Raport 4

Aleksander Milach

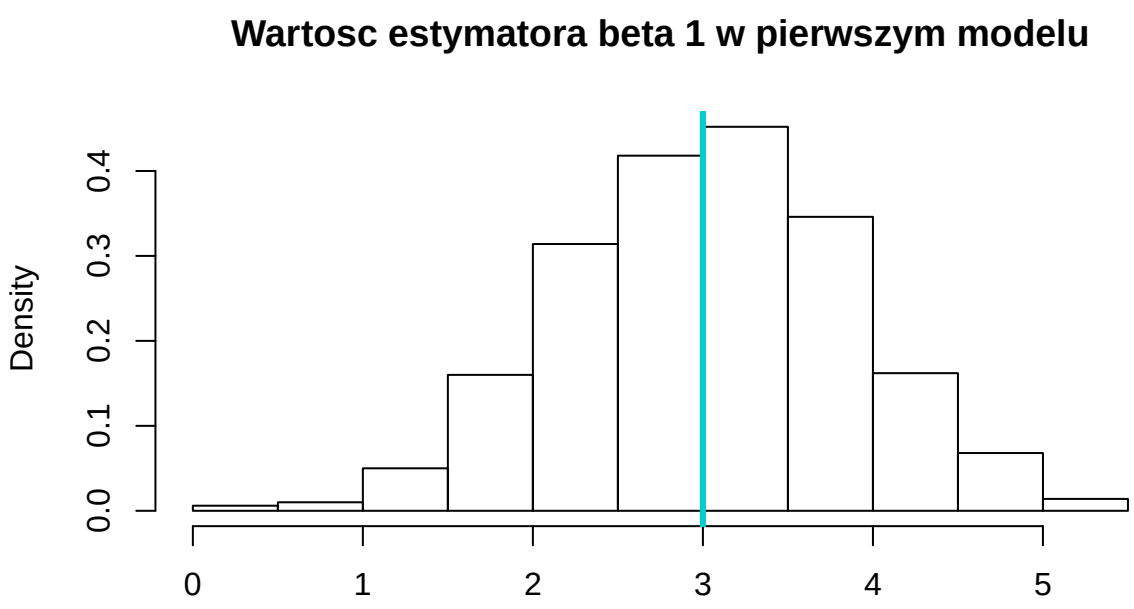
5 January 2019

Zadanie 3

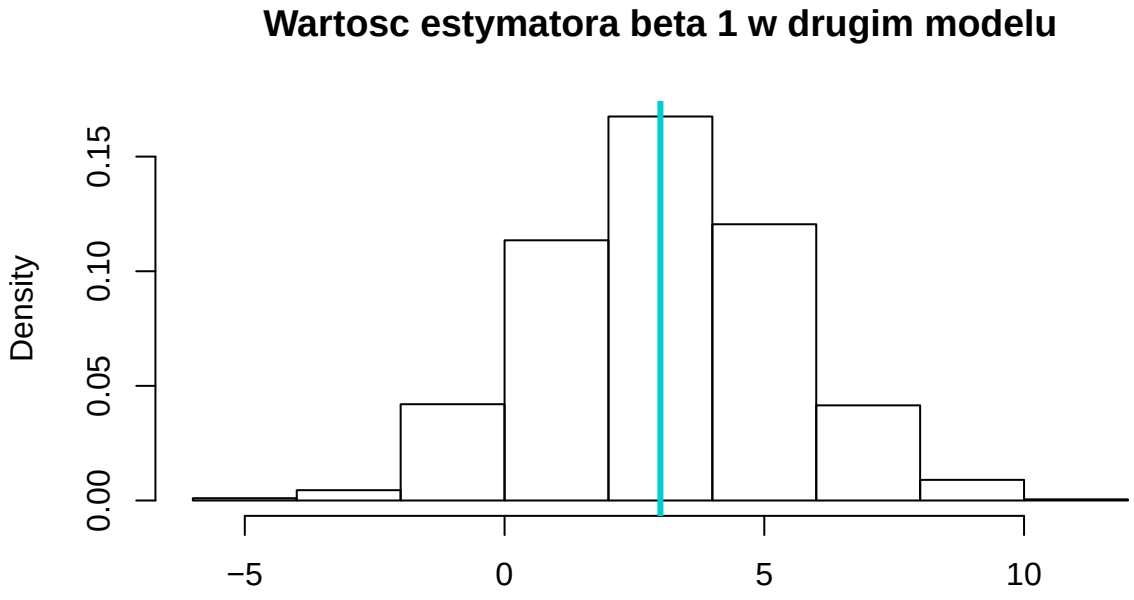
```
l1=numeric(1000)
l2=numeric(1000)
l3=numeric(1000)
l4=numeric(1000)
l5=numeric(1000)
l6=numeric(1000)

 X1=mvrnorm(100,c(0,0),matrix(c(.01,.009,.009,.01),2,2))
 X2=cbind(rep(1,100),X1)

 Y=3*X1[,1]+rnorm(100)
 m0=lm(Y~X1[,1])
 pu3bl=confint(m0)[2,1]
 pu3bp=confint(m0)[2,2]
 pwar=summary(m0)$coefficients[2,4]

 sdM1=sqrt(1*solve(t(X2[,1:2])%*%X2[,1:2])[2,2])
 sdM2=sqrt(1*solve(t(X2)%*%X2)[2,2])
 tc=qt(.975,98)
 powerM1=1-pt(tc,98,3/sdM1)+pt(-tc,98,3/sdM1)
 powerM2=1-pt(tc,98,3/sdM2)+pt(-tc,98,3/sdM2)
```

