Modele Liniowe
Raport nr 4

Michał Kubica

12 marca 2019

1 Zadanie 1

1.1 a)

```
alfa = 0.05
df = 10
tc = qt(1-alfa/2, df)
```

1.2 b)

```
Fc = qf(1-alfa, 1, df)
```

1.3 c)

```
abs(tc^2-Fc)
```

$8.88178410^{-16} \approx 0$

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{residualy_i - \hat{y}_i}{1-h_i}\right)^2$$

2 Zadanie 2

	df	*SS*
Model	1	100
Error	20	400

2.1 a)

20+2 = 22 obseracje

2.2 b)

$s^2 = \frac{SST}{dfT} = \frac{500}{21} \approx 23.8$

2.3 c)

H_0 wariancje obu populacji są równe

$F = \frac{MSM}{MSE} = \frac{100}{20} = 5$ $F \sim F(1, 20)$ $F_{kryt}(0.05, 1, 20) = 4.351$ Skoro $F > F_{kryt}(0.05, 1, 20)$ zatem odrzucamy hipotezę zerową o niezależności.

2.4 d)

$R^2 = \frac{SSM}{SST} = 0.2 = 20\%$

2.5 e)

$|r| = \sqrt{\frac{SSM}{SST}} = \sqrt{0.2}$

3 Zadanie 3

3.1 a)

Dopasowany model regresji:

$$Y = 0.10102X - 3.55706$$

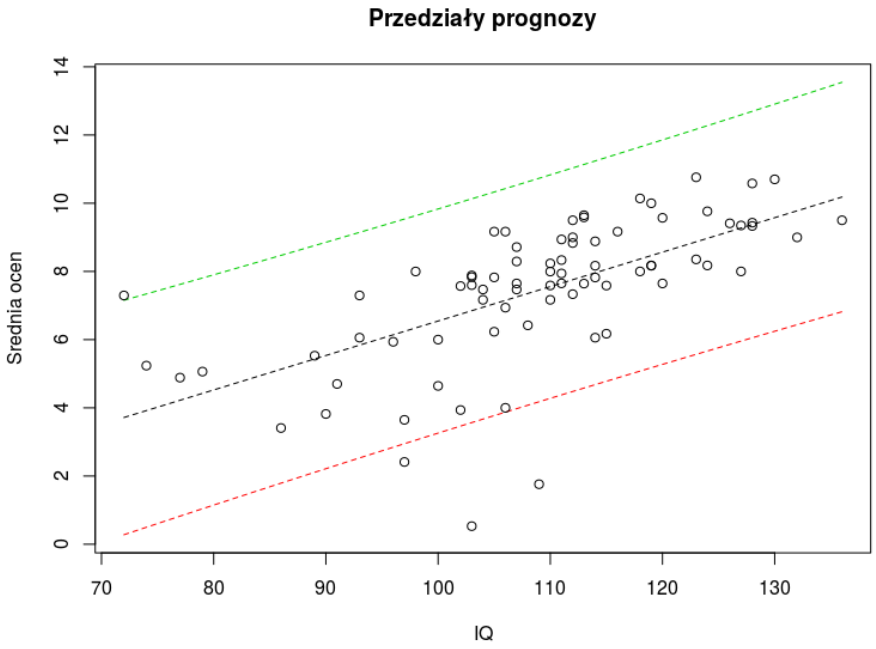
, gdzie X to zmienna objaśniająca(IQ) a Y zmienna objaśniana(GPA)
$R^2 = 0.4016146$

Przetestowano hipotezę H_0 : średnia ocen nie zależy od IQ, tzn $b_0 = 0$ wartość statystyki $t = 7.142$, p-wartość $= 4.7410^{-10}$. Wyciągnięto wniosek, że nawet przy małym poziomie istotności odrzucamy hipotezę zerową. To znaczy średnia ocen zdecydowanie zależy od poziomu inteligencji uczniów.

