Lista 3 - Raport

Erwin Jasic

7 grudnia 2020

Zadanie 1

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

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

W pierwszym zadaniu mieliśmy sprawdzić, czy statystyka t_c podniesiona do kwadratu będzie równa statystyce f_c. Zgodnie z teorią z wykładu zobaczyliśmy, że taka równość zachodzi. $t_c^2 = f_c$.

Zadanie 2

```
## [1] 22
```

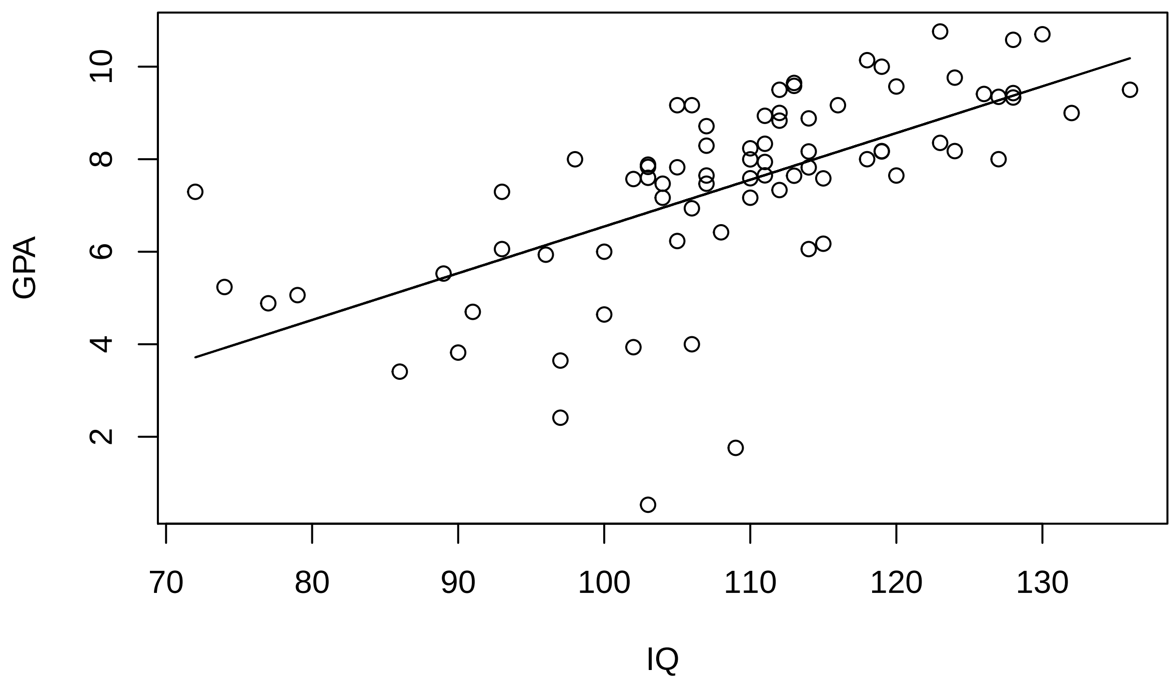
```
## [1] 4.472136
```

```
## [1] 0.2
```

```
## [1] 0.4472136
```

W drugim zadaniu mamy policzyć pewne wartości na podstawie fragmentu tabeli analizy wariancji. W podpunkcie a) oraz b) wykonywaliśmy podstawowe obliczenia do policzenia wszystkich obserwacji jak i estymator sigmy. W podpunkcie c) testowaliśmy czy β_1 jest równa 0. Po obliczeniach wyszło nam, że MSM srednio wieksze niz MSE, więc zachodzi H_1, $F > F_c$, odrzucamy H_0. W podpunkcie d) obliczamy R^2 a w e) pierwiastek z R^2. Nie możemy stwierdzić jaki znak jest przy slope (β_1), ponieważ mamy podane sumy kwadratów.

Zadanie 3



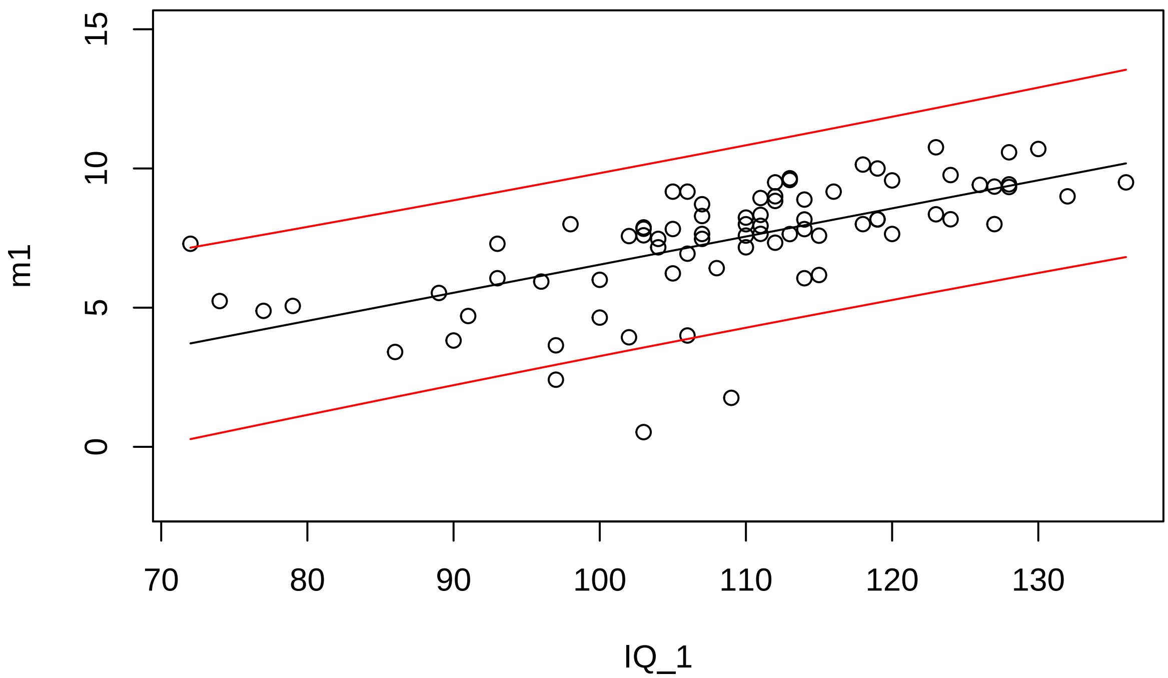

```
##                   2.5 %     97.5 %
## (Intercept) -6.6476600 -0.4664517
## IQ           0.0728501  0.1291933
```

```
##
## Call:
## lm(formula = GPA ~ IQ, data = dane_3_4)
##
## 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 0.000000000474 ***
## ---
## 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: 0.0000000004737
```

```
##        1
## 6.545114
```

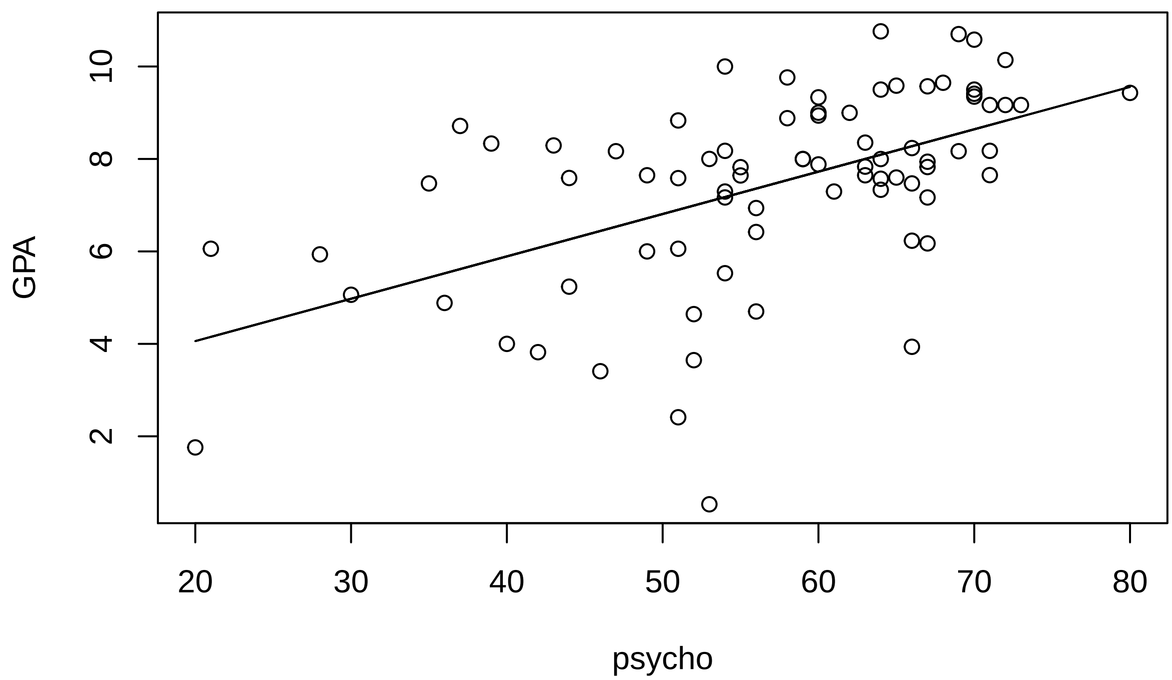
```
##        fit     lwr      upr
## 1 6.545114 3.79753 9.292698
```

```
## Warning in predict.lm(model_a, interval = "prediction"): predictions on current data refer to _future
```



W trzecim zadaniu analizowaliśmy dane dotyczące uczniów pewnej szkoły w USA. W tym zadaniu mieliśmy porównać ze sobą zależność między GPA a IQ. W podpunkcie a) mieliśmy policzyć kilka wartości. Statystyka testowa dla β_0 (intercept) wynosi -2.292 a p-wartosc 0.0247 zatem odrzucamy H_0 dla 95% przedziału ufności, ale dla 99% już nie. Statystyka dla $beta_1$ (IQ) wynosi 7.142 a p-wartosc jest bliska 0 zatem odrzucamy H_0. Multiple R-squared error wynosi 0.4016, a Adjusted R-squared 0.3937. Oznacza to, że model jest słabo dopasowany. $R^2 = SSM/SST = 1 - SSE/SST$ mówi o tym jaka część Y jest wyjaśniona przez model. Adj R^2 bierze średnie kwadratów tzn. $R_a^2 = 1 - MSE/MST$. W podpunkcie b) liczymy predykcje dla $IQ = 100$. W podpunkcie c) rysujemy przedziały predykcyjne. 4 obserwacje wypadają za ten przedzial predykcyjny.

