Kamil Marciniak

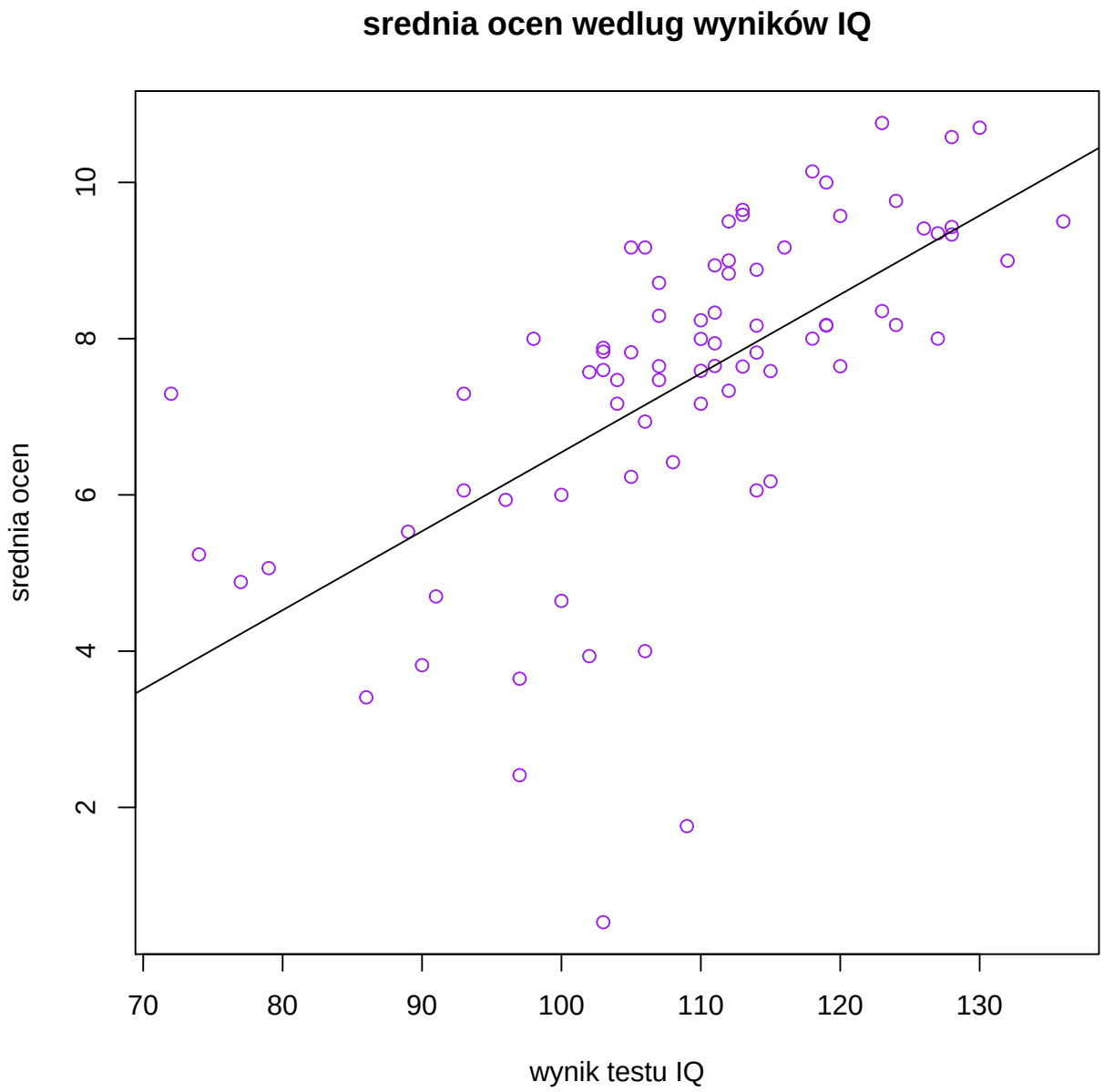
Raport 3 modele liniowe

17 grudnia 2021

Spis treści

1. Zadanie 3

Poniżej przedstawiono wykres rozrzutu wyników testu IQ do średnich ocen.




```
## [1] TRUE
```

```
summary(model_iq)
##
```

```
## Call:
## lm(formula = srednia ~ iq, data = tabela_wynikow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3182 -0.5377  0.2178  1.0268  3.5785
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.55706    1.55176  -2.292   0.0247 *
## iq           0.10102    0.01414   7.142 4.74e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.635 on 76 degrees of freedom
## Multiple R-squared:  0.4016,Adjusted R-squared:  0.3937
## F-statistic: 51.01 on 1 and 76 DF,  p-value: 4.737e-10
```

```
R_squared
```

```
## [1] 0.4016146
```

```
Fstat
```

```
## [1] 51.00845
```

```
pvalue
```

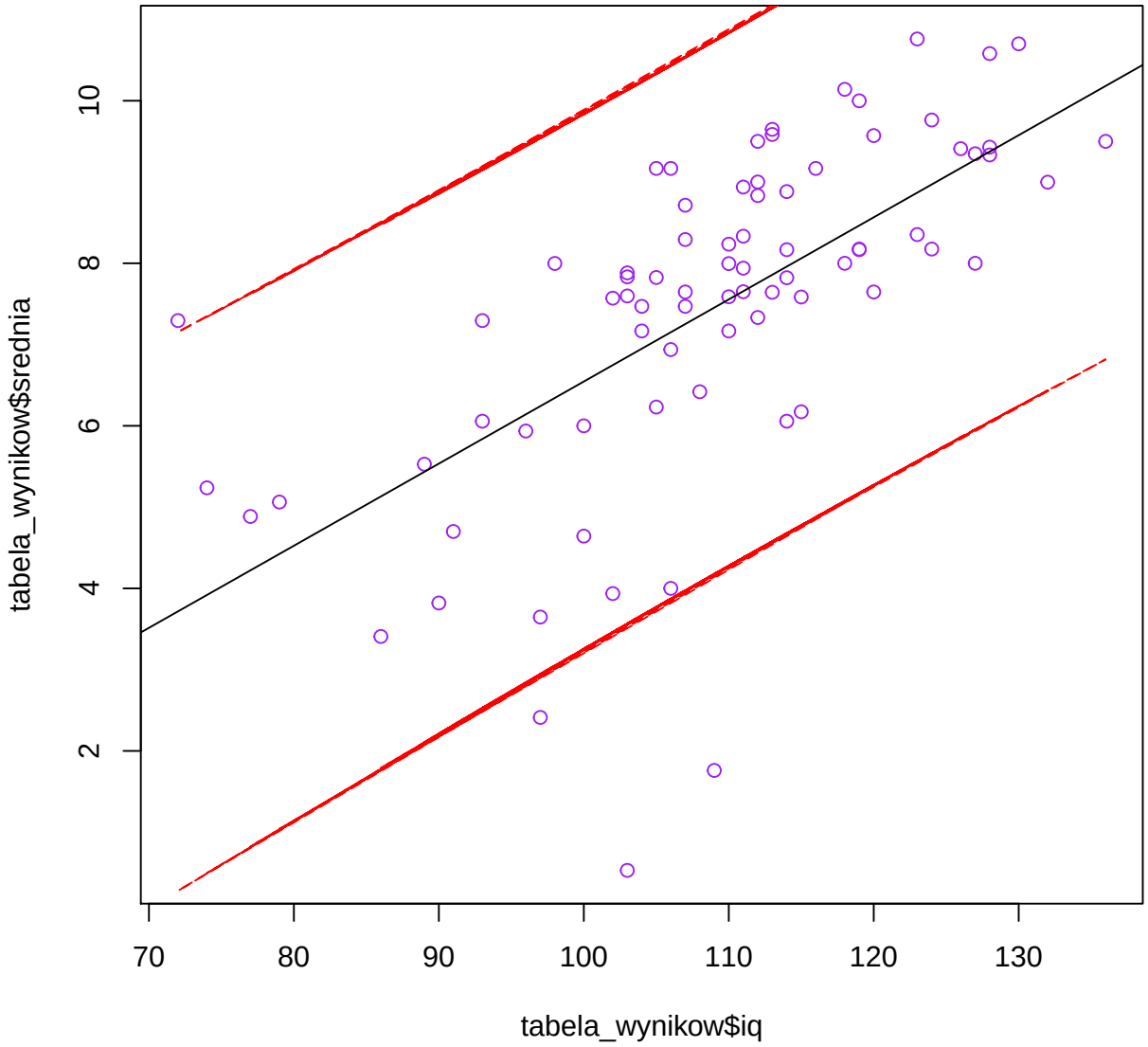
```
## [1] 4.737341e-10
```

Równanie regresji ma postać y =-3.5570558+0.1010217$\cdot x$ Badana hipoteza zerowa to $\beta_1 = 0$ Wartość R^2 wynosi 0.4016146, statystyka F wynosi 51.0084526 a p-wartości wynosi $4.7373405 \times 10^{-10}$. I jak widać wartości policzone samodzielnie zgadzają się z tymi z polecenia summary. P-wartość wynosi prawie 0 więc na poziomie istotności 0.05 odrzucamy hipotezę zerową. Czyli wniosek jest taki że według tego testu średnia ocen jest istotnie statystycznie skorelowana z iq.

```
#podpunkt b
predict(model_iq , data.frame(iq=100) ,interval='prediction',level=0.9)
```

```
##        fit     lwr      upr
## 1 6.545114 3.79753 9.292698
```

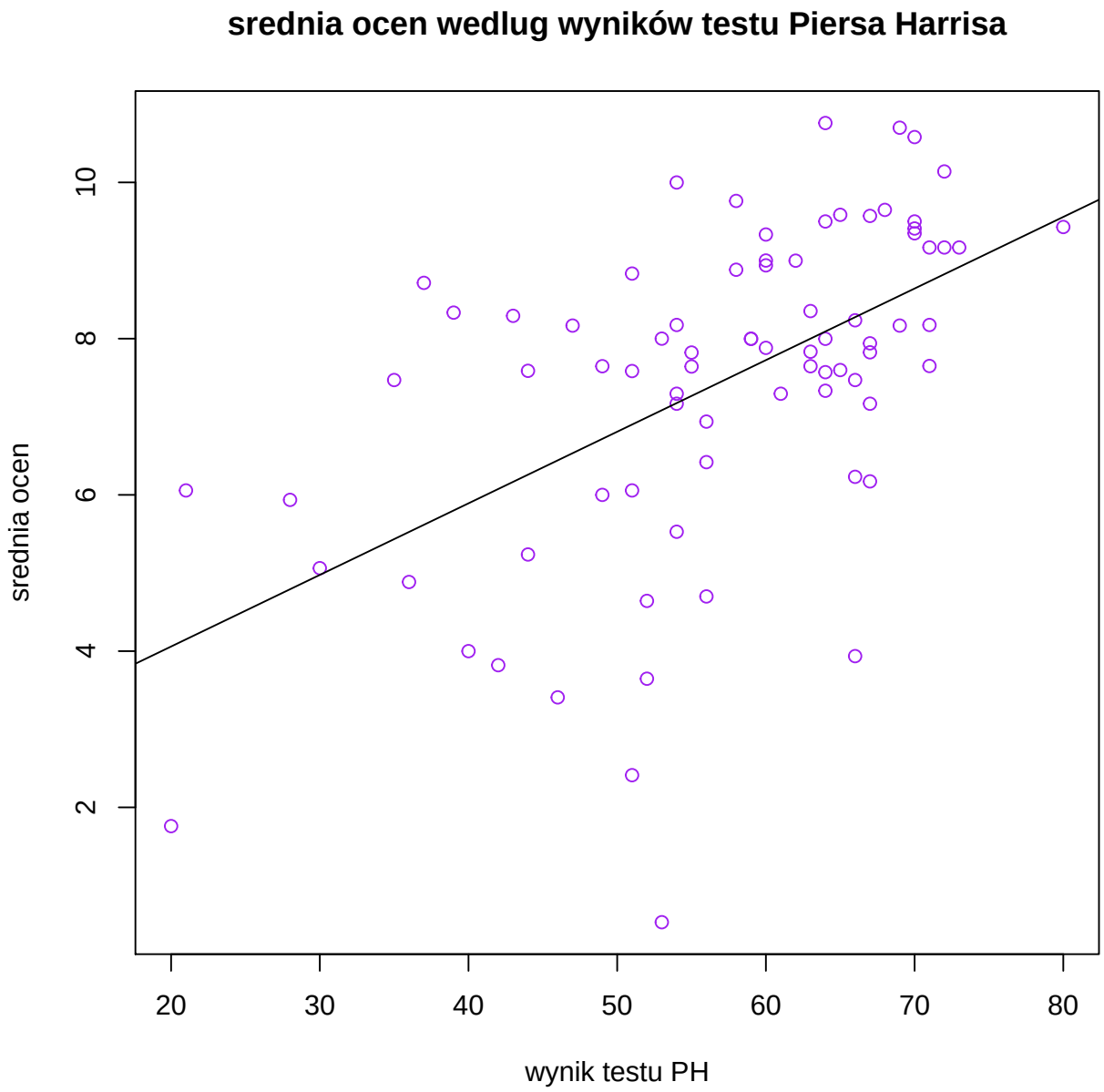
Powyżej wyprintowałem przewidywaną średnią dla ucznia o $iq = 100$ wraz z 90% przedziałem ufności.



