab="
        main="Pasmo predykcyjne dla GPA w zaleznosci od IQ")
points(V2~V3,col='cyan3',pch=19)
```



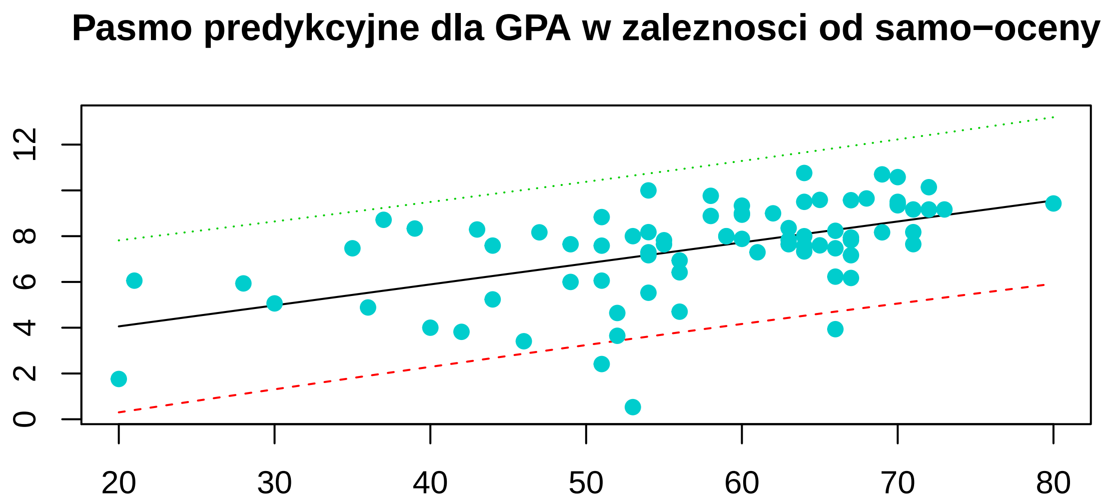
P-wartość jest bardzo mała, zatem dla dowolnego rozsądnego poziomu istotności odrzucamy hipotezę o braku korelacji między IQ a GPA. 90% przedział predykcyjny dla wartości GPA, gdy IQ wynosi 100 to [3.7975299;9.2926982]. Poza pasmo predykcyjne wypadają 4 co wynosi około 5% z 78 obserwacji.

Zadanie 4

```r
z3ab(V2,V5)
```

```
##           r2           b0           b1            t    p-wartosc
## 2.935829e-01 2.225883e+00 9.165230e-02 5.620068e+00 3.006416e-07
##    wariancja
## 3.154960e+00
```

```r
P60l=predict(lm(V2~V5),data.frame(V5=60),interval="prediction",level=.9)[2]
P60p=predict(lm(V2~V5),data.frame(V5=60),interval="prediction",level=.9)[3]

