Zadanie 3

For this and the next problem you will use the data set table1 6.txt, containing the grade point average (GPA) [second column], score on a standard IQ test [third column], gender and a score on the Piers-Harris Childrens Self-Concept Scale (a psychological test, fifth column) for 78 seventh-grade students.

```
data <- read.table("tabela1_6.txt", header = FALSE)[, -1]
colnames(data) <- c("GPA", "IQ", "Gender", "PHCSCS")
```

a) Use a simple regression model to describe the dependence of gpa on the results of iq test. Report the fitted regression equation and R2 . Test the hypothesis that gpa is not correlated with iq : give the test statistic, p-value and the conclusion in words

Wyznaczam współczynniki regresji, korzystając z metody *lm*:

```
model = lm(GPA ~ IQ, data)
```

Dopasowana prosta jest postaci $Y = 0.10102X - 3.557$

R^2, dane wzorem SSM/SST, wyniosło:

```
[1] 0.4016146
```

Wartość ta oznacza, że 40,1614% wariancji GPA jest wyjaśnione poprzez IQ.

Następnie, przetestuję hipotezę, że GPA nie jest skorelowane z IQ, to znaczy przetestuję: $H_0 : \beta_1 = 0 \; vs \; H_1 : \beta_1 \neq 0$

Statystyka testowa F, dana wzorem MSM/MSE wyniosła

```
[1] 51.00845
```

Co jest większe niż zmienna pochodząca z rozkładu Fishera-Snedecora o 1, 76 stopniach swobody:

```
qf(1 - 0.05, 1, 76)
```

```
[1] 3.96676
```

Możemy również zauważyć, że p-wartość jest dużo mniejsza niż 0.05, wyniosła:

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

Oznacza to, że możemy odrzucić hipotezę zerową o braku zależności liniowej pomiędzy GPA a IQ.

b) Predict gpa for a student whose iq is equal to 100. Report 90% prediction interval.

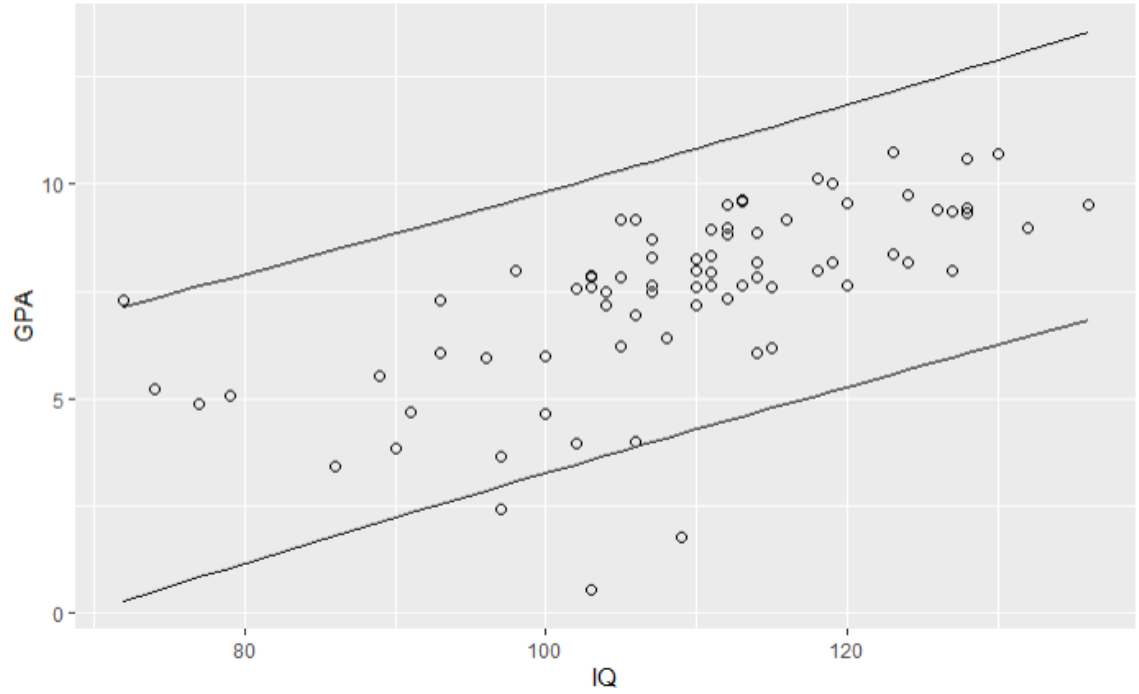
Dokonuję predykcji GPA dla osoby o IQ równym 100. W tym celu skorzystam z funkcji *predict*:

```
test_sample = data.frame(IQ=100)
predict(model, test_sample, interval="prediction", level=0.9)
```