4 obserwacje wypadają poza te przedziały predykcyjne, czyli jest to odsetek 4/78 co wynosi około 0.05128 co jest bardzo blisko zakładanego poziomu ufności $\alpha = 0.05$

2. Zadanie 4

Poniżej przedstawiono wykres rozrzutu wyników testu Piersa-Harrisa oraz średniej ocen.




```
## [1] TRUE
```

```
summary(model_ph)
```

```
##
```

```
## Call:
## lm(formula = srednia ~ ph, data = tabela_wynikow)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5535 -0.7482  0.2037  1.2108  3.0970
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.22588    0.95045   2.342   0.0218 *
## ph           0.09165    0.01631   5.620 3.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.776 on 76 degrees of freedom
## Multiple R-squared:  0.2936,Adjusted R-squared:  0.2843
## F-statistic: 31.59 on 1 and 76 DF,  p-value: 3.006e-07
```

```
R_squaredp
```

```
## [1] 0.2935829
```

```
Fstatph
```

```
## [1] 31.58517
```

```
pvaluep
```

```
## [1] 3.006416e-07
```

Równanie regresji ma postać $y = 2.2258827 + 0.0916523 \cdot x$

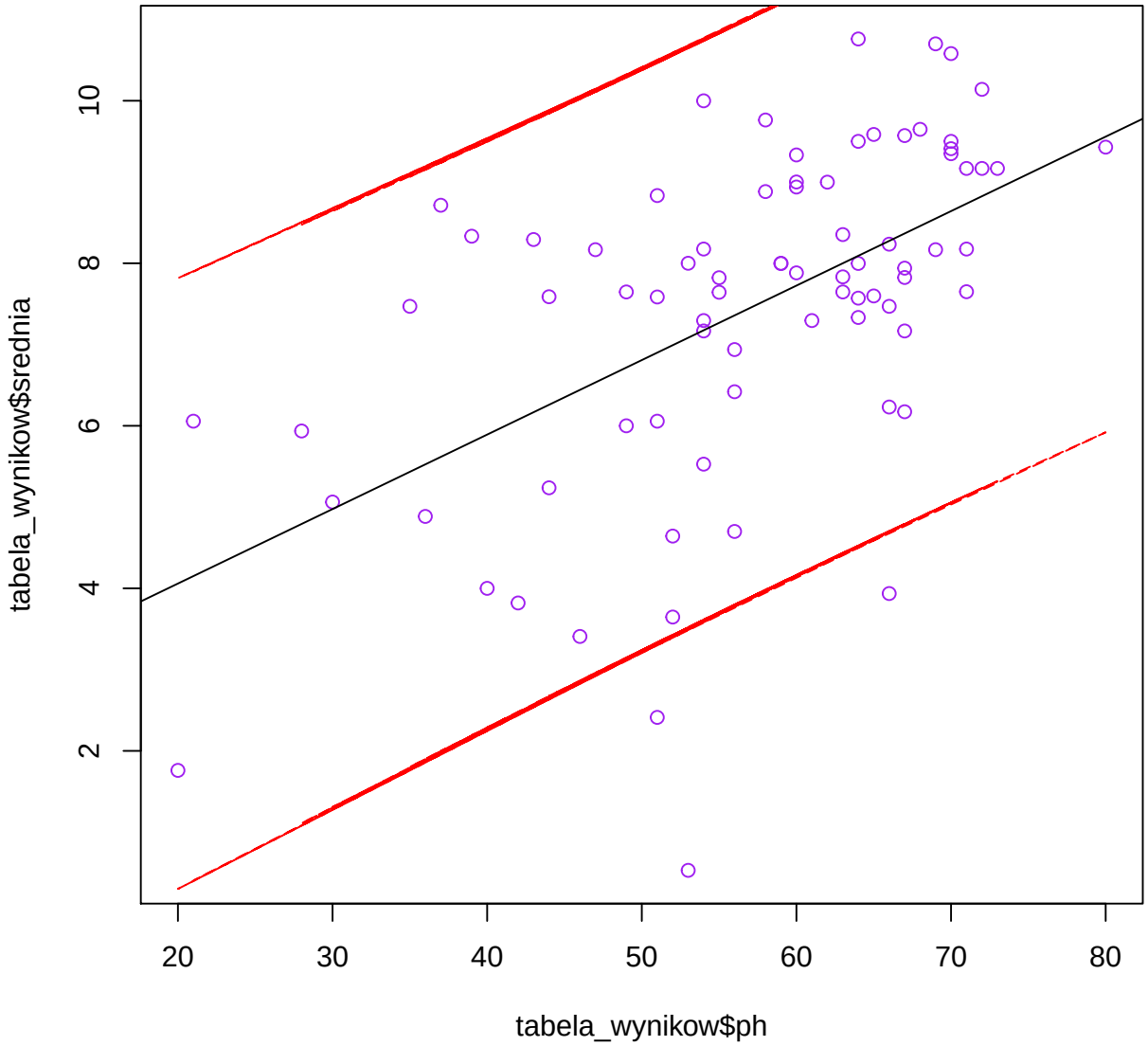
Badana hipoteza zerowa to $\beta_1 = 0$ Wartość R^2 wynosi 0.2935829, statystyka F wynosi 31.585165 a p-wartość wynosi 3.0064163×10^{-7}. I jak widać wartości policzone samodzielnie zgadzają się z tymi z polecenia summary. P-wartość wynosi prawie 0 więc na poziomie istotności 0.05 odrzucamy hipotezę zerową.Czyli wniosek jest taki że według tego testu średnia ocen jest istotnie statystycznie skorelowana z wynikami testu Piersa-Harrisa.

```
predict(model_ph, data.frame(ph=60) ,interval='prediction',level=0.9)
```

```
##       fit      lwr      upr
## 1 7.72502 4.747302 10.70274
```

Powyżej została podana przewidywana średnia dla studenta który na teście Piersa-Harrisa zdobył 60 punktów wraz z 90% przedziałem ufności.

```
plot(tabela_wynikow$ph,tabela_wynikow$srednia,col='purple')

abline(model_ph)
lines(tabela_wynikow$ph ,predict(model_ph, data.frame(ph=tabela_wynikow$ph ) ,interval='

lines(tabela_wynikow$ph,predict(model_ph, data.frame(ph=tabela_wynikow$ph) ,interval='pr
```



3 obserwacje wypadają poza przedziały predykcyjne. Czyli jest to odsetek 3/78 a więc około 0.038 czyli dośc blisko do poziomu istotności 0.05

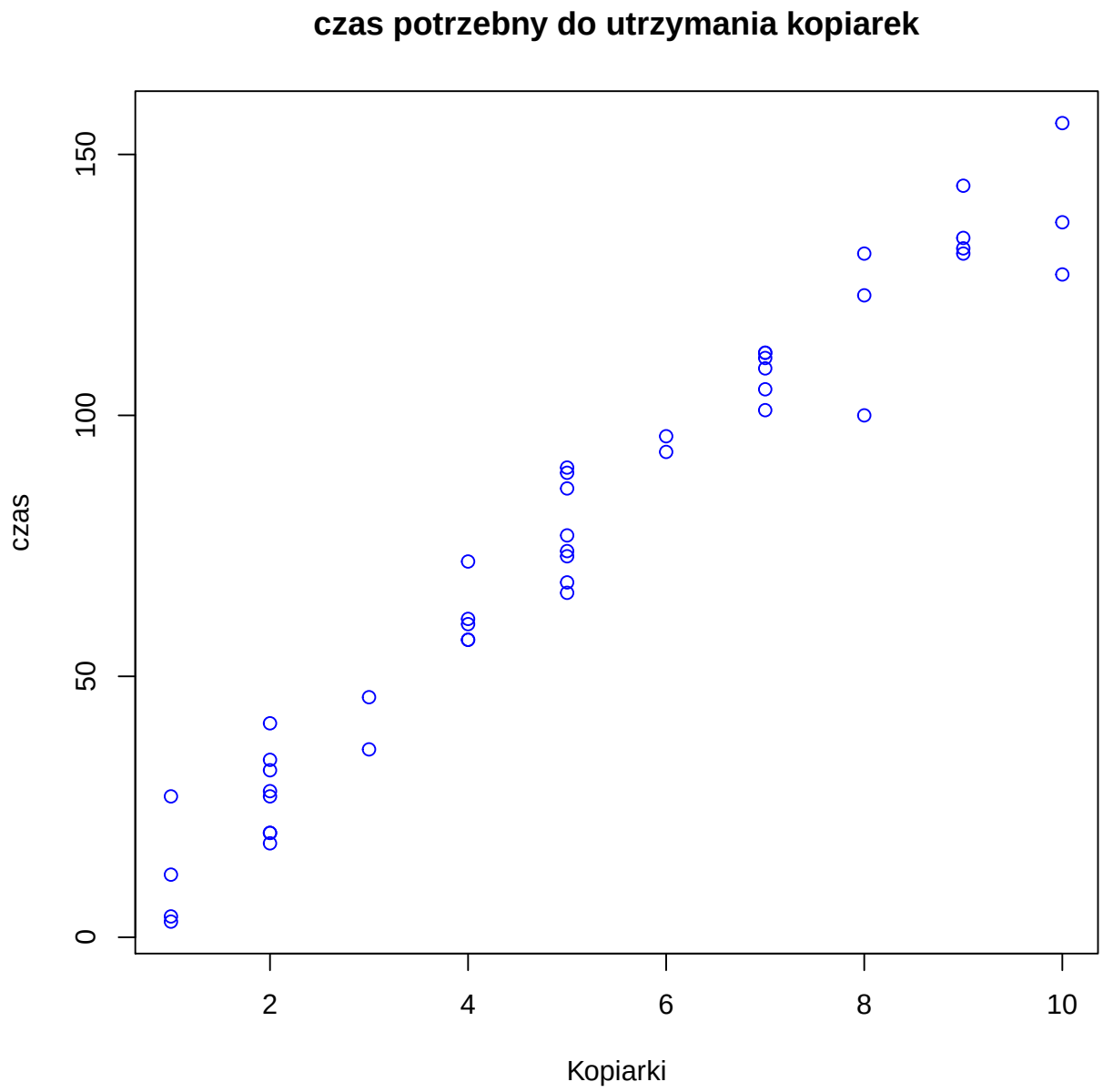
2.1. porównanie modeli

Aby odpowiedzieć na pytanie która ze zmiennych wynik IQ czy wynik testu PH jest lepszym predyktorem średniej porównam współczynniki R^2 tych modeli liniowych. Dla modelu $lm(gpa\ iq)$ ten współczynnik wynosi 0.4016146 a dla modelu $lm(gpa\ ph)$ ten współczynnik wynosi 0.2935829. Wybieramy model który wyjaśnia więcej zmienności wektora Y czyli ten który ma większy współczynnik R^2. Tak więc lepszym predyktorem gpa jest wynik IQ.

3. Zadanie 5

Najpierw przypomnijmy jak wygląda wykres rozrzutu danych.

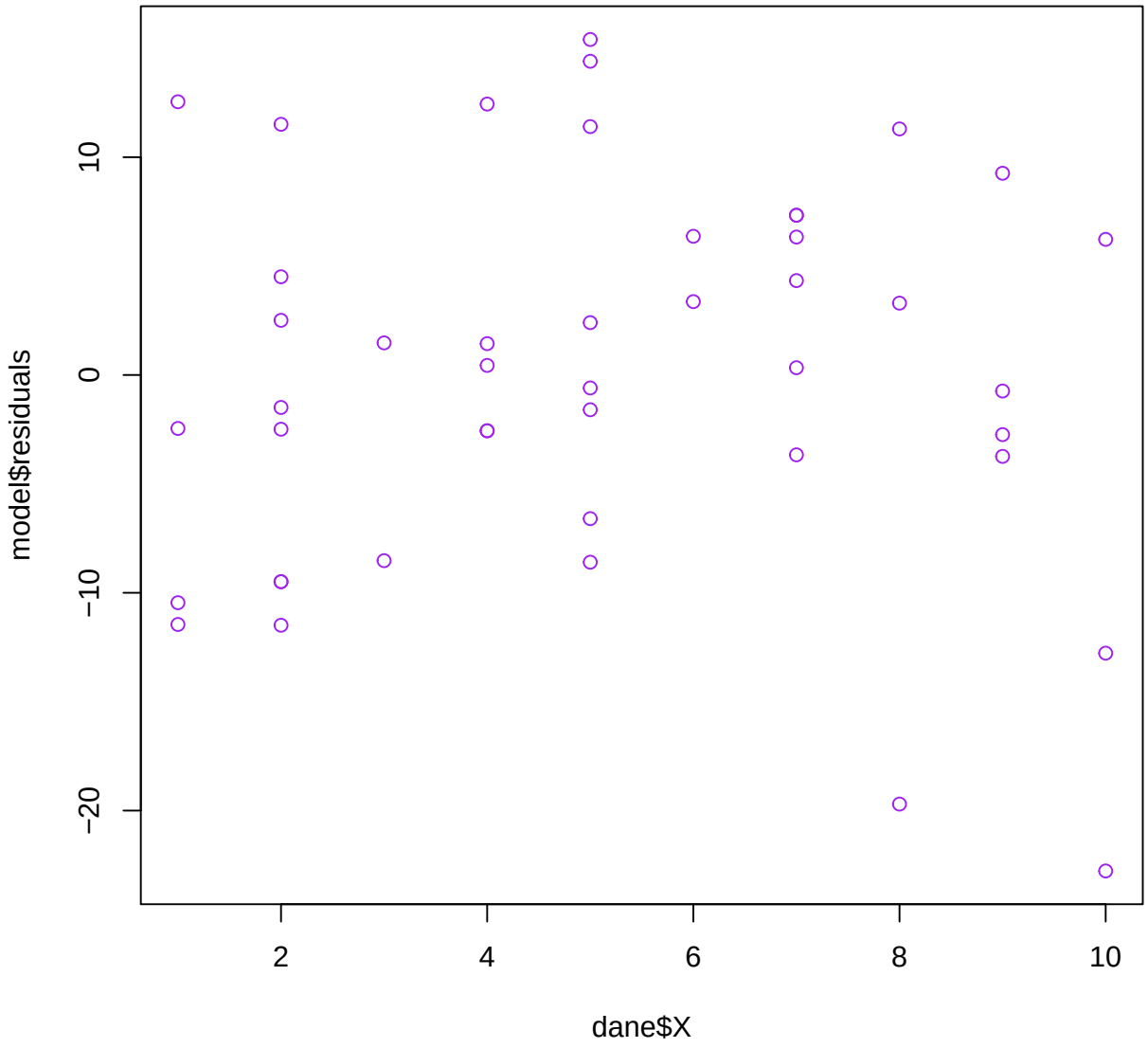
```
dane=read.table("http://www.math.uni.wroc.pl/~mkos/Modele_liniowe/CH01PR20.txt",col.name

plot(dane$X,dane$Y,ylab="czas",xlab="Kopiarki",col='blue',main='czas potrzebny do utrzym
```



```
model=lm(Y~X,dane)

sum(model$residuals)
## [1] -1.176836e-14
```