matplot(20:80,predict(lm(V2~V5),data.frame(V5=20:80),interval="prediction",level=.95),type="l",xlab="",
        main="Pasmo predykcyjne dla GPA w zaleznosci od samo-oceny")
points(V2~V5,col='cyan3',pch=19)
```



P-wartość jest bardzo mała, zatem dla dowolnego rozsądnego poziomu istotności odrzucamy hipotezę o braku korelacji między samo-oceną a GPA. 90% przedział predykcyjny dla wartości GPA, gdy samo-ocena wynosi 60 to [4.7473017;10.7027392]. Poza pasmo predykcyjne wypadają 3 obserwacje co wynosi około 5% z 78 obserwacji.

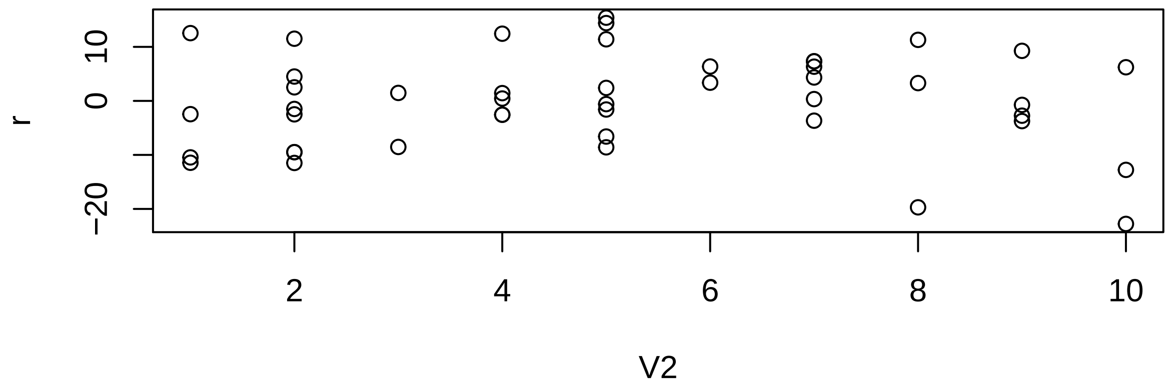
GPA lepiej wyjaśnia IQ, świadczy o tym między innymi większe R^2 i mniejsza wariancja.

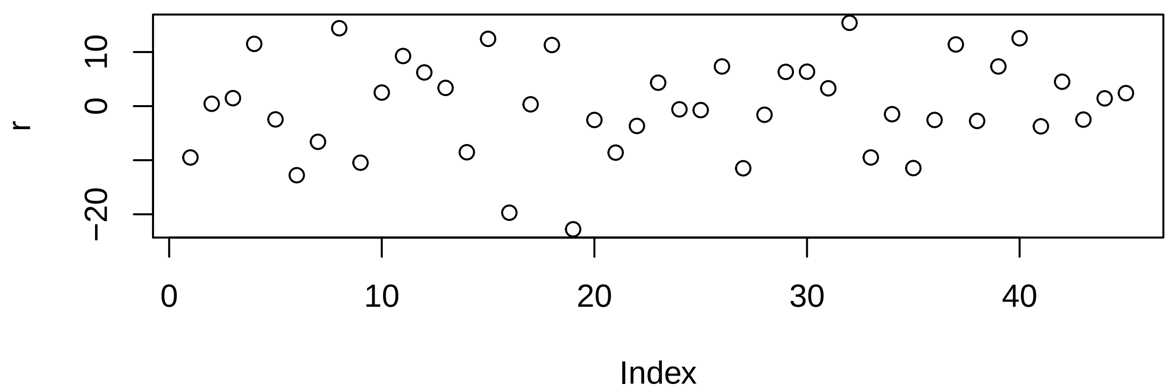
Zadanie 5

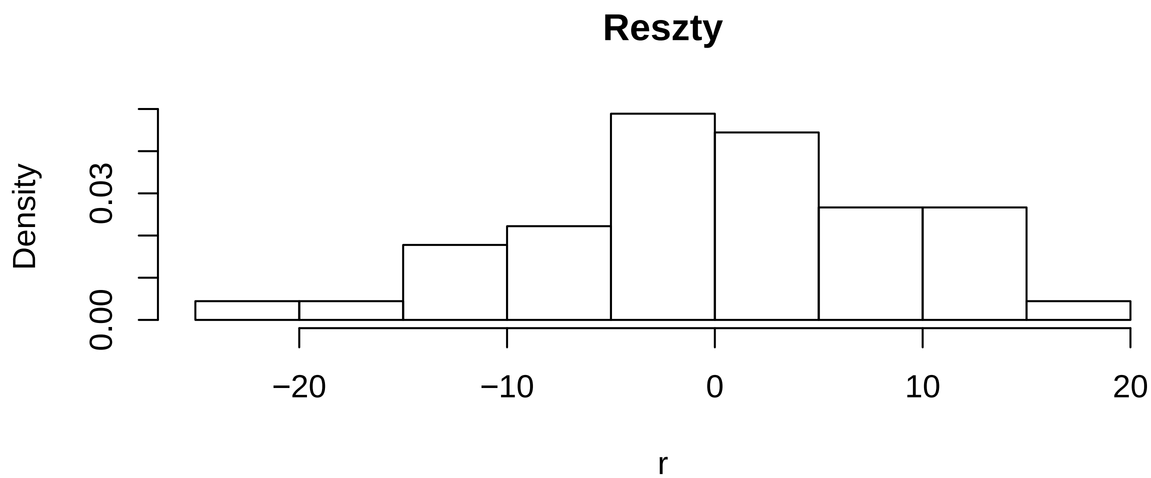
```
s=read.table("http://math.uni.wroc.pl/~mbogdan/Modele_Liniowe/Dane/CH01PR20.txt")
detach(t)
attach(s)