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

Wyznaczony przez nas model, dla IQ wynoszącego 100, dopasował wartość 6.545114. Z prawdopodobieństwem 0.9, prawdziwa wartość GPA dla IQ wynoszącego 100, znajduje się w przedziale [3.79753, 9.292698].

c) Draw a band for 95% prediction intervals (i.e. join the limits of the prediction intervals with the smooth line). How many observations fall outside this band ?



Wyłącznie 4 na 78 obserwacje znajdują się poza przedziałami predykcyjnymi, co oznacza, że model jest dobrze dopasowany.

Zadanie 4

a) Use a simple regression model to describe the dependence of gpa on the results of PiersHarris test. Report the fitted regression equation and R2 .

Ponownie skorzytam z funkcji *lm*. Zmienną objaśnianą pozostaje GPA, natomiast zmienną objaśniająca tym razem stanie się wynik Piers-Harris Childrens Self-Concept Scale.

```
model = lm(GPA ~ PHCSCS, data)
summary(model)
```

```
Call:
lm(formula = GPA ~ PHCSCS, data = data)

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 *  
PHCSCS       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
```

Dopasowana prosta jest postaci $Y = 0.09165X + 2.22588$

R^2 wyniosło:

```
[1] 0.2935829
```

Co oznacza, że tym razem jedynie 29% wariancji GPA jest wyjaśniane przez zmienną objaśniającą - test PHCSCS.

b) Test the hypothesis that gpa is not correlated with Piers-Harris score : give the test statistic, p-value and the conclusion in words.

Ponownie, testujemy: $H_0 : \beta_1 = 0 \; vs \; H_1 : \beta_1 \neq 0$

Statystyka testowa F wyniosła:

```
[1] 31.58517
```

Co jest wartością większą niż obliczona w poprzednim zadaniu wartość zmiennej losowej z rozkładu Fishera-Snedecora dla 1, 76 stopni swobody.

Wartość-p wyniosła:

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

Co jest wartością dużo mniejszą niż 0.05.

Podsumowując, możemy odrzucić hipotezę zerową o braku zależności liniowej pomiędzy GPA a PHCSCS.

c) Predict gpa for a student whose Piers-Harris score is equal to 60. Report 90% prediction interval .

Dokonuję predykcji GPA dla osoby o wyniku testu PHCSCS równym 60. W tym celu skorzystam z funkcji *predict*:

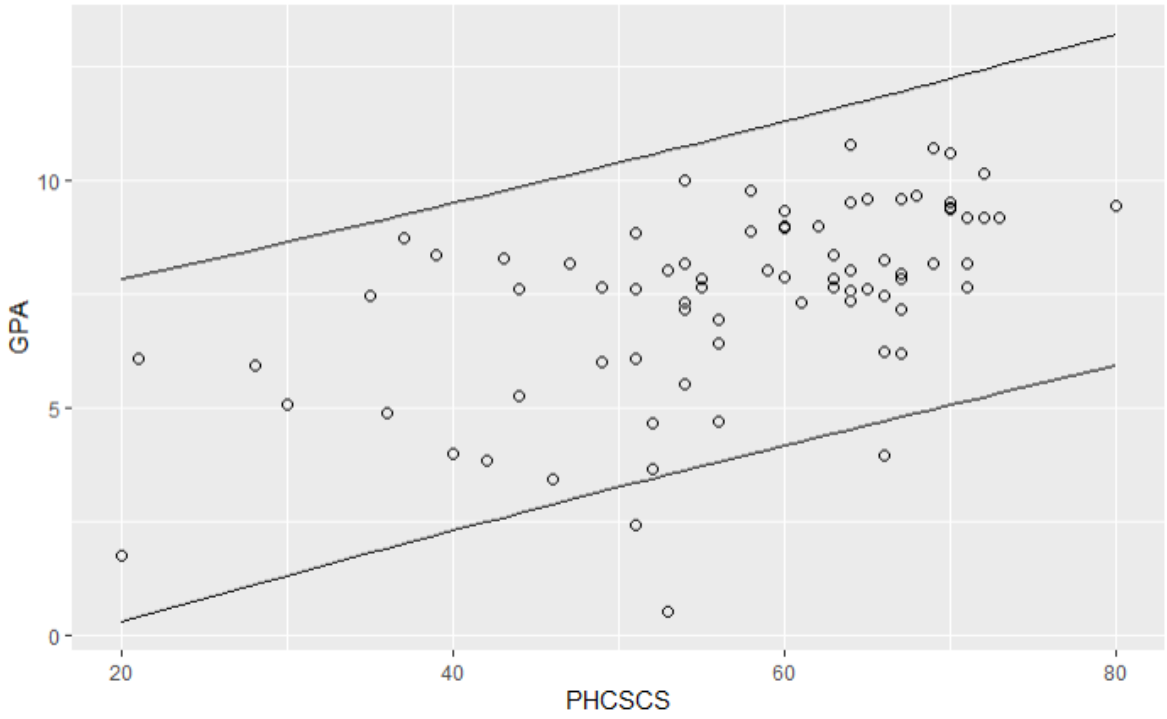
```
test_sample = data.frame(PHCSCS=60)
predict(model, test_sample, interval="prediction", level=0.9)
```

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

Wyznaczony przez nas model, dla PHCSCS wynoszącego 60, dopasował wartość 7.72502. Z prawdopodobieństwem 0.9, prawdziwa wartość GPA dla PHCSCS wynoszącego 60, znajduje się w przedziale [4.747302, 10.70274].

d) Draw a band for 95% prediction intervals. How many observations fall outside this band?

e) Which of the two variables : result of iq test or result of Piers-Harris test, is a better predictor of gpa ?



Po zaznaczeniu 95-procentowych przedziałów predykcyjnych, możemy zauważyć, że tym razem 3 obserwacje znajdują się poza nimi. Jest to mniej niż 4 obserwacje z poprzedniego zadania. Nie oznacza to jednak, że dla GPA PHCSCS jest lepszym predyktorem od IQ. Statystyka testowa F dla pierwszego modelu była o wiele większa niż dla obecnego, co oznacza, że zależność pomiędzy IQ a GPA jest silniejsza niż pomiędzy PHCSC a GPA. Ponadto, wartość R^2 dla modelu zbudowanego dla PHCSC wyniosło jedynie 0.29, a dla modelu zbudowanego dla IQ, wyniosło 0.40, co oznacza, że IQ wyjaśnia większą część wariancji GPA niż PHCSC.

Zadanie 5.

For the next two problems you will use the copier maintenance data,ch01pr20.txt, discussed in class.

```
data = read.table("CH01PR20.txt", header=FALSE, col.names=c("time", "copiers"))
model = lm(time ~ copiers, data)
```

a) Verify that the sum of the residuals is zero.