Zadanie 4

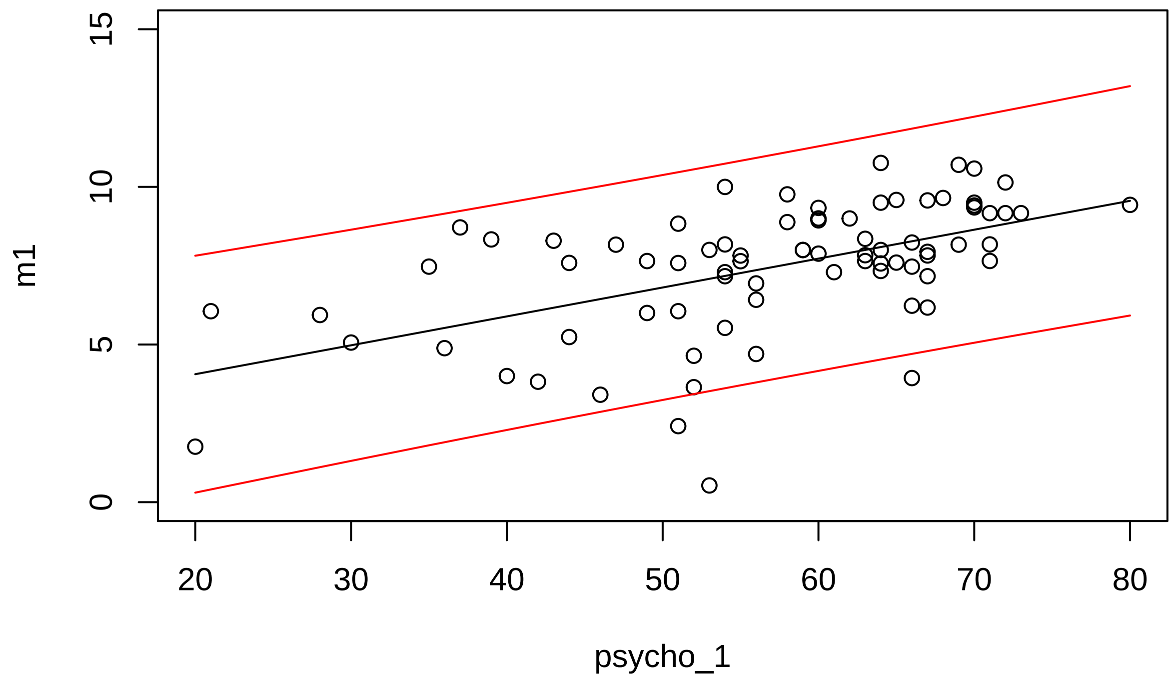



```
##                  2.5 %    97.5 %
## (Intercept) 0.33289123 4.1188741
## psycho      0.05917202 0.1241326
```

```
##
## Call:
## lm(formula = GPA ~ psycho, data = dane_3_4)
##
## 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 *
## psycho       0.09165    0.01631   5.620 0.000000301 ***
## ---
## 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: 0.0000003006
```

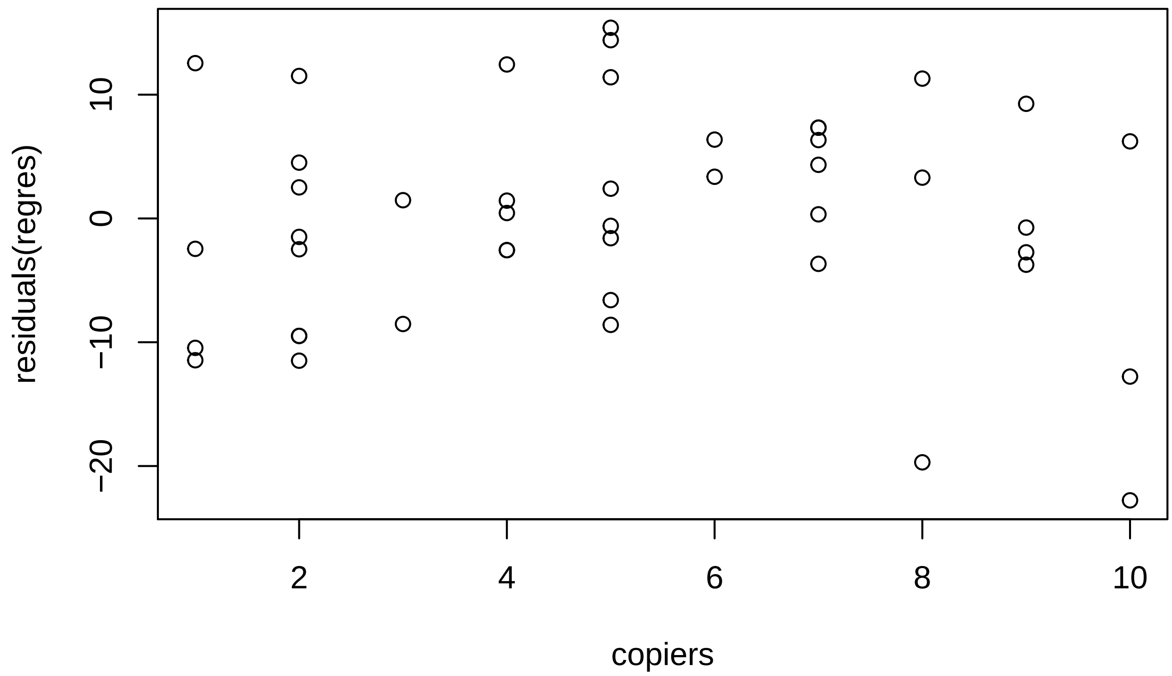
```
##       1
## 7.72502
```

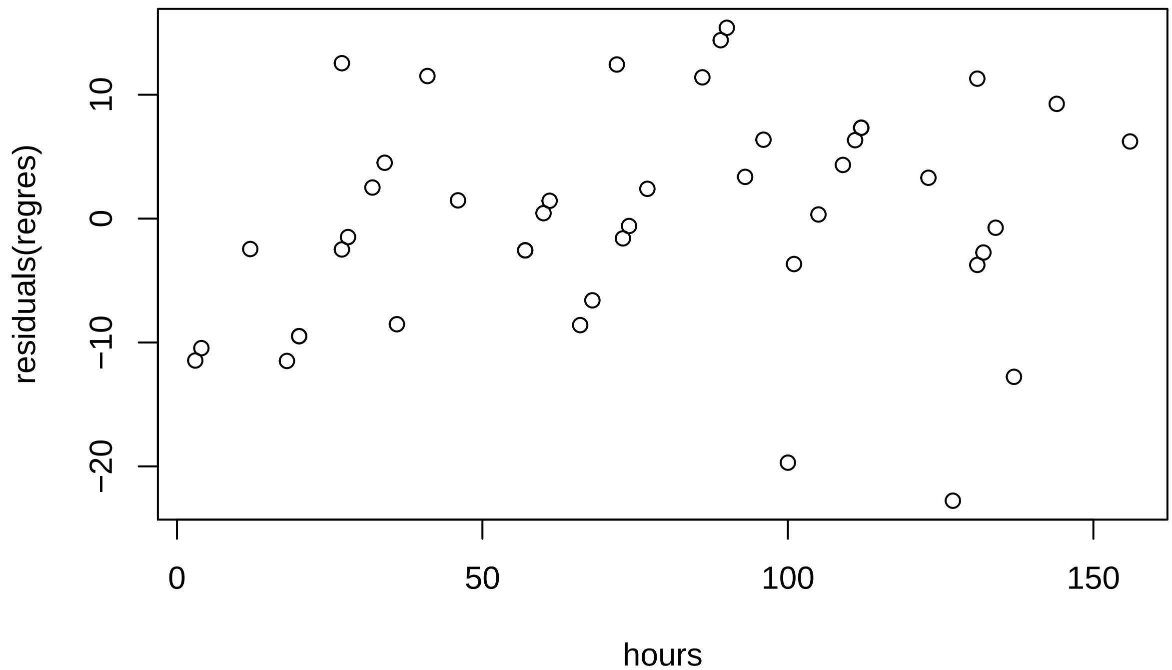
```
##       fit      lwr      upr
## 1 7.72502 4.747302 10.70274

## Warning in predict.lm(model_b, interval = "prediction"): predictions on current data refer to _future
```