r=summary(lm(V1~V2))$residuals
sum(summary(lm(V1~V2))$residuals)
```

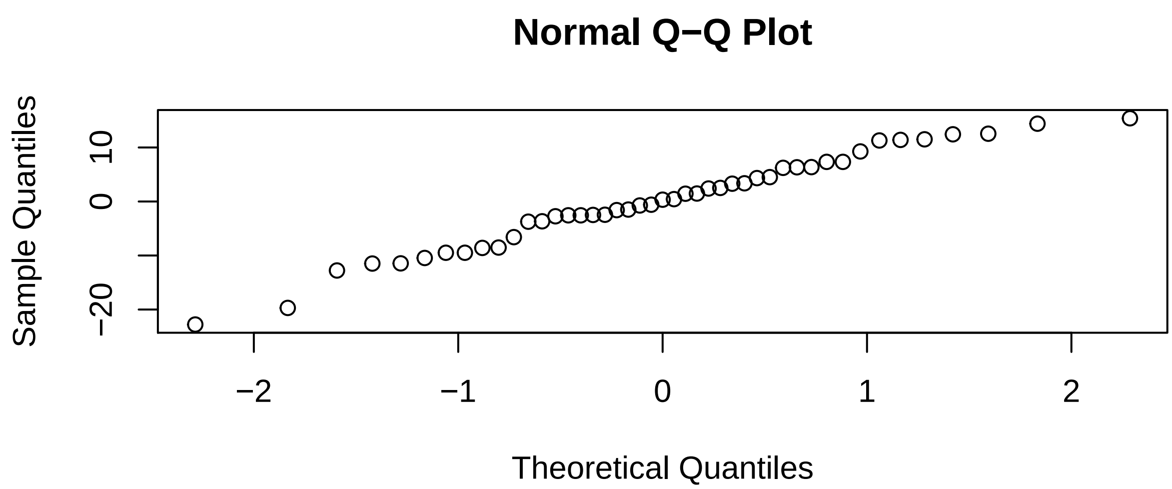
```
## [1] -1.176836e-14
```

```
z5bcd=function(r){
plot(r~V2)
plot(r)
hist(r,freq=F,main="Reszty")
qqnorm(r)
}
z5bcd(r)
```


r
−20
0
10
2
4
6
8
10
V2


r
−20
0
10
0
10
20
30
40
Index





Suma reszt jest, zgodnie z oczekiwaniami bardzo bliska 0. Z wykresów reszt wynika brak wyraźnych zależności między resztami. Nie mamy podstaw do odrzucenia założenia, że są one niezależne i z tego samego rozkładu normalnego.

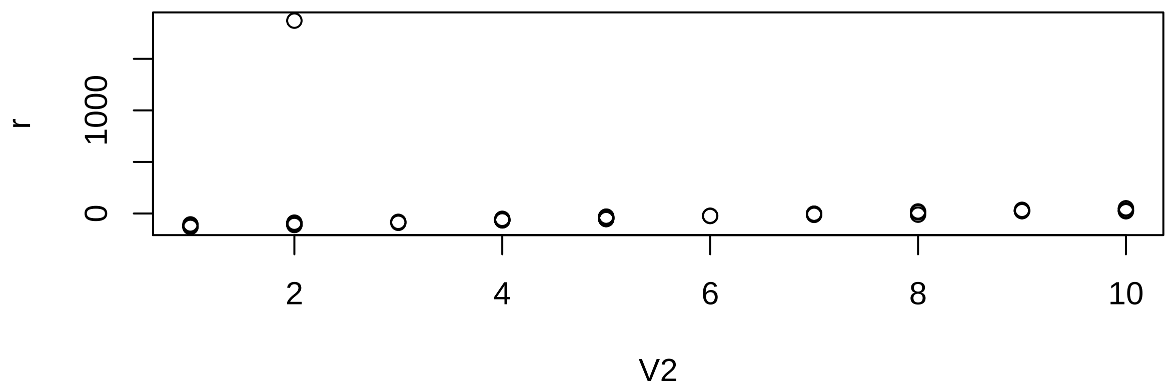
Zadanie 6

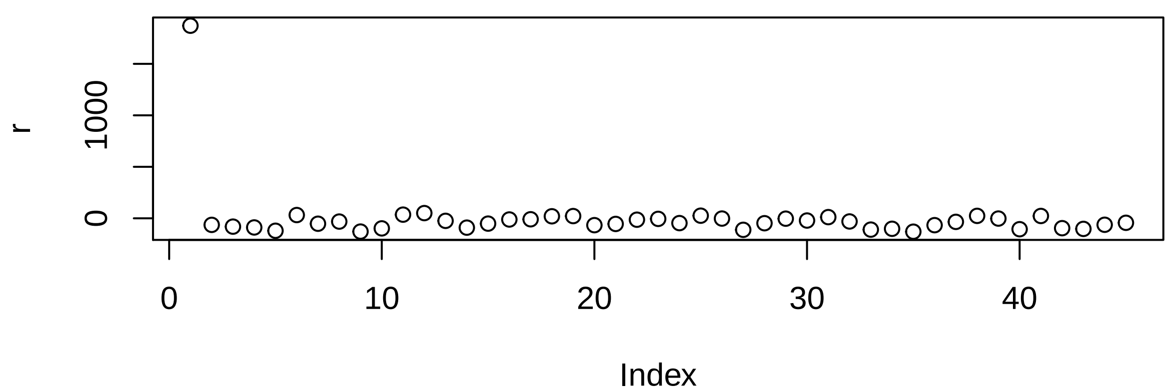
```
q=s
q[1,1]=2000

