Lista 3

Aneta Przydróżna

Kwantyl dwustronnego testu T z $\alpha = 0.05$ i 10 stopniami swobody: tc i kwantyl testu F z $\alpha = 0.05$, jednym stopniem swobody w liczniku i 10 w mianowniku: Fc

```
n=10
alfa=0.05
tc=qt(1-alfa/2,n)
Fc=qf(1-alfa,1,n)
c(Fc, tc^2)
```

```
## [1] 4.964603 4.964603
```

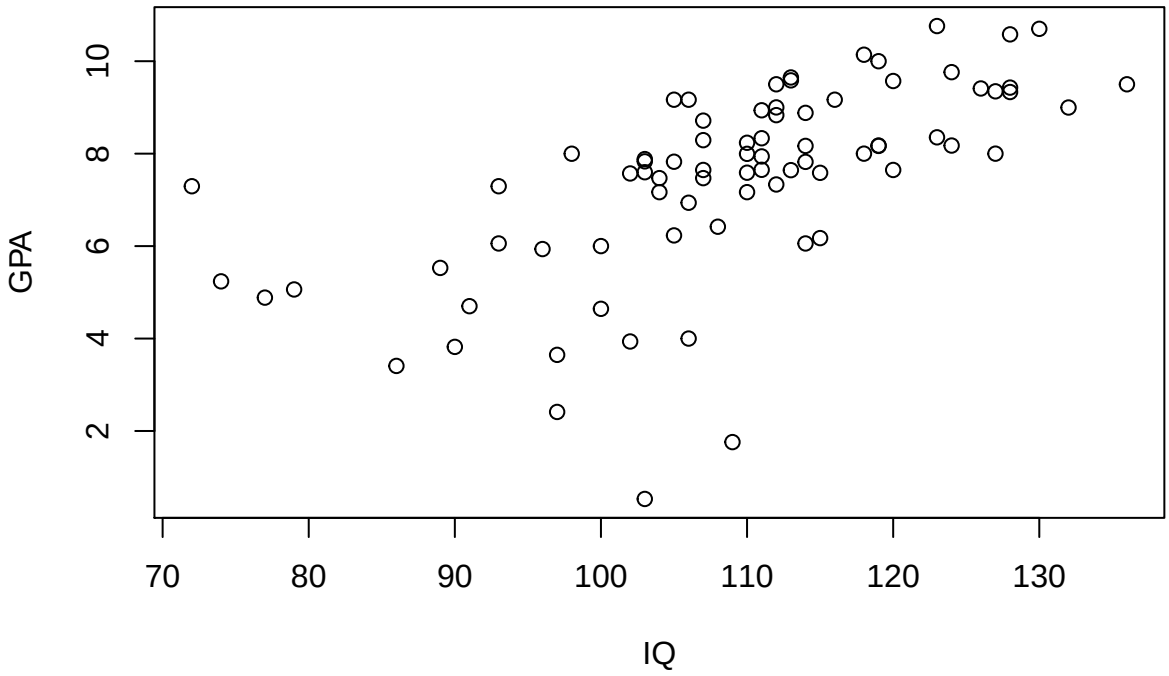
Jak widać oba testy mają ten sam wynik w szczególnym wypadku.

W tabeli ANOVA:

- Są 22 obserwacje
- Estymator σ: 4,47
- Odrzucamy hipotezę, że $\beta_1 = 0$, bo $F = 5 \in (4.35, \infty)$
- Y jest wyjaśniony w 20% przez X, czyli chaotycznie
- Współczynnik korelacji wynosi 0.45

Dane dotyczące średniej ocen, wyniku z testu na IQ, płci oraz wyniku z testu Piers-Harrisa uczniów z prwnej szkoły.

Prosty model regresji zależności GPA od wyników testu IQ




```
model<-lm(GPA~IQ,dane)
summary(model)
```

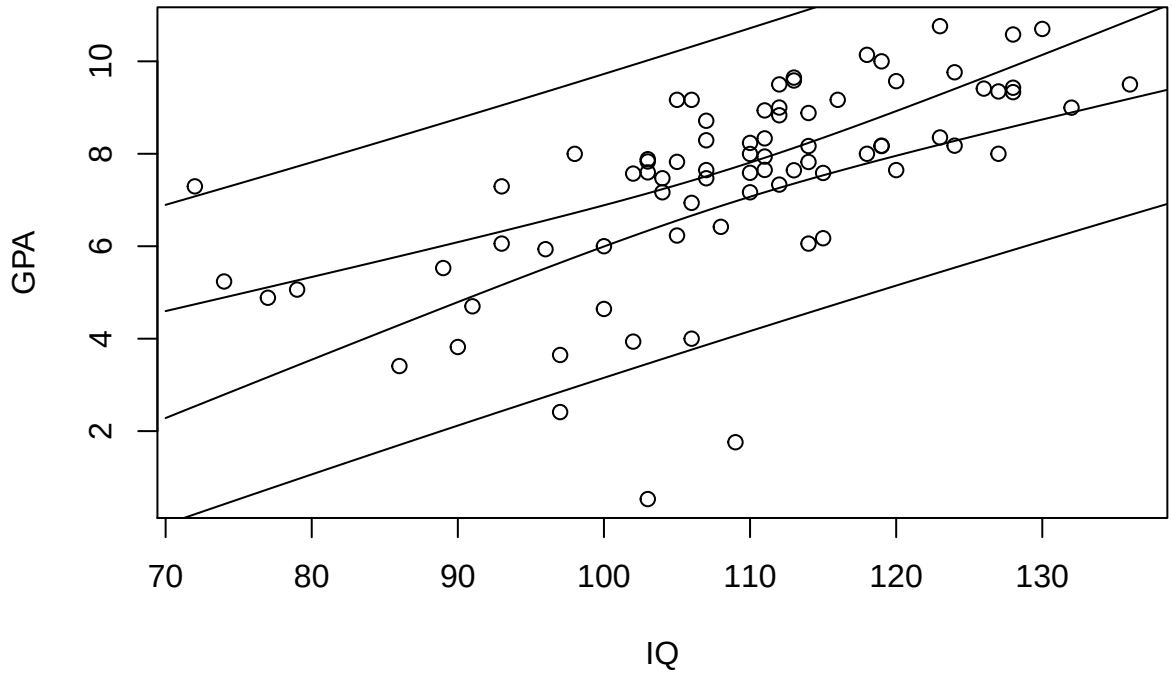
```
##
## Call:
## lm(formula = GPA ~ IQ, data = dane)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3182 -0.5377  0.2178  1.0268  3.5785
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.55706    1.55176  -2.292   0.0247 *
## IQ           0.10102    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
```

Można wyczytać, że $R^2 = 0.4$, zatem zmienna objaśniana(GPA) nie jest dobrze wyjaśniana przez zmienną objaśniającą(IQ).