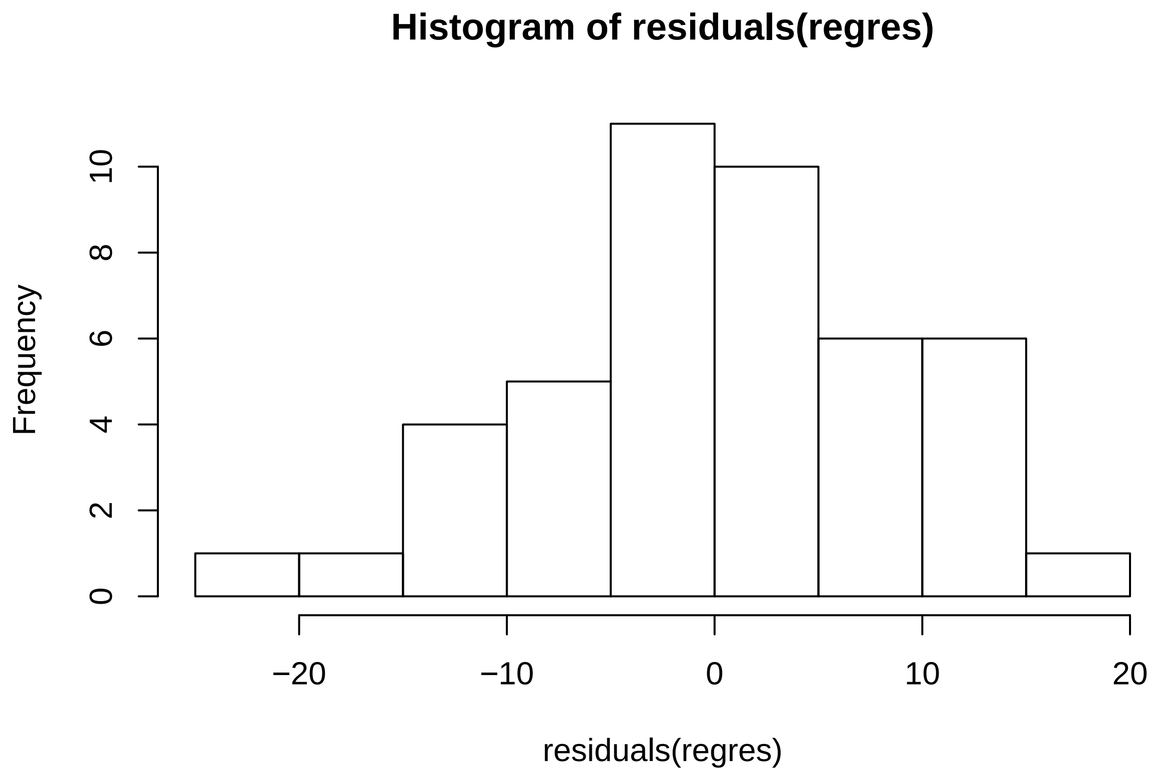
W zadaniu czwartym analizuyjemy te same dane, ale teraz porównujemy GPA i test Piers-Harris'a (u nas psycho). Analizując summary od tych danych widzimy, że po wartosciach R^2 i $AdjR^2$, że to dopasowanie jest jeszcze gorsze. $R^2 = 0.2936$, $AdjR^2 = 0.2843$. Statystyka testowa dla β_0 (intercept) wynosi 2.342 a p-wartosc 0.0218 zatem odrzucamy H_0 dla 95% przedziału ufności, ale dla 99% już nie. Statystyka dla β_1 (psycho) wynosi 5.620 a p-wartosc jest bliska 0 zatem odrzucamy H_0. 3 obserwacje wypadają za przedzial predykcyjny wyznaczony w podpunkcie d).

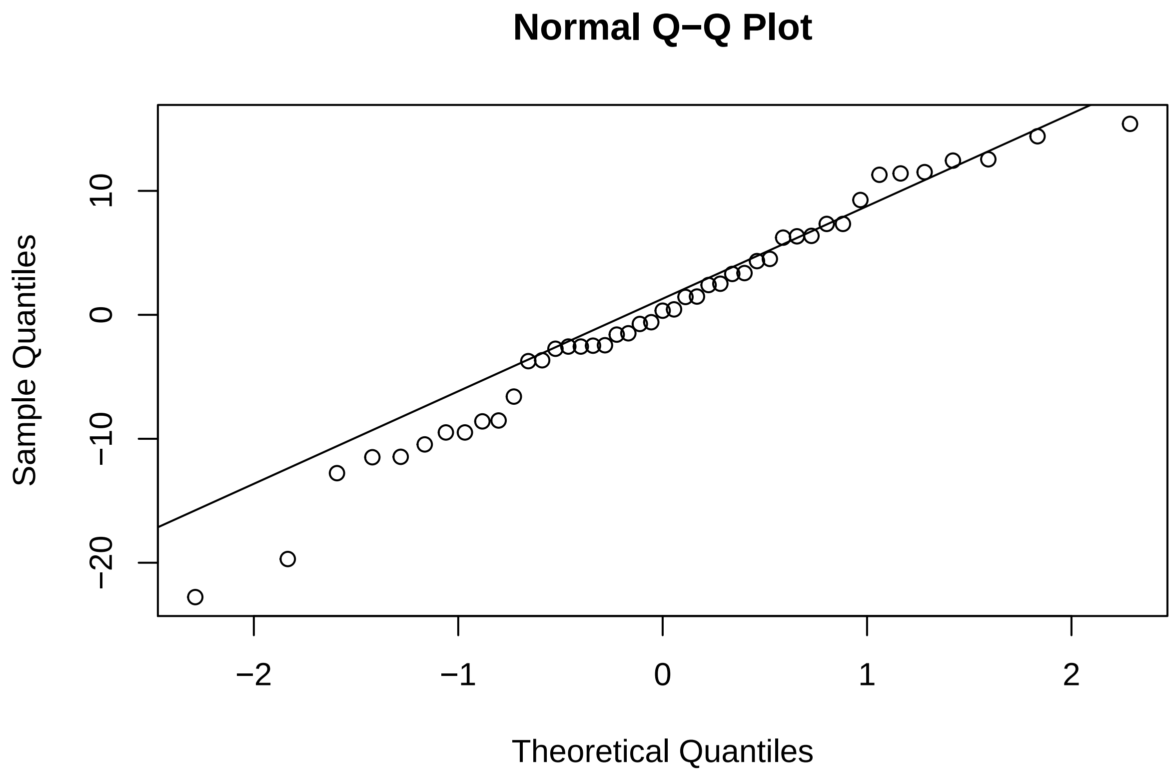
```
## [1] -0.00000000000001176836
```


residuals(regres)
10
0
−10
−20
2
4
6
8
10
copiers



residuals(regres)
10
0
−10
−20
0
50
100
150
hours


Histogram of residuals(regres)
Frequency
0
2
4
6
8
10
−20
−10
0
10
20
residuals(regres)



W zadaniu piątym analizujemy residua dotyczące zależności między czasem a liczbą kopii. Widzimy, że suma ich wynosi 0 oraz możemy zauważyć ze wykresy wskazują na rozkład normalny.

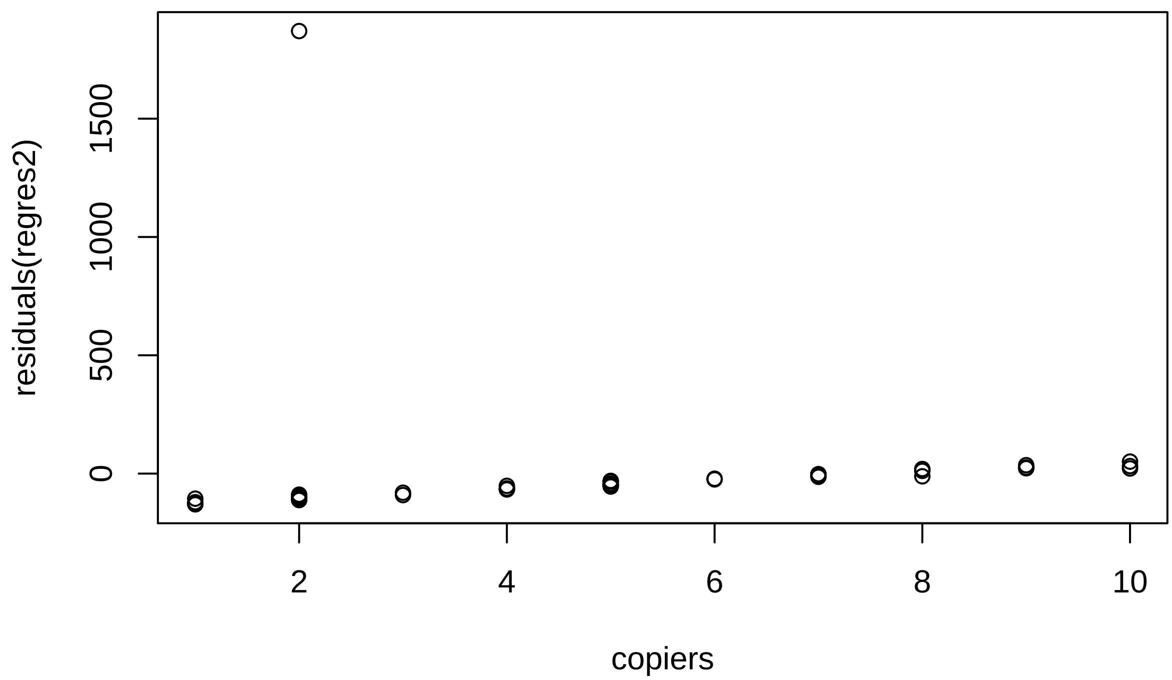
Zadanie 6

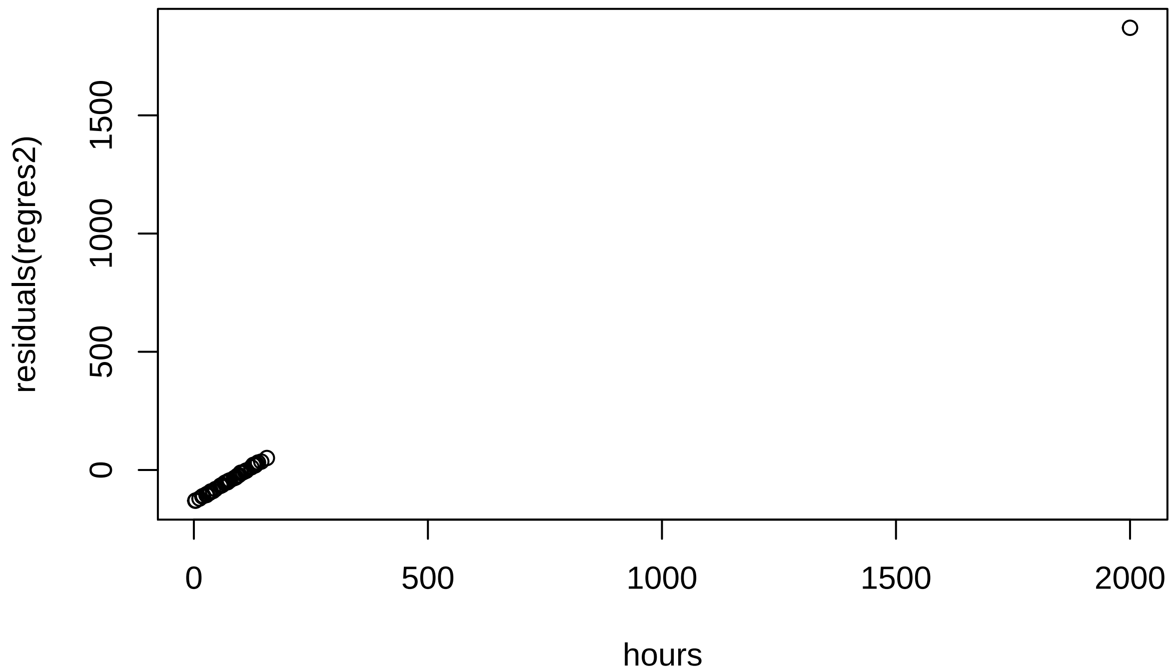
```
##
## Call:
## lm(formula = hours ~ copiers, data = dane_5_6)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207               0.837
## copiers      15.0352     0.4831  31.123 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 0.00000000000000022
```