detach(s)
attach(q)

z3ab(V1,V2)
```

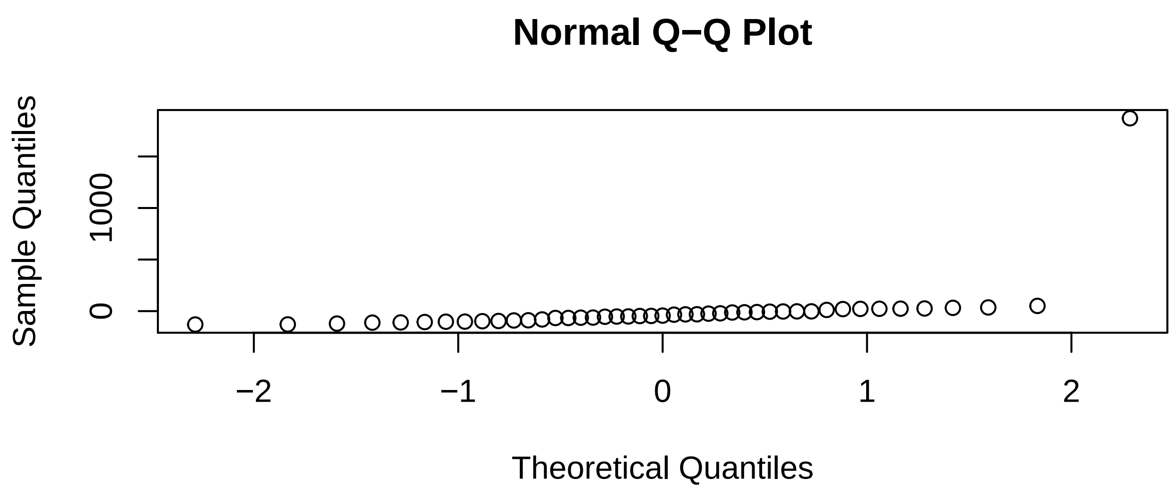
```
##             r2            b0            b1             t     p-wartosc
##   8.629944e-04  1.359003e+02 -3.058747e+00 -1.927195e-01  8.480860e-01
```

