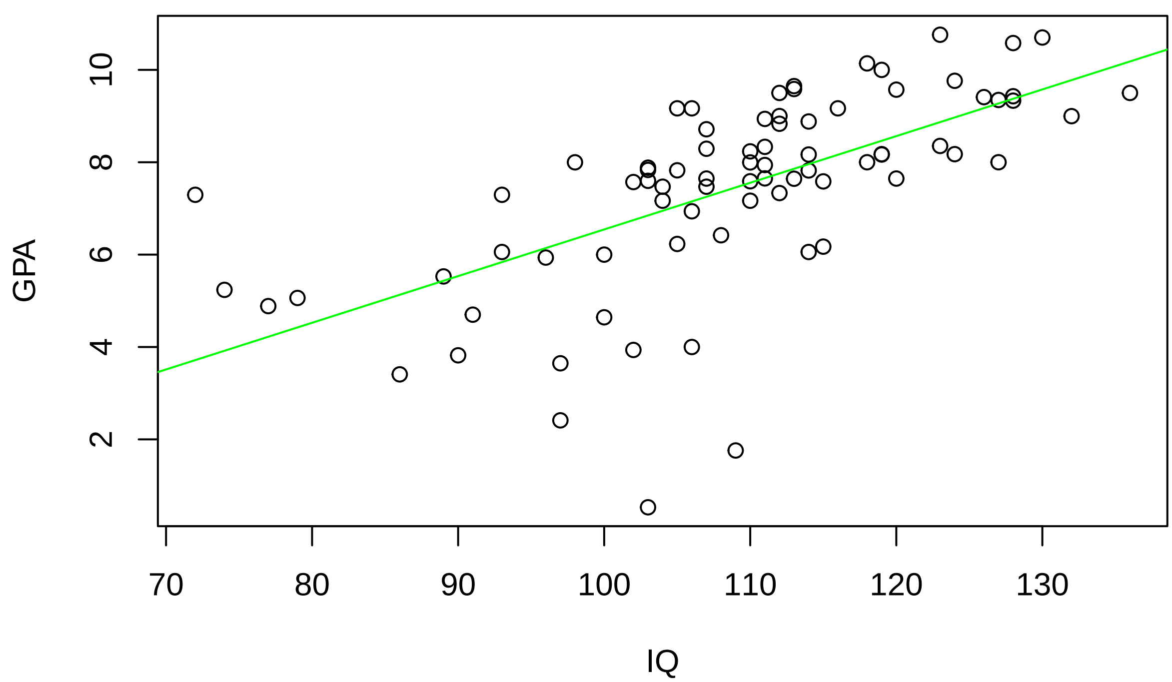
Sprawozdanie Modele liniowe 3

Katarzyna Stasińska

2023-11

Zadanie 1

a)




```
## Równanie regresji to Y = 0.101X + -3.5571
```

```
## [1] "Współczynnik determinacji polecenia wbudowane:  0.401614629007563"
```

```
## [1] "Współczynnik determinacji wzory teoretyczne:  0.401614629007563"
```

Współczynnik determinacji R^2 jest miarą jakości dopasowania modelu. Na podstawie statystyki R^2 możemy powiedzieć, że zmienność wyjaśniona przez model stanowi 40,2% zmienności całkowitej Y.

b)

Aby przetestować hipotezę, że GPA nie jest skorelowane z IQ, należy rozważyć hipotezę: $H_0 : \beta_1 = 0$. Korzystając z funkcji wbudowanych mamy:

```
## Statystyka testowa  51.0084525528291 z 1 i 76 stopniami swobody.
```

```
## p-value wynosi  4.73734072401733e-10
```

Korzystając ze wzorów teoretycznych mamy:

```
## Statystyka testowa  51.0084525528291 z 1 i 76 stopniami swobody.
```

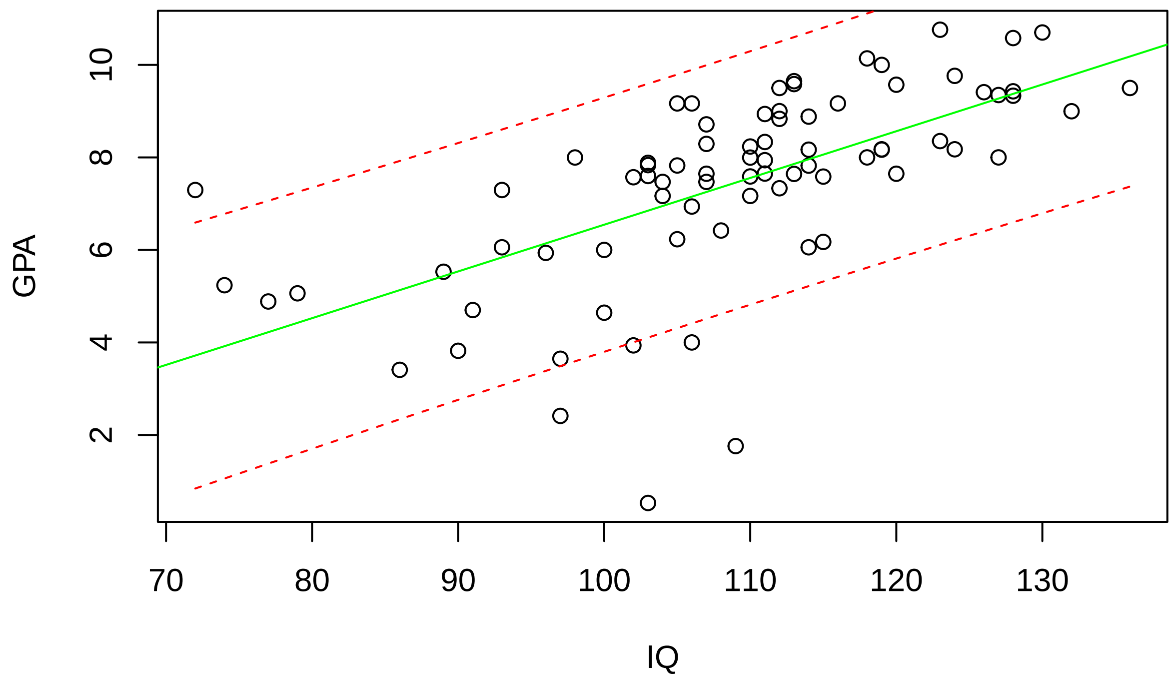
```
## p-value wynosi  4.73734051986696e-10
```

Ustalając standardowy poziom istotności $\alpha = 0.05$ mamy $p < \alpha$, odrzucamy hipotezę zerową. GPA i IQ są ze sobą skorelowane.

c)

```
## Dla k= 75  Przewidywane GPA: 4.0195715931079
## Przedział predykcyjny [ 1.16590067798596 , 6.87324250822984 ]
##
## Dla k= 100  Przewidywane GPA: 6.54511406984245
## Przedział predykcyjny [ 3.79752989821741 , 9.29269824146748 ]
##
## Dla k= 140  Przewidywane GPA: 10.5859820326177
## Przedział predykcyjny [ 7.75034990235693 , 13.4216141628785 ]
```