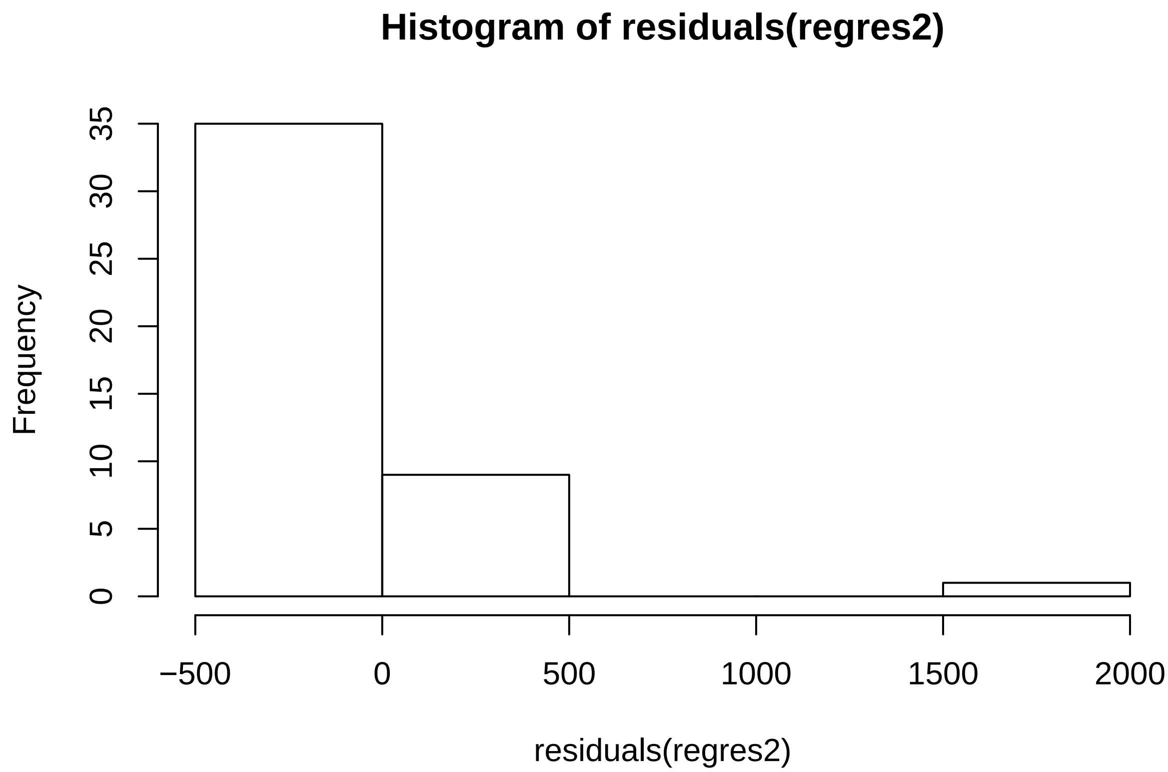
```
##
## Call:
## lm(formula = hours ~ copiers, data = dane_6)
```

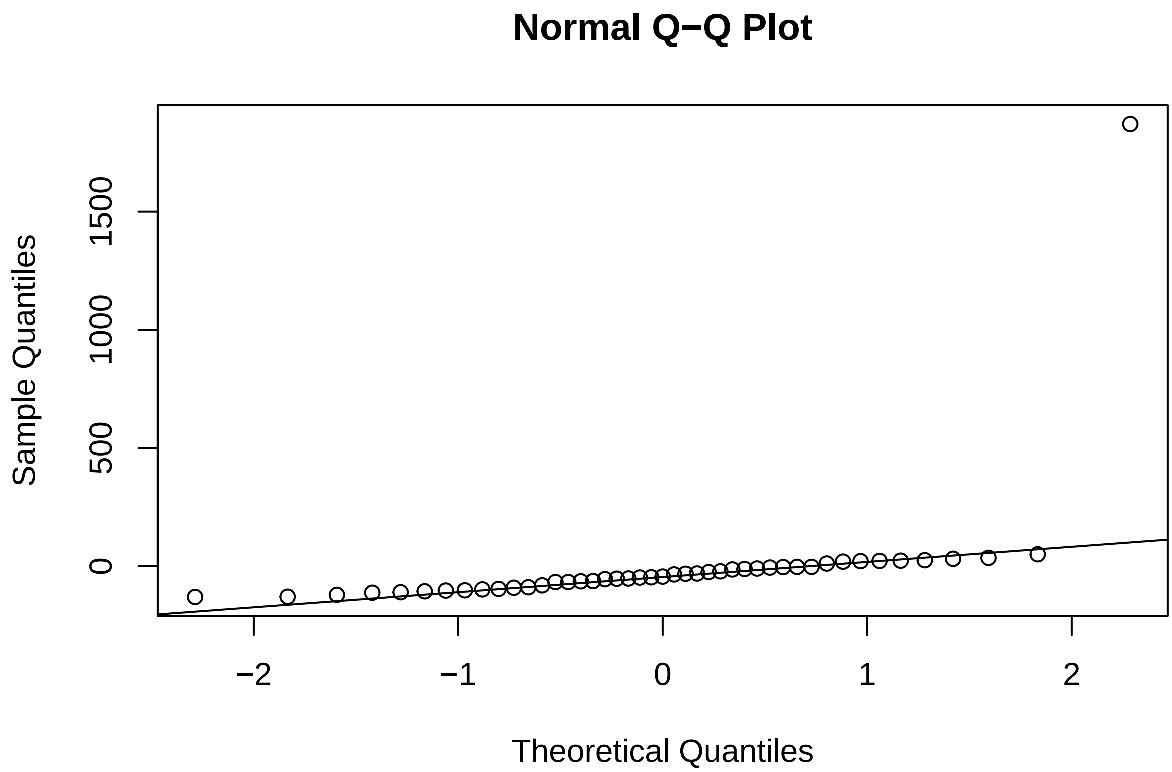
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -129.84  -88.78  -43.61   -2.49 1870.22
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  135.900     92.122   1.475    0.147
## copiers       -3.059     15.871  -0.193    0.848
##
## Residual standard error: 292.8 on 43 degrees of freedom
## Multiple R-squared:  0.000863,   Adjusted R-squared:  -0.02237
## F-statistic: 0.03714 on 1 and 43 DF,  p-value: 0.8481
```