3.2 b)

Prognoza: 3.79753
Uzyskany 90% przedział prognozy: $(6.545114, 9.292698)$

3.3 c)



Rysunek 1

Cztery obserwacje znajdują się poza zynaczonymi przedziałami.

4 Zadanie 4

4.1 a)

Dopasowany model regresji:

$$Y = 0.09165X + 2.22588$$

, gdzie X to zmienna objaśniająca(wyniki testu Piersa-Harrisa) a Y zmienna objaśniana(GPA - średnia ocen)
$R^2 = 0.2935829$

4.2 b)

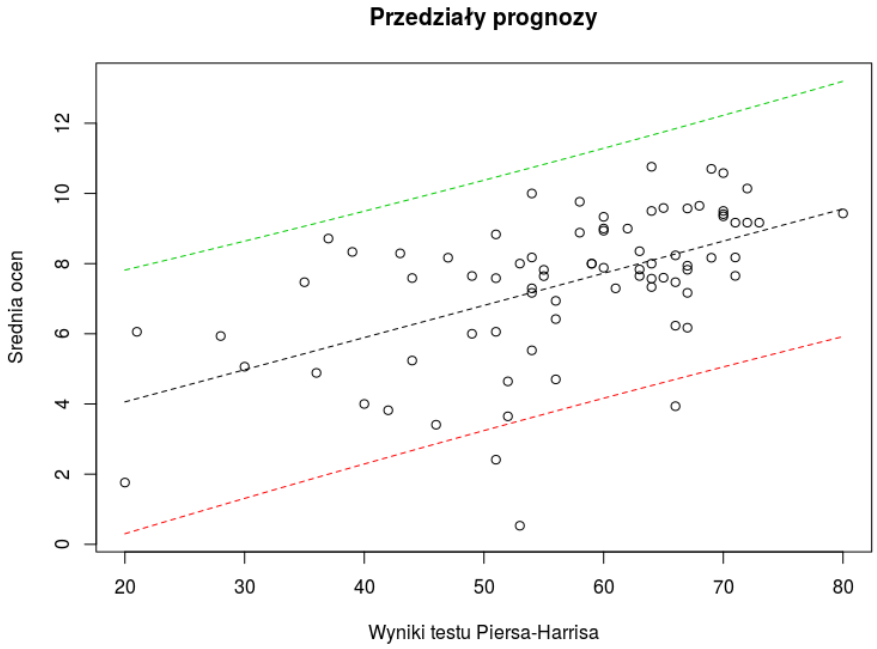
Przetestowano hipotezę H_0 : średnia ocen nie zależy od wyników testu Piersa-Harrisa, tzn $b_0 = 0$

wartość statystyki $t = 5.620$, p-wartość $= 3.0110^{-7}$. Wyciągnięto wniosek, że nawet przy małym poziomie istotności odrzucamy hipotezę zerową. To znaczy średnia ocen zdecydowanie zależy od wyników testu Piersa-Harrisa uczniów.

4.3 c)

Prognoza: 4.747302
Uzyskany 90% przedział prognozy: $(7.72502, 10.70274)$

4.4 d)



Rysunek 2

Trzy obserwacje znajdują się poza zynaczonymi przedziałami.

4.5 e)

Lepszym predyktorem średniej ocen jest poziom inteligencji.