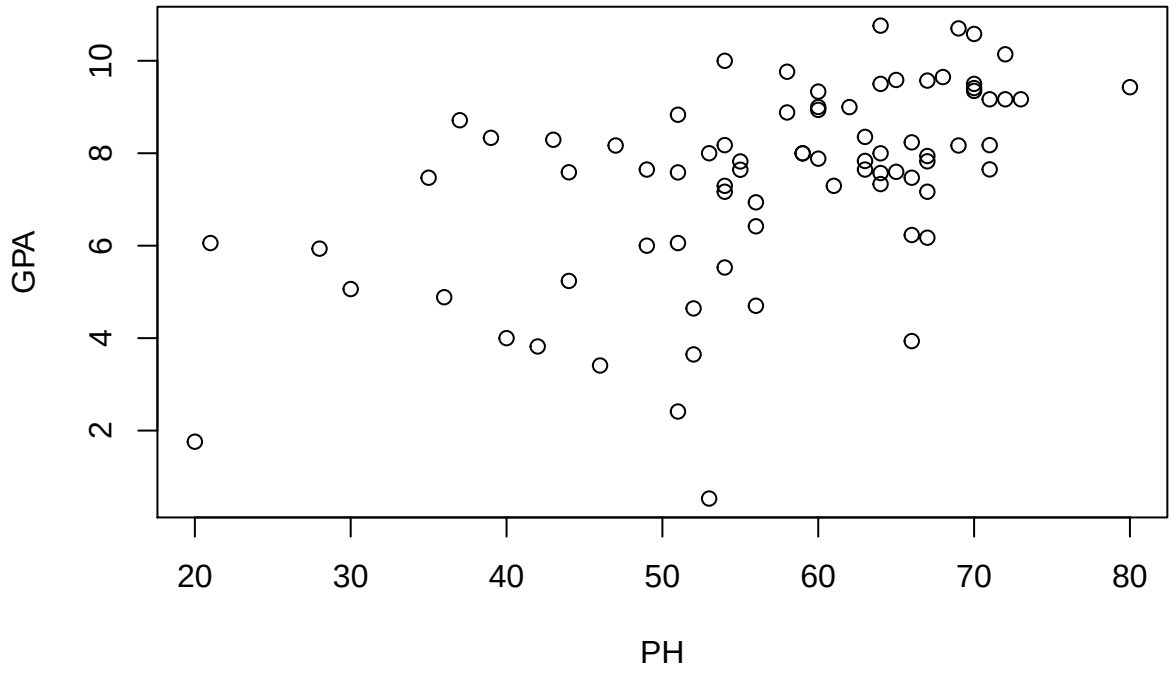
Sprawdzam hipotezę zerową,że GPA jest nieskorelowane z IQ

Poziom ufności:(3.25, 9.83).Statystyka F wynosi 51. P-value jest mniejsza niż 0.05,zatem można odrzucić hipotezę zerową i przyjąć że GPA jest skorelowane z IQ.



Dwaj uczniowie z IQ równym 100 mają GPA równe 4.64 i 6.Przewidywana wartość GPA w naszym modelu to 6.54 ze wzoru na regresję liniową.Przedział ufności (90 procentowy) dla GPA=100 : (4.4 , 8.7). Natomiast poza ograniczenia 95% obszaru przewidywań, wypadają 3 punkty.

Prosty model regresji liniowej opisujący zależność GPA i testu Piersa Harrisa



Statystyka F wynosi 31.6.Statystyka R^2 wynosi 0.28. Oznacza to że P-H w jeszcze mniejszym stopniu niż IQ wyjaśnia GPA.

Testowanie H_0: GPA jest nieskorelowane z wynikiem testu Piersa Harrisa

```
cor(PH,GPA)
```

```
## [1] 0.5418329
```

P-value jest mniejsza niż 0.05 (niemalże zerowe),zatem można odrzucić hipotezę zerową i przyjąć że GPA jest skorelowane z PH.

Typowanie GPA dla uczniów, którzy mają z testu Pierca Harrisa 60 pkt.

Otrzymane z danych wyniki GPA przy PH=60 to: 9.3, 8.9, 7.9, 9.Poza 90% przedział ufności: (-0.4 , 4.1) wypada około 7 obserwacji. Poza pasmo dla 95-procentowych przedziałów predykcji wystają 3 obserwacje.

Okazuje się, że IQ lepiej wyjaśnie GPA niż test PH.W pierwszym rzypadku statystyka F wynosi 51, a w drugim 31.6. A im większe F tym β jest mniej zerowa.Również statystyka R^2,która jest większa przy testach ze zmienną IQ, jest bardziej istotna.