Przedział ufności dla β_1 wynosi [1.3646395,4.4793772]. P-wartość dla β_1 wynosi 0.001020683. Jest mniejsza od 0,05, zatem odrzucamy hipotezę o $\beta_1 = 0$. Zero jest w przedziale ufności wtedy, i tylko wtedy, gdy brak podstaw do odrzucenia H na rzecz K.

```
for(i in 1:1000){

  e=rnorm(100)
  Y1=3*X1[,1]+e

  m1=lm(Y1~X1[,1])
  m2=lm(Y1~X1[,1]+X1[,2])

  l1[i]=summary(m1)$coefficients[2,4]<.05
  l2[i]=summary(m2)$coefficients[2,4]<.05
  l3[i]=summary(m1)$coefficients[2,1]
  l4[i]=summary(m2)$coefficients[2,1]
  l5[i]=summary(m1)$coefficients[2,2]
  l6[i]=summary(m2)$coefficients[2,2]
}


hist(l3,freq=F,xlab="",main="Wartosc estymatora beta 1 w pierwszym modelu")
abline(v=3,col='cyan3',lwd=3)
```




```
hist(l4,freq=F,xlab="", main="Wartosc estymatora beta 1 w drugim modelu")
abline(v=3,col='cyan3',lwd=3)
```




```
M=matrix(c(mean(l1),mean(l2),mean(l5),mean(l6),powerM1,powerM2,sdM1,sdM2),2,4,byrow=T)
colnames(M)=c("Moc w M1","Moc w M2","SD w M1","SD w M2")
```

```
rownames(M)=c("Tyle wyszlo","Teoretyczna wartosc")
kable(M,format='markdown')
```

	Moc w M1	Moc w M2	SD w M1	SD w M2
Tyle wyszlo	0.9380000	0.2460000	0.8820353	2.352668
Teoretyczna wartosc	0.9195169	0.2428328	0.8835874	2.356345

Wartości teroetyczne i wyestymowane są bliskie sobie.

Zadanie 4

Podpunkt a

```
X=matrix(rnorm(950000,0,.1),1000,950)
eps=rnorm(1000)
beta=c(rep(3,5),rep(0,945))
Y=X%*%beta+eps

podpa=function (k,X){
    m=lm(Y~X[,1:k])
    if(k==1){
      w2=sum((m$fitted.values-X[,1:k]*beta[1:k])^2)
      w5=NA
    }
    else
    {
      w2=sum((m$fitted.values-X[,1:k]%*%beta[1:k])^2)
      w4=summary(m)$coefficients[2,4]
      w5=summary(m)$coefficients[3,4]
    }

  v=c(anova(m)[2,2],
  w2,
  AIC(m),
  summary(m)$coefficients[2,4],
  w5,
  sum(summary(m)$coefficients[-(1:5),4]<0.05))
  v
}

M=matrix(0,6,8)

numerki=c(1,2,5,10,50,100,500,950)
for(i in 1:8)
M[,i]=podpa(numerki[i],X)

strnumerki=c('1','2','5','10','50','100','500','950')
colnames(M)=strnumerki
rownames(M)=c('Resztowa SS','MSE','AIC','P-wartosc dla 1','P-wartosc dla 2','Falszywe odkrycia')
```

```
bestmodelA=which.min(M[3,])

kable(M,format='markdown')
```

	1	2	5	10	50	100	500	950
Resztowa SS	1344.472215	1264.309808	1004.940708	997.8248	972.66767	913.03277	526.9862711	42.2650076
MSE	5.262803	2.977045	4.883889	11.9998	37.15693	96.79183	482.8383260	967.5595896
AIC	3139.878597	3080.403433	2856.805609	2859.6995	2914.16426	2950.89356	3201.2962846	1578.0812865
P-wartosc dla 1	0.000000	0.000000	0.000000	0.0000	0.00000	0.00000	0.0000000	0.0011275
P-wartosc dla 2	NA	0.000000	0.000000	0.0000	0.00000	0.00000	0.0000001	0.0198420
Falszywe odkrycia	0.000000	0.000000	1.000000	1.0000	1.00000	6.00000	21.0000000	92.0000000

Kryterium AIC uznaje model 8 za najlepszy.

Podpunkt b

```
podpb=function(k,X)
{
mb=lm(Y~X)
K=abs(summary(mb)$coefficients[,1])
names(K)=0:950
L=sort(K,decreasing=T)
nKolej=as.integer(names(L)[-which(as.integer(names(L))==0)])
Xb=matrix(0,1000,950)
betab=numeric(950)
for (i in 1:950){
    Xb[,i]=X[,nKolej[i]]
    betab[i]=beta[nKolej[i]]
}
m=lm(Y~Xb[,1:k])
if(k==1){
  w2=sum((m$fitted.values-Xb[,1:k]*betab[1:k])^2)
}
else
{
  w2=sum((m$fitted.values-Xb[,1:k]%*%betab[1:k])^2)
}

if(which(nKolej==1)>k)
  w4=NA
else
  w4=summary(m)$coefficients[which(nKolej==1)+1,4]

if(which(nKolej==2)>k)
  w5=NA
else
  w5=summary(m)$coefficients[which(nKolej==2)+1,4]
```

```
v=c(anova(m)[2,2],w2,AIC(m),w4,w5,
sum(summary(m)$coefficients[-(which(as.integer(names(L))>0 & as.integer(names(L))<6)+1),4]<0.05))
v
}

N=matrix(0,6,8)
colnames(N)=strnumerki
rownames(N)=c('Resztowa SS','MSE','AIC','P-wartosc dla 1','P-wartosc dla 2','Falszywe odkrycia')
for(i in 1:8)
N[,i]=podpb(numerki[i],X)

kable(N,format='markdown')
```