Suma residuów powinna wynieść 0. Wynika to wprost ze wzoru:

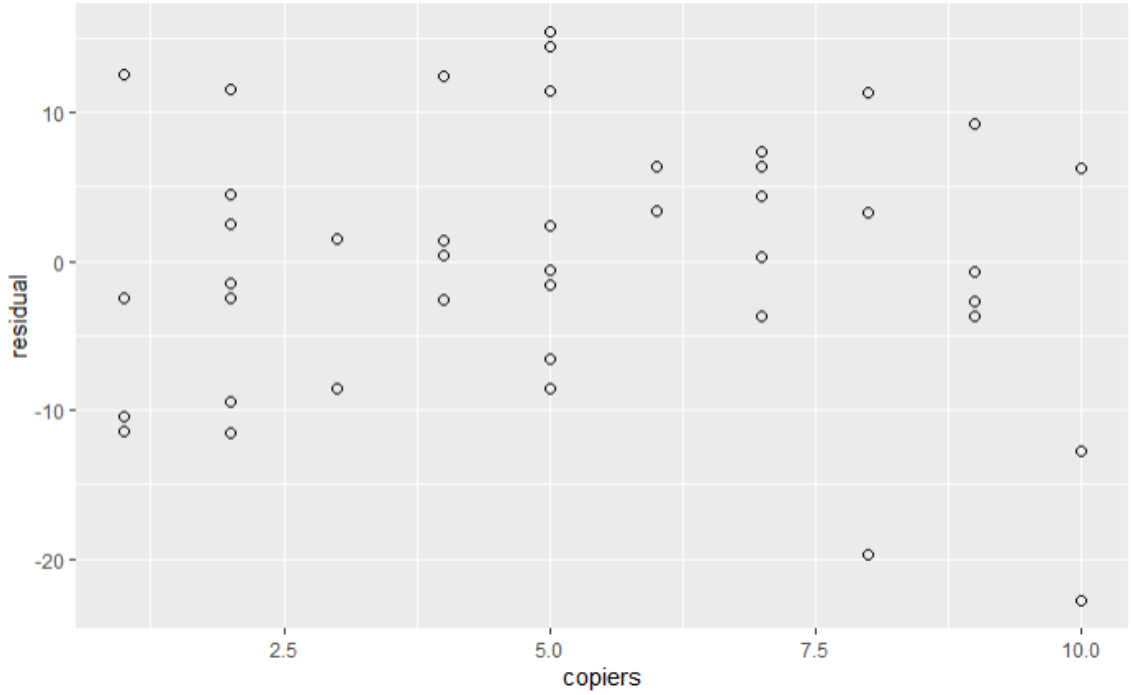
$$\sum_i (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)) = \sum_i (Y_i - \bar{Y} + \hat{\beta}_1 \bar{X} - \hat{\beta}_i X_i) = \sum_i (Y_i - \bar{Y} + \hat{\beta}_1 (\bar{X} - X_i)) = \sum_i (Y_i) - n\bar{Y} + \hat{\beta}_1 (n\bar{X} - \sum_i X_i) = 0$$

```
sum(model$residuals)
```