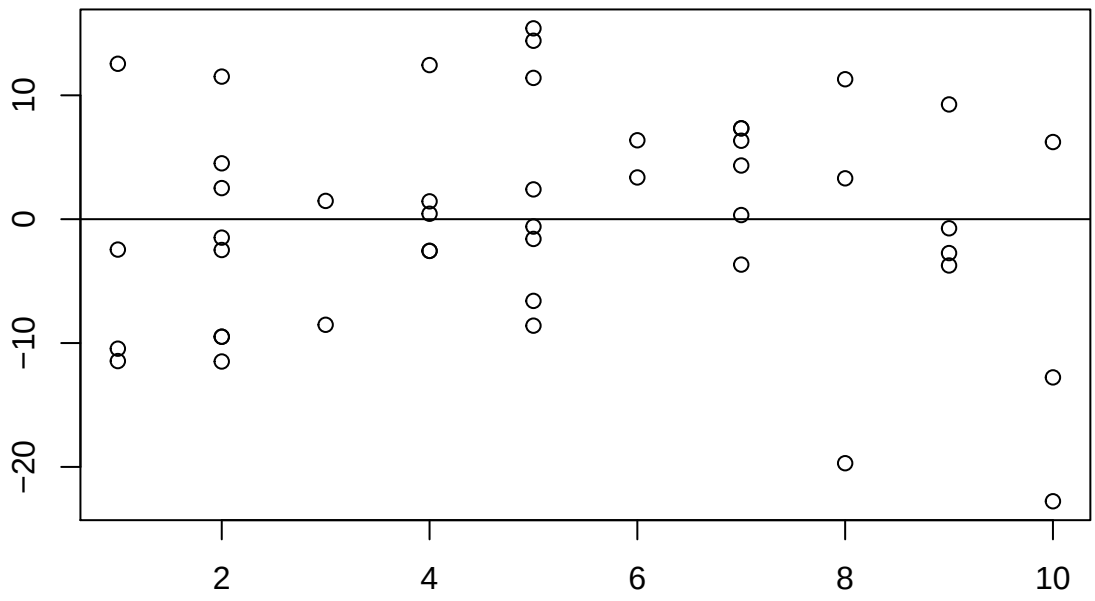
Analiza danych dotyczących ilości obsługiwanych kopiarek w czasie.

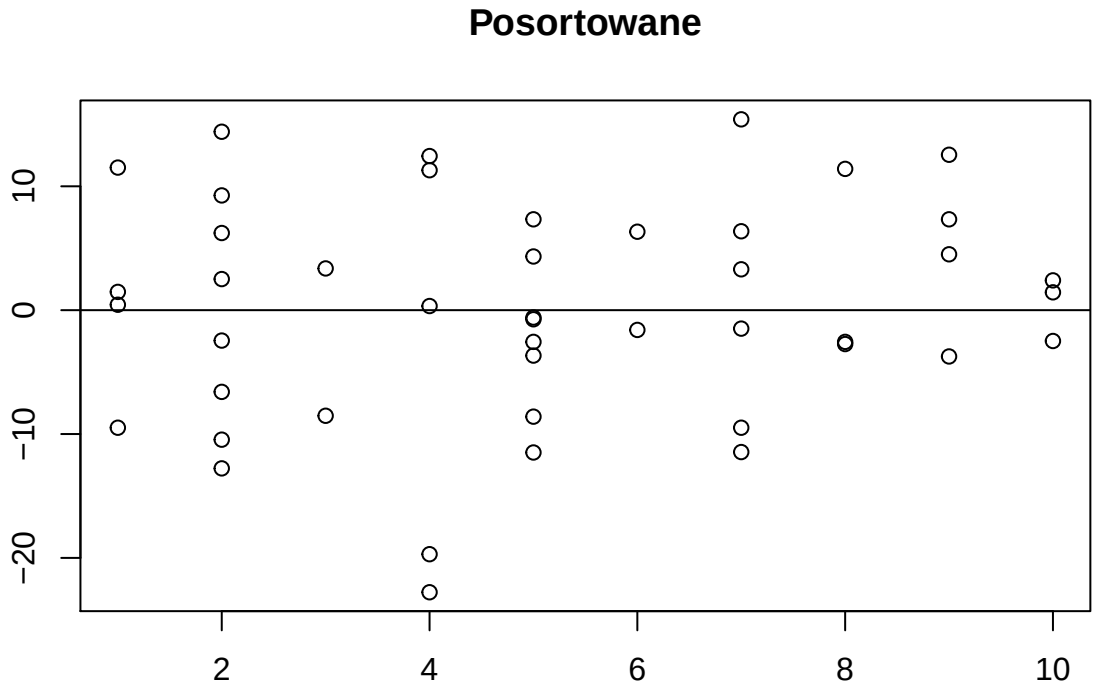
Suma reszt w modelu wynosi zero.Reszty względem X są porozrzucane.

```
model<-lm(Y~X,dane)
sum(residuals(model))
```

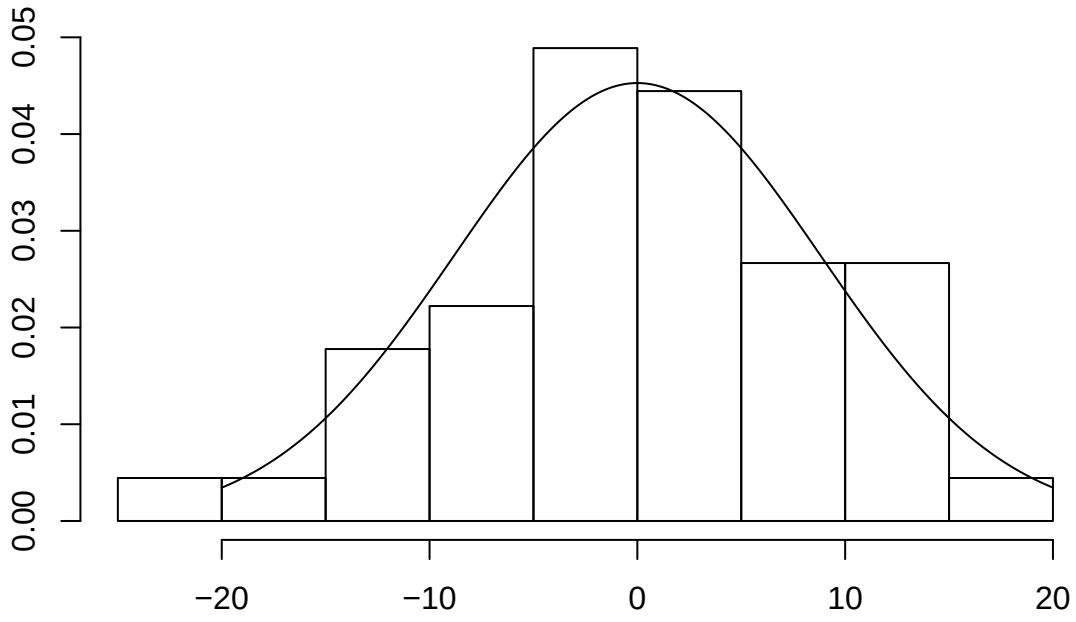
```
## [1] -1.176836e-14
```

Błędy występują przy estymacji poszczególnych wartości względem zmiennej objaśniającej.Dane rozkładają się bez określonego schematu.Mamy zależność liniową niezależnych reszt.Również na kolejnym wykresie dane układają się w miare symetrycznie nad i pod zerem obciętych.Wykres posortowany według objaśniającej zachowuje się podobnie.