	1	2	5	10	50	100	500	950
Resztowa SS	1483.064968	1479.631390	1396.809424	1247.53978	1096.99889	977.6260	339.5792	42.2650076
MSE	1.152955	4.586533	8.635452	21.71285	79.30132	117.7904	670.2454	967.5595896
AIC	3237.987937	3237.670062	3186.067719	3083.05051	3034.45524	3019.2490	2761.8290	1578.0812865
P-wartosc dla 1	NA	NA	NA	0.00000	0.00000	0.0000	0.0000	0.0011275
P-wartosc dla 2	NA	NA	NA	NA	NA	0.0000	0.0000	0.0198420
Falszywe odkrycia	0.000000	0.000000	0.000000	2.00000	4.00000	10.0000	260.0000	92.0000000

```
bestmodelB=which.min(N[3,])
```

Kryterium AIC uznaje model 8 za najlepszy.

Podpunkt d

```
pix1a=matrix(0,1000,8)
pix1b=matrix(0,1000,8)
foa=matrix(0,1000,8)
fob=matrix(0,1000,8)
aica=matrix(0,1000,8)
aicb=matrix(0,1000,8)

for (i in 1:1000){

  X=matrix(rnorm(950000,0,.1),1000,950)
  eps=rnorm(1000)
  Y=X%*%beta+eps

  for (j in 1:8){

    m=lm(Y~X[,1:numerki[j]])

    aica[i,j]=AIC(m)
    pix1a[i,j]=summary(m)$coefficients[2,4]<0.05
    foa[i,j]=sum(summary(m)$coefficients[-(1:5),4]<0.05)
```

```
  }

  fullm=lm(Y~X)
  K=abs(summary(fullm)$coefficients[,1])
  names(K)=0:950
  L=sort(K,decreasing=T)
  nKolej=as.integer(names(L)[-which(as.integer(names(L))==0)])
  Xb=matrix(0,1000,950)
  betab=numeric(950)
  for (p in 1:950){
      Xb[,p]=X[,nKolej[p]]
      betab[p]=beta[nKolej[p]]
  }

  for (j in 1:8){

    przestawm=lm(Y~Xb[,1:numerki[j]])

    aicb[i,j]=AIC(przestawm)
    pix1b[i,j]=summary(przestawm)$coefficients[2,4]<0.05
    fob[i,j]=sum(summary(przestawm)$coefficients
                 [-(which(as.integer(names(L))>0 & as.integer(names(L))<6)+1),4]<0.05)
  }

}

O=rbind(apply(pix1a,2,mean),
apply(foa,2,mean),
apply(pix1b,2,mean),
apply(fob,2,mean))

rownames(O)=c('Moc identyfikacji X1 w p-cie A','Falszywe odkrycia w p-cie A',
              'Moc identyfikacji X1 w p-cie B','Falszywe odkrycia w p-cie B')

colnames(O)=c('I','II','III','IV','V','VI','VII','VIII')

avgbestmodelA=mean(apply(aica,1,which.min))
avgbestmodelB=mean(apply(aicb,1,which.min))

kable(O,format='markdown')
```

	I	II	III	IV	V	VI	VII	VIII
Moc identyfikacji X1 w p-cie A	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.542
Falszywe odkrycia w p-cie A	0.000	0.000	1.000	1.274	3.199	5.750	25.652	47.209
Moc identyfikacji X1 w p-cie B	0.474	0.475	0.475	0.481	0.511	0.563	1.000	0.995
Falszywe odkrycia w p-cie B	0.120	0.195	0.445	0.904	4.974	11.518	284.024	46.750

Średni numer modelu, który w podpunkcie A kryterium AIC uznaje model 8 za najlepszy. Średni numer modelu, który w podpunkcie B kryterium AIC uznaje model 8 za najlepszy.

Zadanie 5

```
t=read.table("http://math.uni.wroc.pl/~mbogdan/Modele_Liniowe/Dane/CH06PR15.txt")
n=dim(t)[1]
p=dim(t)[2]

m3=lm(t[[4]]~t[[1]]+t[[2]]+t[[3]])

r2=summary(m3)$r.squared
wsp=summary(m3)$coefficients[,1]
fstat=summary(m3)$fstatistic
```

Równanie regresji ma postać Y=1.0532451+-0.0058605X1 +0.001928X2 +0.0301477X3. Wartość R^2 wynosi 0.5415482. H: $\beta_1 = 0 \wedge \beta_2 = 0 \wedge \beta_3 = 0$ K:$\exists_{i\in\{1,2,3\}}\ \beta_i \neq 0$ Wartość statystyki testowej wynosi 16.5375621, przy H ta statystyka ma rozkład F-Snedecora z (1,42) stopniami swobody, p-wartość wynosi 3.04e-07, toteż odrzucamy H na rzecz K.