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

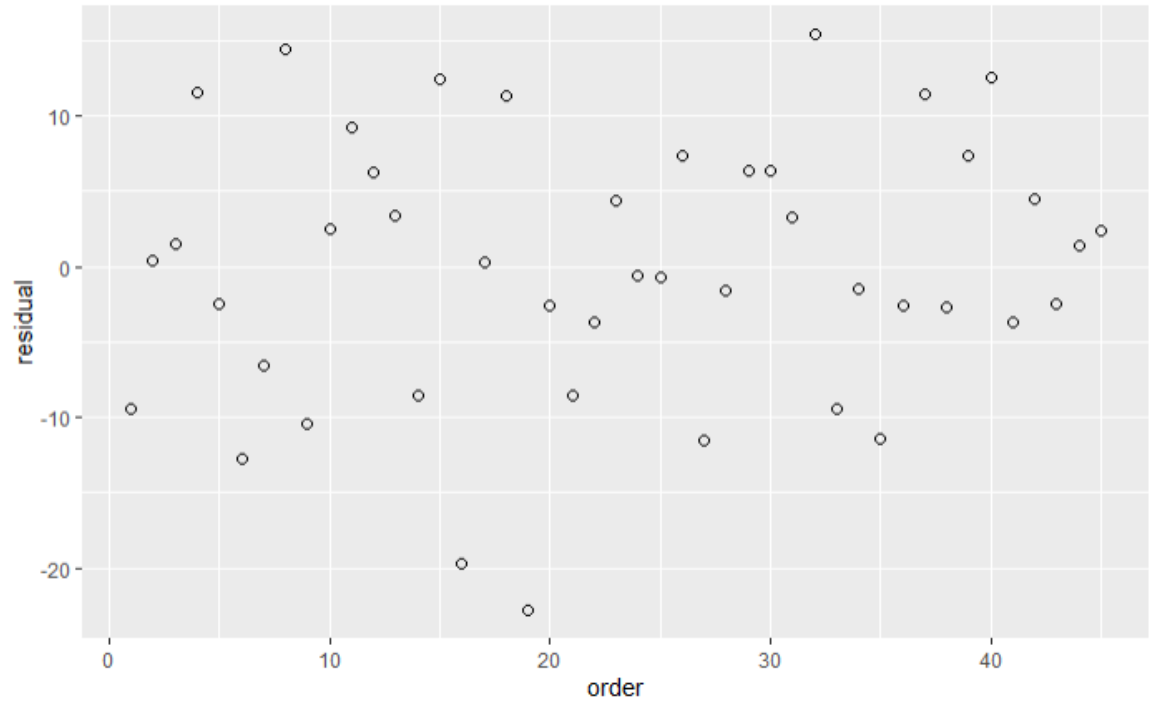
Po zweryfikowaniu możemy stwierdzić, że suma residuów jest bardzo bliska 0.

b) Plot the residuals versus the explanatory variable and briefly describe the plot noting any unusual patterns or points.



Poza paroma obserwacjami odstającymi, wariancja residuów dla poszczególnych wartości zmiennej copiers, jest podobna.

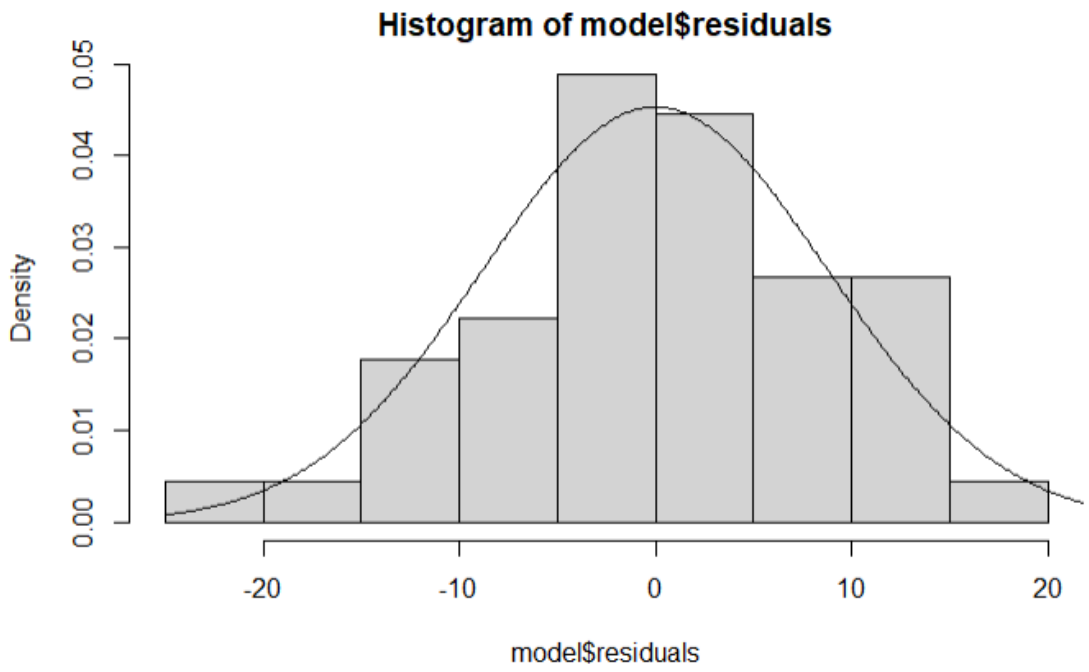
c) Plot the residuals versus the order in which the data appear in the data file and briefly describe the plot noting any unusual patterns or points.