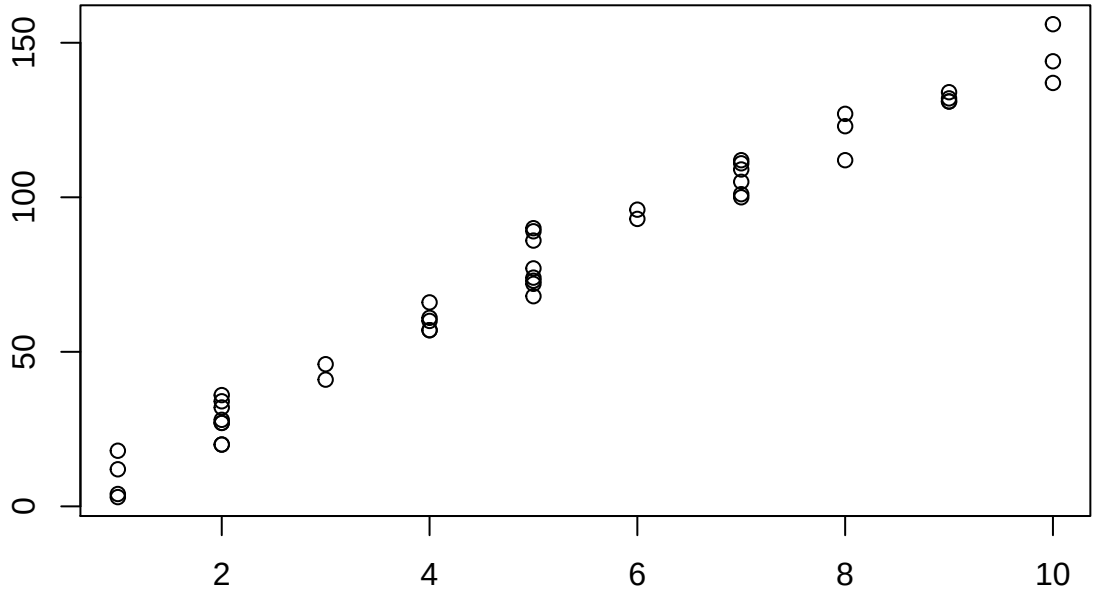

Posortowane
10
0
−10
−20
2
4
6
8
10

Rozkład błędów na histogramie.Największa częstotliwość w okolicy zera,spadająca w miarę symetrycznie z obydwóch stron.



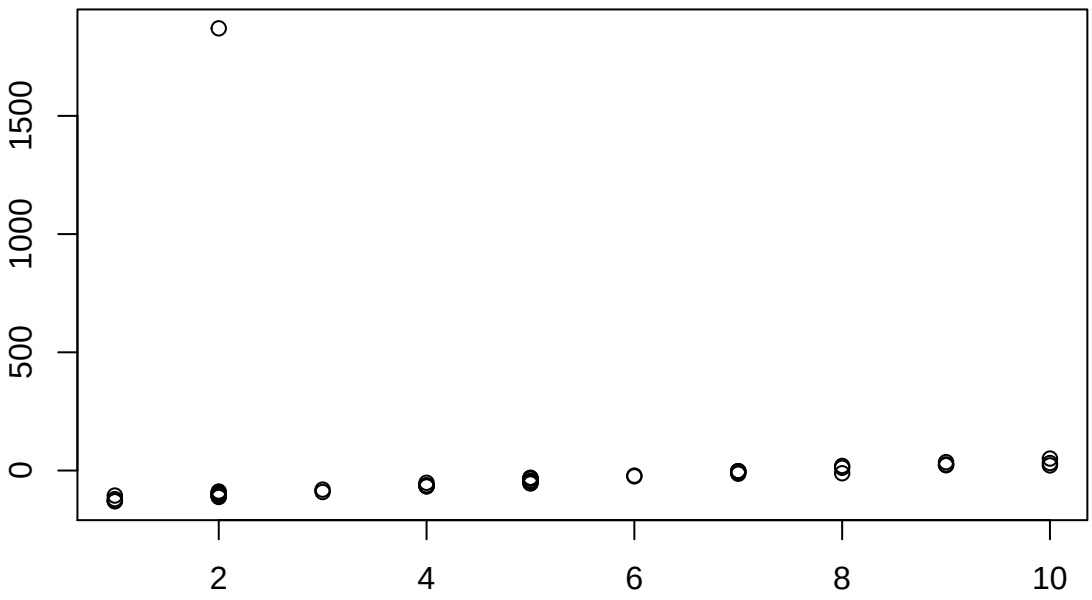
Statystyka R^2 jest na poziomie 96% , czyli model dobrze przedstawia zależność między zmiennymi.P-wartość dla b1 wynosi prawie 0, więc zmienne są od siebie zależne.