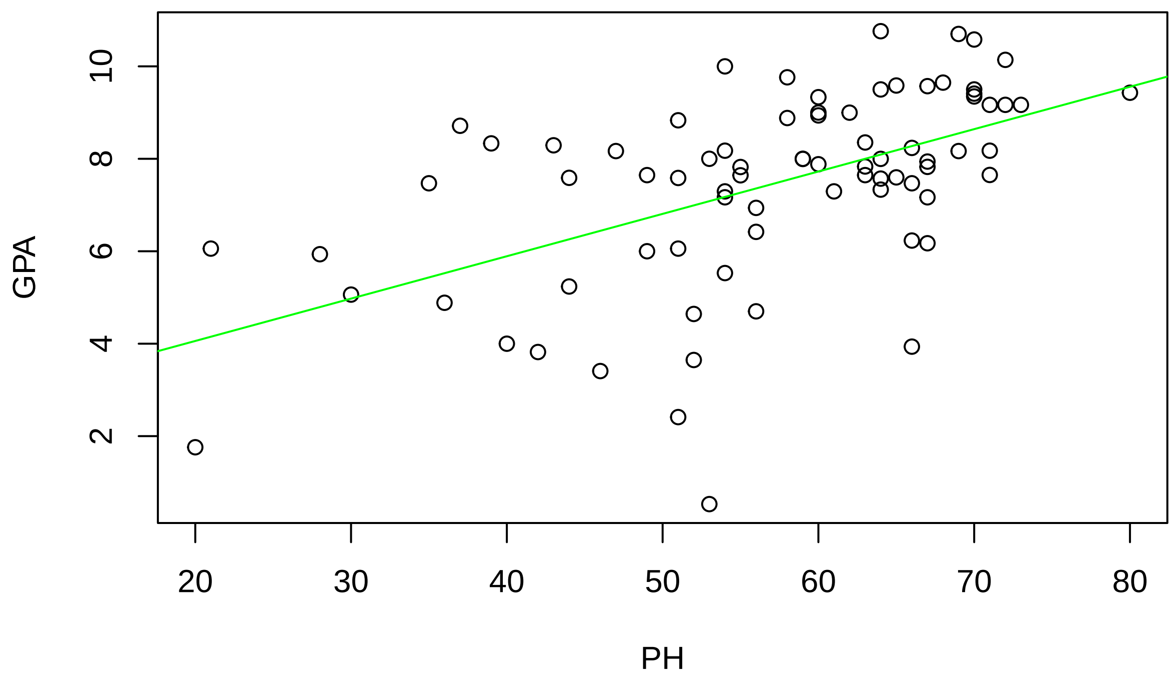
d)



6 obserwacji znajduje się poza tymi przedziałami, liczba ta podzielona przez wszystkie inne obserwacje powinna wynosić około 10%. $6/78 = 0.07692308$

Zadanie 2

a)




```
## Równanie regresji to Y = 0.0917X + 2.2259
```

```
## [1] "Współczynnik determinacji:  0.293582902749107"
```

Współczynnik determinacji R^2 jest miarą jakości dopasowania modelu. Na podstawie statystyki R^2 możemy powiedzieć, że zmienność wyjaśniona przez model stanowi 29,4% zmienności całkowitej Y.

b)

Aby przetestować hipotezę, że GPA nie jest skorelowane z PH, należy rozważyć hipotezę: $H_0 : \beta_1 = 0$.

```
## Statystyka testowa  31.5851650473398 z 1 i 76 stopniami swobody.
```

```
## p-value wynosi  3.00641629030262e-07
```

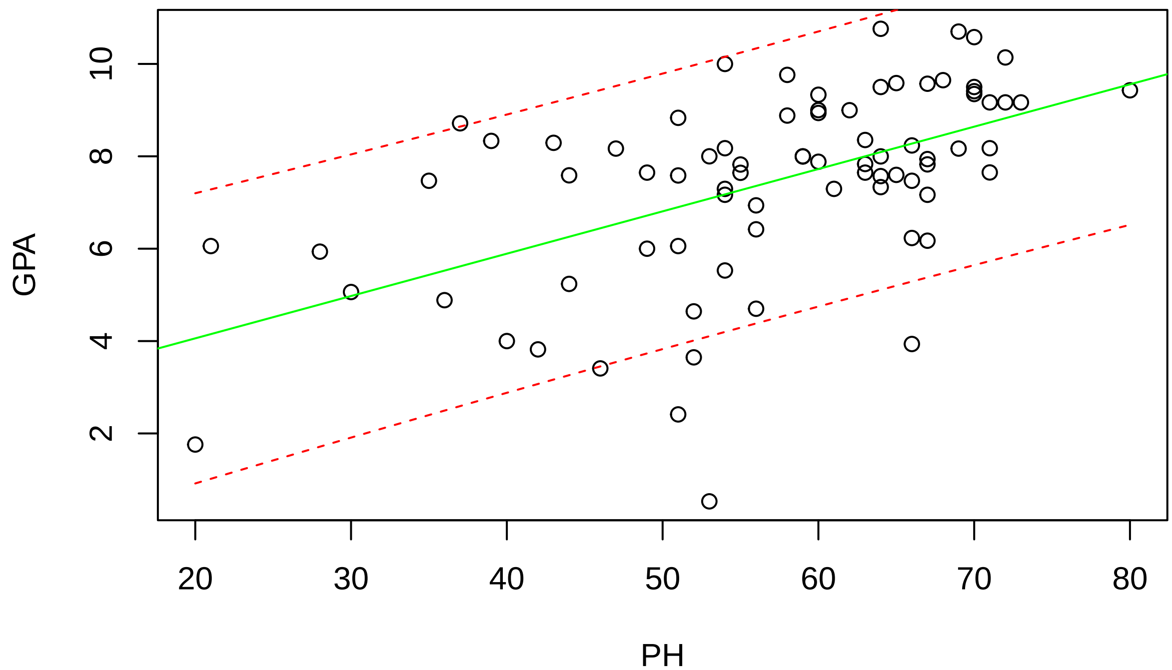
Ustalając standardowy poziom istotności $\alpha = 0.05$ mamy $p < \alpha$, odrzucamy hipotezę zerową. GPA i PH są ze sobą skorelowane.

c)

```
## Dla k= 25  Przewidywane GPA: 4.51719006922042
## Przedział predykcyjny [ 1.41665814236331 , 7.61772199607753 ]
##
## Dla k= 55  Przewidywane GPA: 7.26675895731678
## Przedział predykcyjny [ 4.28970699878095 , 10.2438109158526 ]
```

```
##
## Dla k= 85  Przewidywane GPA: 10.0163278454131
## Przedział predykcyjny [ 6.94391470941984 , 13.0887409814064 ]
```

d)



W tym przypadku również 6 obserwacji znajduje się poza tymi przedziałami, liczba ta podzielona przez wszystkie inne obserwacje powinna wynosić około 10%. $6/78 = 0.07692308$

e)

Wynik testu IQ jest lepszym predyktorem GPA, ponieważ przedziały predykcyjne mają mniejszą szerokość. Poza tym współczynnik determinacji przyjmuje większą wartość.

zadanie 3

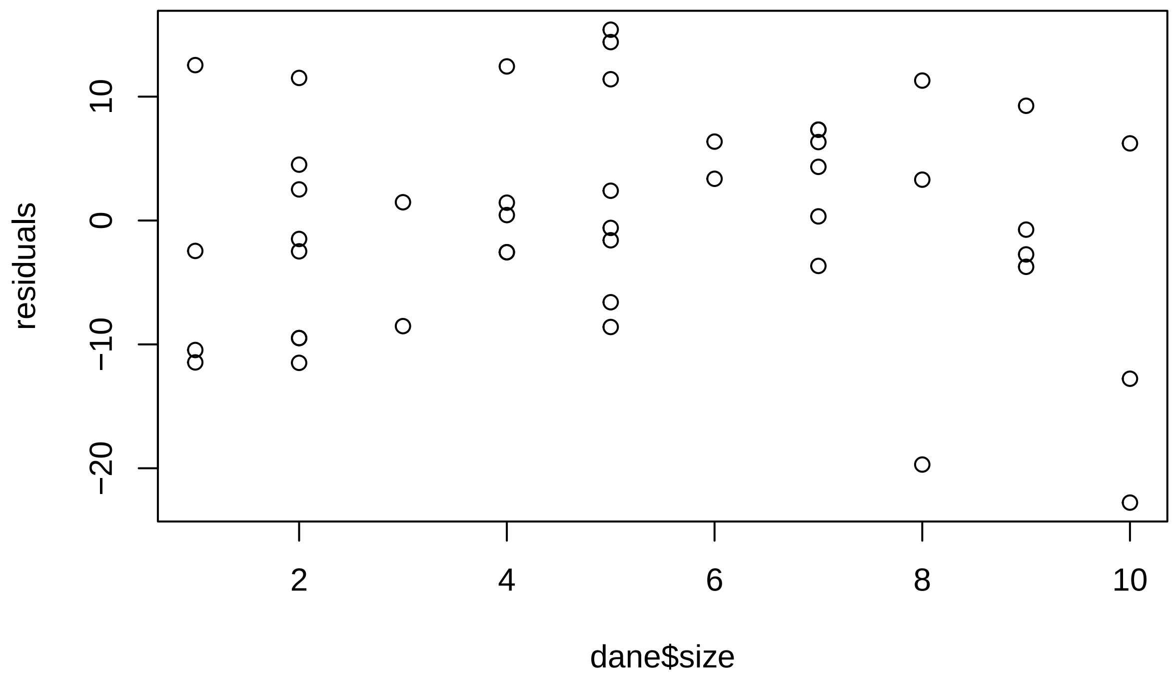
a)

```
reg1 = lm(hours~size, dane)
residuals = reg1$residuals
sum(residuals)
```

```
## [1] -6.217249e-15
```