```
## [1] 0.000000000001051603
```



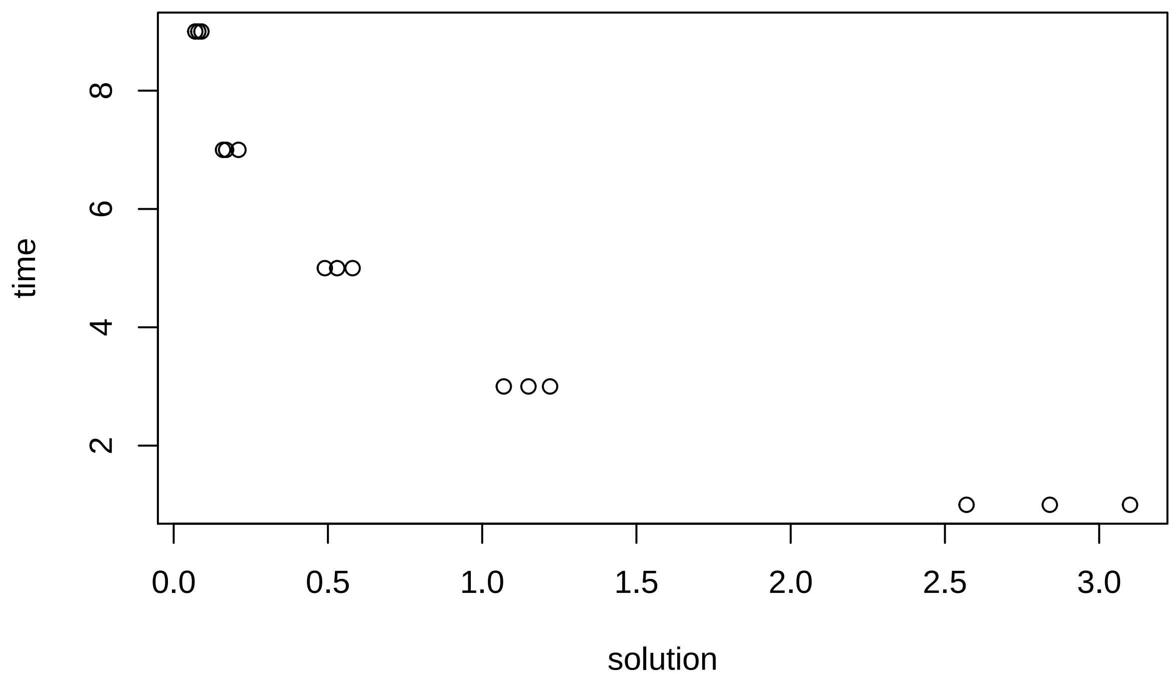

residuals(regres2)
0
500
1000
1500
0
500
1000
1500
2000
hours


Histogram of residuals(regres2)
Frequency
0
5
10
15
20
25
30
35
−500
0
500
1000
1500
2000
residuals(regres2)

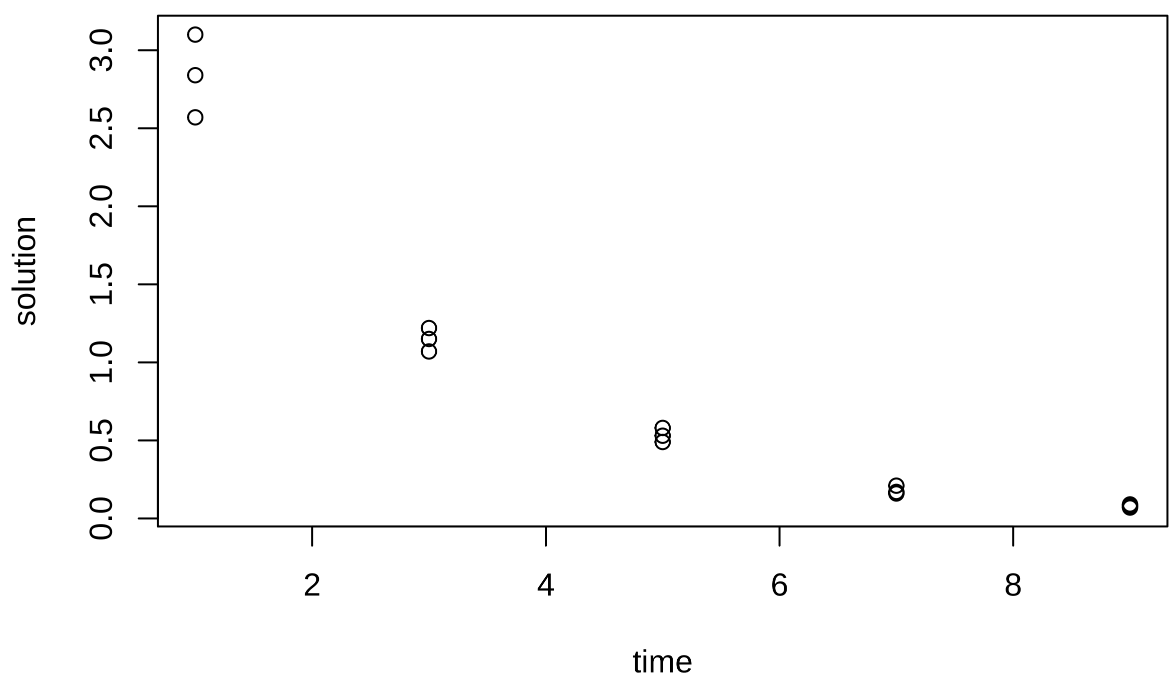


W zadaniu szóstym mieliśmy zmienić pierwszą obserwację z pierwszej kolumny z 20 na 2000 i zanalizować tak zmienione dane w podobny sposób jak w zadaniu piątym. Widzimy, że każdy wykres pokazuje nam obserwację odstającą i czasem bardzo duży wpływ na całość jaki generuje. Porównując summary od danych w zadaniu piątym i szóstym widzimy znaczne różnice. Na przykład R^2 dla danych z zadania piątego wynosi 0.9575 a dla danych z zadania szóstego praktycznie 0. Tak samo jest z testem istotności dla slope. Dla danych z zadania piątego odrzucamy H_0 a dla danych z zadania szóstego już nie.

Zadanie 7 i 8



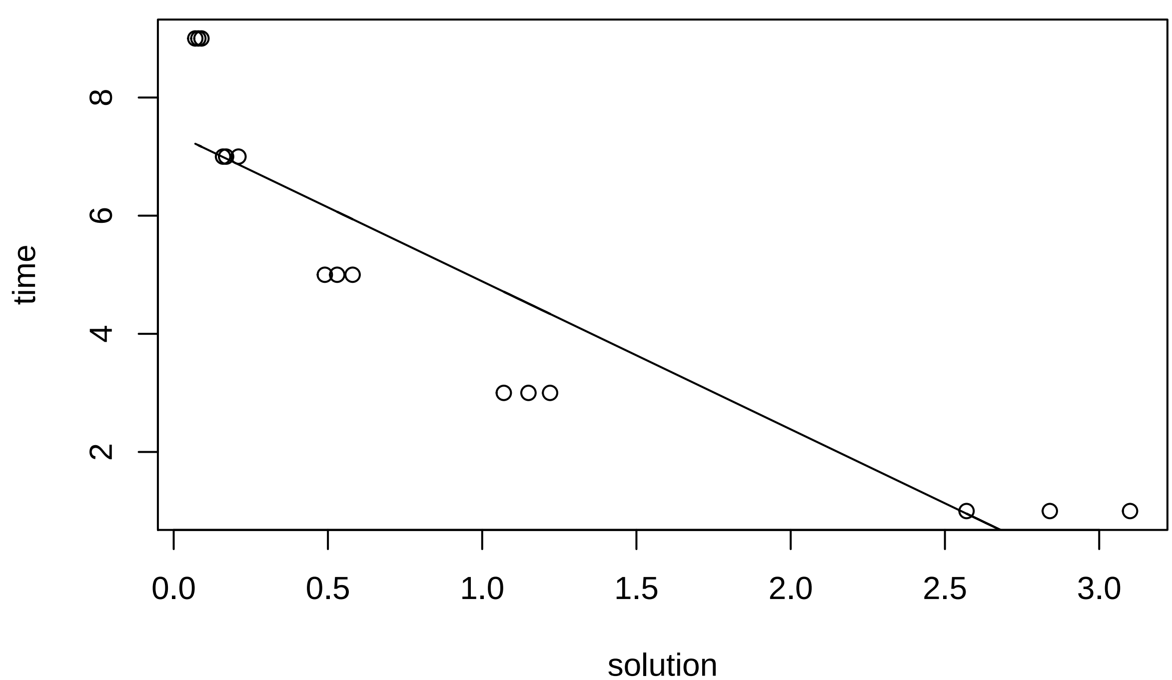
time
8
6
4
2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
solution



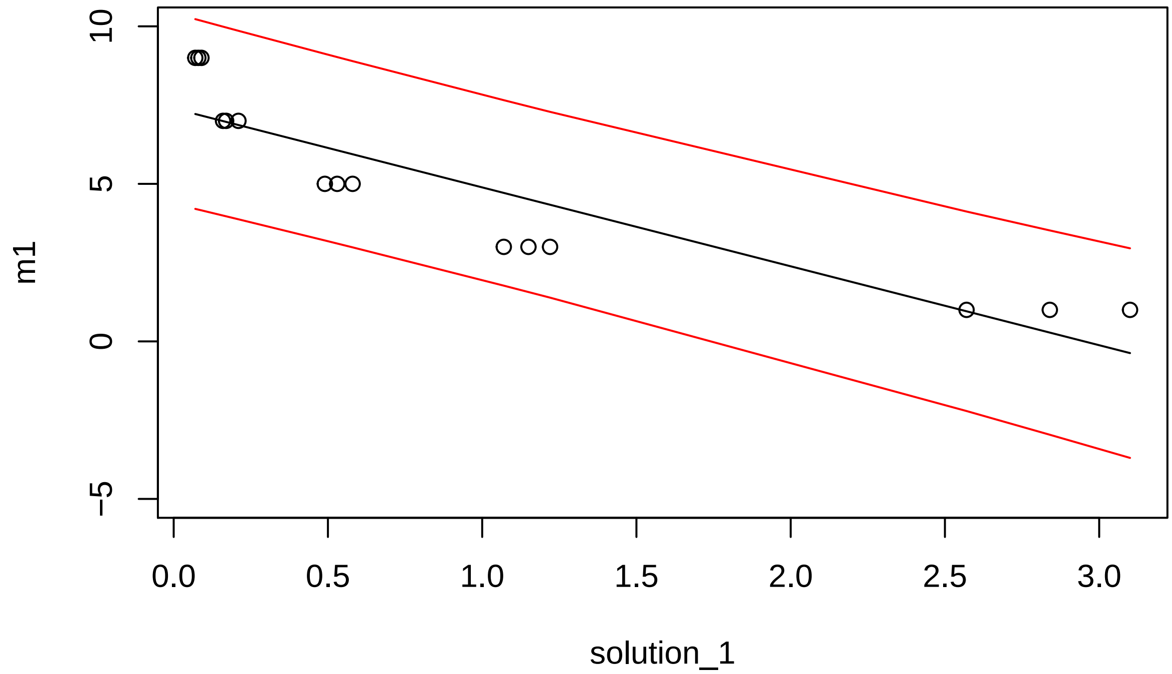
solution
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2
4
6
8
time

```
##
## Call:
## lm(formula = time ~ solution, data = dane_7_12)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71278 -1.11550  0.03284  1.04647  1.83245
##
## Coefficients:
##             Estimate Std. Error t value       Pr(>|t|)
## (Intercept)   7.3930     0.4671  15.826 0.000000000711 ***
## solution     -2.5049     0.3347  -7.483 0.000004611199 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.319 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: 0.000004611
```

```
## Analysis of Variance Table
##
## Response: time
##           Df Sum Sq Mean Sq F value      Pr(>F)
## solution   1 97.389  97.389  55.994 0.000004611 ***
## Residuals 13 22.611   1.739
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