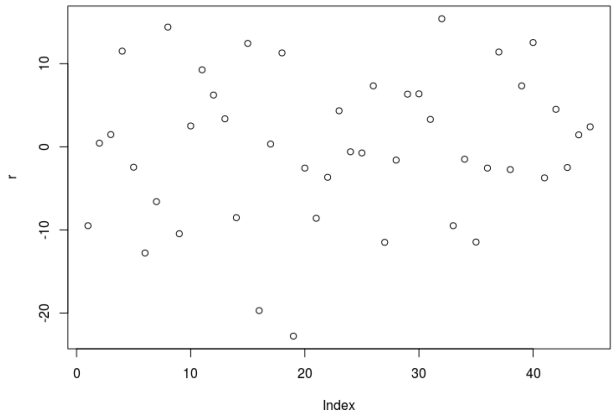
5 Zadanie 5

5.1 a)

$6.883383e-15$

5.2 b)

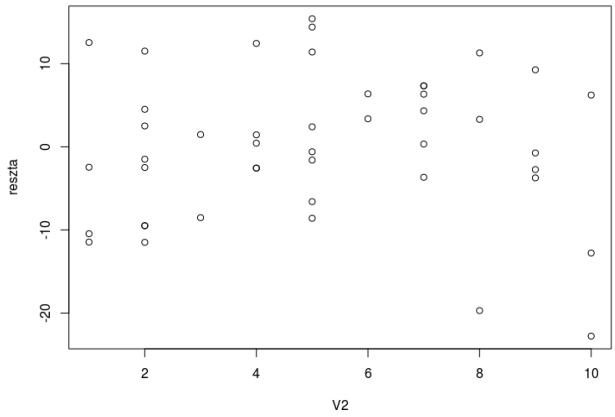
Na poniższym wykresie nie widać żadnych zależności między zmienną objaśniającą a resztami.



Rysunek 3

5.3 c)

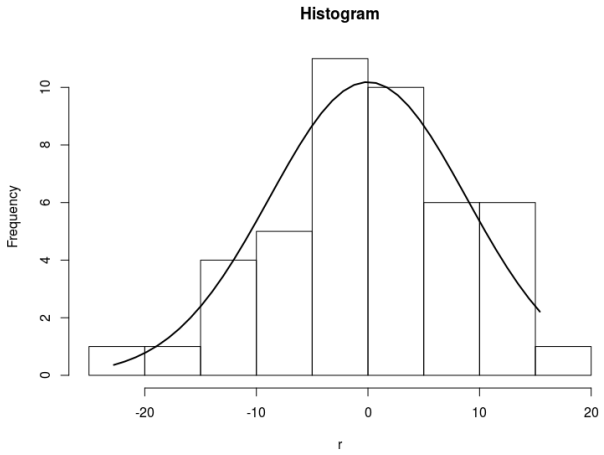
Na poniższym wykresie nie widać żadnych zależności między zmienną objaśniającą a resztami.



Rysunek 4

5.4 d)

Wyznaczono histogram reszt oraz na jego tle wyrysowano wykres gęstości rozkładu normalnego. Ze względu na widoczne pdobieństwo można stwierdzić, że reszty pochodzą z rozkładu normalnego, co zgadza się z założeniami modelu regresji liniowej.



Rysunek 5

6 Zadanie 6

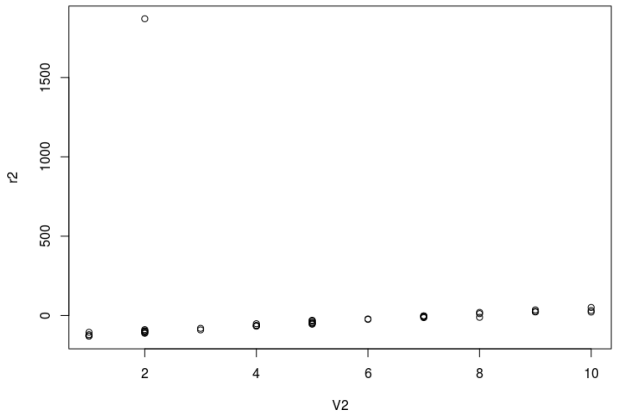
6.1 a)

Tabela 1: Porównanie uzyskanych modeli

	b_0	b_1	p-wartość	R^2	σ^2
Oryginalny model	-0.5801567	15.035248	$4.009032 \cdot 10^{-31}$	0.9574954835	79.45063
Zmieniony model	135.9002611	-3.058747	0.8480860	0.0008629944	85759.43314