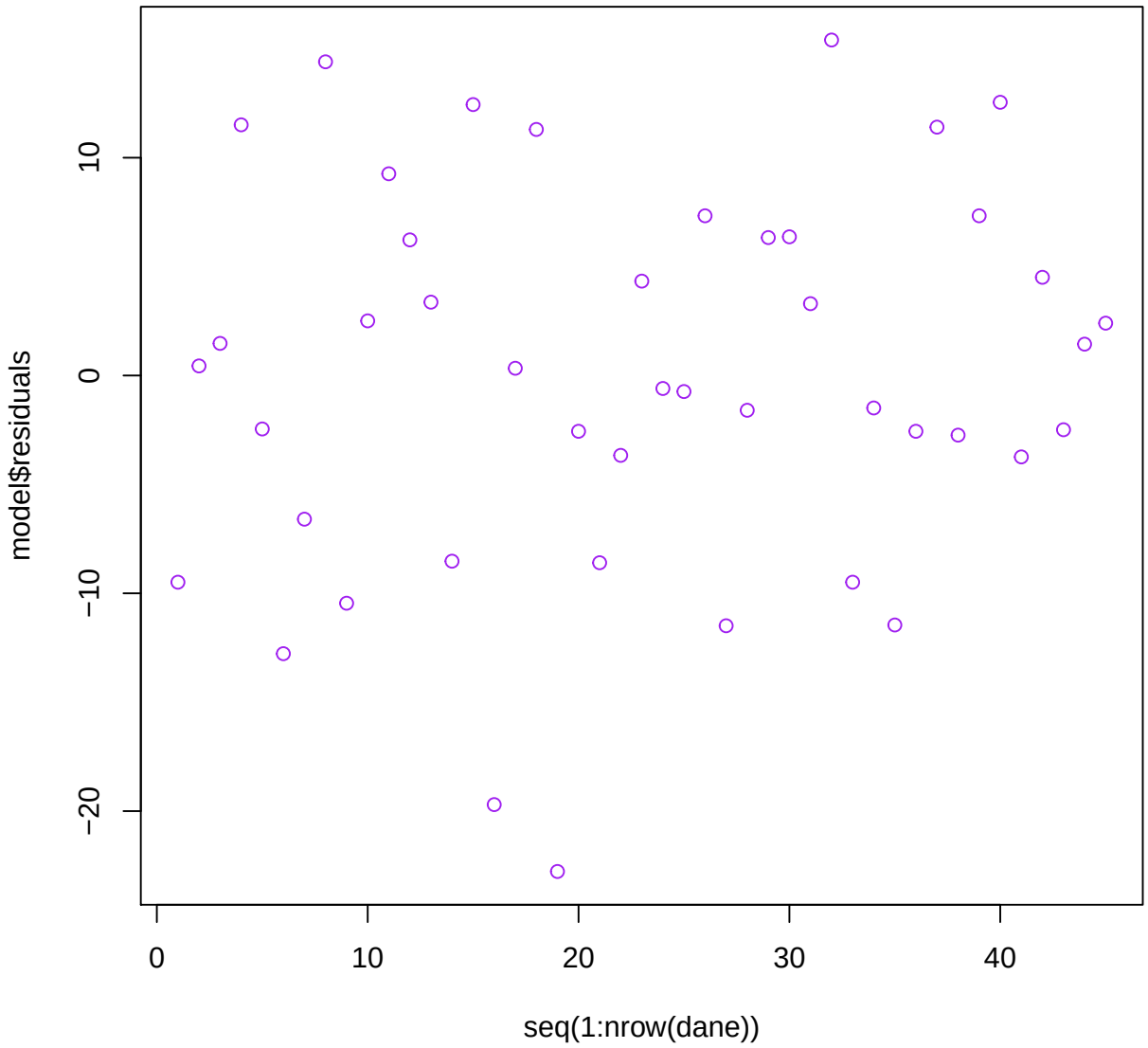
Rzeczywiście suma residuów wynosi 0

```
plot(model$residuals~dane$X,col='purple')
```



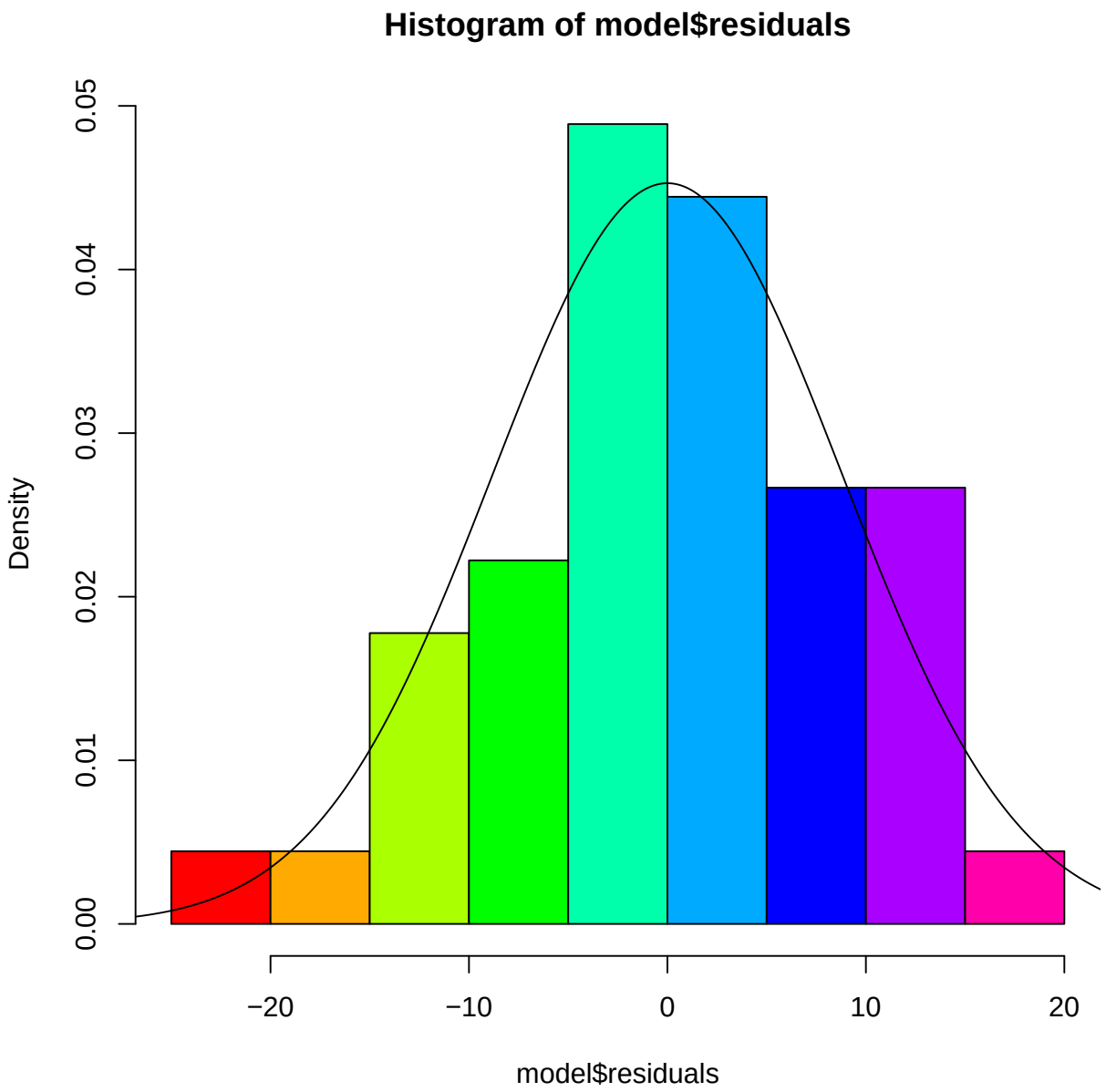
Dla wszystkich x residua są dość losowo rozmieszczone. Nie widać żadnych szczególnie odstających punktów ani też jakiejś wyraźnej zależności która by świadczyła o łamaniu założeń.

```
plot(model$residuals~seq(1:nrow(dane)),col='purple')
```



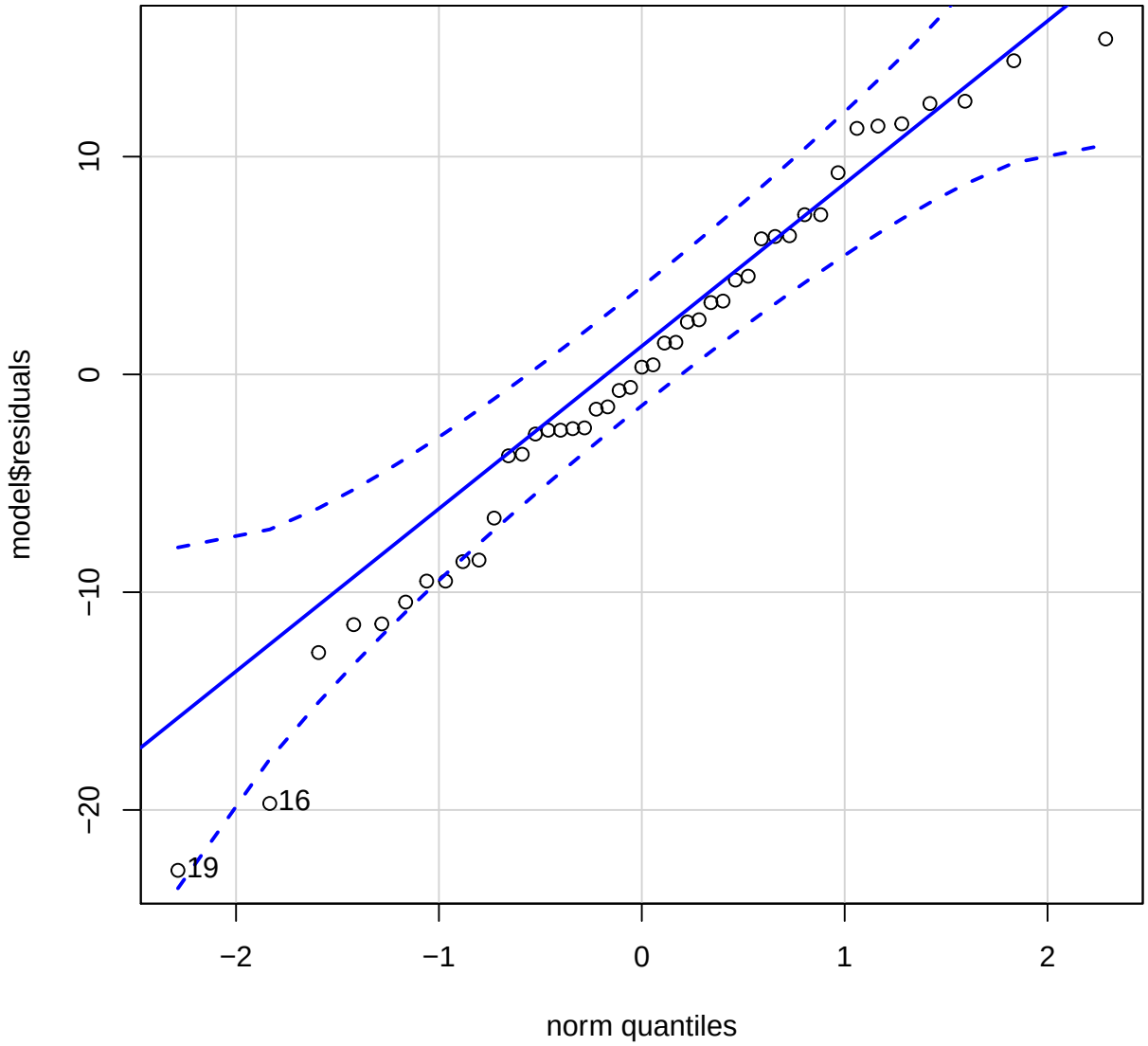
Podobnie jak w poprzednim wykresie Dla kolejnych x residua są dość losowo rozmieszczone. Nie widać żadnych szczególnie odstających punktów ani też jakiejś wyraźnej zależności która by świadczyła o łamaniu założeń.

```
hist(model$residuals,probability=TRUE,col=rainbow(9))
xnormalny <- seq(-30,30,by=0.01)
ynormalny <- dnorm(xnormalny,mean=mean(model$residuals),sd=sd(model$residuals))
lines(xnormalny,ynormalny)
```



Histogram dość dobrze dopasowuje się do krzywej określającej gęstość rozkładu normalnego. Co jest argumentem za tym że faktycznie residua mają rozkład normlany

```
library(car)

qqPlot(model$residuals)
```




```
## [1] 19 16
```

Prawie wszystkie punkty mieszczą się między przerywanymi liniami więc to również argument za tym że residua pochodzą z rozkładu normalnego