Zadanie 6

```
P=matrix(c(summary(m3)$coefficients[2,1]-summary(m3)$coefficients[2,2]*qt(.975,n-p),
  summary(m3)$coefficients[2,1]+summary(m3)$coefficients[2,2]*qt(.975,n-p),
  summary(m3)$coefficients[3,1]-summary(m3)$coefficients[3,2]*qt(.975,n-p),
  summary(m3)$coefficients[3,1]+summary(m3)$coefficients[3,2]*qt(.975,n-p),
  summary(m3)$coefficients[4,1]-summary(m3)$coefficients[4,2]*qt(.975,n-p),
  summary(m3)$coefficients[4,1]+summary(m3)$coefficients[4,2]*qt(.975,n-p)),2,3)

P=rbind(P,summary(m3)$coefficients[2:4,3])
P=rbind(P,summary(m3)$coefficients[2:4,4]>0.05)
P=rbind(P,apply(P,2,function(v){v[1]<0 & v[2]>0}))

rownames(P)=c("Lewy koniec PU","Prawy koniec PU",
              "Statystyka testowa","Przyjmujemy H?","Czy 0 jest w PU?")
kable(P,format='markdown')
```

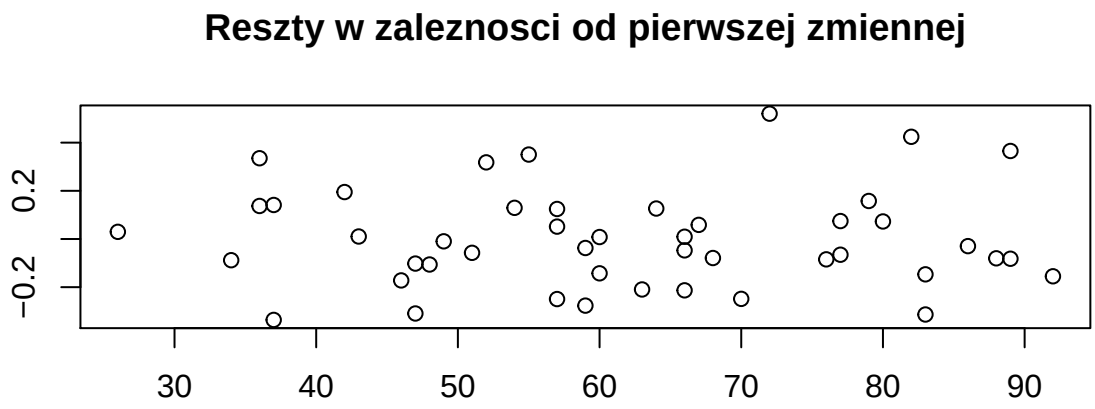
	t[[1]]	t[[2]]	t[[3]]
Lewy koniec PU	-0.0120941	-0.0097499	0.0114672
Prawy koniec PU	0.0003731	0.0136060	0.0488283
Statystyka testowa	-1.8972967	0.3331876	3.2568922
Przyjmujemy H?	1.0000000	1.0000000	0.0000000
Czy 0 jest w PU?	1.0000000	1.0000000	0.0000000

Dla i=1,2,3: H: $\beta_i = 0$ K:$\beta_i \neq 0$; Przy H statystyka testowa ma rozkład t-Studenta z 42 stopniami swobody, odrzucamy H na rzecz K dla i=3. Dla i=1,2 brak podstaw do odrzucenia H na rzecz K.

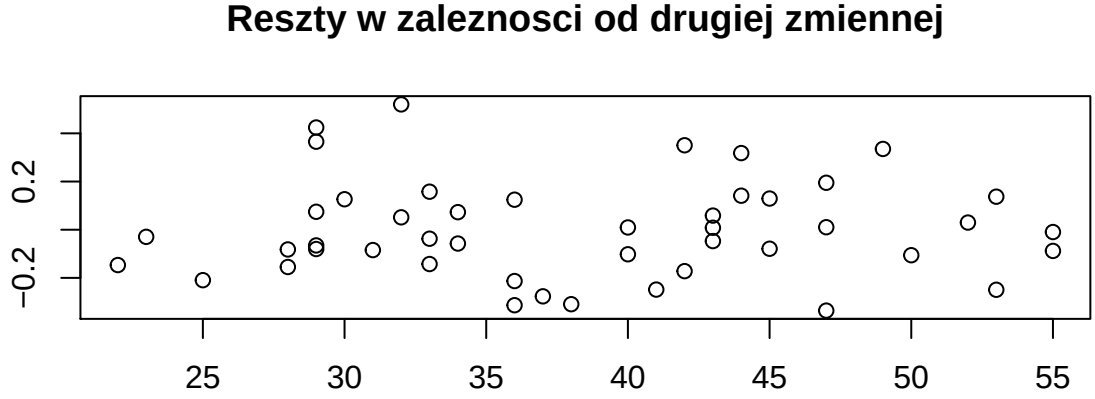
Zero jest w przedziale ufności wtedy, i tylko wtedy, gdy brak podstaw do odrzucenia H na rzecz K.

zadanie 7

```
plot(summary(m3)$residuals~t[[1]],xlab="",ylab="",
     main="Reszty w zależności od pierwszej zmiennej")
```




```
plot(summary(m3)$residuals~t[[2]],xlab="",ylab="",
     main="Reszty w zależności od drugiej zmiennej")
```