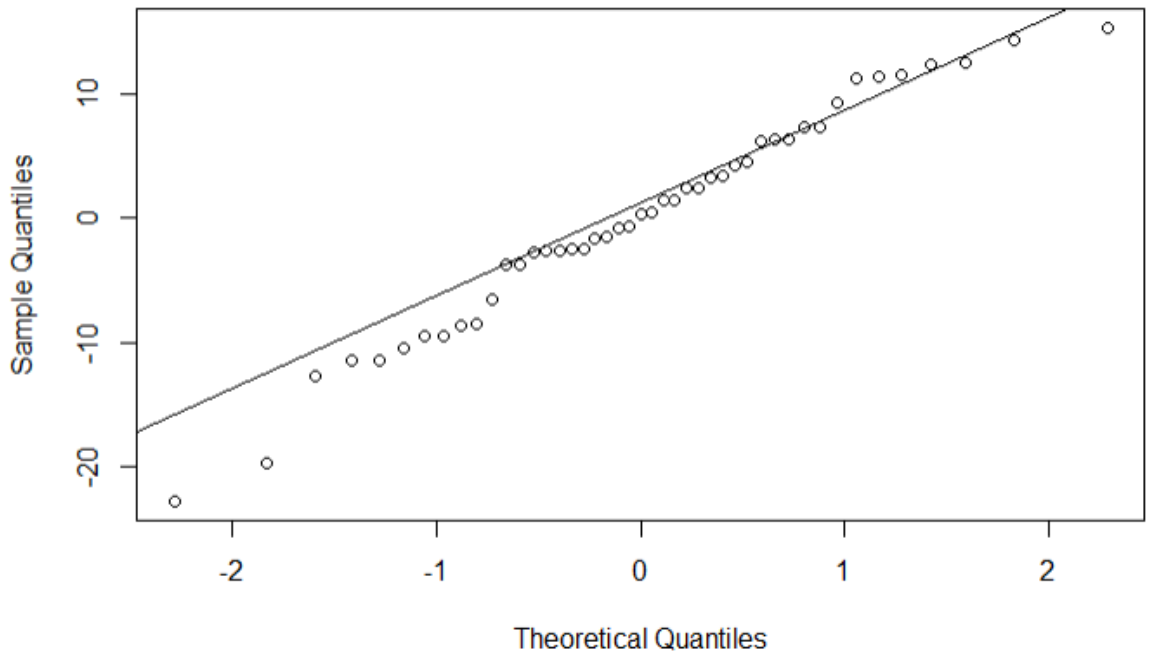
Rozmieszczenie residuów na wykresie residua vs kolejność obserwacji, jest losowe.

d) Examine the distribution of the residuals by getting a histogram and a normal probability plot. What do you conclude ?

Histogram wartości residuów wraz z naniesionym na nim wykresem gęstości rozkładu normalnego o średniej 0 i standardowym odchyleniu równym standardowemy odchyleniu residuów:



Wykres kwantylowo kwantylowy:



Powyższe wykresy sugerują, że residua są z rozkładu normalnego, co oznacza, że dane spełniają założenia dla regresji liniowej.

Zadanie 6

Change the data set by changing the value of service time for the first observation from 20 to 2000.

```
changed_data = data
changed_data$time[1] = 2000
```

a) Run the regression with changed data and make a table comparing the results of this analysis with the results of the analysis of the original data. Include in the table the following: fitted equation, t-test for the slope with P-value, R2 , and the estimate of σ 2 . Briefly summarize the differences.