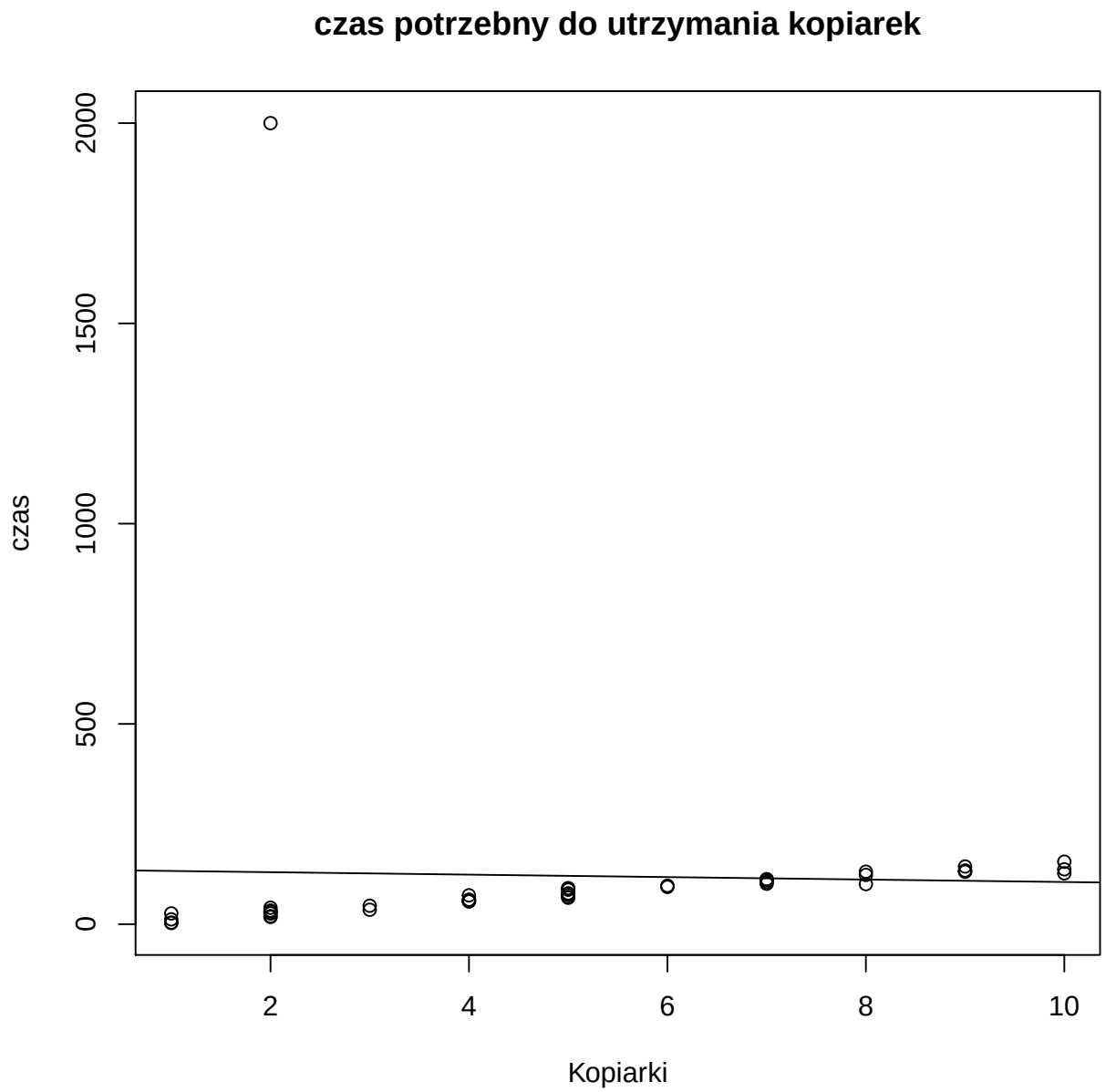
```
shapiro.test(model$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.97583, p-value = 0.4614
```

W tym teście hipoteza zerowa jest taka że dane pochodzą z rozkładu normalnego. p-wartosc jest większa od 0.05 więc nie ma podstaw do odrzucenia hipotezy zerowej. Czyli test Shapiro-Wilka również wskazuje na normalnośc.

Podsumowując faktycznie pokazane tutaj metody wskazują na to że residua pochodzą z rozkładu normalnego

4. zadanie 6

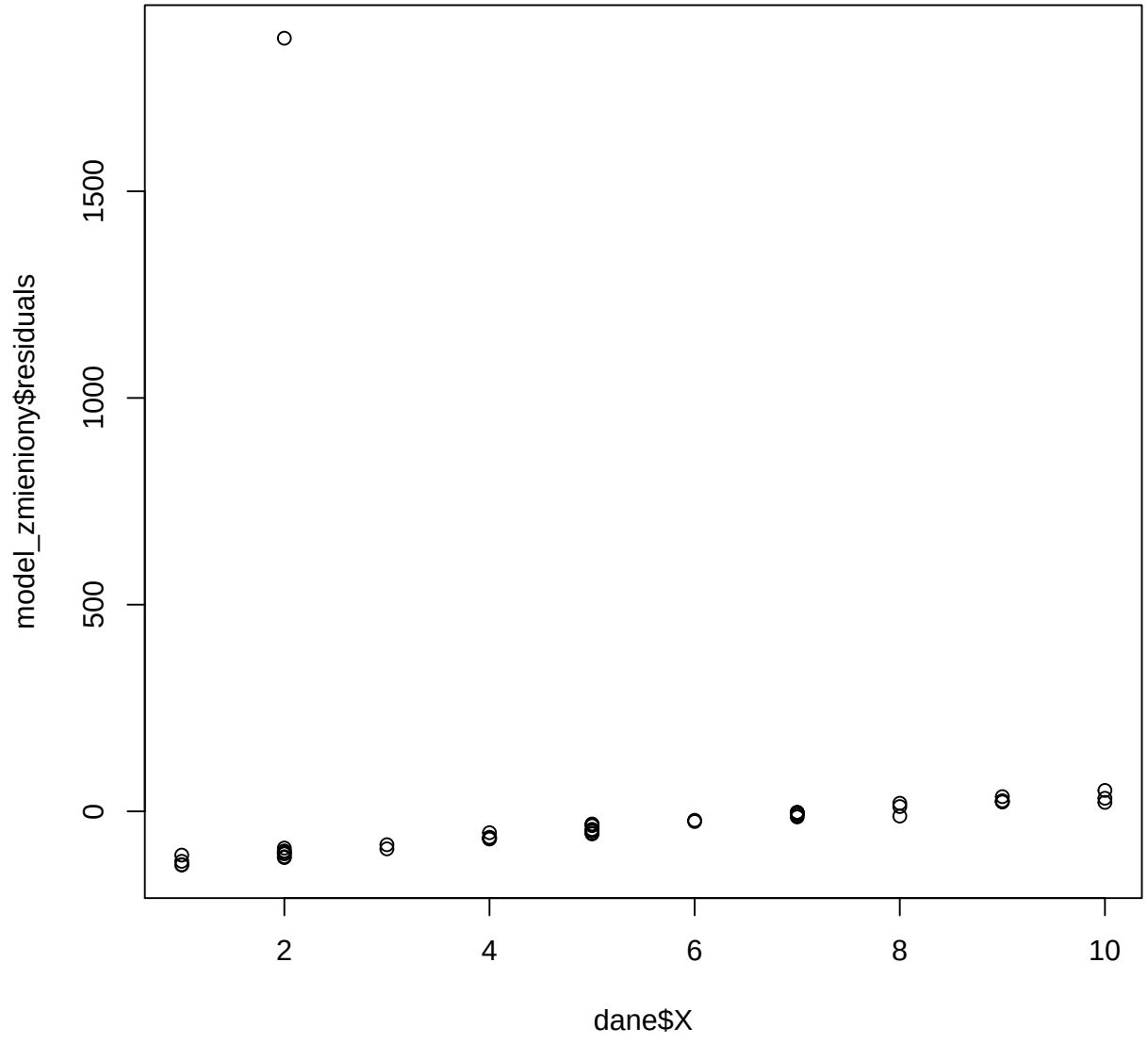
Wczytuję dane zmieniając tę jedną wartość.



Powyższy wykres to naniesiona linia regresji na nasza dane po zmianie jednej wartości z 20 do 2000. Już wizualnie można ocenić że prosta w ogóle nie odpowiada temu jak zachowują się dane ponieważ przez ten jeden ekstremalny punkt linia regresji maleje wraz ze wzrostem x a wszystkie niezmienione wartości y rosną wraz ze wzrostem x.

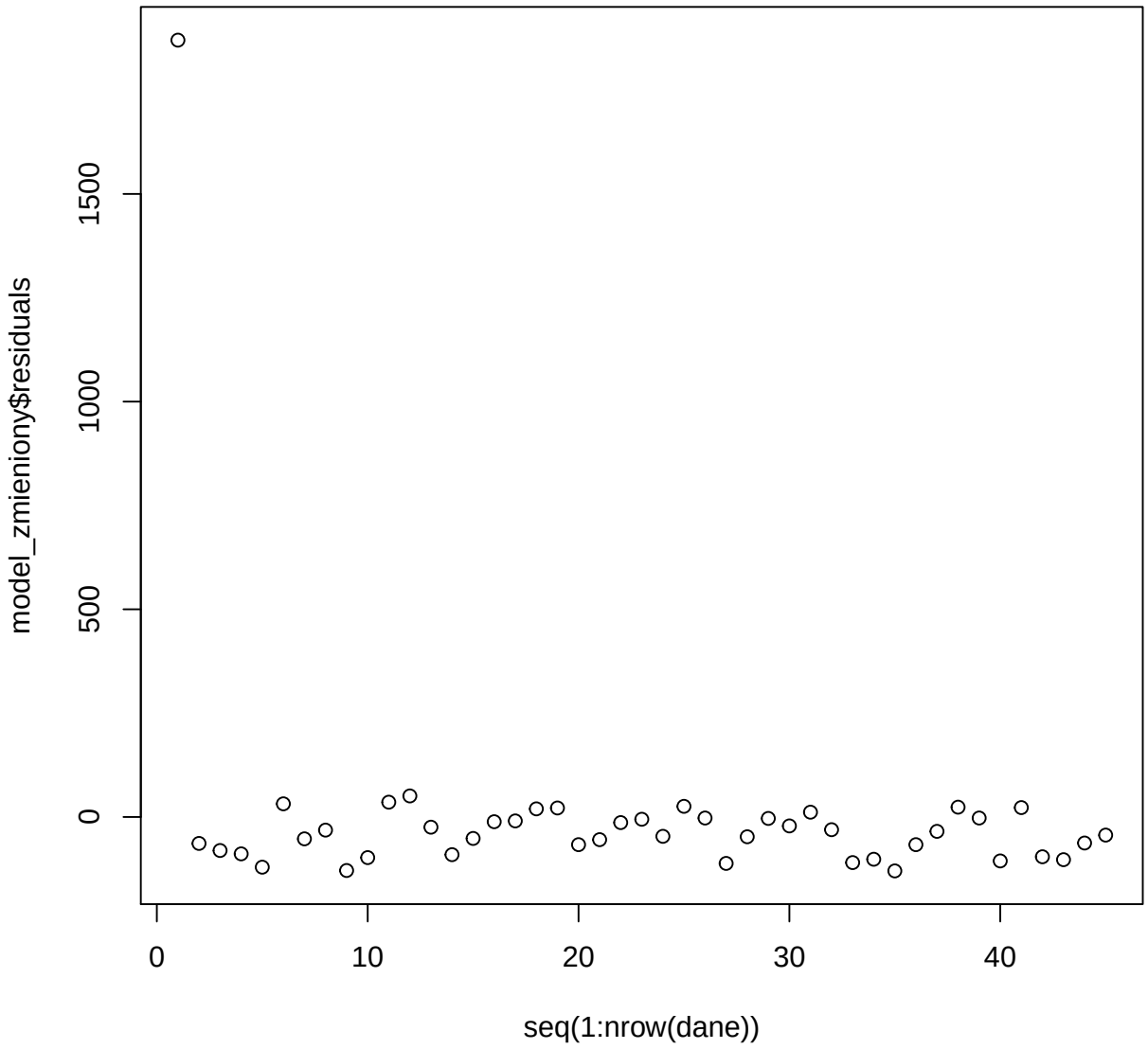
	pierwotny model	zmieniony model
intercept	-0.5801567	135.9002611
slope	15.0352480	-3.0587467
statystyka T	31.1232581	-0.1927195
p-wartość	0.0000000	0.8480860
R^2	0.9574955	0.0008630
estimate of sigma^2	8.9135082	292.8471156

W tabelce widzimy że wszystkie przedstawione parametry uległy drastycznej zmianie. Intercept się bardzo zwiększył a slope bardzo wyraźnie zmniejszył aż na tyle że zmienił znak. Statystyka T bardzo zbliżyła się do 0. Co za tym idzie teraz już nie ma podstaw do odrzucenia hipotezy zerowej bo p-wartość jest bardzo duża. Współczynnik R^2 spadł właściwie do 0 a estymator σ^2 bardzo się zwiększył. Tak więc wyraźnie widać że jedna obserwacja odstająca może sprawić że model jest kompletnie bezużyteczny.