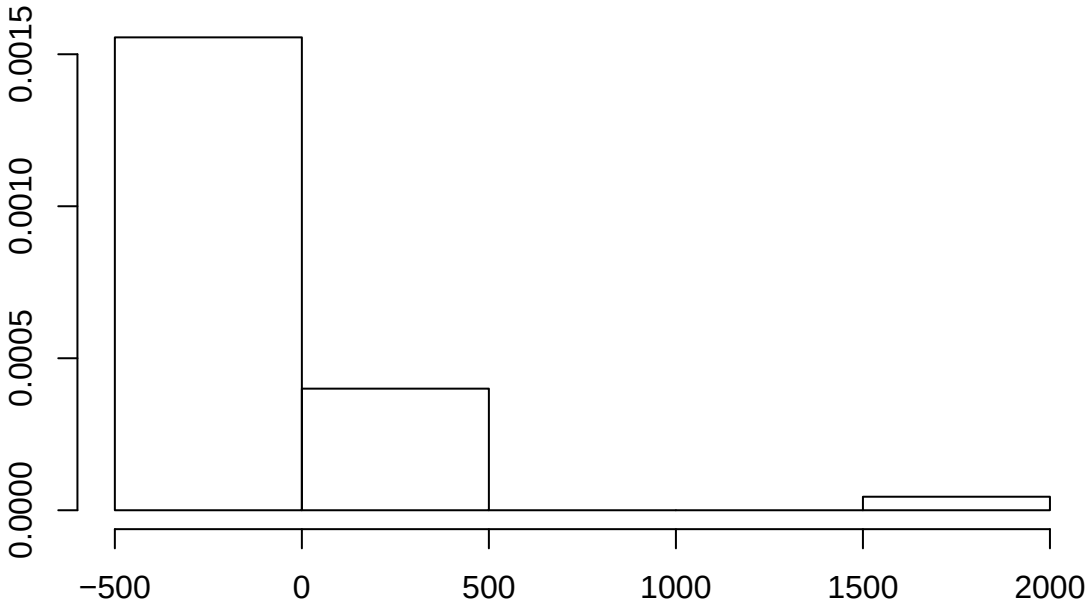
Mamy podstawy do twierdzenia że to rozkład normalny patrząc na jego qqplot(bez dużych ogonów).

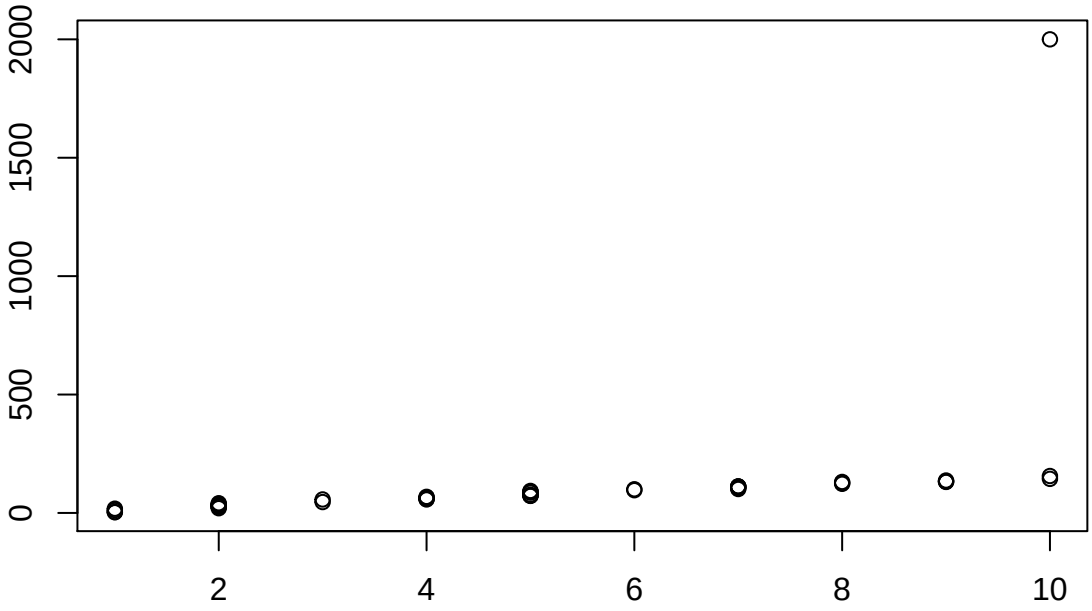


Zmiana czasu serwisowania pierwszej maszyny z 20 na 2000 godzin.

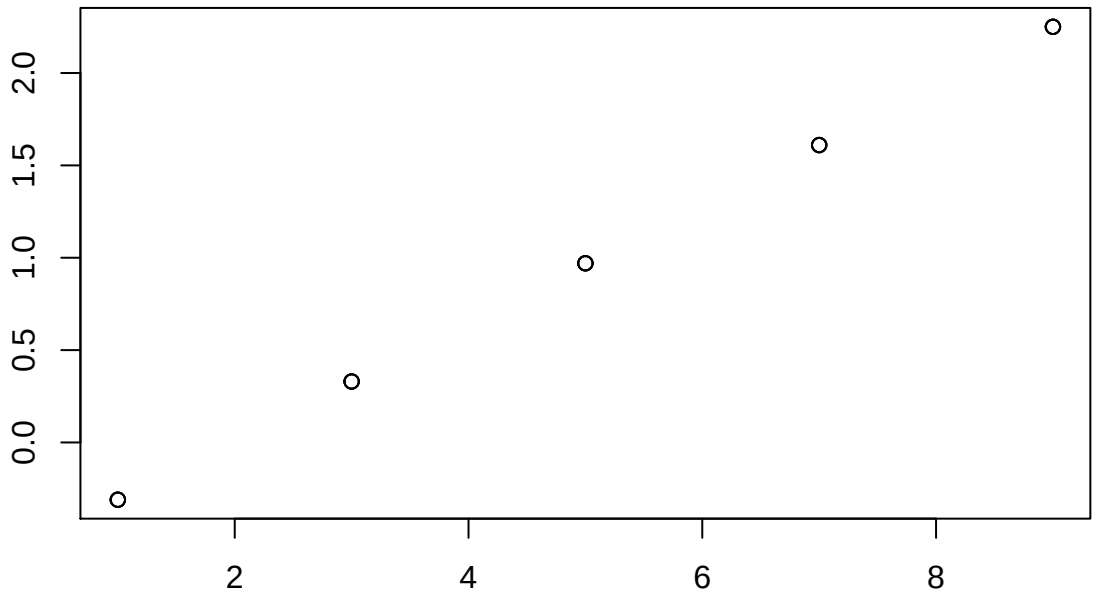
Statystyka F wynosi 0.037 więc jest nieistotna .Nie ma zależności X i Y.Jedna zamieniona obserwacja na to wpłynęła.P-wartość wynosi 0.84. Stworzenie pasującego modelu jest mało prawdopodobne. Wykres reszt znacząco się zmieniły,dzieląc się na pierwszą obserwacje odstającą i resztę blisko siebie..Histogram ma za to bardzo ciężki prawy ogon.Żaden model regresji nie będzie pasował.