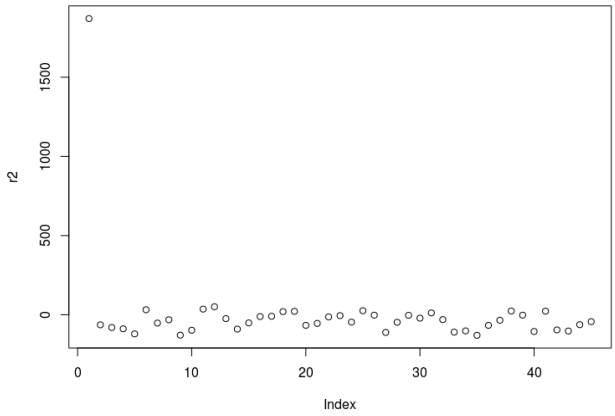
Na podstawie tabeli stwierdzono, że wprowadzenie do danych błędu grubego wpłynęło na wszystkie powyższe parametry. Dopasowana prosta regresji została odchylona. Estymator wariancji(miary rozrzutu danych) wzrósł bardzo mocno. P-wartość także wzrosła, do takiego poziomu, że przy testowaniu hipotezy zerowej o niezależności danych, musialibyśmy ją przyjąć. Podczas gdy w modelu oryginalnym jesteśmy pewni, że między danymi istnieje liniowa zależność. W związku z tym także drastycznie zmalał współczynnik korelacji R^2.

6.2 b)



Rysunek 6

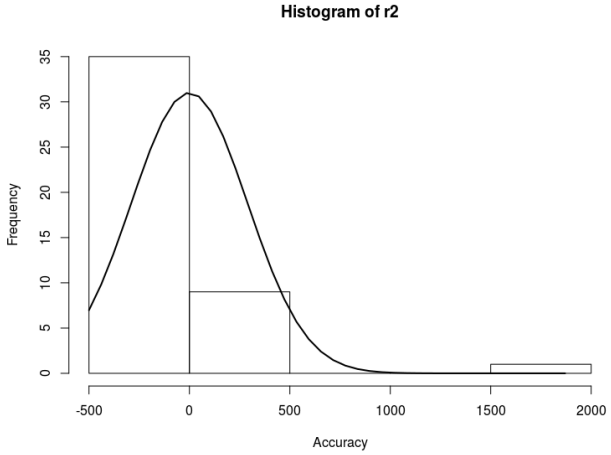
6.3 c)



Rysunek 7

6.4 d)

Wyznaczono histogram reszt oraz na jego tle wyrysowano wykres gęstości rozkładu normalnego.



Rysunek 8

Na każdym z powyższych wykresów widać wyraźnie jedną obserwację odstającą.

7 Zadanie 7

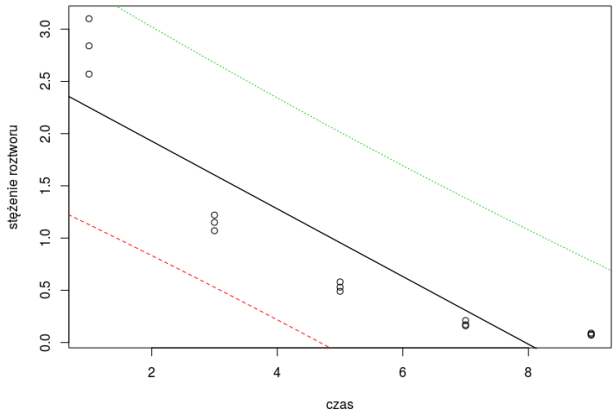
Dopasowany model regresji:

$$Y = -0.3240 + 2.5753$$

$R^2 = 0.7971$
Przetestowano następującą hipotezę:
H_0: stężenie roztworu nie zalezy od czasu
H_A: stężenie roztworu zależy od czasu
Otrzymana p-wartość: $4.61 \cdot 10^{-6}$.
Zatem przy nawet bardzo małym poziomie istotności odrzucamy hipotezę zerową na rzecz alternatywnej. Stąd wyciągamy wniosek, że stężenie roztworu zależy od czasu.