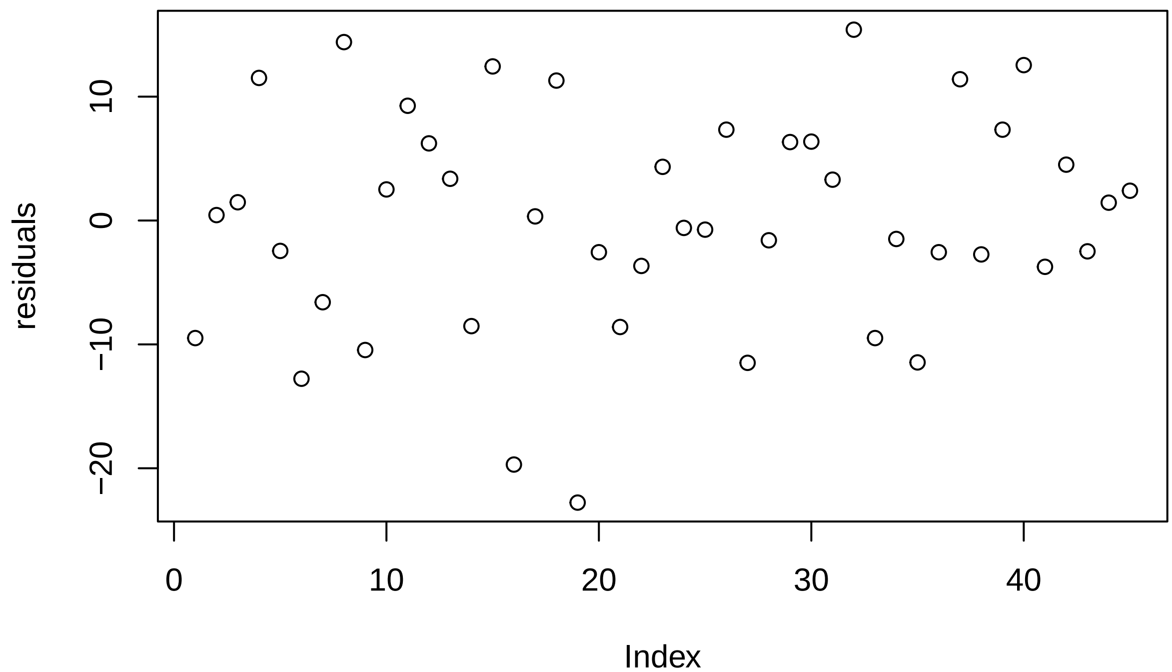
Na podstawie powyższych obliczeń możemy zauważyć, że suma residuów jest prawie równa zero.

b)



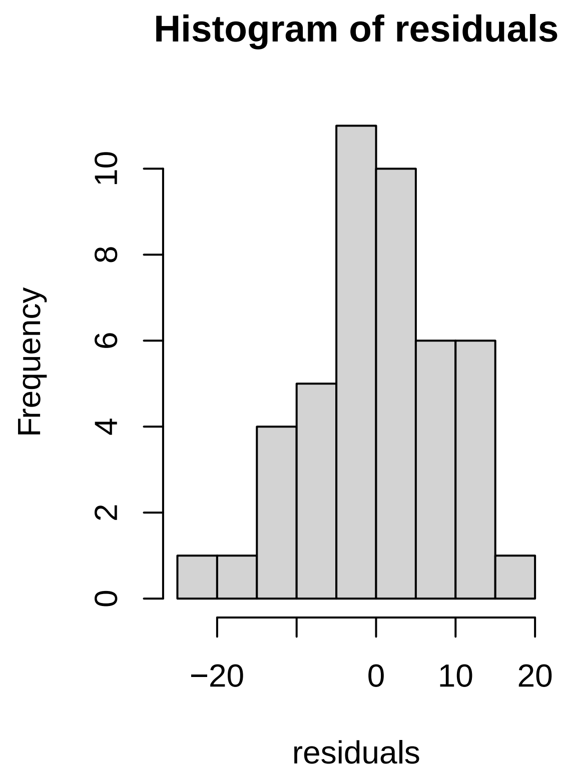
Dwie wartości w okolicach -20, dla rozmiaru równego 8 i rozmiaru równego 10 mogą świadczyć o jakiejś tendencji wraz ze wzrostem rozmiaru. Wartości residuuów oscylują wokół zera.

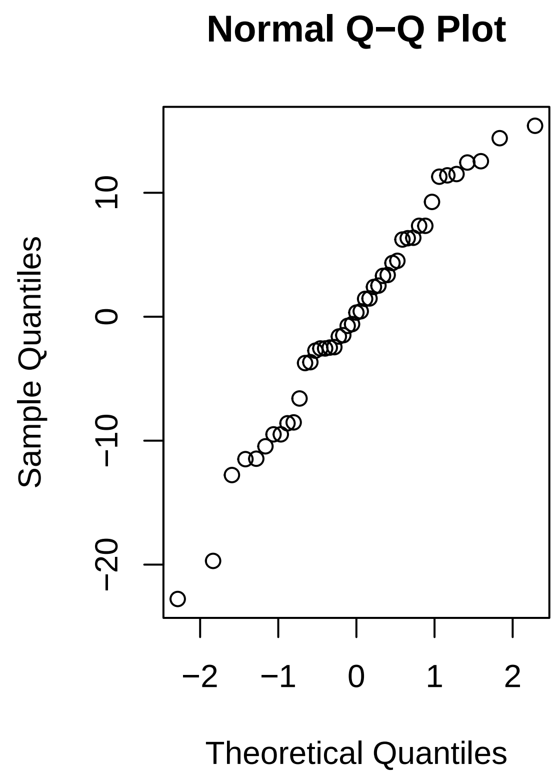
c)



Nie zauważam żadnych nietypowych wzorców.

d)





Na podstawie powyższych wykresów można stwierdzić, że residua z dużym prawdopodobieństwem nie pochodzą z rozkładu normalnego. Na obu wykresach ogony są niedopasowane.