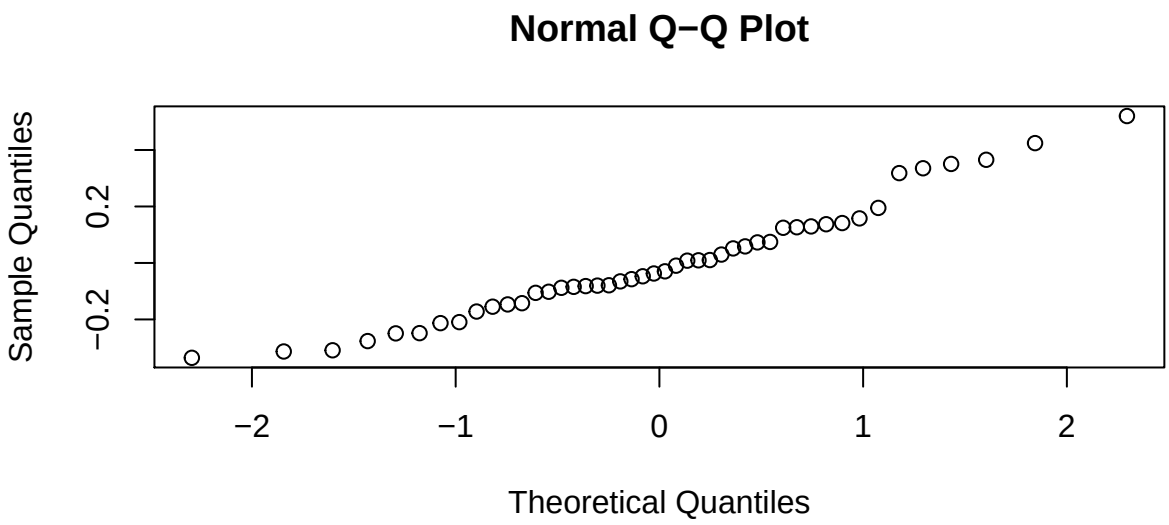

```
plot(summary(m3)$residuals~t[[3]],xlab="",ylab="",
     main="Reszty w zależności od trzeciej zmiennej")
```


```
plot(summary(m3)$residuals~m3$fitted.values,xlab="",ylab="",
     main="Reszty w zależności od predykcji")
```

Na żadnym z czterech wykresów nie ma wyrażnych obserwacji odstających. Pierwsze cztery wykresy wskazują na brak zależności reszt od czegokolwiek, ale na czwartym wykresie już taką zależność dostrzegamy. Wartość resztowa rośnie wraz z wartością zmiennej wyjaśnianej (satysfakcji pacjenta).

Zadanie 8

```
shapiropwart=shapiro.test(summary(m3)$residuals)$p.value
qqnorm(summary(m3)$residuals)
```



P-wartość dla testu Shapiro-Wilka wynosi 0.1481172, jest większa od 0,05 zatem możemy przyjąć, że reszty mają rozkład normalny.

Zadanie 9

```
s=read.table("http://math.uni.wroc.pl/~mbogdan/Modele_Liniowe/Dane/csdata.dat")

m4=lm(s[[2]]~s[[3]]+s[[4]]+s[[5]])
m5=lm(s[[2]]~s[[6]]+s[[7]]+s[[3]]+s[[4]]+s[[5]])

F1spos=(anova(m4)[4,2]-anova(m5)[6,2])/(anova(m4)[4,1]-anova(m5)[6,1])/anova(m5)[6,3]
F2spos=anova(m4,m5)[2,5]
pwart=anova(m4,m5)[2,6]
ndf=anova(m4,m5)[2,3]
ddf=anova(m4,m5)[2,1]
```

H: $\beta_4 = 0 \wedge \beta_5 = 0$ K:$\beta_4 \neq 0 \vee \beta_5 \neq 0$ Wartość statystyki testowej wynosi 0.9503276 w pierwszym sposobie i 0.9503276 w drugim, przy H ta statystyka ma rozkład F-Snedecora z (2,218) stopniami swobody, p-wartość wynosi 0.38821, stąd brak podstaw do odrzucenia H na rzecz K.

Zadanie 10

```
sst1=anova(m5)[1:5,2]
sst2=numeric(5)

colnumbers=c(6,7,3,4,5)
for (i in 1:5){
  templm=lm(s[[2]]~as.matrix(s[colnumbers[-i]]))
  sst2[i]=anova(templm,m5)[2,4]
}
Q=rbind(sst1,sst2)
colnames(Q)=c('SATM','SATV','HSM','HSE','HSS')
```

```
rownames(Q)=c("SS typu I", "SS typu II")
kable(Q,format='markdown')
```

	SATM	SATV	HSM	HSE	HSS
SS typu I	8.5829336	0.0009055	17.726470	1.3765322	0.956804
SS typu II	0.9279988	0.2326519	6.772431	0.4421433	0.956804

Dla ostatniej ze zmiennych wyjaśniających wartości SS typu I i II są równe ponieważ obie wartości to wartość sumy kwadratów objaśnianej przez model bez ostatniej zmiennej.