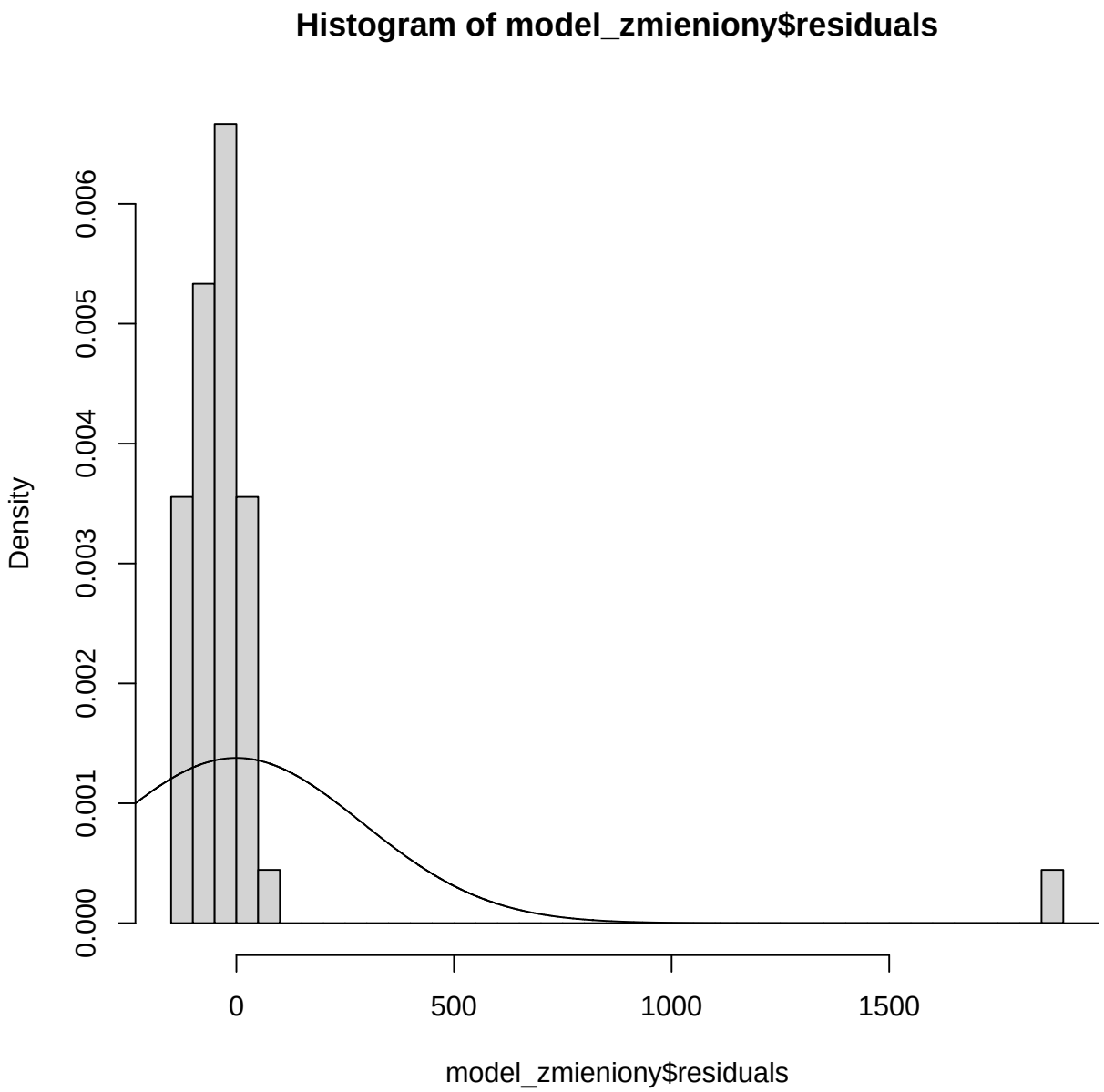


Residua powinny być dość losową chmurą punktów oscylującą wokół 0 a zdecydowanie

tak nie jest. Obecnie residua układają się w wyraźny rosnący trend liniowy. Oczywiście dla obserwacji odstającej wartość residuum jest bardzo bardzo duża.

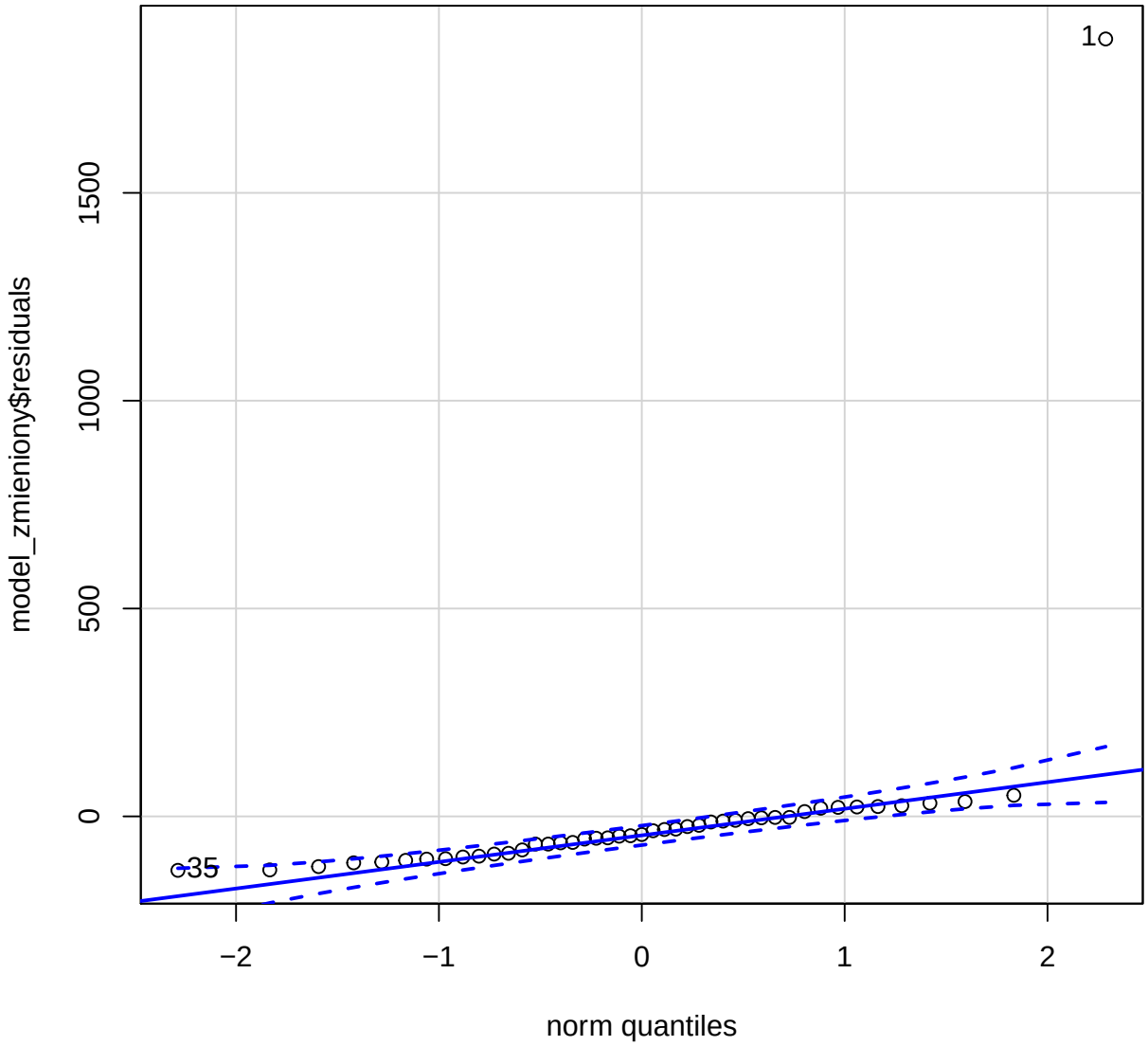


Oczywiście pierwsza wartość na wykresie bardzo odstaje od reszty danych. Ale już pozostałe dane akurat w tym wypadku układają się dość losowo.



Ta jedna bardzo duża obserwacja sprawia że histogram jest mało czytelny. Postanowiłem mocno zwiększyć liczbę klas aby przynajmniej w jakiś sposób było widać kształt tych regularnych niezmienionych danych. Ale z racji tego że ta jedna odstająca obserwacja wpływa również na średnią i wariancję rozkładu normalnego którego gęstość również rysuje na wykresie to całość nie jest już dobrze dopasowana do tej linii gęstości. W takim razie przez tę jedną obserwację należy stwierdzić, że rozkład residuów nie jest normalny.

```
library(car)
qqPlot(model_zmieniony$residuals)
```




```
## [1]  1 35
```

Nie jest zaskoczeniem że również na qq-plot obserwacja odstająca daje o sobie znać w postaci punktu kompletnie odległego od pozostałych w prawym górnym rogu.

```
shapiro.test(model_zmieniony$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  model_zmieniony$residuals
## W = 0.27107, p-value = 1.276e-13
```

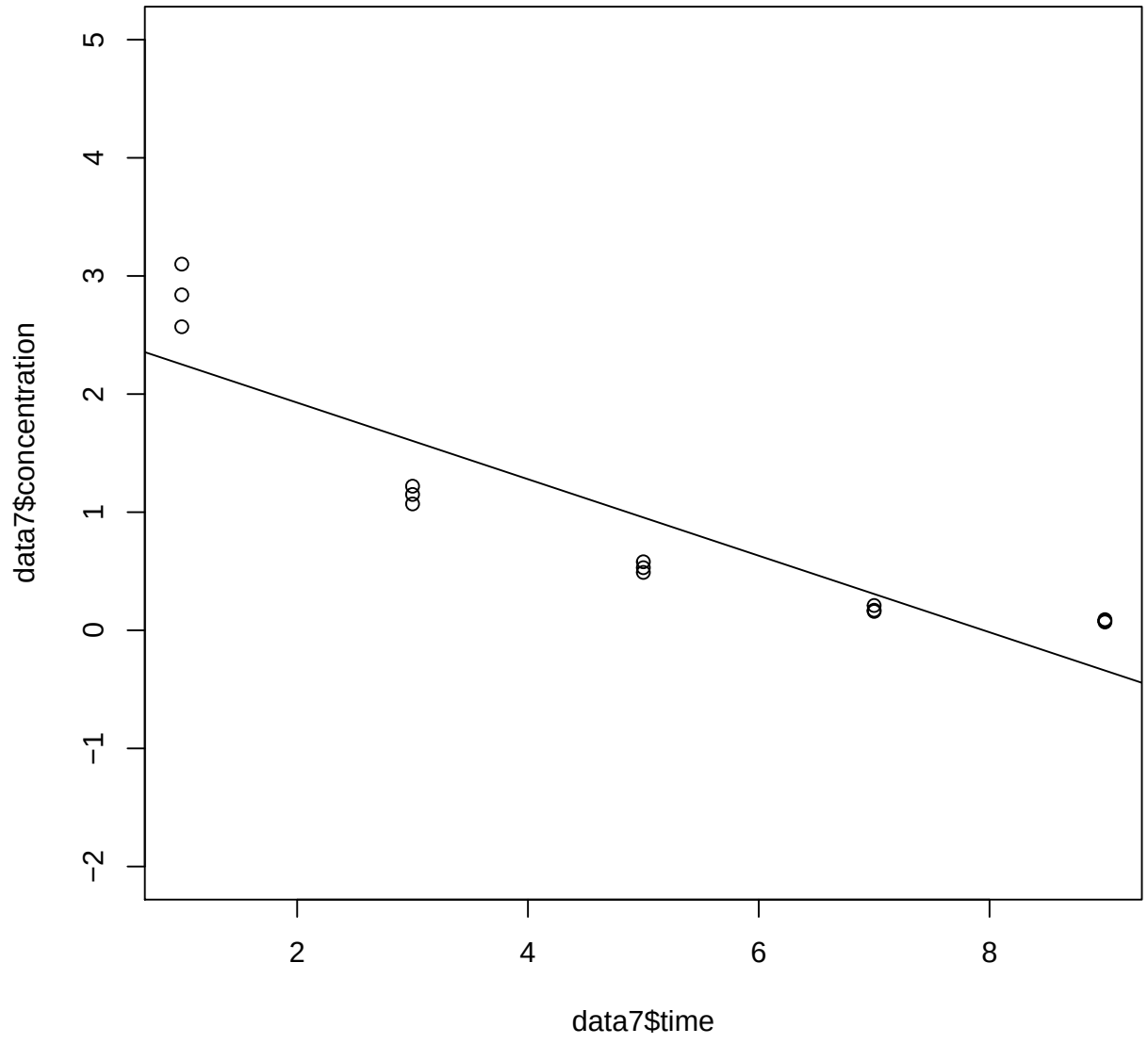
Również test Shapiro-Wilka daje nam bardzo małą p-wartość czyli odrzucamy hipotezę zerową o normalności rozkładu.

4.1. Podsumowanie

Jak widzimy zmiana jednej obserwacji może drastycznie wpłynąć na model regresji liniowej sprawiając chociażby że łamane jest założenie o normalności rozkładu residuów ale też tak skrajna wartość po prostu wyraźnie wpływa na wszystkie liczone współczynniki zaciemniając obraz sytuacji.

5. Zadanie 7

W tym zadaniu zajmujemy się już innymi danymi. Poniżej przedstawiłem już dane wraz z prostą regresji. Kształt danych niespecjalnie układa się w linię prostą tak więc użycie modelu liniowego bez żadnych środków zaradczych nie jest tutaj idealnym pomysłem.