zadanie 4

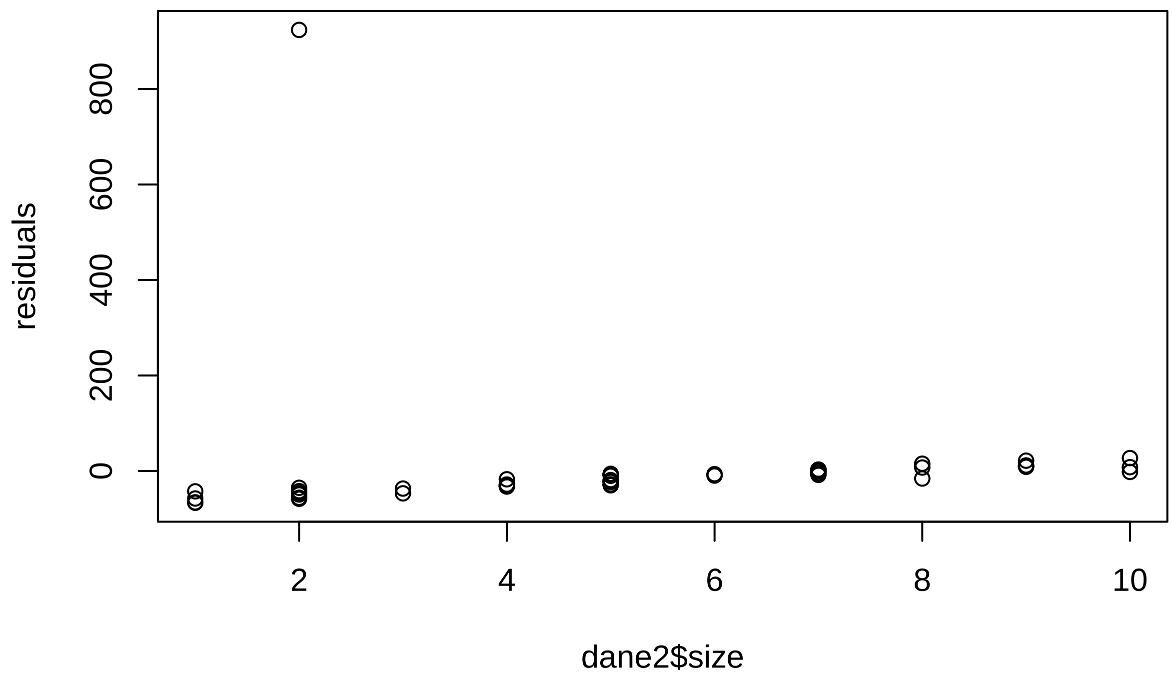
a)

```
##      nazwa                  reg1               reg2
## 1 równanie   Y=15.0352*X-0.5802 Y=6.5939*X+63.0915
## 2  t value     31.1232581195971  0.862491738582135
## 3  p value 4.00903211860461e-31  0.393093684635158
## 4      R^2     0.95749548347908 0.0166255541457373
## 5  sigma^2      79.450628453458   20452.1867431773
```

W przypadku zmodyfikowanych danych p-wartość jest na tyle duża, że przy standardowym $\alpha = 0.05$ nie możemy odrzucić hipotezy zerowej. Sam model nie jest zbyt dobrze określony, wartość R^2 jest mała, a wariancja błędów bardzo wysoka. W przeciwieństwie do początkowych danych, które wykluczają hipotezę zerową, a pozostałe parametry świadczą o dość dobrym dobraniu modelu.

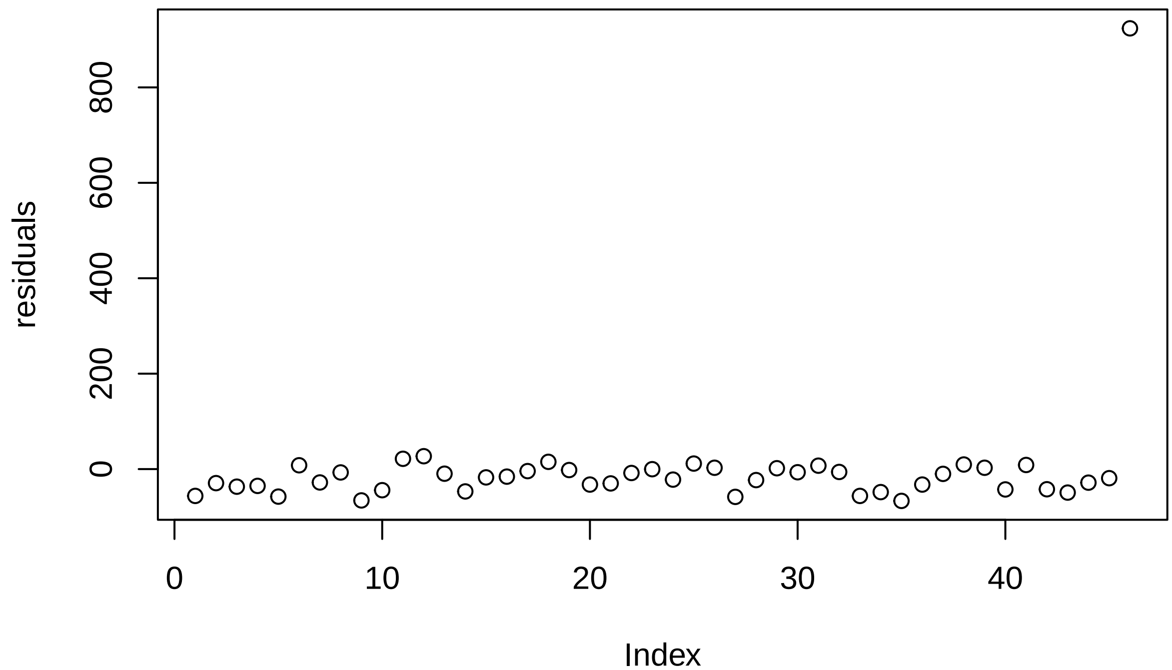
b)

b)



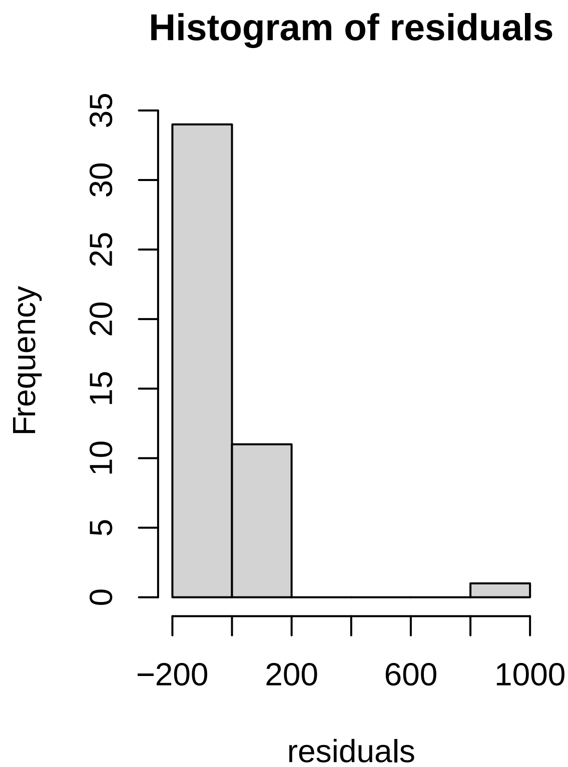
Nietypową obserwacją jest ta obserwacja, którą dodaliśmy. Umieszczenie jej na wykresie sprawia, że wartości pozostałych residuuów są niewidoczne, przez co ciężko wyciągać z nich jakiekolwiek wnioski.

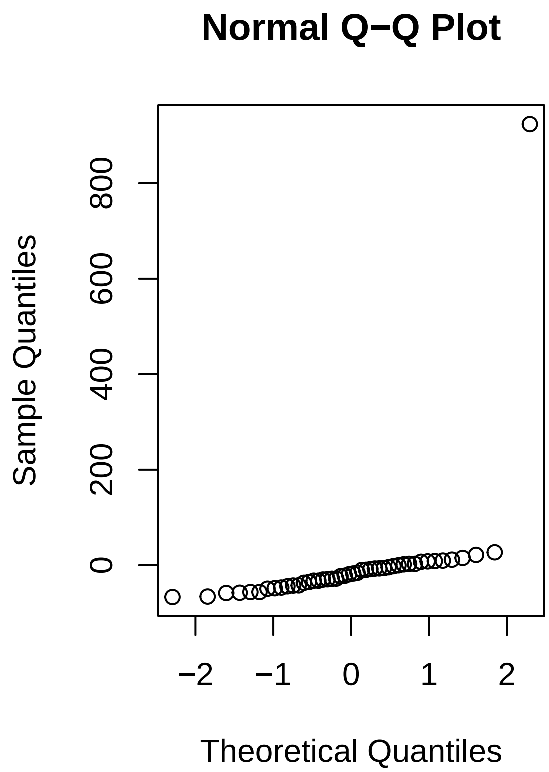
c)



Ponownie nietypową obserwacją jest ta obserwacja, którą dodaliśmy. Umieszczenie jej na wykresie sprawia, że wartości pozostałych residuuów są niewidoczne, przez co ciężko wyciągać z nich jakiekolwiek wnioski.

d)





Na podstawie powyższych wykresów nie można stwierdzić, czy residua pochodzą z rozkładu normalnego. Dodanie obserwacji o znacznie większej wartości niż pozostałe sprawiło, że odczytanie z wykresów jak zachowują się residua jest niemożliwe.