Zadanie 11

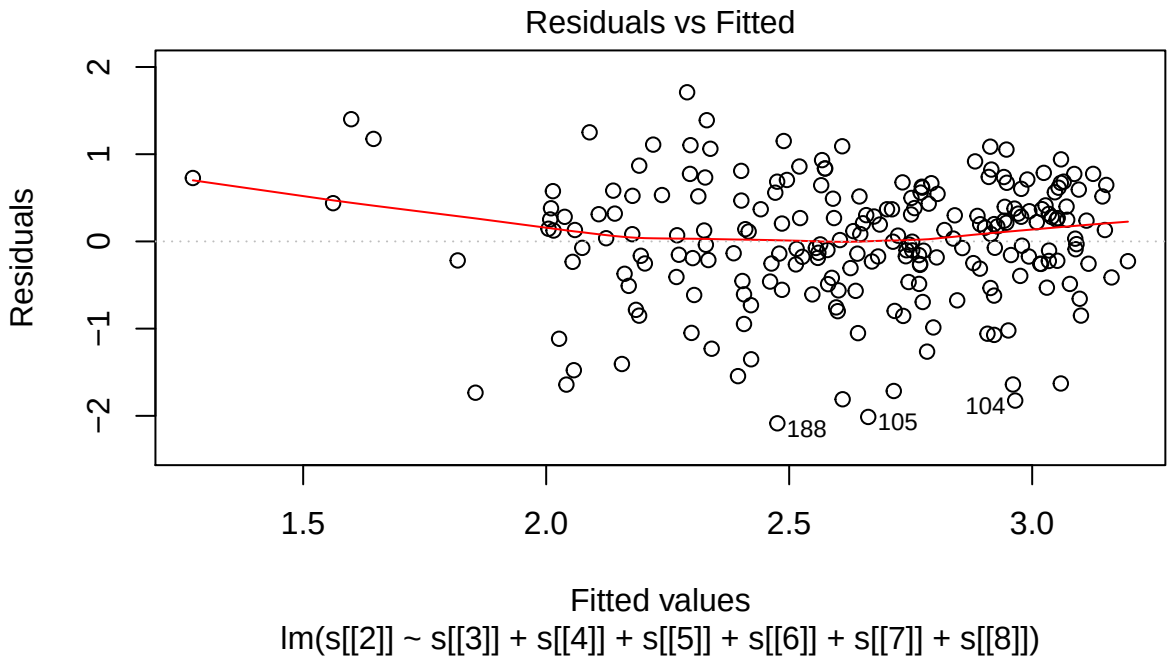
```
R=as.matrix(cbind(s[[6]],s[[7]],s[6]+s[7]))
m6=lm(s[[2]]~R)
print(summary(m6)$coefficients)
```

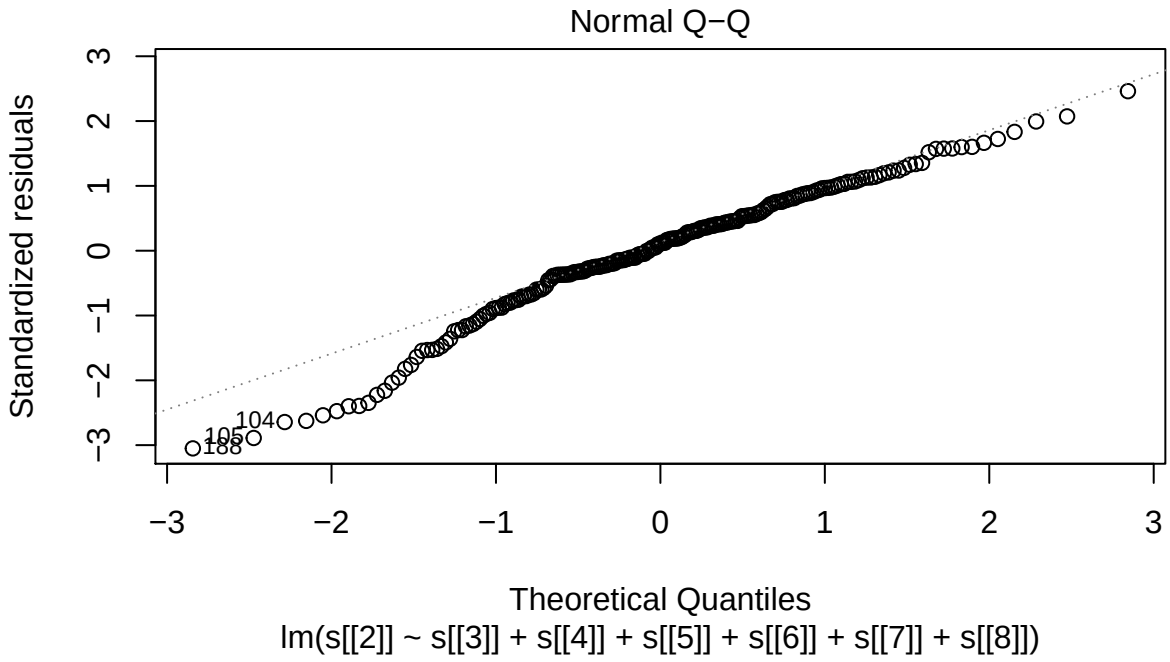
```
##                  Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)  1.288677e+00 0.3760368379  3.42699767 0.0007276956
## Rs[[6]]      2.282834e-03 0.0006629143  3.44363406 0.0006865143
## Rs[[7]]     -2.456193e-05 0.0006184696 -0.03971405 0.9683569529
```