```
##      wariancja
##  8.575943e+04
```

```r
r2=summary(lm(V1~V2))$residuals
z5bcd(r2)
```







W porównaniu do poprzedniego zadania znacznie zmniejszyło się R^2 i wartość statystyki testowej, zauważalnie wzrosły p-wartość i wariacja. Zmienił się znak estymatora b1; nie mamy podstaw do odrzucania hipotezy o braku korelacji między zmiennymi.

Wartość reszty rośnie wraz ze wzrostem wartości zmiennej objaśniającej. Histogram nie przypomnia kształtem gęstości rozkładu normalnego. Na każdym z wykresów widzimy obserwację odstającą, która jest przyczyną niepożądanych wyników.

Zadanie 7

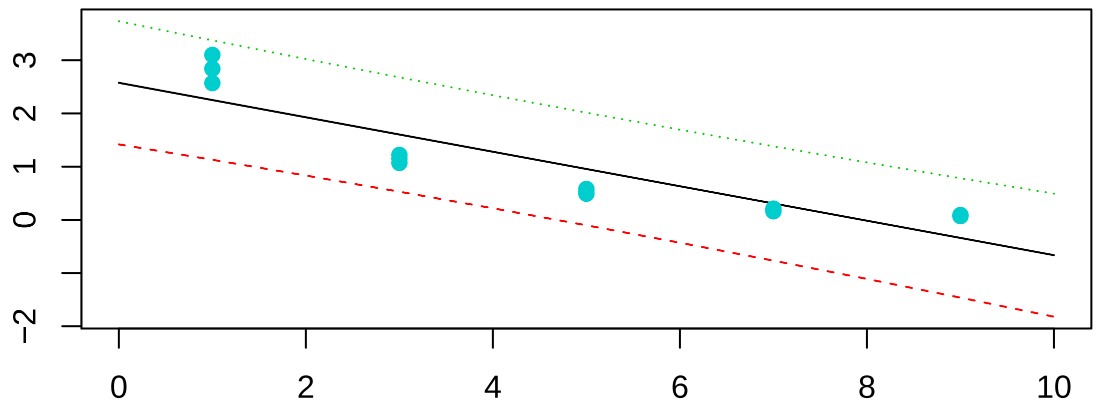
```
u=read.table("http://math.uni.wroc.pl/~mbogdan/Modele_Liniowe/Dane/CH03PR15.txt")

detach(q)
attach(u)
n=length(V1)
z3ab(V1,V2)
```

```
##            r2           b0           b1            t    p-wartosc
##  8.115774e-01  2.575333e+00 -3.240000e-01 -7.482903e+00  4.611199e-06
##     wariancja
##  2.249733e-01
```

Załóżmy, że $V1 = \beta_0 + \beta_1 V2 + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, ϵ_i są niezależne. Rozważmy hipotezę H:$\beta_1 = 0$ przeciw K:$\beta_1 \neq 0$. Przy H, statystyka testowa t ma rozkład Studenta z 13 stopniami swobody. Wartość statystyki jest duża i p-wartość jest mała, zatem dla dowolnego rozsądnego poziomu istotności odrzucamy H na rzecz K. Duża wartość R^2, stosunkowo mała p-wartość i wariancja (jak się poniżej okaże), pozornie świadczą o ładnej linowej zależności między zmiennymi.

Zadanie 8

```r
matplot(0:10,predict(lm(V1~V2),data.frame(V2=0:10),interval="prediction",level=.95),type="l",xlab="",yla
points(V1~V2,col='cyan3',pch=19)
```

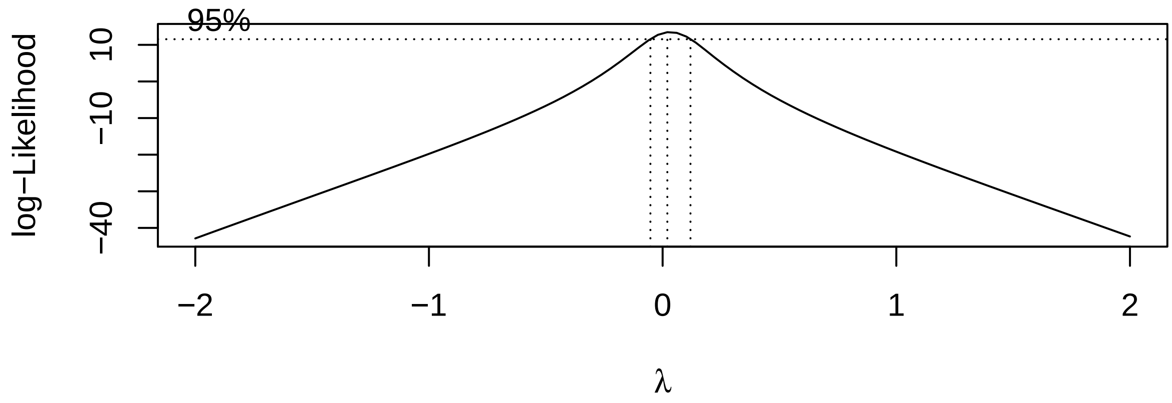



```r
wspkor1=cor(V1,predict(lm(V1~V2)))
```

Pasmo predykcyjne zawiera wszystkie obserwacje, ale obserwacje nie układają się w prostą, reszty maleją ze wzrostem V2. Wartość współczynnika korelacji między wyestymowanymi wartościami V1, a wartościami rzeczywistymi wynosi 0.9008759.

Zadanie 9