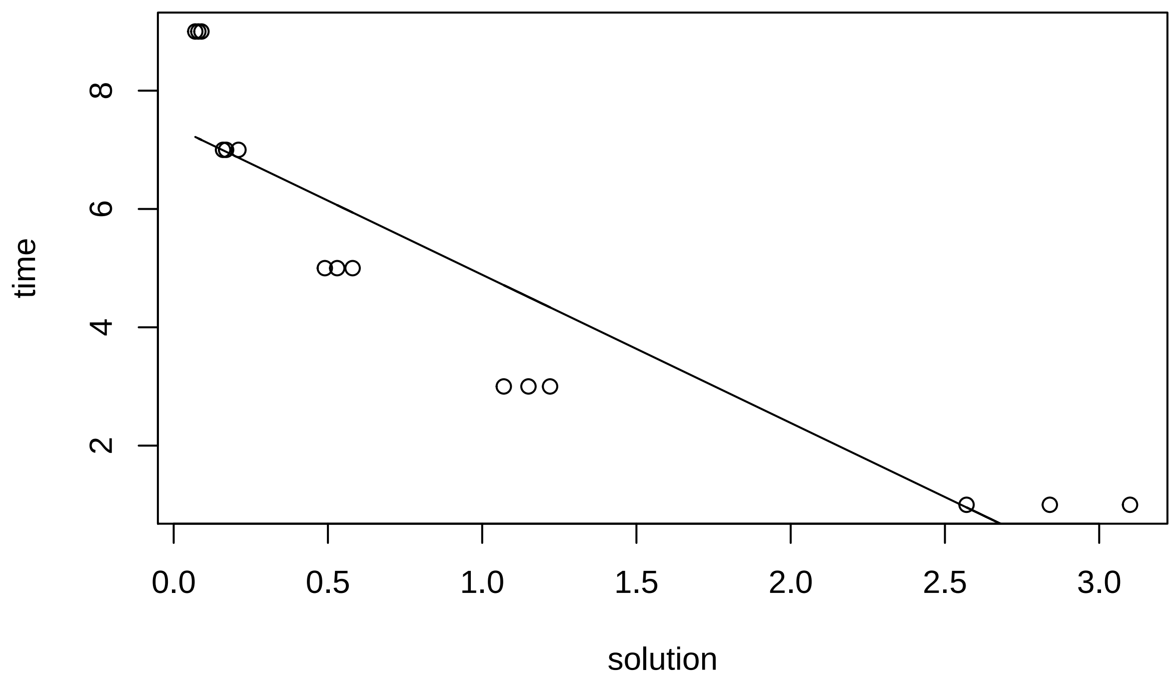


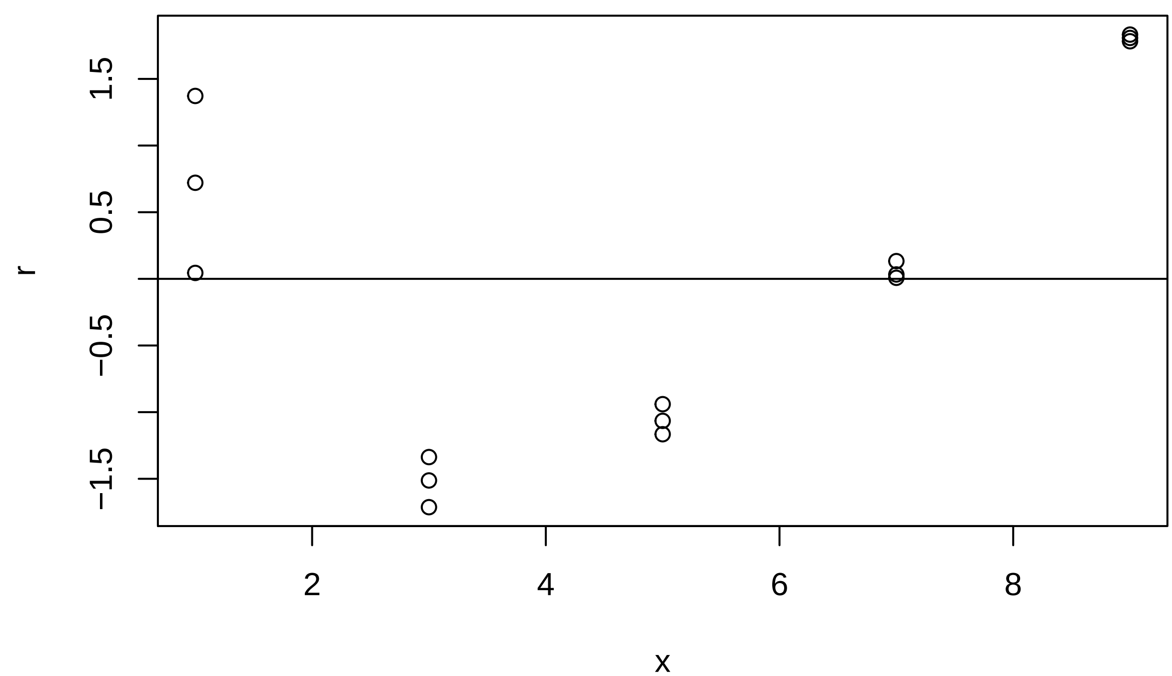
```
## [1] -0.9008759

## Warning in predict.lm(regres7, interval = "prediction"): predictions on current data refer to _futur
```

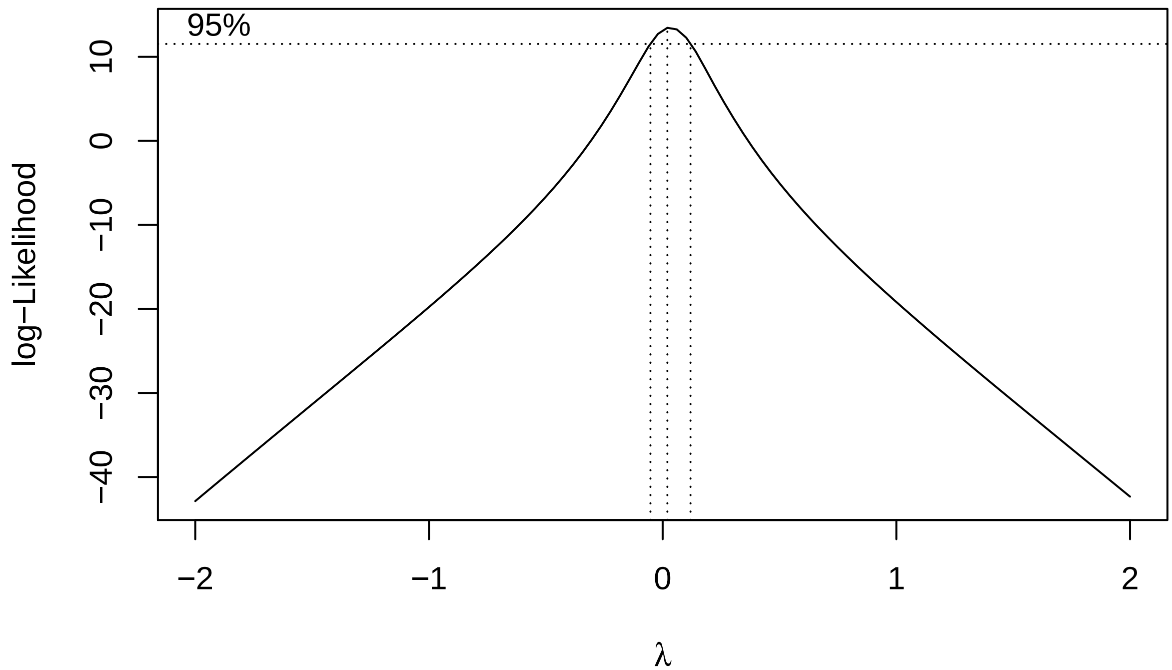


W zadaniu 7 mamy przeanalizować dane dotyczące stężenia roztworu (u nas solution) od czasu (time). Analizując summary od modelu liniowego mamy $R^2 = 0.8116$, a $AdjR^2 = 0.7971$. Odrzucamy H_0 dla Interceptu, czyli time, ponieważ p-wartość jest bliska 0, a statystyka T wynosi 15.826. Podobnie dla slope. Odrzucamy H_0 dla slope, czyli solution, ponieważ p-wartość jest bliska 0, a statystyka T wynosi -7.483. Rysujemy dodatkowo przedział predykcyjny dla tych danych w zadaniu ósmym. Wszystkie obserwacje mieszczą się w przedziale predykcujnym, ale jest on bardzo szeroki co wskazuje na słabe dopasowanie modelu do danych.