0
500
1000
1500
2
4
6
8
10


0.0000
0.0005
0.0010
0.0015
−500
0
500
1000
1500
2000


0
500
1000
1500
2000
2
4
6
8
10

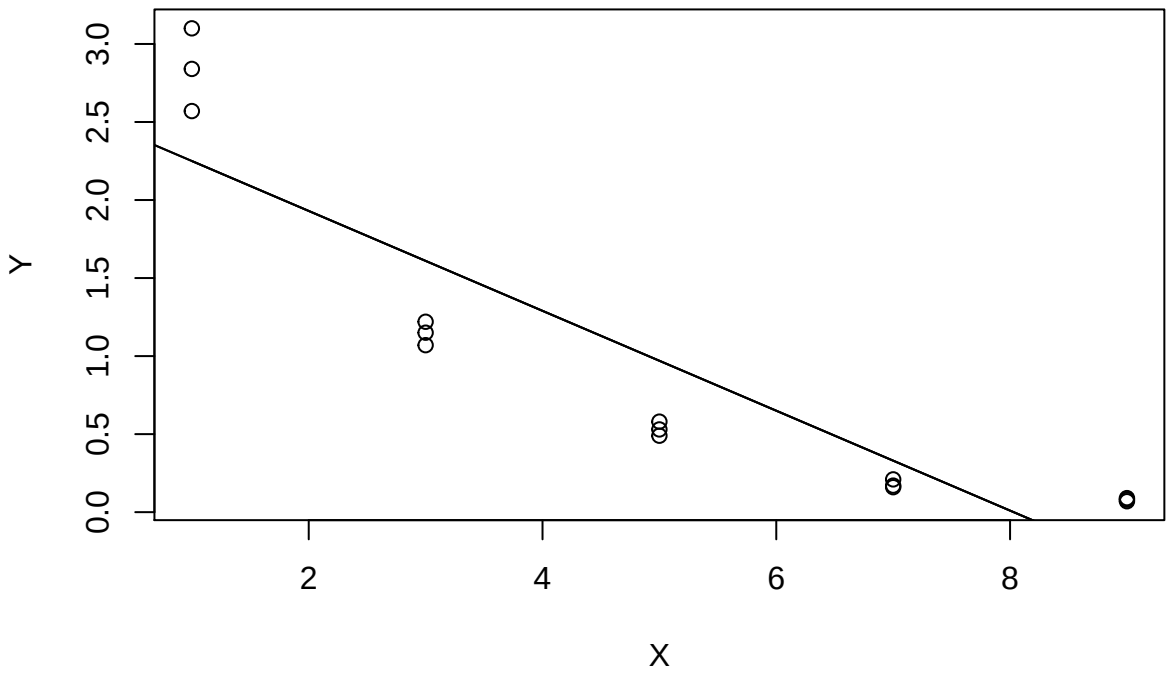
Dane dotyczące wartości stężenia roztworu,w zależności od czasu.



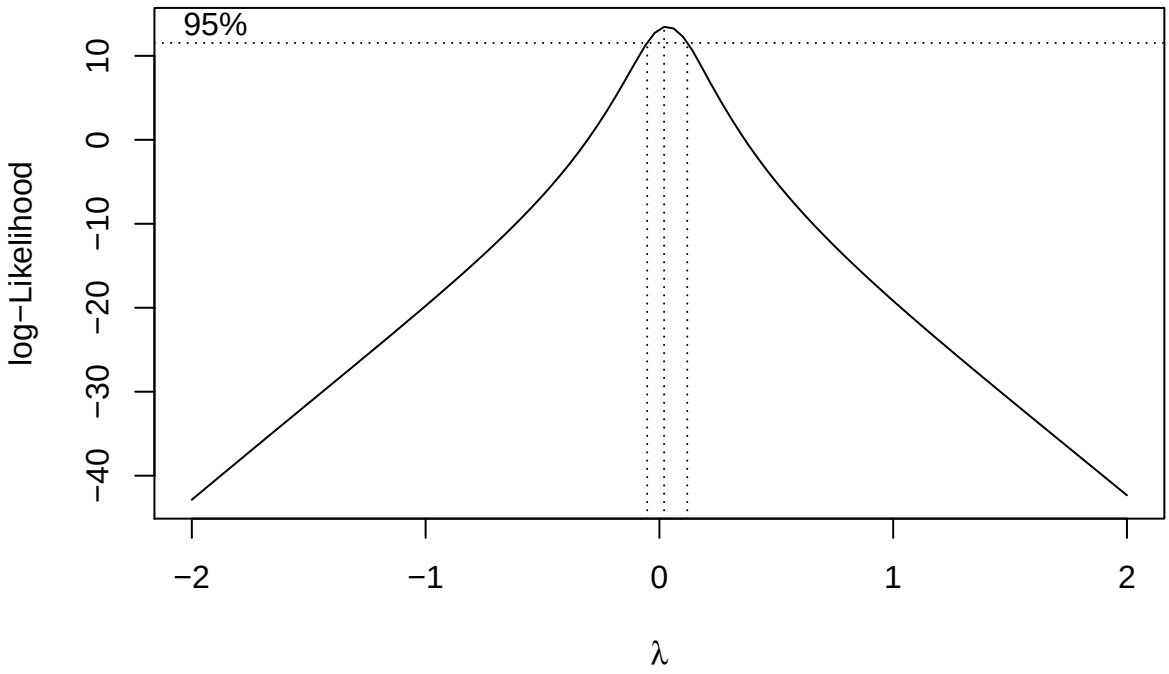
Statystyka R^2 wynosi 0.8 czyli stężenie jest bardzo dobrze wyjaśnione przez czas.H_0 mówiące, że β jest zerowa odrzucamy przez p-value bliskie zeru.Statystyka F Wynosi 56, więc spokojnie przekracza wartość 4.6 dla 95% ufności, wpadając do obszaru krytycznego.Zatem przyjmujemy H_A mówiące,że przynajmniej jeden ze współczynników β jest niezerowy.