```r
library(MASS)
boxcox(lm(V1~V2))
```



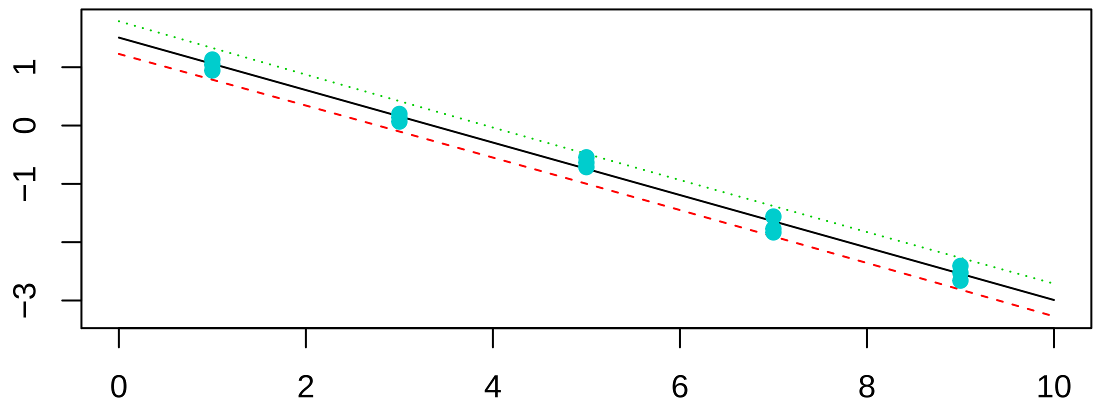
Maksimum funkcji pokazanej na wykresie jest bardzo blisko zera, zatem użyjemy $\lambda = 0$.

Zadanie 10

```
logy=log(V1)

z3ab(logy,V2)
```

```
##            r2            b0            b1             t     p-wartosc
##  9.929776e-01  1.507916e+00 -4.499258e-01 -4.287453e+01  2.188252e-15
##     wariancja
##  1.321492e-02
```

```
matplot(0:10,predict(lm(logy~V2),data.frame(V2=0:10),interval="prediction",level=.95),type="l",xlab="",y
points(logy~V2,col='cyan3', pch=19)
```




```
cor(logy,predict(lm(logy~V2)))
```

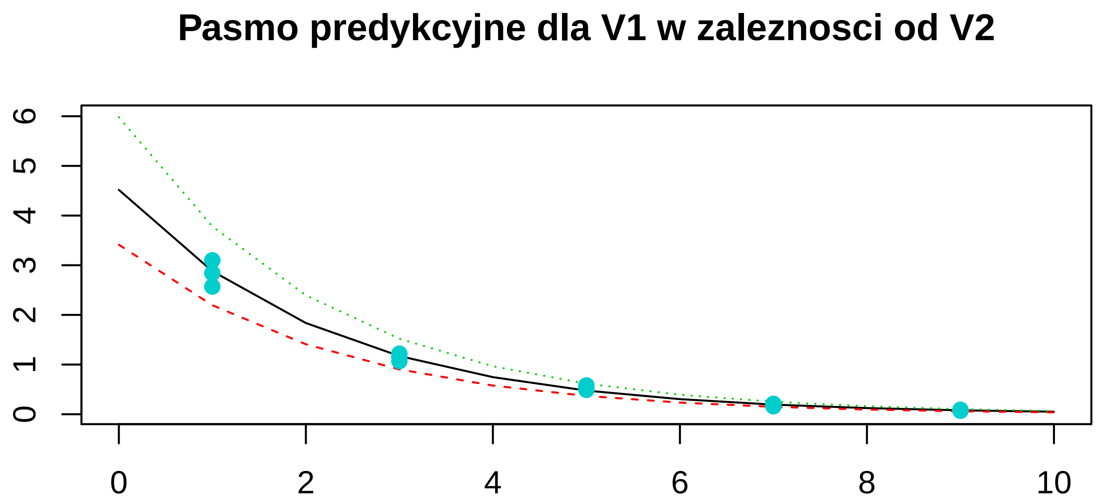
```
## [1] 0.9964826
```

Załóżmy, że $V1 = \beta_0 + \beta_1 V2 + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, ϵ_i są niezależne. Rozważmy hipotezę H:$\beta_1 = 0$ przeciw K:$\beta_1 \neq 0$. Przy H, statystyka testowa t ma rozkład Studenta z 13 stopniami swobody. Wartość statystyki jest duża i p-wartość jest mała, zatem dla dowolnego rozsądnego poziomu istotności odrzucamy H na rzecz K.

Wartość R^2 oraz wartość współczynnika korelacji między estymatorami a obserwacjami przekraczają 0.99, dane układają się w prostą, pasmo predykcyjne zawiera wszystkie obserwacje.

Zadanie 11