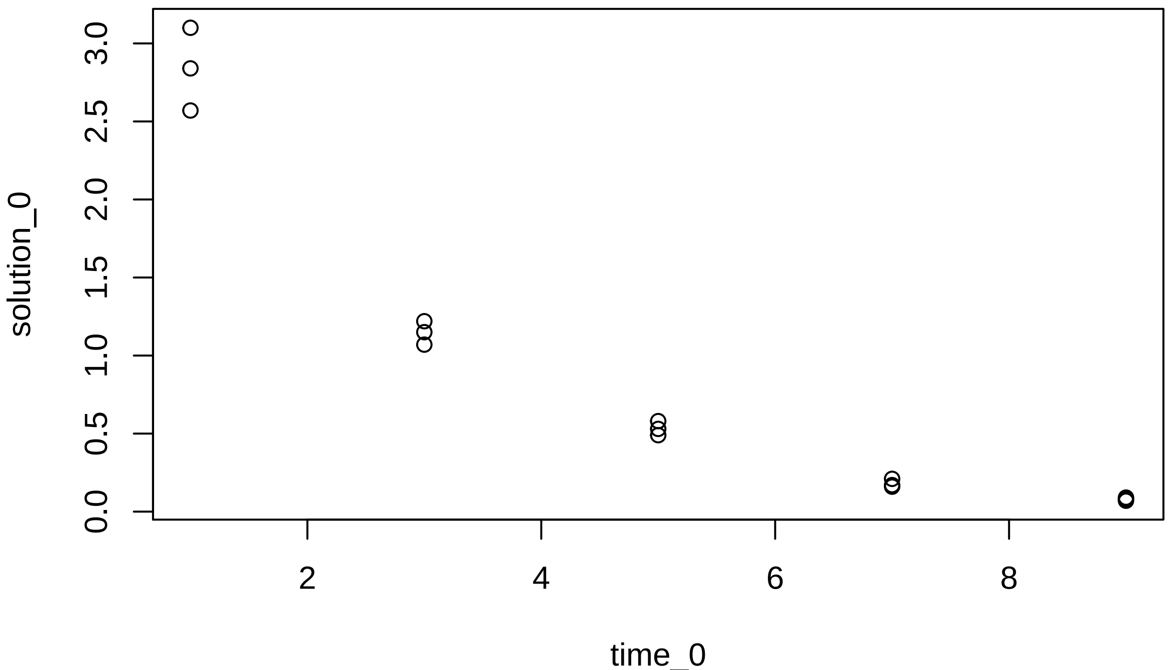
Zadanie 9 i 10



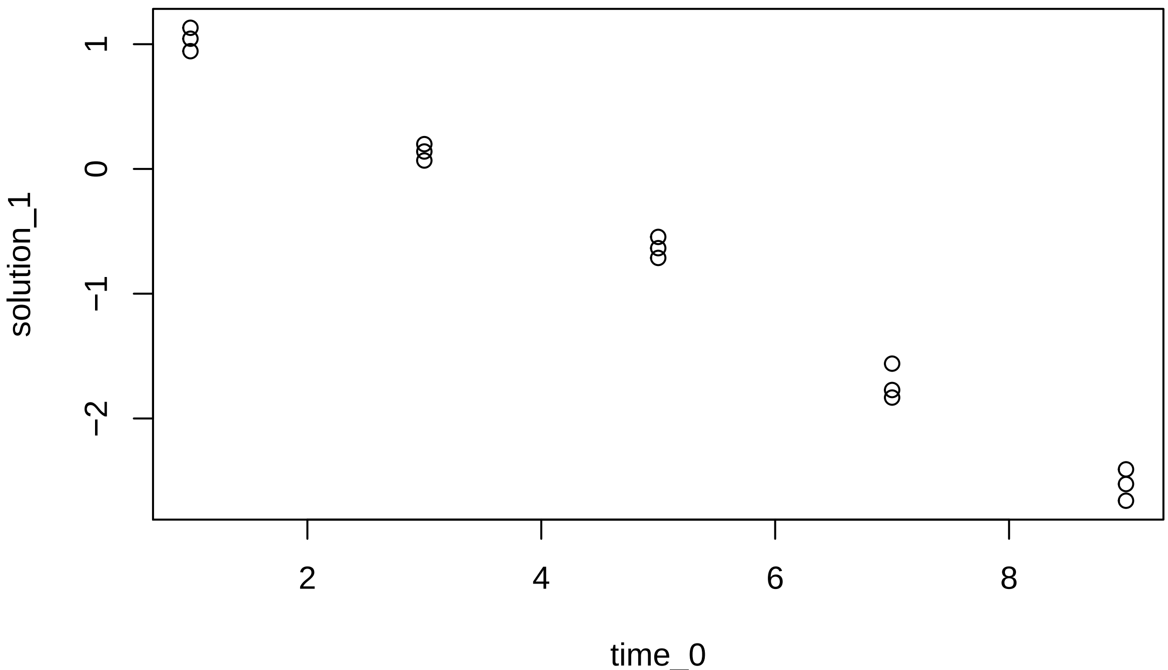




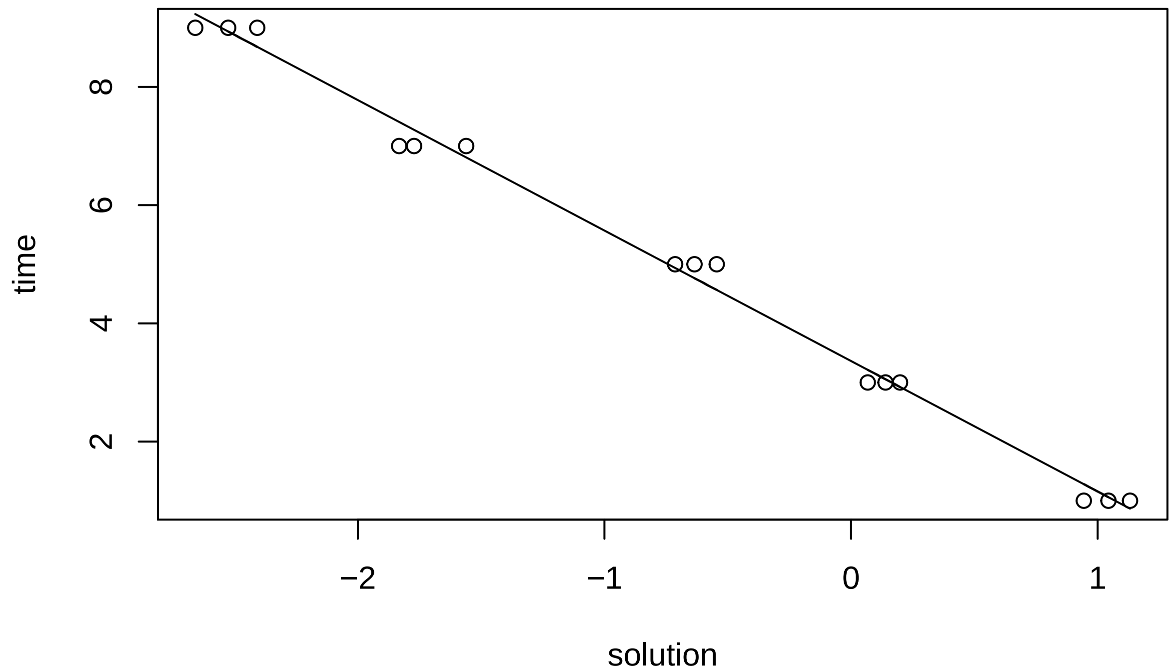
95%
log−Likelihood
10
0
−10
−20
−30
−40
−2
−1
0
1
2
λ