8 Zadanie 8

Narysowano dane, dodano regresję oraz przedziały prognozy.

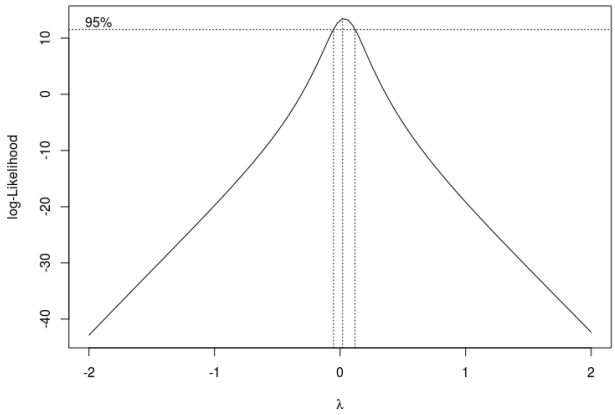


Rysunek 9

Oraz wyznaczono współczynnik korelacji między obserwacją a prognozą stężenia roztworu: 0.9008759

9 Zadanie 9

Na wykresie z Zadania 8 można zauważyć, że możemy poszukać transformacji zmiennej objaśnianej tak, zeby lepiej dopasować prostą regresji. Do tego służy transformacja Boxa-Coxa.



Rysunek 10

Na podstawie wykresu zauważono, że najlepszą transformacją będzie funkcją logarytm (współczynnik $\lambda \approx 0$)

10 Zadanie 10

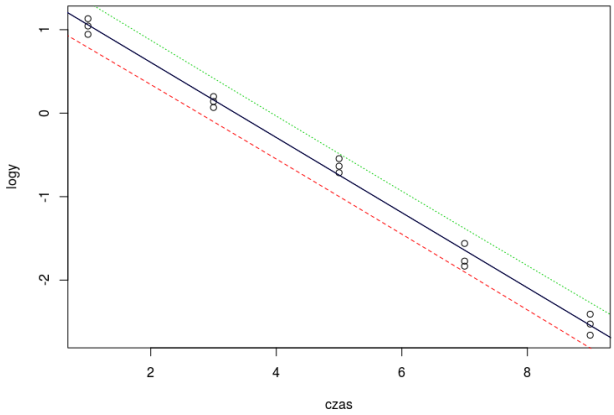
Przekształcono zmienną objaśnianą a następnie wykonano analizę podobną jak w zadaniu 7.

Dopasowany model regresji:

$$Y = -0.44993X + 1.50792$$

$R^2 = 0.9924$
Przetestowano następującą hipotezę:
H_0: logarytm wartości stężenia roztworu nie zalezy od czasu
H_A: logarytm wartości stężenia roztworu zależy od czasu
Otrzymana p-wartość: $2.19 \cdot 10^{-15}$.
Zatem przy nawet bardzo małym poziomie istotności odrzucamy hipotezę zerową na rzecz alternatywnej. Stąd wyciągamy wniosek, że stężenie roztworu zależy od czasu.

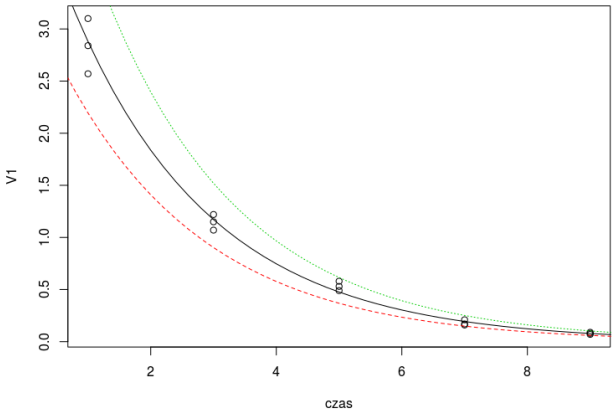
Narysowano dane, dodano regresję oraz przedziały prognozy.