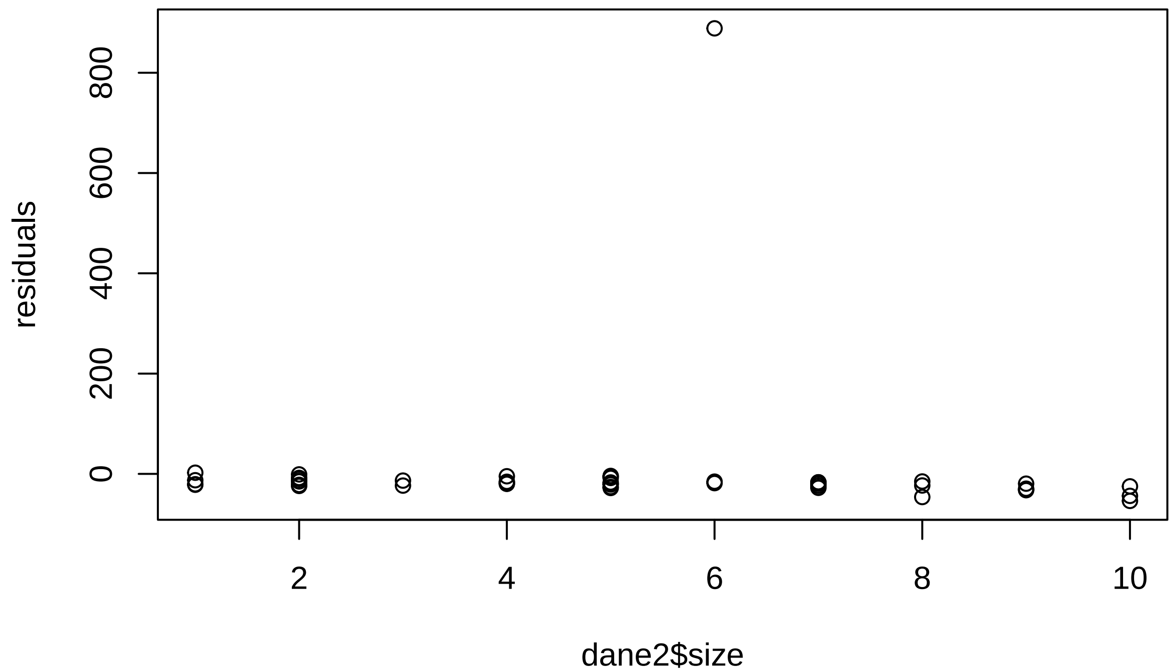
c)

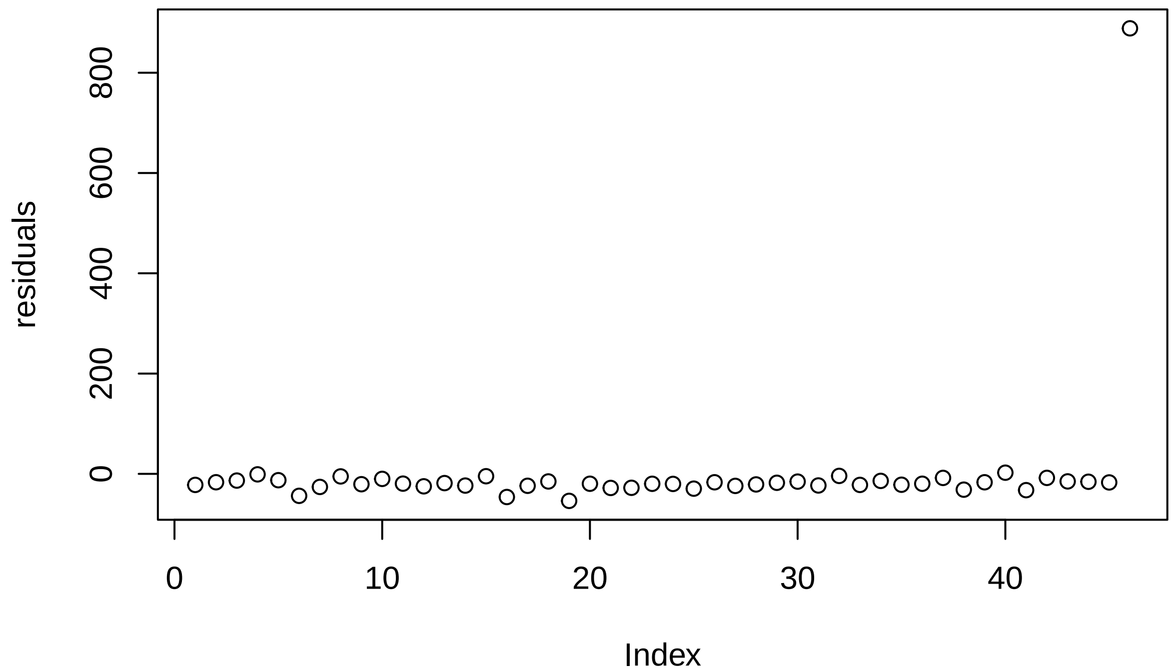
a)

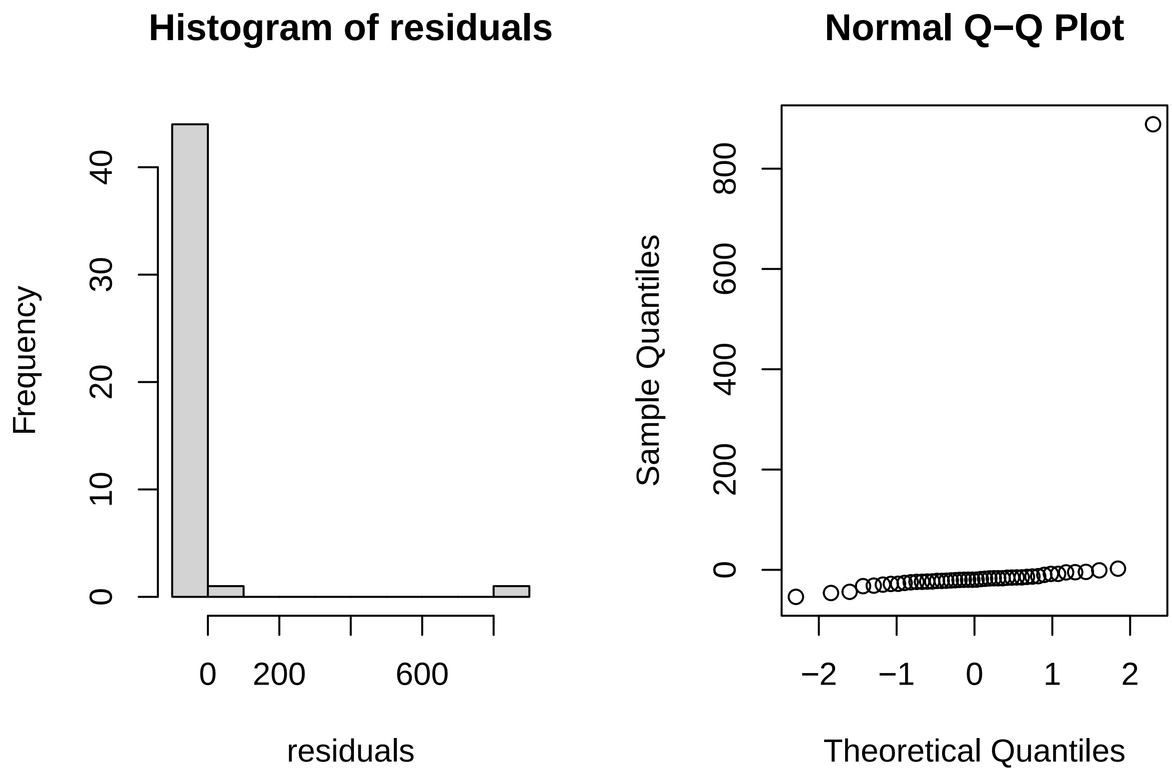
```
##      nazwa                 reg1              reg2
## 1 równanie   Y=15.0352*X-0.5802 Y=17.3552*X+7.3078
## 2  t value     31.1232581195971   2.35942093294602
## 3  p value 4.00903211860461e-31 0.0228060517139825
## 4      R^2     0.95749548347908   0.11231024795722
## 5  sigma^2      79.450628453458   18462.1398850894
```