Dopasowanie linii regresji i pasmo dla 95% interwałów predykcyjnych.

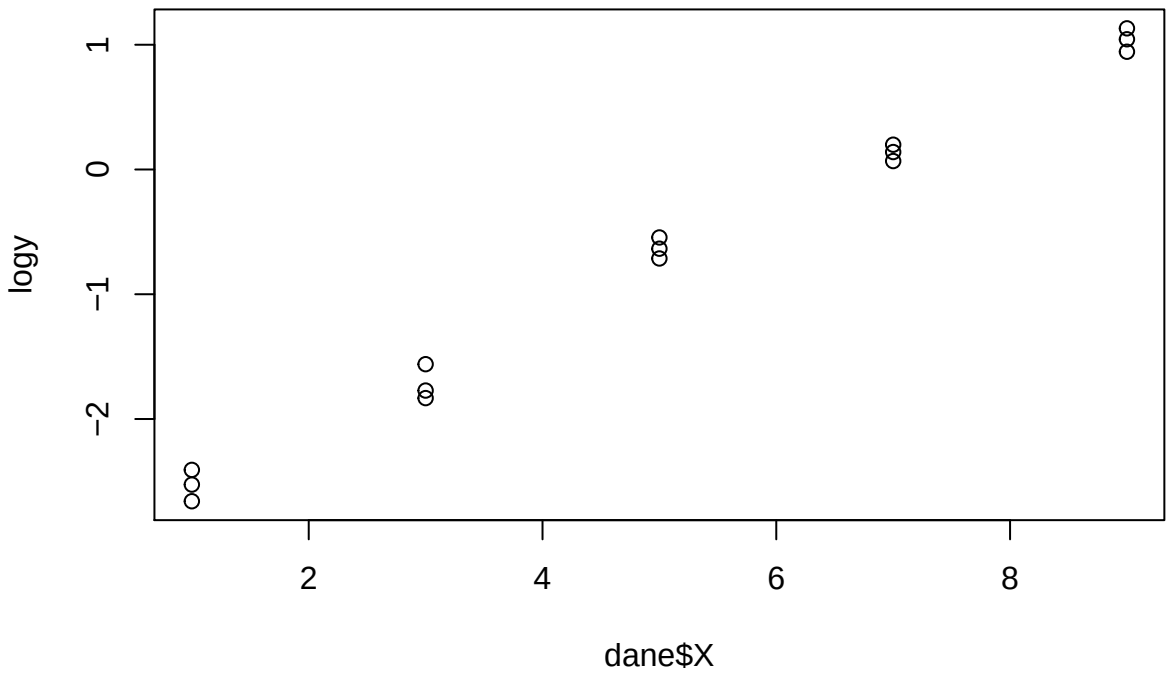
Korelacja zmiennych wynosi 0.98.Rozkład błędów nie jest normalny.Regresja wychodzi malejąca.Model nie do końca dobrze przewiduje dane.

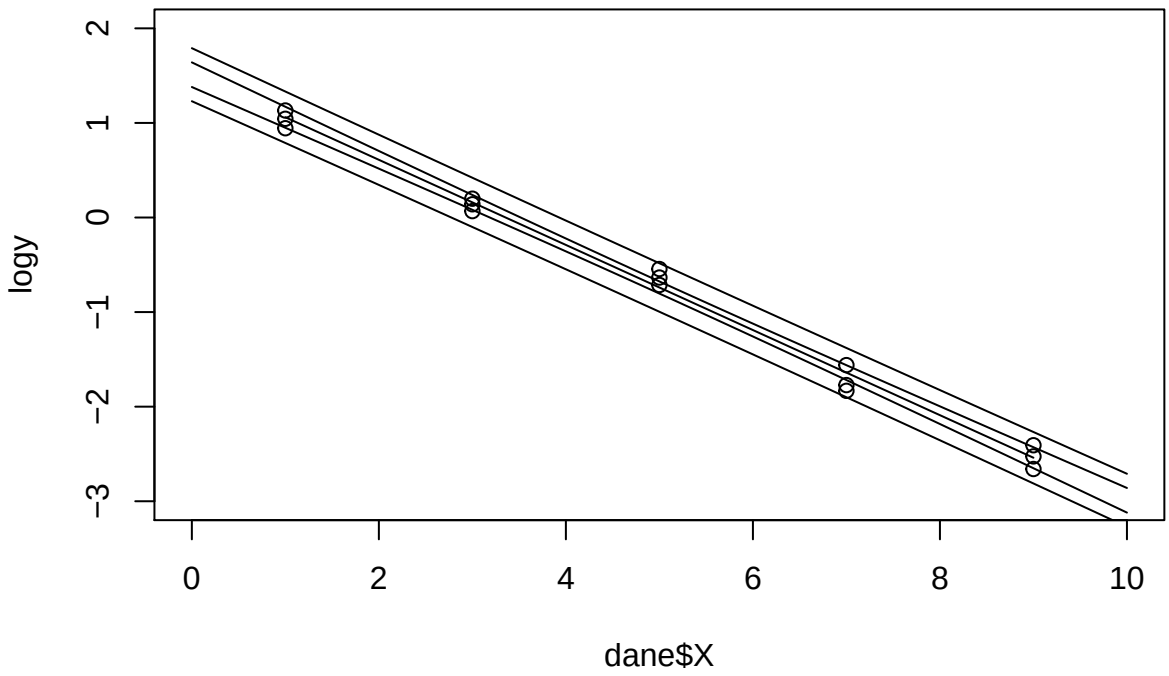

Y
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
X

Transformacja modelu metodą Box-Cox.Rozkład zbliżony do normalnego.