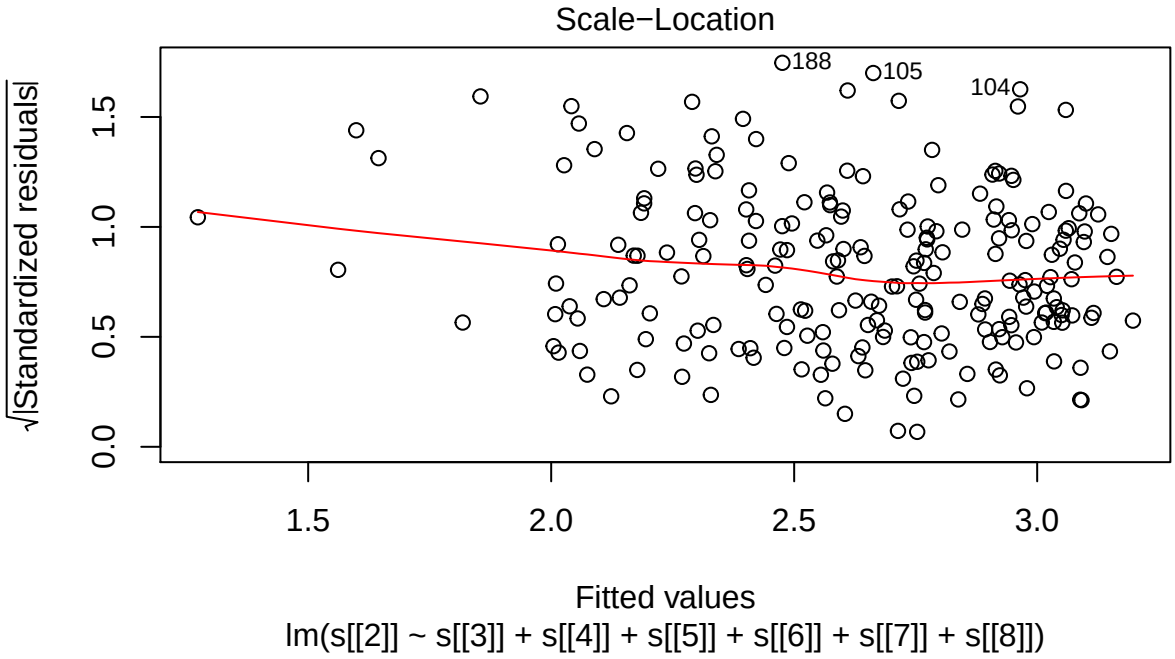
Dla zmiennej SAT wartości wynoszą NA, R rozpoznał, że nowa zmienna jest liniowo zależna od dwóch pozostałych zmiennych przez co macierz X'X jest osobliwa.

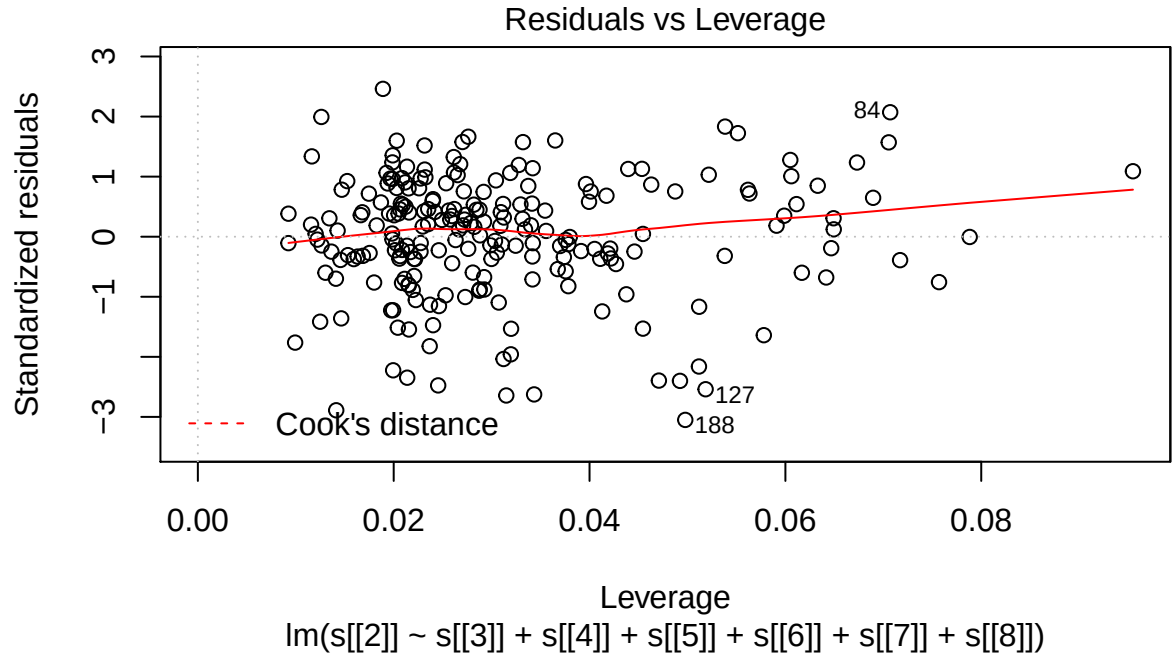
Zadanie 12

```
colnumbers2=3:8
m7=lm(s[[2]]~s[[3]]+s[[4]]+s[[5]]+s[[6]]+s[[7]]+s[[8]])
plot(m7)
```


Residuals vs Fitted
Residuals
Fitted values
lm(s[[2]] ~ s[[3]] + s[[4]] + s[[5]] + s[[6]] + s[[7]] + s[[8]])
188
105
104


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(s[[2]] ~ s[[3]] + s[[4]] + s[[5]] + s[[6]] + s[[7]] + s[[8]])
104
105
188


Scale−Location
188
105
104
√|Standardized residuals|
0.0
0.5
1.0
1.5
1.5
2.0
2.5
3.0
Fitted values
lm(s[[2]] ~ s[[3]] + s[[4]] + s[[5]] + s[[6]] + s[[7]] + s[[8]])