Ponownie w przypadku zmodyfikowanych danych p-wartość jest na tyle duża, że przy standardowym $\alpha = 0.05$ nie możemy odrzucić hipotezy zerowej. Sam model nie jest zbyt dobrze określony, wartość R^2 jest mała, a wariancja błędów bardzo wysoka. W przeciwieństwie do początkowych danych, które wykluczają hipotezę zerową, a pozostałe parametry świadczą o dość dobrym dobraniu modelu.

b)




residuals
0
200
400
600
800
0
10
20
30
40
Index

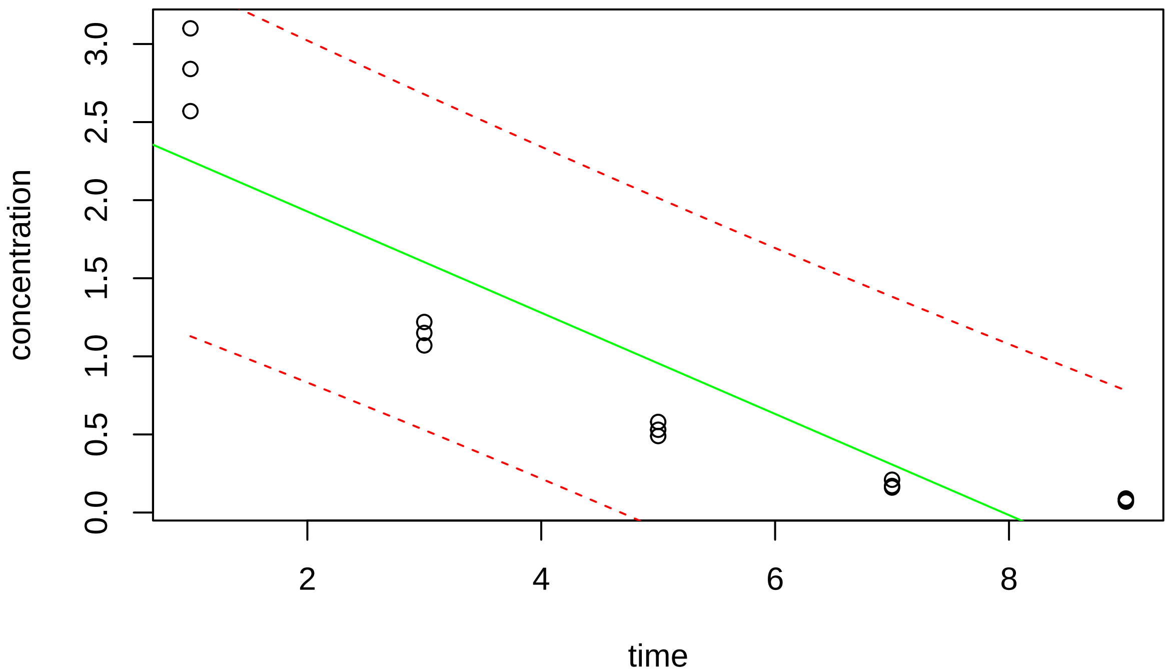


W przypadku wszystkich wykresów można wysnuć dokładnie takie same wnioski, co w przypadku dodania obserwacji $(1000; 2)$.

zadanie 5

a)

```
## Równanie regresji to Y = -0.324X + 2.5753
```



Czas jest dobrym predyktorem, ale to nie jest zależność liniowa.

b)

```
## [1] "Współczynnik determinacji:  0.811577371314103"
```

Rozważmy hipotezę: $H_0 : \beta_1 = 0$ i $H_1 : \beta_1 \neq 0$.

```
## Statystyka testowa  55.9938363065253 z 1 i 13 stopniami swobody.
```

```
## p-value wynosi  4.61119948957169e-06
```

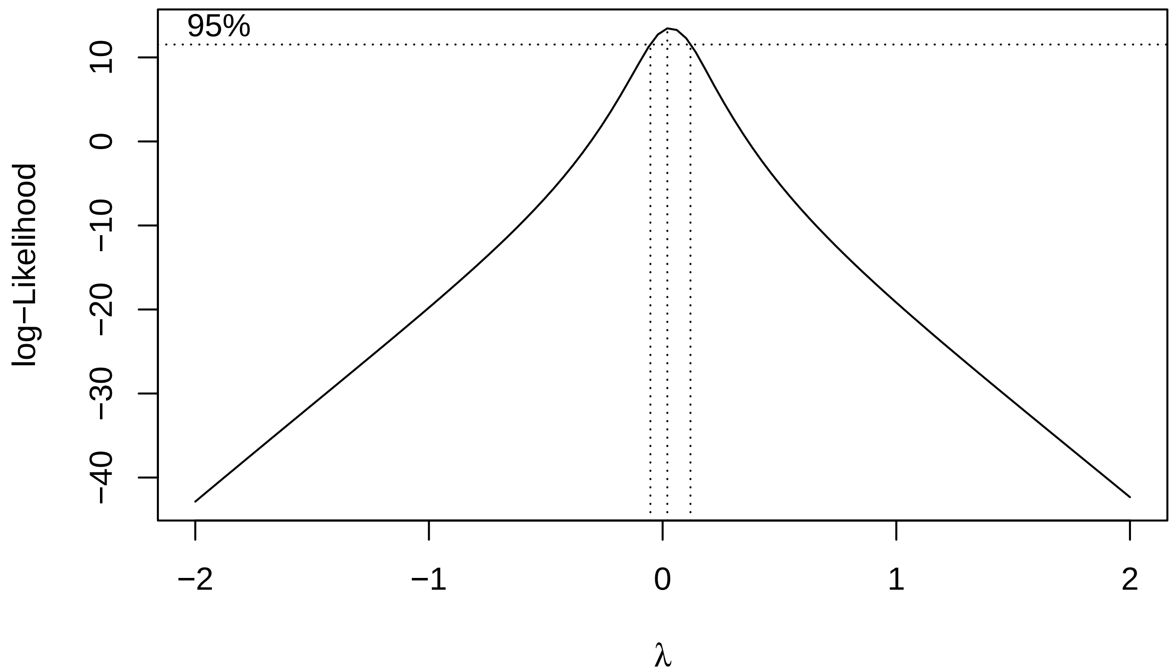
Ustalając standardowy poziom istotności $\alpha = 0.05$ mamy $p < \alpha$, odrzucamy hipotezę zerową. Stężenie i czas są ze sobą skorelowane.

c)

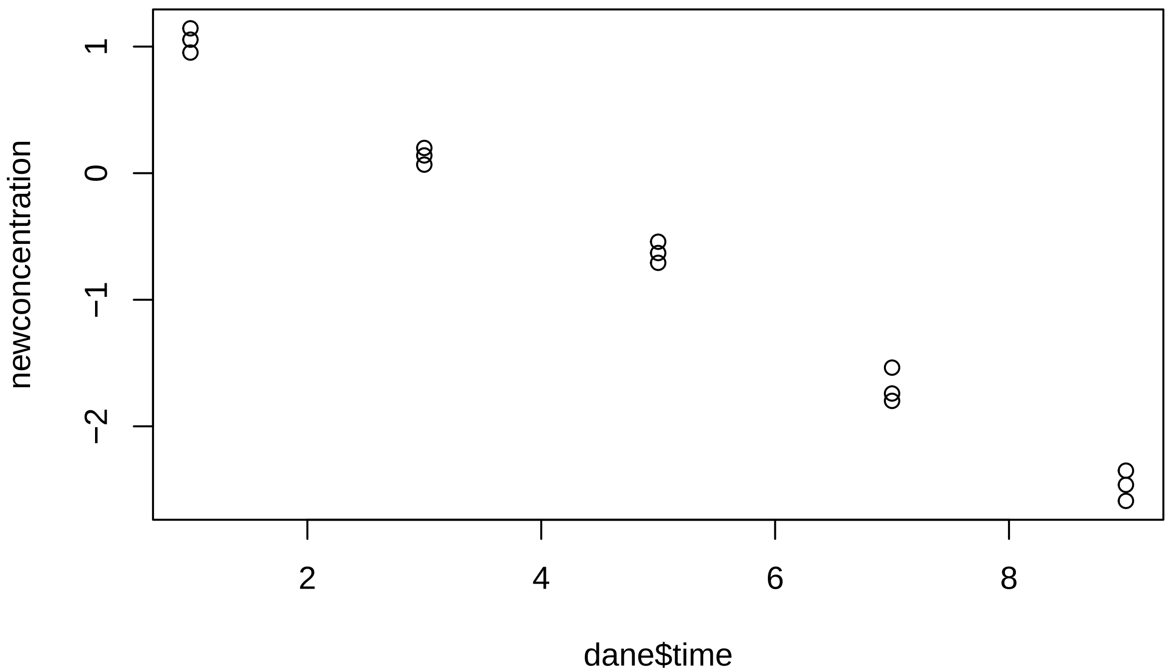
```
## [1] "Współczynnik korelacji: 0.900875891182633"
```