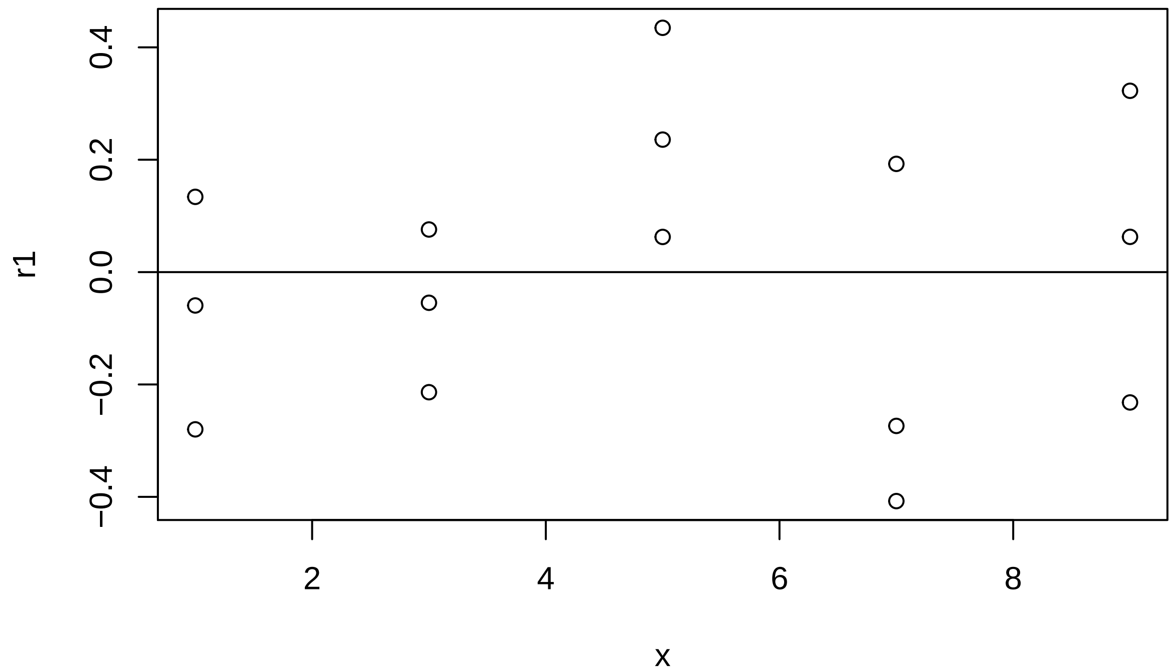


solution_0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
2
4
6
8
time_0



1
0
−1
−2
solution_1
2
4
6
8
time_0


time
8
6
4
2
−2
−1
0
1
solution



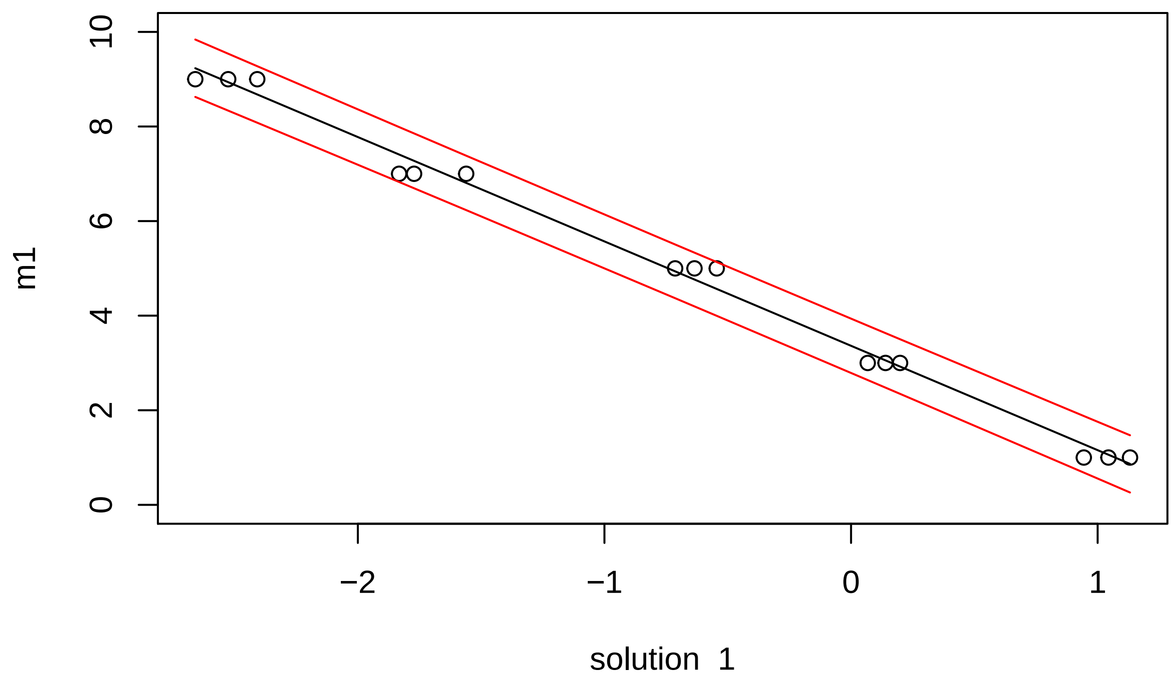

```
##
## Call:
## lm(formula = dane_7_12_boxcox)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4075 -0.2229  0.0626  0.1633  0.4347
##
## Coefficients:
##             Estimate Std. Error t value           Pr(>|t|)
## (Intercept)  3.36305    0.07602   44.24 0.00000000000000146 ***
## solution    -2.20698    0.05148  -42.88 0.00000000000000219 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2546 on 13 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.9924
## F-statistic:  1838 on 1 and 13 DF,  p-value: 0.000000000000002188
```

w zadaniu dziewiątym dopasowujemy model liniowy za pomocą transformaty boxa-coxa. Analizując wykresy widzimy, że model liniowy tutaj nie zadziala, ponieważ dane nie układają się liniowo. Po zastosowaniu boca-coxa widzimy, że najlepiej będzie nałożyc logarytm na solution. Po tym analizujemy tak uzyskane dane, widać że uzyskaliśmy już liniową zależność. W zadaniu dziesiątym bardziej szczegółowo analizujemy nowe dane. $R^2 = 0.993$, a $AdjR^2 = 0.9924$ co jest znakomitym wynikiem i wskazuje na silne dopasowanie prostej do danych. Intuicyjnie wiemy, że p-wartość musi być mała i H_0 bedzie odrzucona dla slope oraz dla interceptu (potwierdzenie dostajemy w summary).

Zadanie 11

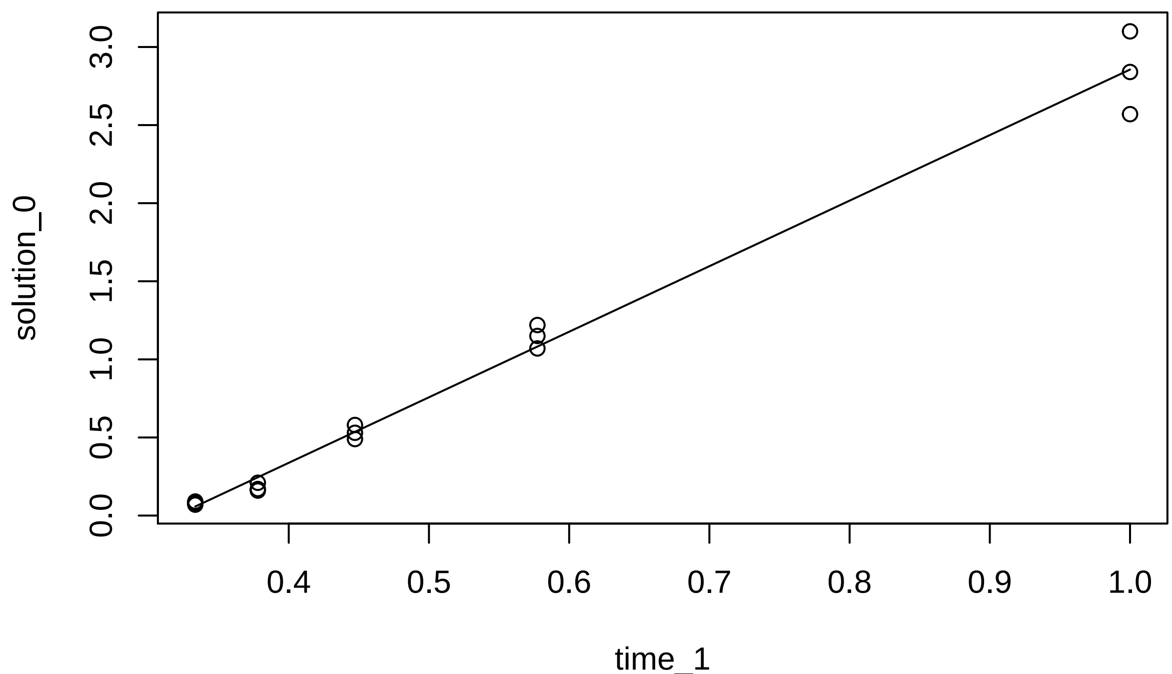
```
## [1] -0.9964826
```

```
## Warning in predict.lm(regres_boxcox, interval = "prediction"): predictions on current data refer to _
```

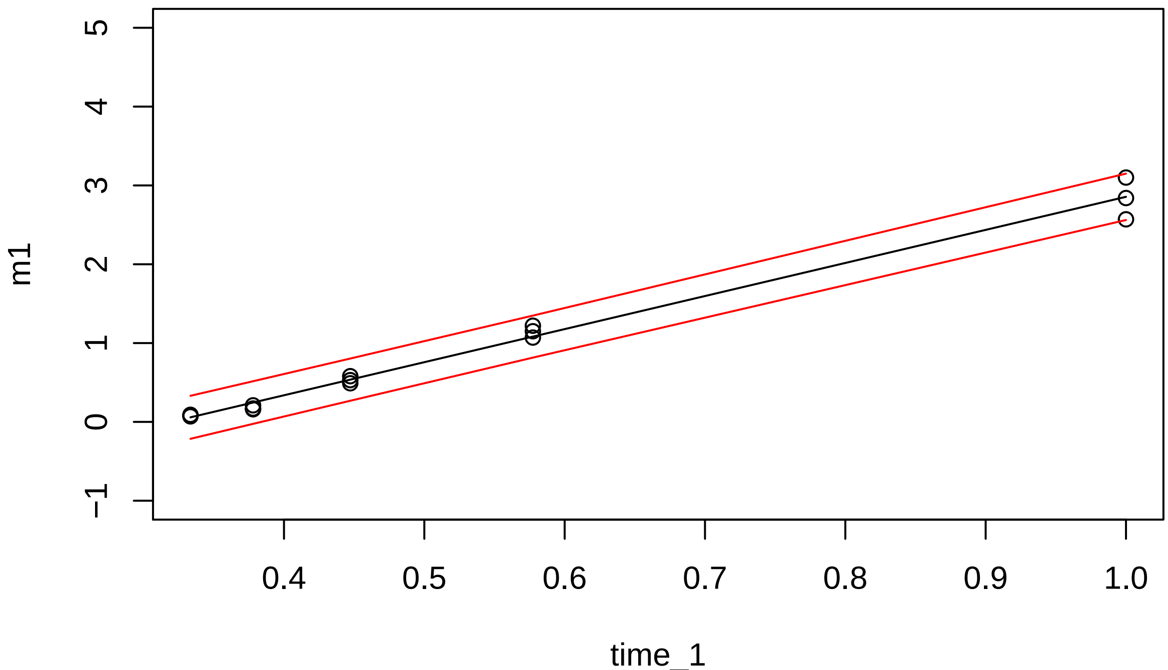


W zadaniu jedenastym tworzymy przedział predykcyjny dla nowych danych. Widzimy, że współczynnik korelacji wynosi prawie 1 co wskazuje na bardzo silną zależność. Widzimy, że przedział predykcyjny jest bardzo wąski, a i tak wszystkie dane się w nim znajdują co wskazuje na dobrze dobrany model liniowy.

Zadanie 12




```
##
## Call:
## lm(formula = solution_0 ~ time_1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.285543 -0.040579 -0.005875  0.038064  0.244457
##
## Coefficients:
##             Estimate Std. Error t value          Pr(>|t|)
## (Intercept) -1.34078    0.07648  -17.53 0.000000000198978 ***
## time_1       4.19632    0.12792   32.80 0.000000000000069 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1194 on 13 degrees of freedom
## Multiple R-squared:  0.9881, Adjusted R-squared:  0.9871
## F-statistic:  1076 on 1 and 13 DF,  p-value: 0.00000000000006898

## [1] 0.9940136

## Warning in predict.lm(regres_12, interval = "prediction"): predictions on current data refer to _futu
```



W ostatnim zadaniu mieliśmy zastąpić zmienną time przez time podniesione do potęgi -0.5. $R^2 = 0.9881$, a $AdjR^2 = 0.9871$ co jest również bardzo dobrym wynikiem i wskazuje na silne dopasowanie prostej do danych, ale trochę słabsze niż poprzednio. Intuicyjnie wiemy, że p-wartość musi być mała i H_0 bedzie odrzucona dla slope oraz dla interceptu, dokładnie tak samo jak poprzednio (potwierdzenie dostajemy w summary). Analizując wykresy i summary od modelu dochodzimy do wniosku, że lepsze wyniki dostaliśmy dzięki transformacie boxa-coxa. Zgodnie z oczekiwaniami najlepszy model jest po trasformacji boxa-coxa co jest zgodne z poznaną teorią na wykładzie.