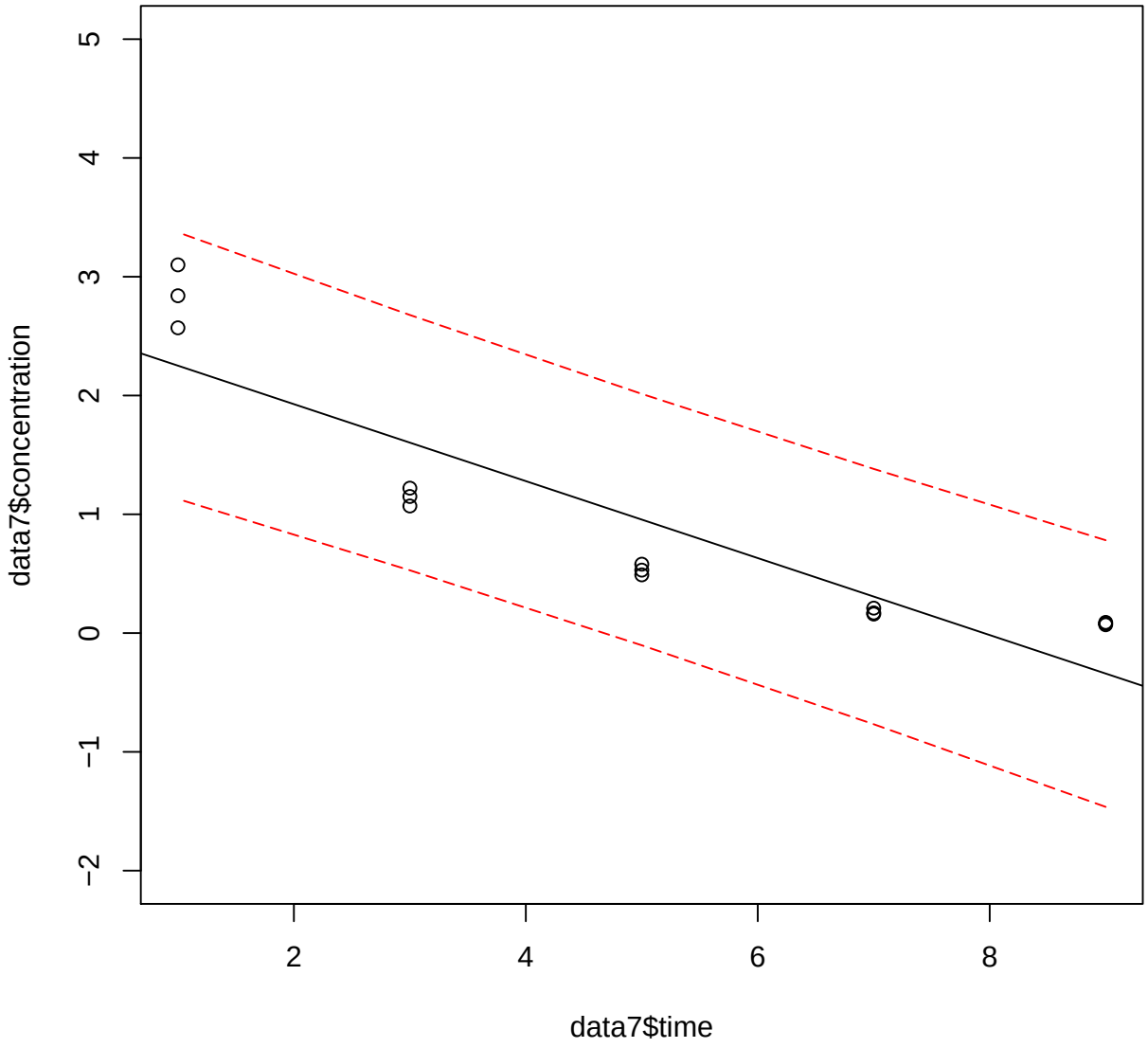

```
##
## Call:
## lm(formula = data7$concentration ~ data7$time)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5333 -0.4043 -0.1373  0.4157  0.8487
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5753     0.2487  10.354 1.20e-07 ***
## data7$time   -0.3240     0.0433  -7.483 4.61e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4743 on 13 degrees of freedom
## Multiple R-squared:  0.8116,Adjusted R-squared:  0.7971
## F-statistic: 55.99 on 1 and 13 DF,  p-value: 4.611e-06
```

	model
intercept	2.5753333
slope	-0.3240000
statystyka T	-7.4829029
p-wartość	0.0000046
R^2	0.8115774

W tabelce pojawiają się potrzebne nam wartości dodatkowo warto jeszcze przypomnieć że hipoteza zerowa to $\beta_1 = 0$ a hipoteza alternatywna to $\beta_1 \neq 0$. Liczba stopni swobody to $15 - 2 = 13$. Na podstawie p-wartości mniejszej od 0.05 odrzucamy hipotezę zerową. A gdybyśmy korzystali ze statystyki F to jej wartość wynosi 55.99 a liczba stopni swobody to 1 i 13 , p-wartość jest taka sama jak dla statystyki T.

6. Zadanie 8



Widać że relacja między zmiennymi nie jest liniowa przez co nad prostą regresji znajdują się obserwacje dla bardzo małych i bardzo dużych x, a pod prostą regresji dla średnich wartości x. A co za tym idzie gdybyśmy badali rozkład residuów to nie spełniałby on założeń o niezależności. Tak się składa że wszystkie obserwacje trafiły wewnątrz przedziałów predykcyjnych. Ale skoro obserwacji jest tylko 15 to nie jest to powód do niepokoju. Gdyby choć jedna wypadała poza to byłby to odsetek 1/15 czyli około 0.0666 a więc trochę więcej niż 0.05. Akurat w naszym przypadku po prostu wszystkie obseracje się mieszczą w przedziałach.

```
person <- cor(data7$concentration,predict(model7))
```

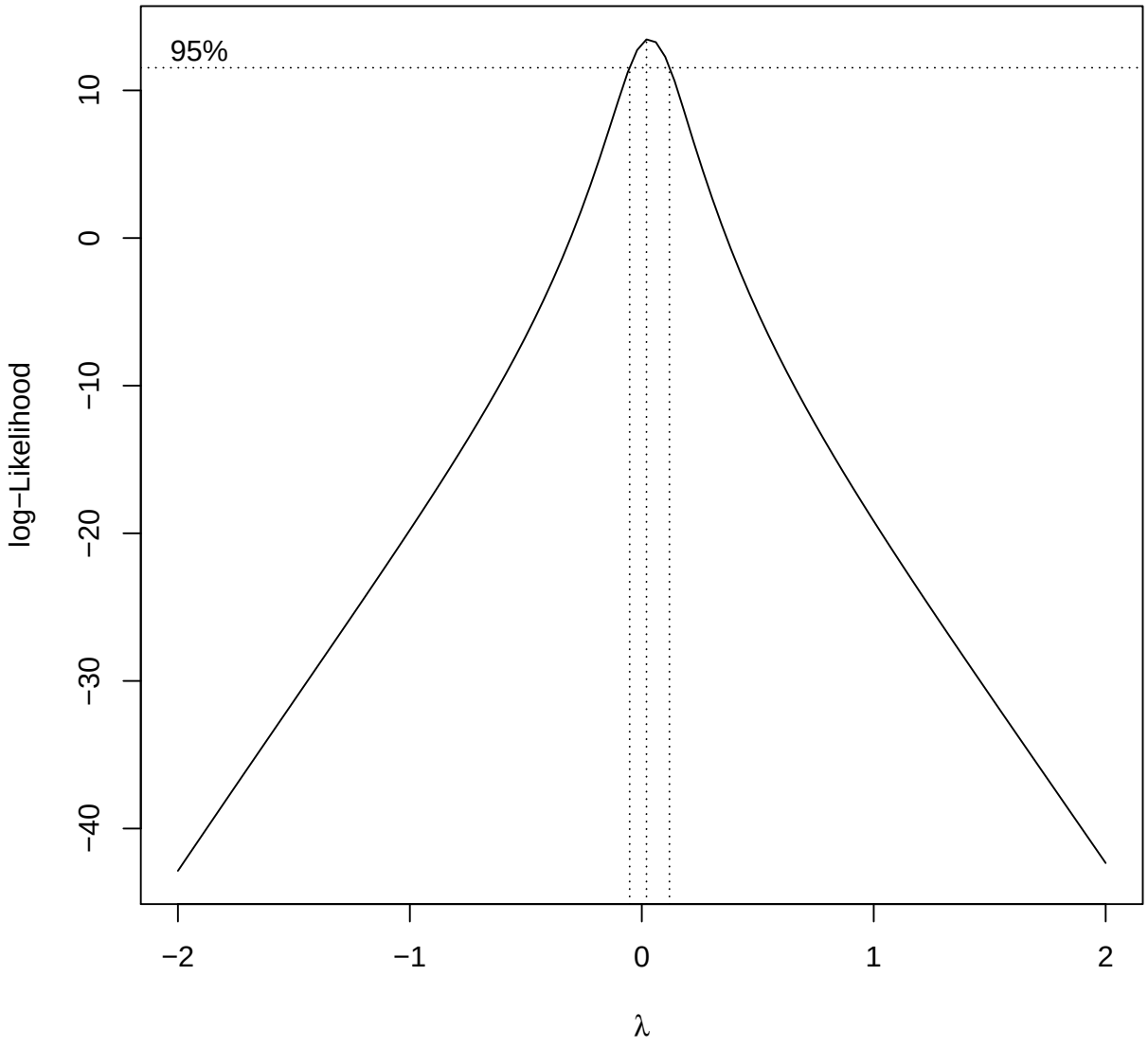
Korelacja wynosi 0.9008759 czyli mimo łamania pewnych założeń regresji liniowej jest całkiem duża.

7. zadanie 9

```
library('MASS')
```

```
## Warning: package 'MASS' was built under R version 4.0.3
```

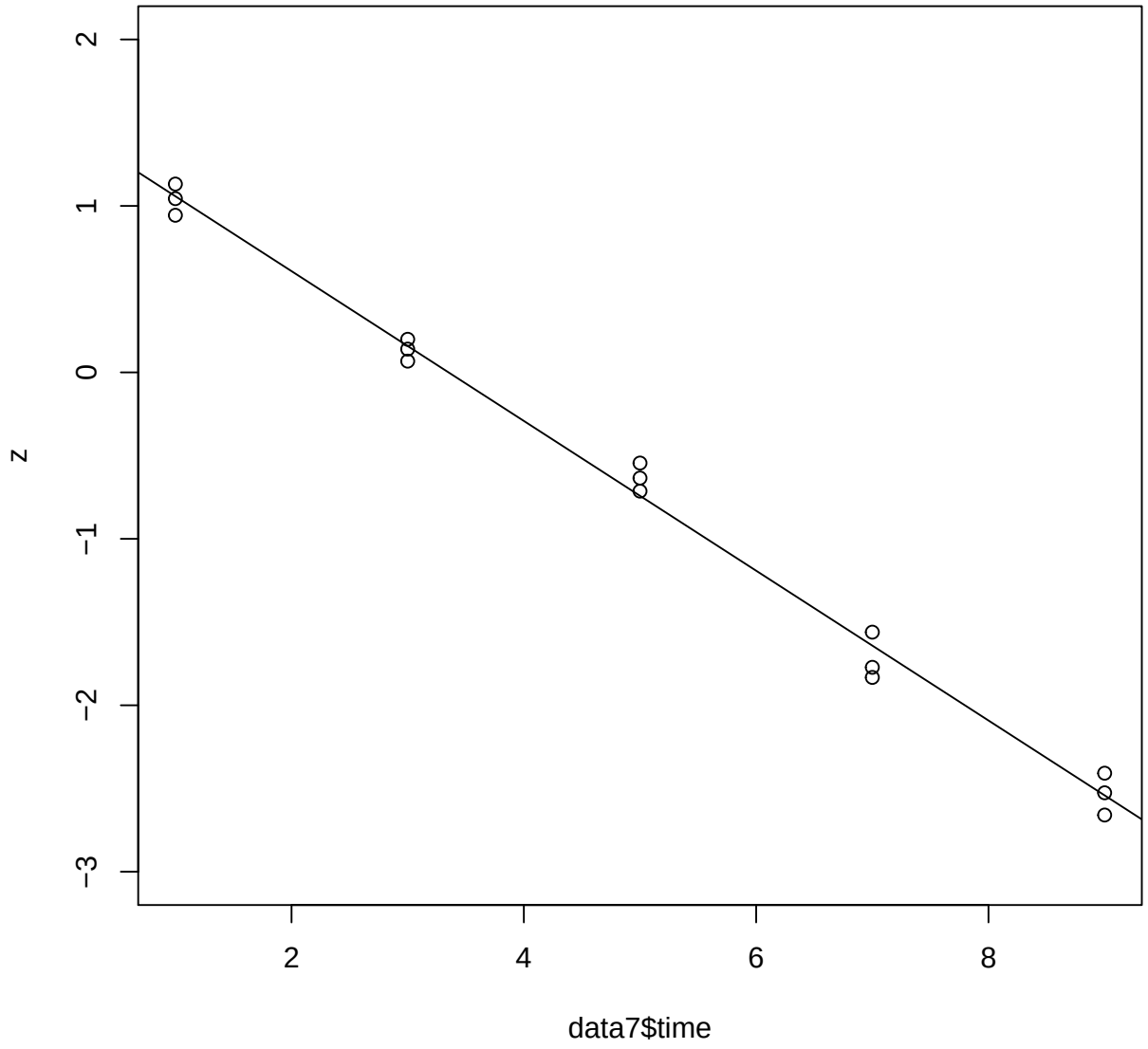
```
boxcox(data7$concentration~data7$time)
 lam <-boxcox(data7$concentration~data7$time)$x[which.max(boxcox(data7$concentration~dat
```



Możemy wyłuskać z komendy boxcox dokładną wartość i wynosi ona λ =0.020202 czyli jest bardzo bliska 0 w takim razie zgodnie z teorią w tym przypadku warto wziąć logarytm zmiennej objaśnianej. Co ma swój dalszy ciąg w kolejnym zadaniu.