zadanie 6

```
library(MASS)
b = boxcox(reg1)
```




```
lambda = b$x[which.max(b$y)]
newconcentration = (dane$concentration^lambda - 1) / lambda
plot(newconcentration~dane$time)
```



zadanie 7

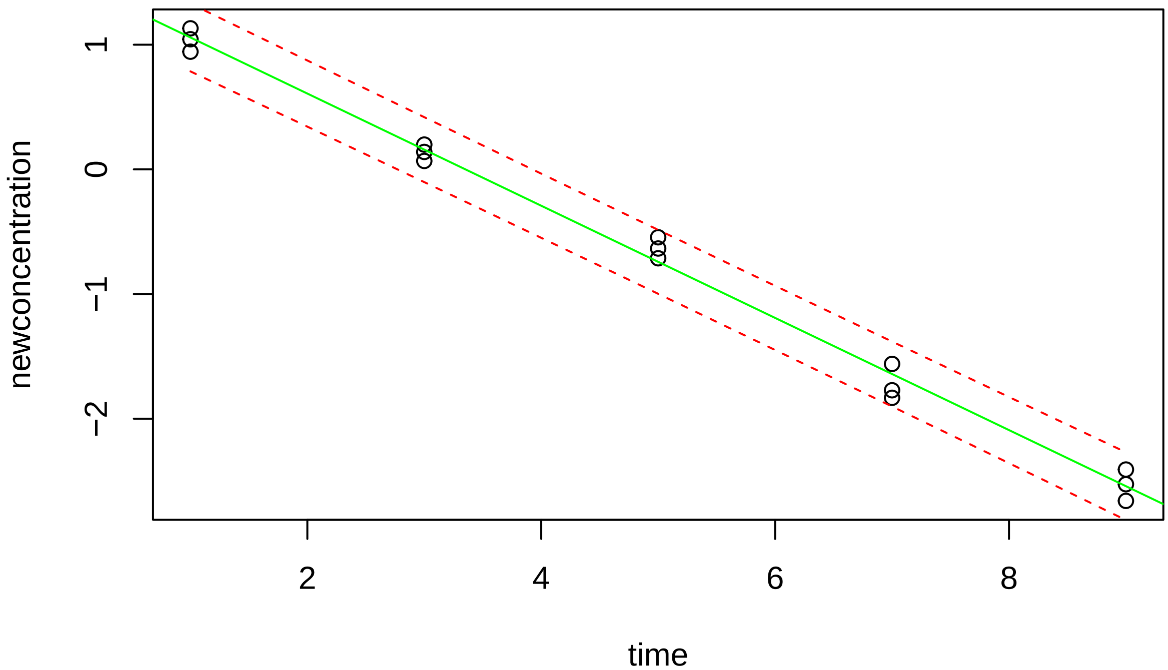
a)

```
##  [1] -2.65926004 -2.40794561 -2.52572864 -1.83258146 -1.77195684 -1.56064775
##  [7] -0.71334989 -0.54472718 -0.63487827  0.19885086  0.13976194  0.06765865
## [13]  1.04380405  0.94390590  1.13140211
```

b)

a)

```
## Równanie regresji to Y = -0.4499X + 1.5079
```



Czas jest dobrym predyktorem, występuje zależność liniowa.

b)

```
## [1] "Współczynnik determinacji:  0.992977623087684"
```

Rozważmy hipotezę: $H_0 : \beta_1 = 0$ i $H_1 : \beta_1 \neq 0$.

```
## Statystyka testowa  1838.22504279147 z 1 i 13 stopniami swobody.
```

```
## p-value wynosi  2.18825207386548e-15
```

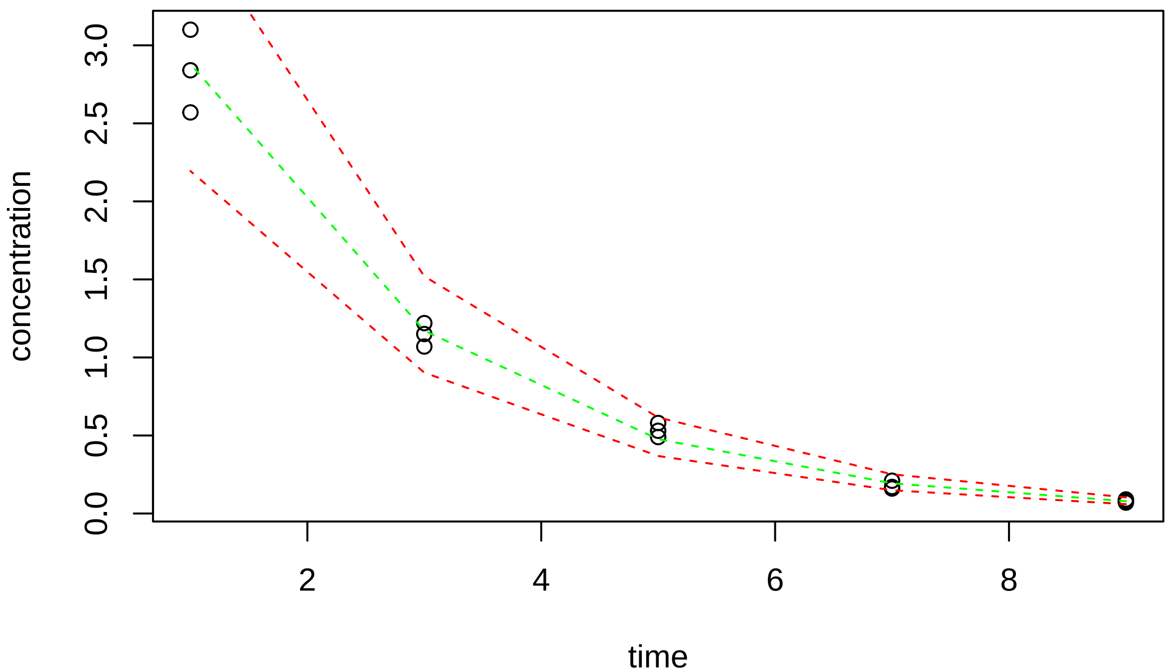
Ustalając standardowy poziom istotności $\alpha = 0.05$ mamy $p < \alpha$, odrzucamy hipotezę zerową. Stężenie i czas są ze sobą skorelowane.

c)

```
## [1] "Współczynnik korelacji: 0.996482625582445"
```

Po transformacji zmiennej objaśnianej, zależność między zmienną objaśnianą a zmienną objaśniającą jest w większym stopniu liniowa. Współczynnik korelacji jest bliski 1 i znacząco większy od tego w zadaniu 5.

c)



Powyższy wykres lepiej opisuje oryginalne dane niż wykres z zadania 5.

d)

```
## [1] "Współczynnik korelacji: 0.994558741278343"
```

Współczynnik korelacji wyszedł zdecydowanie większy niż w zadaniu 5.