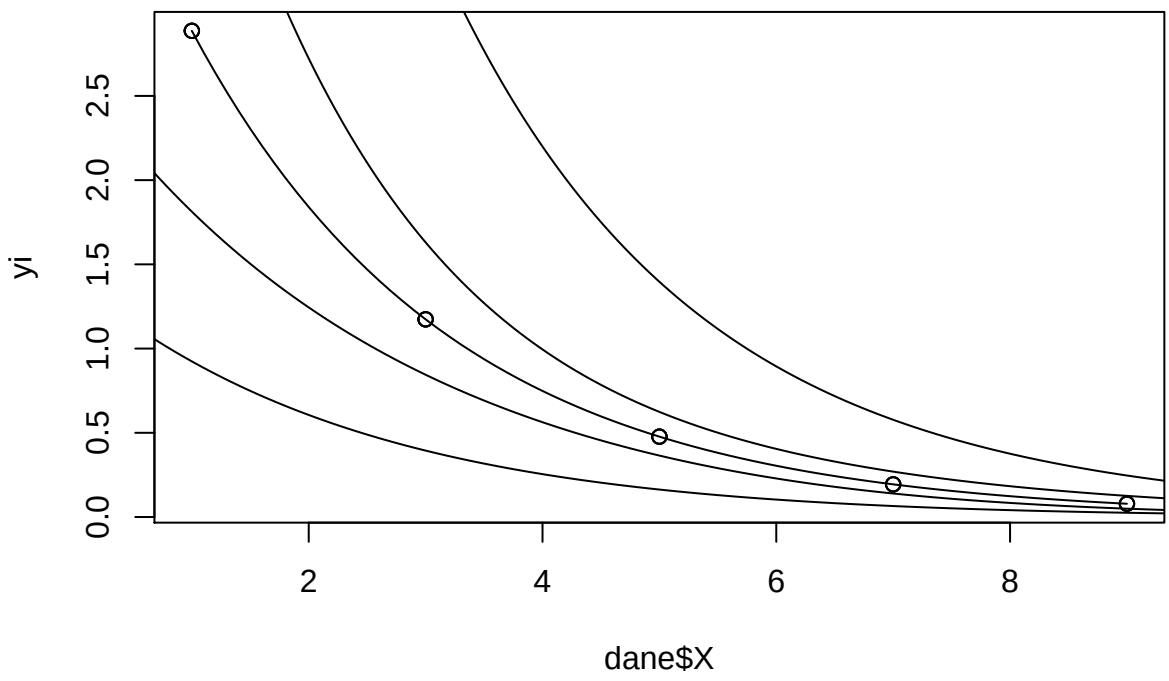
Według wykresu zmienna $\lambda = 0$, więc można nałożyć logarytm na zmienną stężenia.Konstruuję nowy model.


logy
1
0
−1
−2
2
4
6
8
dane$X



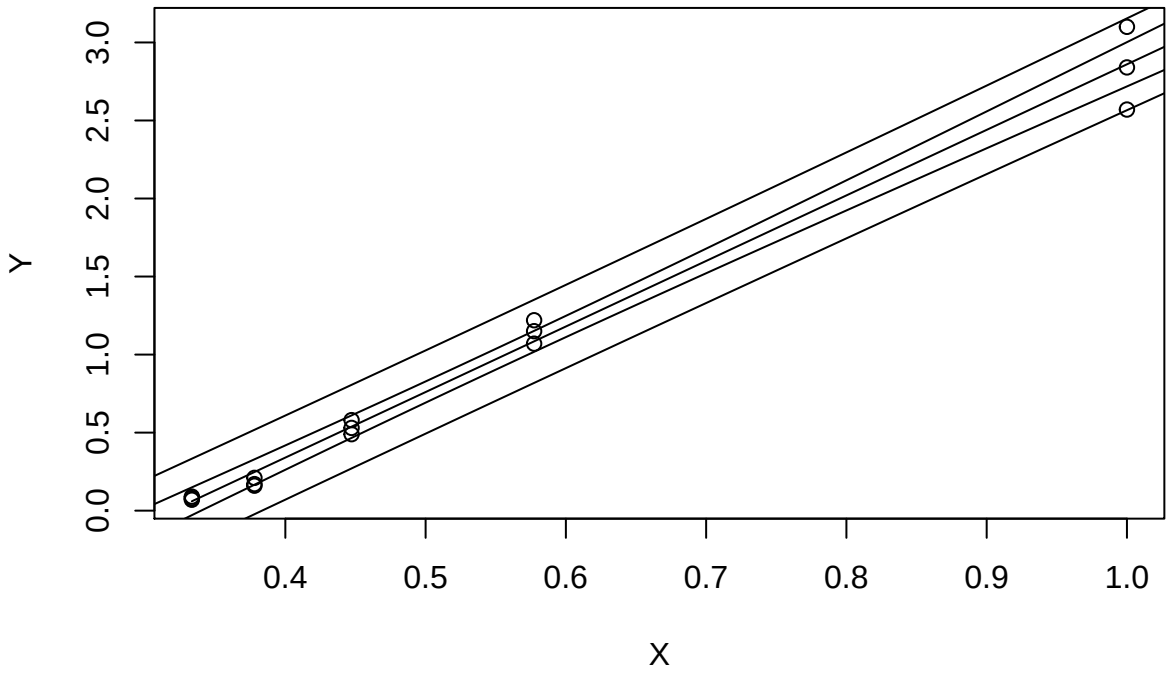
Po nałożeniu na Y logarytmu nasz model znacznie się poprawił.Wszystko wpada do przedziału 95%, dla średniej nie wszystkie,bo 3 obserwacje wypadają. R^2 wynosi prawie 1, p-value 0 natomiast statystyka F 1838.To wszystko świadczy o bardzo dobrym dopasowaniu modelu.Rozkład błędów wygląda lepiej niż w pierwotnej wersji danych-Są one normalne.

Nałożenie exp na y



Punkty ustawiają się dokładnie na jednaj lini.Regresja jest malejąca.

Teraz zmienna czasu wygląda następująco: $X = X^{-\frac{1}{2}}$



Tym razem na wykresie mamy rosnące prostą i przedziały.Przedział predykcyjny obejmuje wszystkie punkty.

Porównując powyższe wykresy, wykonane na tych samych danych róznymi metodami, okazuje się że nr 1 ma za każdym razem inną wariancję błędów (trudno wskazać że dane są iid).Najlepszym modelem jest ten log(Y) ponieważ dane wydają się idealnie pasować do modelu.Również statystyka R^2=0.992 jest tu największa, zaraz za nią z wartością 0.987 jest statystyka w ostatnim prypadku.W pozostałych dwóch przypadkach prawie 0.8.