Residuals vs Leverage
84
127
188
Cook's distance
Standardized residuals
−3
−1
0
1
2
3
0.00
0.02
0.04
0.06
0.08
Leverage
lm(s[[2]] ~ s[[3]] + s[[4]] + s[[5]] + s[[6]] + s[[7]] + s[[8]])

Zadanie 13

```
studentyzowane=rstudent(m7)
sort(abs(studentyzowane),decreasing = T)[1:5]
```

```
##      188      105      104      138      127
## 3.110517 2.940178 2.681093 2.662293 2.572071
```

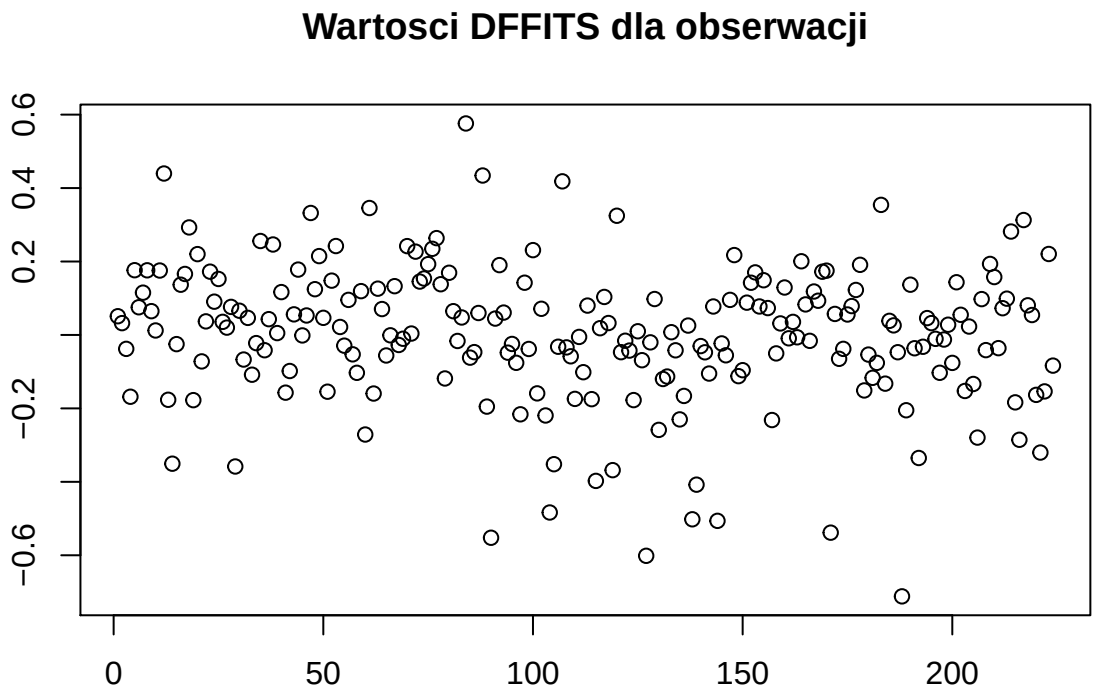
Podejrzane są obserwacje 104, 105 i 188, te same są zazaczone na wykresach w funkcji plot(lm()). Jednak wartości parametrów dla tych obserwacji nie są znaczącą większe od wartości dla pozostałych obserwacji i mogą pozostać w modelu.

Zadanie 14

```
dff=dffits(m7)
sort(abs(dff),decreasing = T)[1:5]
```

```
##       188       127        84        90       171
## 0.7119046 0.6014848 0.5759329 0.5522947 0.5384145
```

```
plot(dff,xlab="",ylab="",main="Wartości DFFITS dla obserwacji")
```



Zgodnie z oczekiwaniami na wykresie obserwujemy chmurkę punktów i brak zależności wartości DFFITS od czegokolwiek. Tym razem z podejrzanych obserwacji pozostała jedynie 188, wciąż jednak odchylenie od pozostałych wartości nie jest bardzo znaczące.

Zadanie 15

```
VIF=vif(m7)
tolerancja=1/VIF
Z=rbind(VIF,tolerancja)
kable(Z,format='markdown')
```

	s[[3]]	s[[4]]	s[[5]]	s[[6]]	s[[7]]	s[[8]]
VIF	1.9272916	1.9653302	1.8417747	1.7404932	1.3678889	1.2915693
tolerancja	0.5188628	0.5088203	0.5429546	0.5745498	0.7310535	0.7742519

Wartości VIF dla każdej z kolumn nie przekracza 2 (niepożądane są wartości wyższe od 10). Tolerancja dla każdej ze zmiennych objaśniających jest wysoka i ma mały rozrzut, wszystkie wartości są pomiędzy 0,5 a 0,8.

Zadanie 16

```
m8aic=step(m7,direction = 'both')
m9bic=step(m7, direction = 'both', k=log(dim(s)[1]))
```

```
summary(m8aic)$coefficients[,1]
```

```
## (Intercept)      s[[3]]      s[[5]]
##  0.62422848  0.18265442  0.06067015
```

```
summary(m9bic)$coefficients[,1]
```

```
## (Intercept)      s[[3]]
##   0.9076782   0.2076020
```

Kryterium AIC za najlepszy model uznaje model objaśniający przy użyciu trzeciej i piątej kolumny, zaś BIC do objaśniania proponuje używać tylko trzeciej kolumny.