zadanie 8

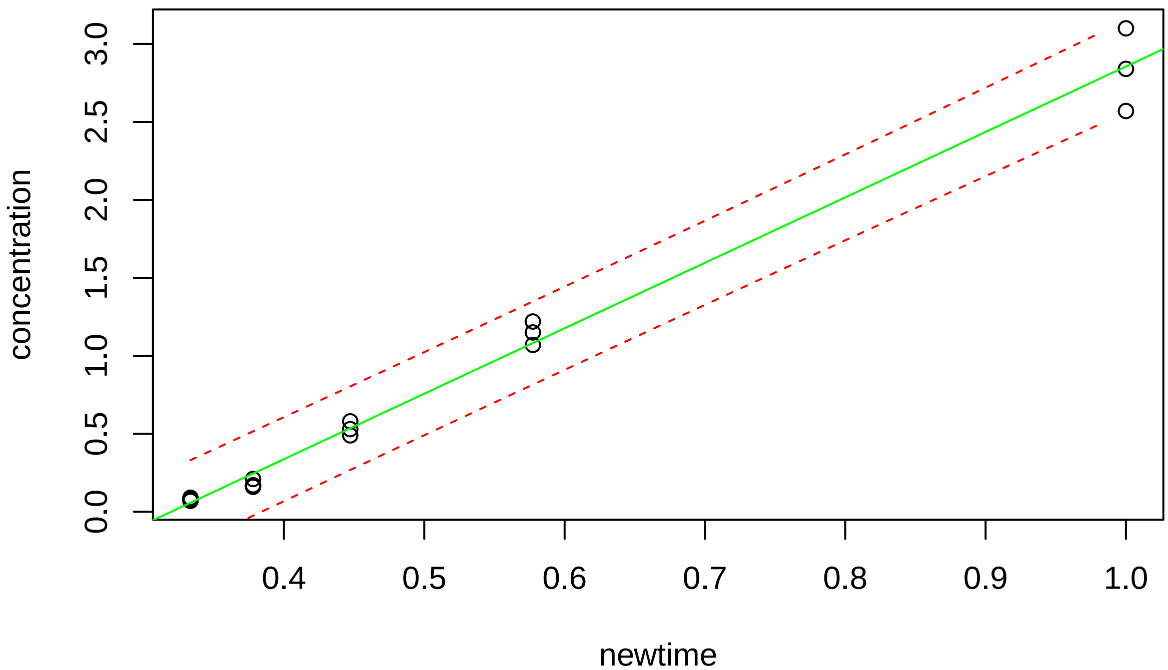
a)

```
##  [1] 0.3333333 0.3333333 0.3333333 0.3779645 0.3779645 0.3779645 0.4472136
##  [8] 0.4472136 0.4472136 0.5773503 0.5773503 0.5773503 1.0000000 1.0000000
## [15] 1.0000000
```

b)

a)

```
## Równanie regresji to Y = 4.1963X + -1.3408
```



Czas jest dobrym predyktorem, występuje zależność liniowa.

b)

```
## [1] "Współczynnik determinacji:  0.988062991398199"
```

Rozważmy hipotezę: $H_0 : \beta_1 = 0$ i $H_1 : \beta_1 \neq 0$.

```
## Statystyka testowa  1076.05006552805 z 1 i 13 stopniami swobody.
```

```
## p-value wynosi  6.89769586080124e-14
```

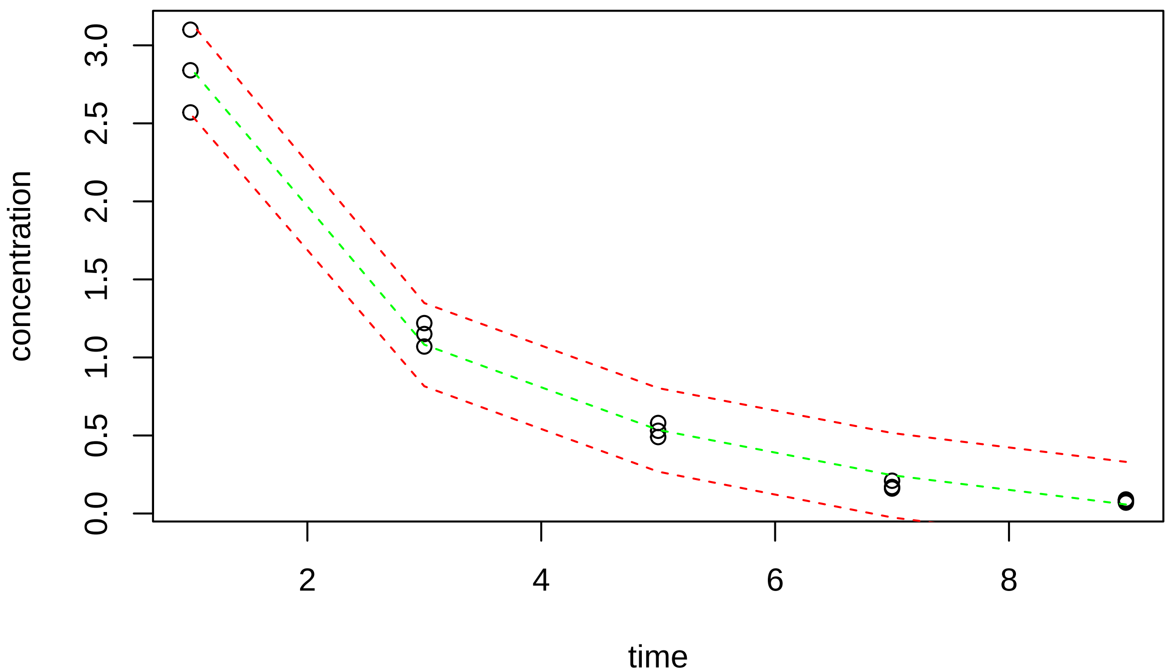
Ustalając standardowy poziom istotności $\alpha = 0.05$ mamy $p < \alpha$, odrzucamy hipotezę zerową. Stężenie i czas są ze sobą skorelowane.

c)

```
## [1] "Współczynnik korelacji: 0.89073740888566"
```

Po transformacji zmiennej objaśniającej współczynnik korelacji zmniejszył się.

c)



Powyższy wykres lepiej opisuje oryginalne dane niż wykres z zadania 5.

d)

```
## [1] "Współczynnik korelacji: 0.994013577069347"
```

Współczynnik korelacji jest znacząco większy niż w zadaniu 5.

Modele z zadania 7 i 8 są zdecydowanie lepsze niż ten z zadania 5. Lepiej opisują dane, bo zależność między oryginalnymi danymi nie jest liniowa. Model z zadania 7 przyda się gdy będziemy chcieli zbadać, co się dzieje w późniejszym czasie, a model z zadania 8 przyda się gdy będziemy chcieli zbadać, co się dzieje we wcześniejszym czasie (patrząc w którą stronę zwężają się przedziały predykcyjne).

Zadania teoretyczne

zadanie 1

a)

```
## [1] 2.570582 2.228139 2.008559
```

b)

```
## [1] 6.607891 4.964603 4.034310
```

c)

```
criticT^2
```

```
## [1] 6.607891 4.964603 4.034310
```

Możemy zauważyć, że kwadrat wartości t_c wynosi F_c. Wynika to z rozkładów z jakich pochodzą te wartości krytyczne.

Zadanie 2

a)

$n - 2 = 20$ zatem w pliku znajdują się 22 obserwacje.

b)

$\sigma^2 = SSE/dfE = 400/20 = 20$ zatem $\sigma = 2\sqrt{5}$.

c)

$F = MSM/MSE = (SSM/dfM)/\sigma^2 = 100/20 = 5$ ponadto $F_c = F^*(1-\alpha, 1, n-2) = 4.351244$. Widzimy, że $F > F_c$, zatem odrzucamy hipotezę zerową, slope nie jest równy zero.

d)

$R^2 = SSM/SST = SSM/(SSM + SSE) = 100/500 = 0.20$ zatem model wyjaśnia 20% zmienności zmiennej odpowiedzi.

e)

Próbkowy współczynnik korelacji między zmienną odpowiedzi a zmienną objaśniającą wynosi $\pm\sqrt{R^2} = \pm 0.4472136$, gdzie znak zależy od nachylenia prostej regresji (to znak współczynnika nachylenia).