```
matplot(0:10,exp(predict(lm(logy~V2),data.frame(V2=0:10),interval="prediction",level=.95)),type="l",xlab
        main="Pasmo predykcyjne dla V1 w zaleznosci od V2")
points(V1~V2,col='cyan3',pch=19)
```




```r
cor(V1,exp(predict(lm(logy~V2))))
```

```
## [1] 0.9945587
```

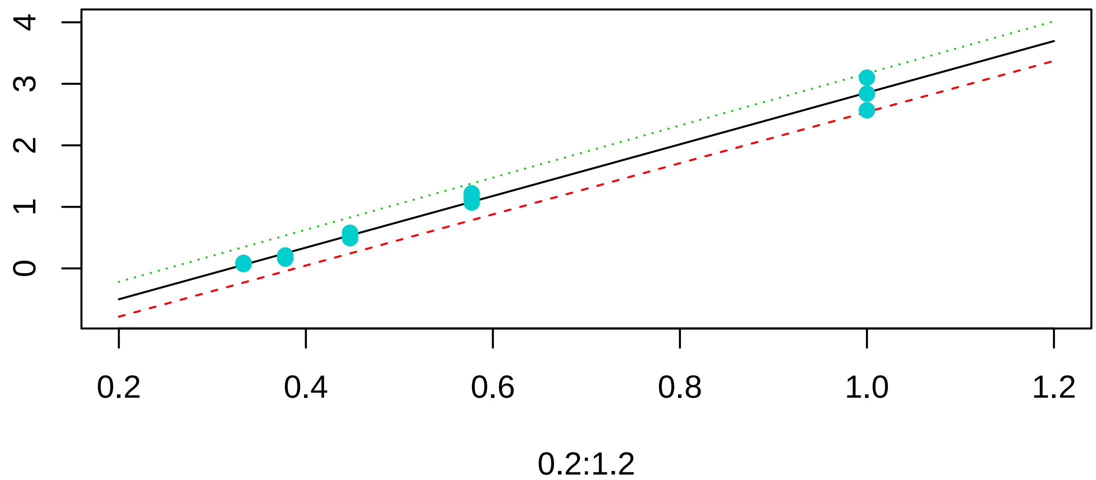
Skoro po zlogarytmowaniu V1 otrzymaliśmy zależność linową, to lepsze pasmo predykcyjne niż w zadaniu 8 uzyskamy przekształcając pasmo z zadania 10 przez funkcję e^x. Wartość współczynnika korelacji przekracza 0.99, pasmo predykcyjne zawiera wszystkie obserwacje, ale co ważniejsze jest o wiele węższe niż w zadaniu 8.

Zadanie 12

```r
t1=V2^(-1/2)
z3ab(V1,t1)
```

```
##            r2            b0            b1             t     p-wartosc
##  9.880630e-01 -1.340777e+00  4.196319e+00  3.280320e+01  6.897696e-14
##     wariancja
##  1.425258e-02
```

```r
matplot(0.2:1.2,predict(lm(V1~t1),data.frame(t1=0.2:1.2),interval="prediction",level=.95),type="l",ylab=
points(V1~t1,col='cyan3',pch=19)
```




```
cor(V1,predict(lm(V1~t1)))
```

```
## [1] 0.9940136
```

Załóżmy, że $V1 = \beta_0 + \beta_1 t1 + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, ϵ_i są niezależne. Rozważmy hipotezę H:$\beta_1 = 0$ przeciw K:$\beta_1 \neq 0$. Przy H, statystyka testowa t ma rozkład Studenta z 13 stopniami swobody. Wartość statystyki jest duża i p-wartość jest mała, zatem dla dowolnego rozsądnego poziomu istotności odrzucamy H na rzecz K.

Wartość R^2 oraz wartość współczynnika korelacji między estymatorami a obserwacjami przekraczają 0.99, dane układają się w prostą, pasmo predykcyjne zawiera wszystkie obserwacje.