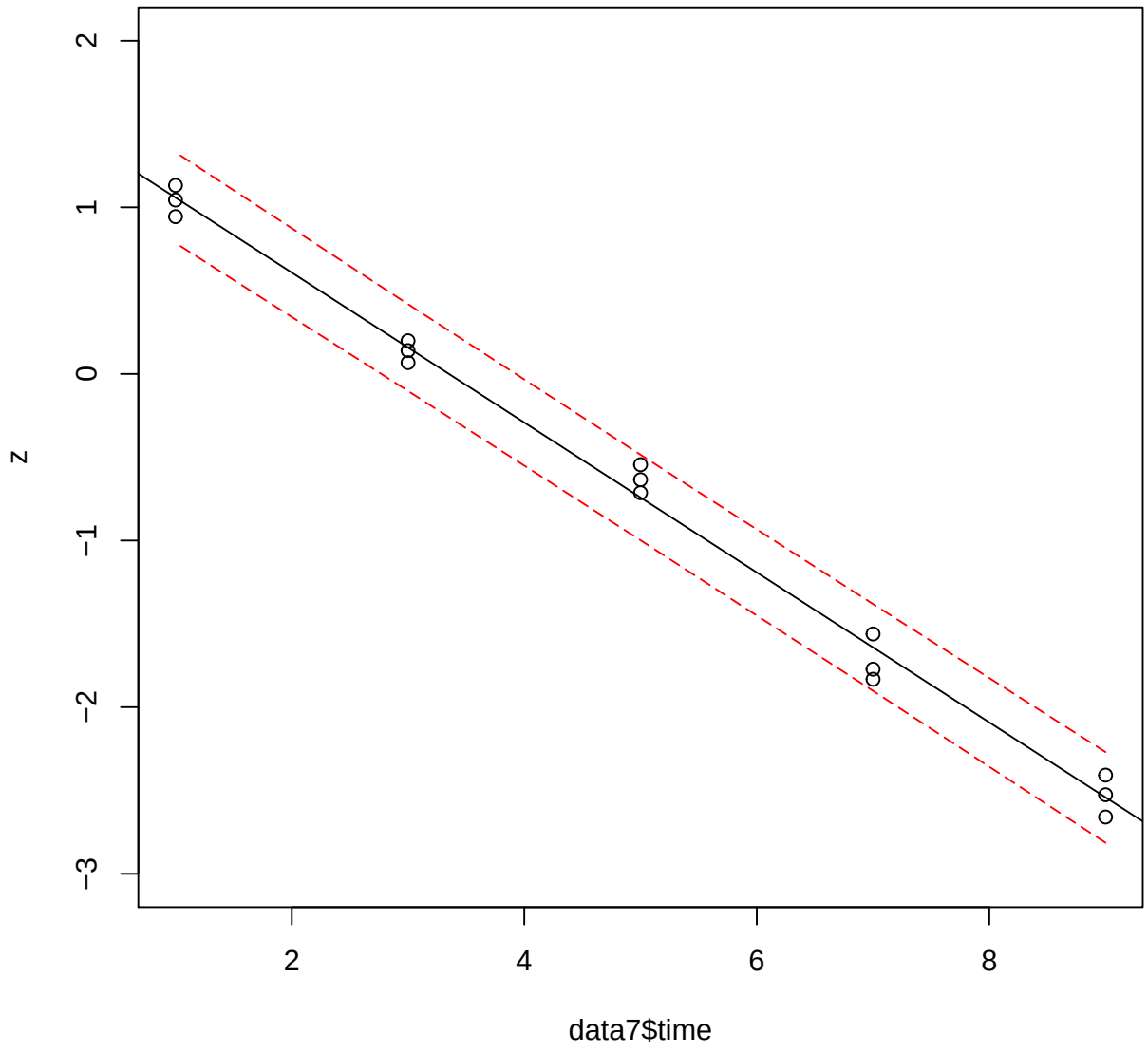
8. zadanie 10

Poniżej przedstawiam wykres ze zlogarytmowanym Y. Wizualnie można ocenić że model dobrze dopasowuje się do danych.



	model
intercept	1.5079164
slope	-0.4499258
statystyka T	-42.8745267
p-wartość	0.0000000
R^2	0.9929776

Powyżej przedstawiono podstawowe parametry modelu.Współczynnik R^2 jest bardzo bliski 1 co oznacza że model jest bardzo dobrze dopasowany do danych.hipoteza zerowa to $\beta_1 = 0$ a hipoteza alternatywna to $\beta_1 \neq 0$. p-wartość jest bliska 0 czyli odrzucamy hipotezę zerową na rzecz hipotezy alternatywnej. Można jeszcze dodatkowo podać wartość statystyki F i wynosi ona 1838. Liczba stopni swobody nie zmienia się ponieważ nadal mamy tyle samo obserwacji tylko troche przekształconych.Czyli dla statystyki T 13 stopni swobody a dla stystyki F 1 i 13 stopni swobody.



Wszystkie 15 obserwacji mieści się wewnątrz przedziałów predykcyjnych.

```
cor10 <- cor(z,predict(model10))
```

Korelacja wynosi 0.9964826 czyli jest wyższa niż w modelu z zadania 7

9. zadanie 11

```
plot(data7$time,data7$concentration)
lines(spline(data7$time,exp(predict(model10))))
```

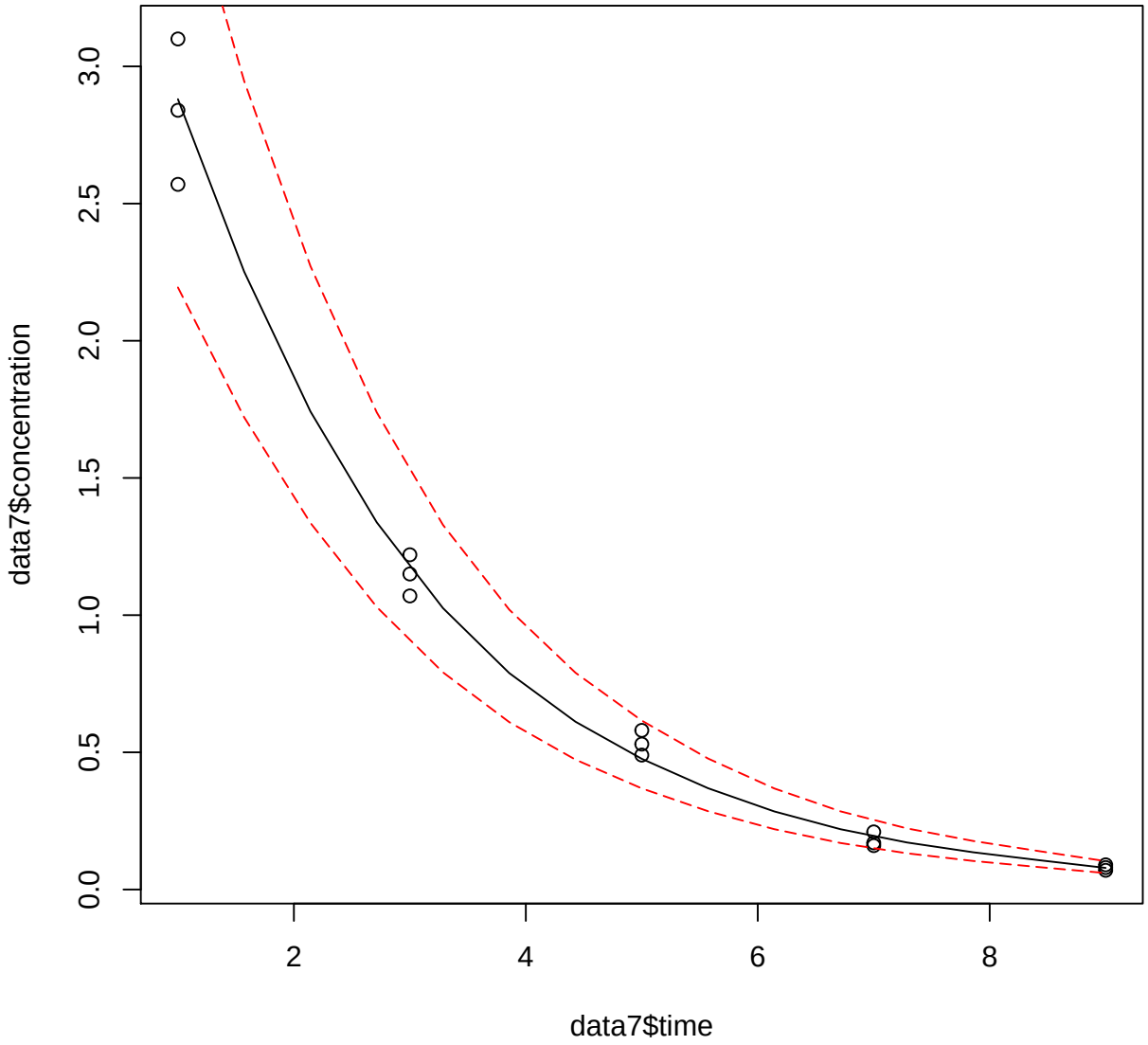
```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
'x' values
```

```
lines(spline(data7$time,exp(predict(model10, data.frame(time=data7$time) ,interval='pred
```

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
'x' values
```

```
lines(spline(data7$time,exp(predict(model10, data.frame(time=data7$time) ,interval='pred
```

```
## Warning in regularize.values(x, y, ties, missing(ties)): collapsing to unique
'x' values
```



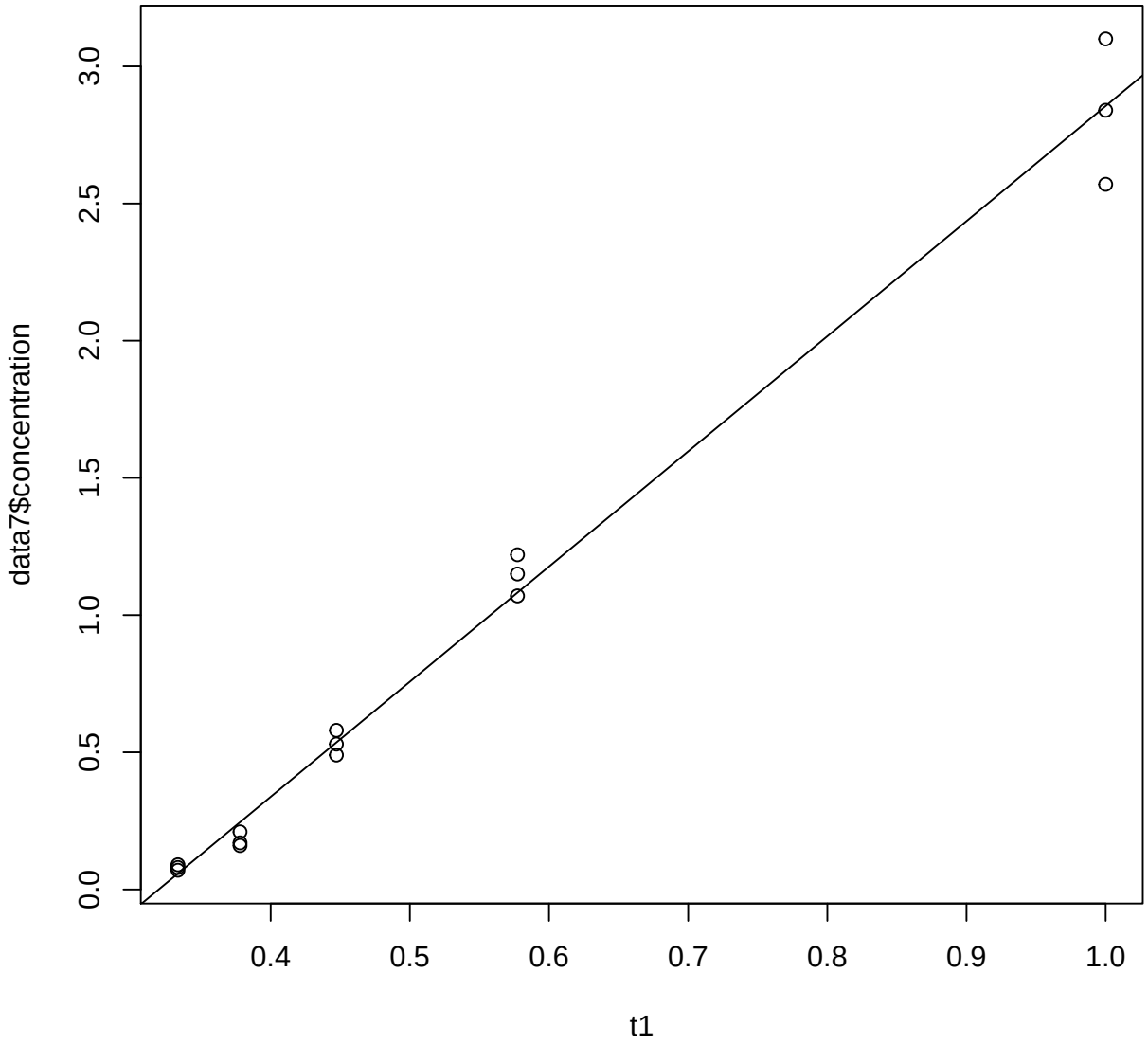
Ten wykres nieco różni się od wykresu z zadania 8 ponieważ tym razem zależność x i y jest wykładnicza a model liniowy został wykorzystany tylko pomocniczo. Nawet oceniając wizualnie widać że model ten jest lepiej dopasowany do danych niż model z zadania 8. Potwierdzają to również współczynniki modelu gdyż tak jak widać w tabelce powyżej R^2 wynosi około 0.99.Widzimy że wszystkie obserwacje trafiają między przedziały predykcyjne. Jednocześnie warto zauważyć że tutaj przedziały predykcyjne zwężają się wraz ze wzrostem x.