Rysunek 11

Oraz wyznaczono współczynnik korelacji między obserwacją a prognozą logarytmem z wartości stężenia roztworu: 0.9964826

11 Zadanie 11



Rysunek 12

Współczynnik korelacji między zaobserwowanym stężeniem a prognozą z zadania 10: 0.9008759

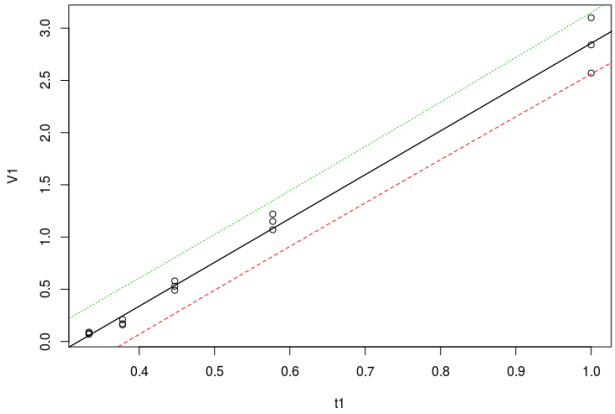
12 Zadanie 12

Przekształcono zmienną objaśniającą ($t1 = t^{-\frac{1}{2}}$)a następnie wykonano analizę podobną jak w zadaniu 7.

Dopasowany model regresji:

$$Y = 4.19632X - 1.34078$$

$R^2 = 0.9871$
Przetestowano następującą hipotezę:
H_0: wartość stężenia roztworu nie zalezy od czasu
H_A: wartość stężenia roztworu zależy od czasu
Otrzymana p-wartość: $6.9 \cdot 10^{-14}$.
Zatem przy nawet bardzo małym poziomie istotności odrzucamy hipotezę zerową na rzecz alternatywnej. Stąd wyciągamy wniosek, że stężenie roztworu zależy od czasu.



Rysunek 13

Współczynnik korelacji: 0.9940136

Porównując R^2 i p-wartości najlepszym modelem okazał się model z podpunktu 10. Niewiele gorszy okazał się model powyższy.

13 Kod R

```
## ZAD 1 ##
X=runif(100)
Y=1+.3*X+rnorm(100)
```

```
summary(lm(Y~X))
abs(summary(lm(Y~X))$coefficients[2,3]^2-
summary(lm(Y~X))$fstatistic[1])

qf(0.05,1,20)

## ZAD 3 ##

# a
t=read.table(url("http://www.math.uni.wroc.pl/~mbogdan/Modele_Liniowe/Dane/table1_6.TXT"))
head(t)
attach(t)
summary(lm(V2~V3))
summary(lm(V2~V3))$r.squared
plot(V2~V3)
abline(lm(V2~V3), col='red')

# b
predict(lm(V2~V3),data.frame(V3=100),interval="prediction",level=0.9)

# c
matplot(min(t$V3):max(t$V3),predict(lm(V2~V3),data.frame(V3=min(t$V3):max(t$V3)),interval="prediction"
    main="Przedziay prognozy", xlab="IQ", ylab="Srednia ocen")
points(V3,V2)

max(t$V5)
min(t$V5)

## ZAD 4 ##

# a i b
summary(lm(V2~V5))
summary(lm(V2~V5))$r.squared

# c
predict(lm(V2~V5),data.frame(V5=60),interval="prediction",level=.9)

# d
matplot(min(t$V5):max(t$V5),predict(lm(V2~V5),data.frame(V5=min(t$V5):max(t$V5)),interval="prediction"
    main="Przedziay prognozy", xlab="Wyniki testu Piersa-Harrisa",
    ylab="Srednia ocen")
points(V5,V2)

# e
summary(lm(V2~V3))
summary(lm(V2~V5))


## ZAD 5 ##
```

```r
t=read.table(url("http://www.math.uni.wroc.pl/~mbogdan/Modele_Liniowe/Dane/CH01PR20.txt"))
head(t)
attach(t)
plot(t[2:1])
m1=lm(V1~V2)
abline(m1,col="red")
str(summary(m1))

# a
summary(m1)$residuals
residuals(m1)
V1-m1$fitted.values
r=V1-m1$coefficients[1]-m1$coefficients[2]*V2
sum(r)
sum(summary(lm(V1~V2))$residuals)

# b
plot(V2,r, ylab="reszta")
# c
plot(r)
# d
h <- hist(r, main= "Histogram")
xfit <- seq(min(r), max(r), length = 40)
yfit <- dnorm(xfit, mean = mean(r), sd = sd(r))
yfit <- yfit * diff(h$mids[1:2]) * length(r)

lines(xfit, yfit, col = "black", lwd = 2)

qqnorm(r)


## ZAD 6 ##

head(t)

s=t
s[1,1]=2000
heas(s)

# a
f=function(m) {v=c(m$coefficients,
summary(m)$coefficients[2,4],
summary(m)$r.squared,
summary(m)$sigma^2); names(v)=c("b0","b1","p","R2","sigma2"); v}

f(m1)

plot(s[2:1],ylim=c(-20,200))
m2=lm(s$V1~s$V2)
abline(m1,col="red")
```

```
f(m2)

View(rbind(f(m1),f(m2)))

plot(V2,residuals(m2))
identify(V2,s$V1)

# b
r2 = summary(m2)$residuals
plot(V2,r2)
plot(r2)

h <- hist(r2, xlab = "Accuracy")
xfit <- seq(-500, max(r2), length = 40)
yfit <- dnorm(xfit, mean = mean(r), sd = sd(r2))
yfit <- yfit * diff(h$mids[1:2]) * length(r2)

lines(xfit, yfit, col = "black", lwd = 2)

qqnorm(r2)

## ZAD 7 ##

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

detach(t)

detach(u)
attach(u)
m3=lm(V1~V2)
summary(m3)

plot(V1~V2, xlab = "czas", ylab="stenie roztworu")
abline(m3,col="black" )

## ZAD 8 ##

p=predict(m3,data.frame(V2=seq(0,10,.2)),interval="prediction")
matplot(seq(0,10,.2),p,add=TRUE,type="l")

cor(V1,predict(m3))
## ZAD 9 ##

library(MASS)
boxcox(m3)


## ZAD 10 ##
```

```
logy=log(V1)
m4=lm(logy~V2)
summary(m4)
plot(logy~V2, xlab="czas")
abline(m4,col="blue")

p=predict(m4,data.frame(V2=seq(0,10,.2)),interval="prediction")
matplot(seq(0,10,.2),p,add=TRUE,type="l")
cor(logy,predict(m4))


## ZAD 11 ##
plot(V1~V2, xlab="czas")
p=predict(m4,data.frame(V2=seq(0,10,.2)),interval="prediction")
matplot(seq(0,10,.2),exp(p),add=TRUE,type="l")
cor(V1,predict(m4))


## ZAD 12 ##
t1 = V2^(-1/2)
m5 = lm(V1~t1)
plot(V1~t1)
summary(m5)
abline(m5,col="black")

p=predict(m5,data.frame(t1=seq(0,1.2,.005)),interval="prediction")
matplot(seq(0,1.2,.005),p,add=TRUE,type="l")
cor(V1,predict(m5))



1-pchisq(2.09, 12)

chisq.test(c(1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0))
13*124/207
```