```
cor11 <- cor(data7$concentration,exp(predict(model10) ))
```

Korelacja między obserwowaną zmienną solution concentration a predykcją wartości na bazie modelu z zadania 10 wynosi 0.9945587

10. zadanie 12

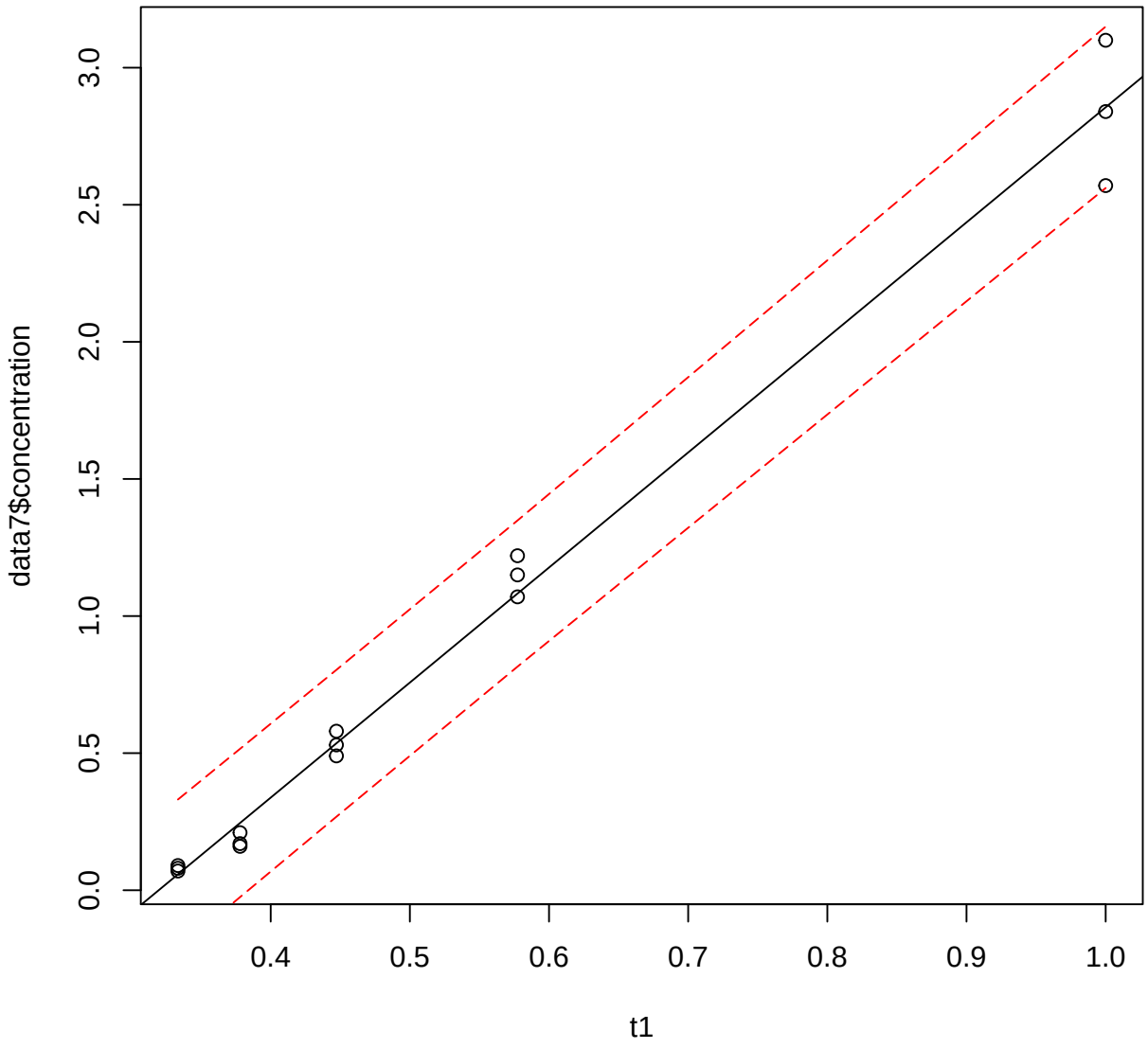
Poniżej widzimy wykres z nową zmienną t1 oraz naniosłem prostą regresji



Jak widzimy to przekształcenie z podniesieniem x do potęgi $-1/2$ również sprawiło że zależność jest liniowa. W poniższej tabeli przedstawiono dokładne parametry modelu

	model
intercept	-1.340777
slope	4.196319
statystyka T	32.803202
p-wartość	0.000000
R^2	0.988063

Z powodu bardzo niskiej p-wartości możemy odrzucić hipotezę zerową o tym, że $\beta_1 = 0$ na rzecz hipotezy alternatywnej na poziomie 0.05. Współczynnik R^2 jest dość bliski 1 co znaczy że model dobrze opisuje dane.



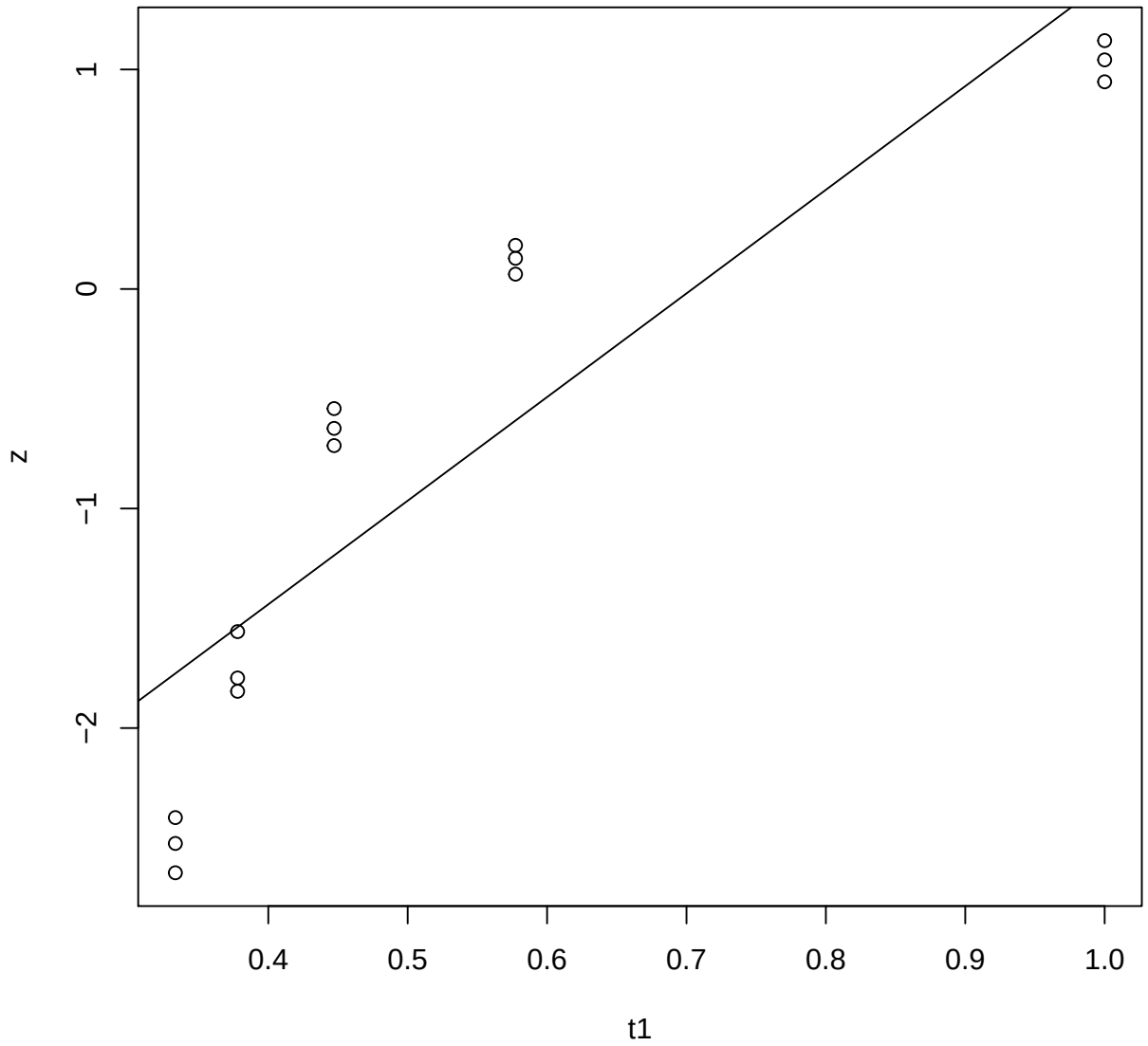
Wszystkie obserwacje wpadają do przedziałów predykcyjnych.

```
cor12<- cor(data7$concentration,predict(model12))
```

Korelacja między zmienną solution concentration a przewidywaniami modelu wynosi 0.9940136

11. kolejny model

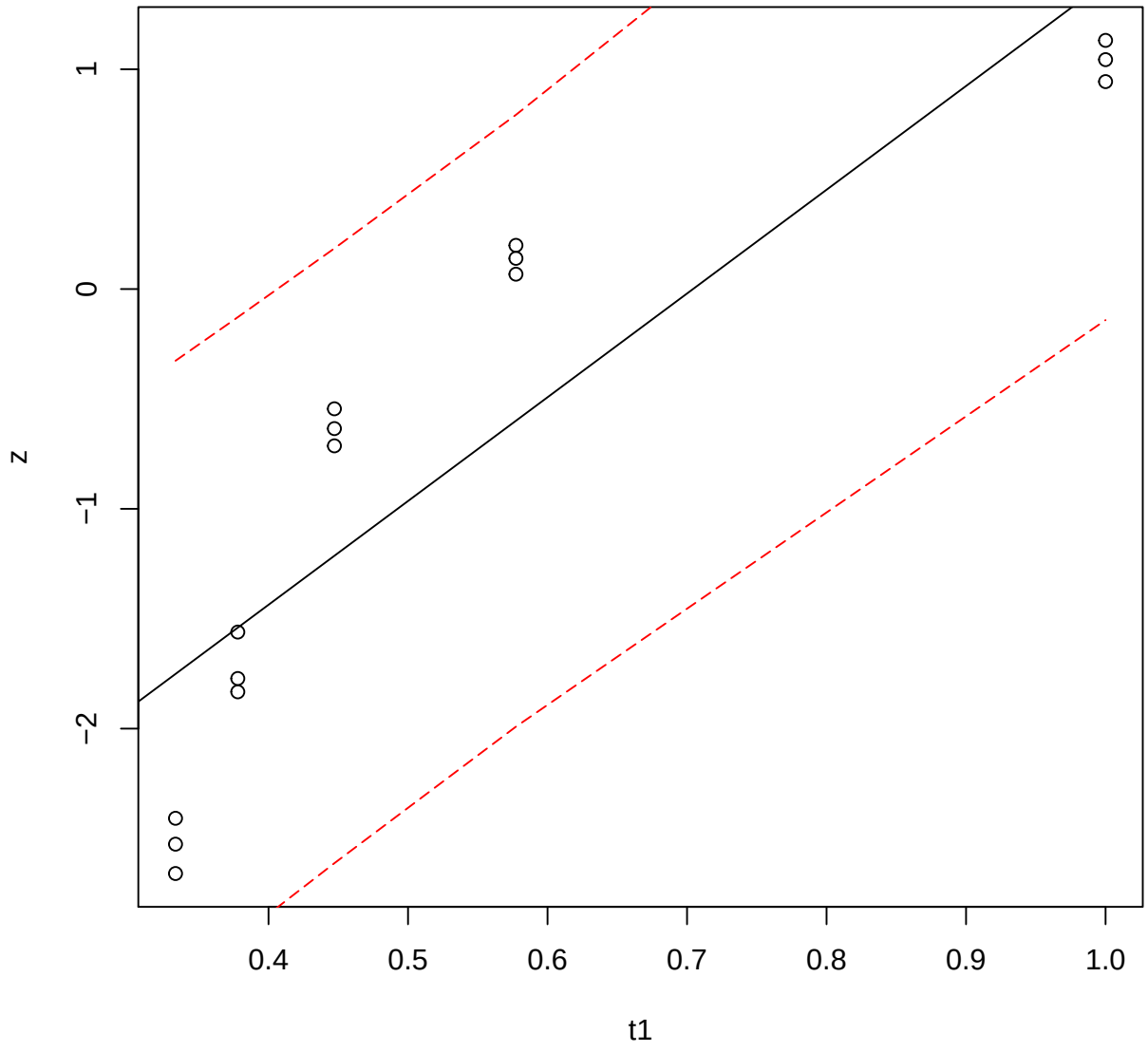
Poniżej widzimy wykres z zmienną t1 oraz log(y) naniosłem również prostą regresji



Jak widzimy po przekształceniu obu zmiennych model nie jest zbyt dobry bo założenie o liniowości jest łamane. Dodatkowo kalibracja obu zmiennych utrudnia interpretację wyników.

	model
intercept	-3.3248127
slope	4.7208160
statystyka T	7.0659411
p-wartość	0.0000085
R^2	0.7934131

Z powodu bardzo niskiej p-wartości możemy odrzucić hipotezę zerową o tym, że $\beta_1 = 0$ na rzecz hipotezy alternatywnej na poziomie 0.05. Współczynnik R^2 wynosi około 0.8 co jest dość dobrym wynikiem ale poprzednie modele były lepsze.



Wszystkie obserwacje wpadają do przedziałów predykcyjnych.

```
cor13<- cor(z,predict(model13))
```

Korelacja między zmienną solution concentration a przewidywaniami modelu wynosi 0.8907374

12. porównanie modeli

	model7	model10	model12	dodatkowy model
intercept	2.5753333	1.5079164	-1.340777	-3.3248127
slope	-0.3240000	-0.4499258	4.196319	4.7208160
statystyka T	-7.4829029	-42.8745267	32.803202	7.0659411
p-wartość	0.0000046	0.0000000	0.000000	0.0000085
R^2	0.8115774	0.9929776	0.988063	0.7934131

Porównując współczynniki R^2 okazuje się, że najlepszym modelem jest model z zadania 10 gdzie wzięliśmy logarytm ze zmiennej Y. Model z zadania 12 gdzie podnieśliśmy X do potęgi $-1/2$ również jest bardzo dobry. Warto również odnotować że to że te dwa modele dobrze wyjaśniają zmienność Y to są one dość różne jeśli chodzi o inne współczynniki. W szcególnośći slope modelu10 jest ujemny tak jak w pierwotnych danych a slope modelu12 już ma dodatni slope. Nie jest to oczywiście żadna magia tylko konsekwencja użytych przekształceń. Gdyż funkcja log zachowuje monotoniczność(jeśli pierwotnie mieliśmy funkcję malejącą to nadal tak będzie) a już odwrócony pierwiastek zmienia tę relację z malejącej na rosnącą. Według mnie jest to jeszcze jeden argument na korzyść modelu z zadania 10 gdyż jest on łatwiejszy do interpetacji. Pozostałe dwa modele są już wyraźnie gorsze.