Wyznaczam model regresji liniowej dla danych ze zmienioną pierwszą obserwacją czasu serwisu z 20 na 2000.

```
new_model = lm(time ~ copiers, changed_data)
```

Poniższa tabela porównuje otrzymany model z modelem otrzymanym w poprzednim zadaniu (dla danych niezmienionych):

	Oryginalny model	**Model dla zmienionych danych**
równanie regresji	$15.0352X - 0.5802$	$-3.059X + 135.9$
statystyka F	968.7	0.03714
wartość-p	$< 2e - 16$	0.848
R^2	0.9575	0.000863
σ^2	8.914	292.8

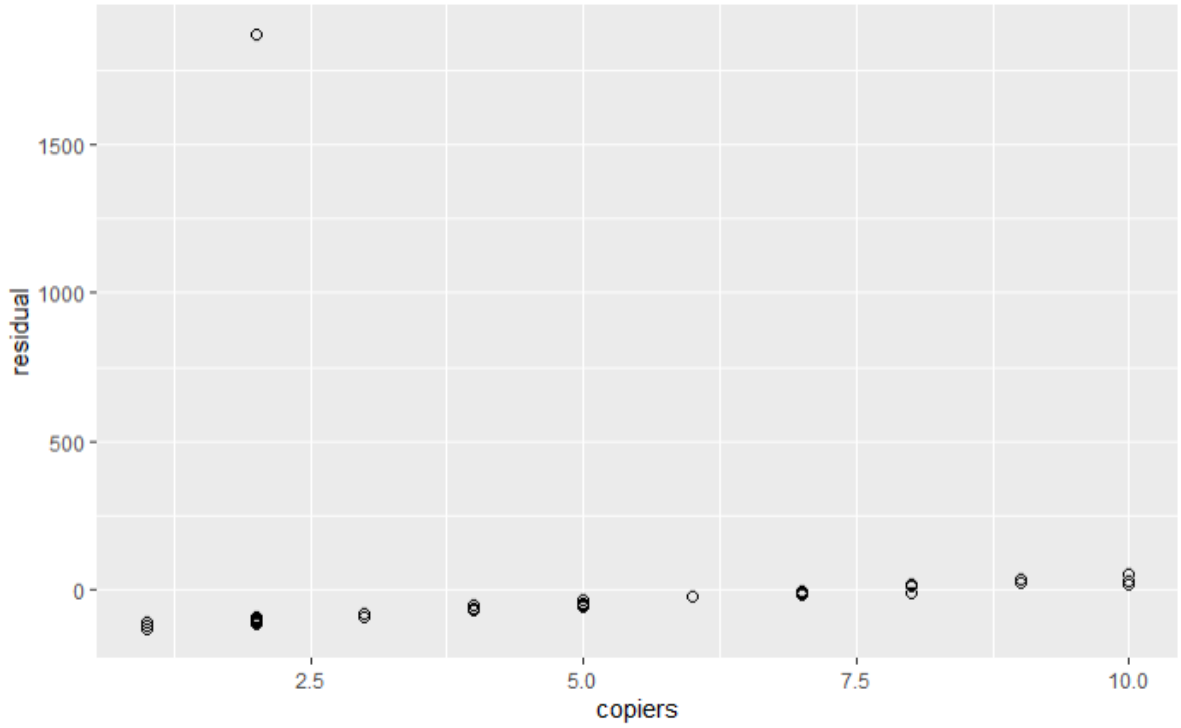
P-wartość dla oryginalnego modelu jest bardzo mały, o wiele mniejsza niż dla modelu ze zmienionych danych, która wynosi aż 0.848, co nie pozwala odrzucić hipotezy o braku zależności liniowej pomiędzy predyktorem a zmienną objaśnianą.

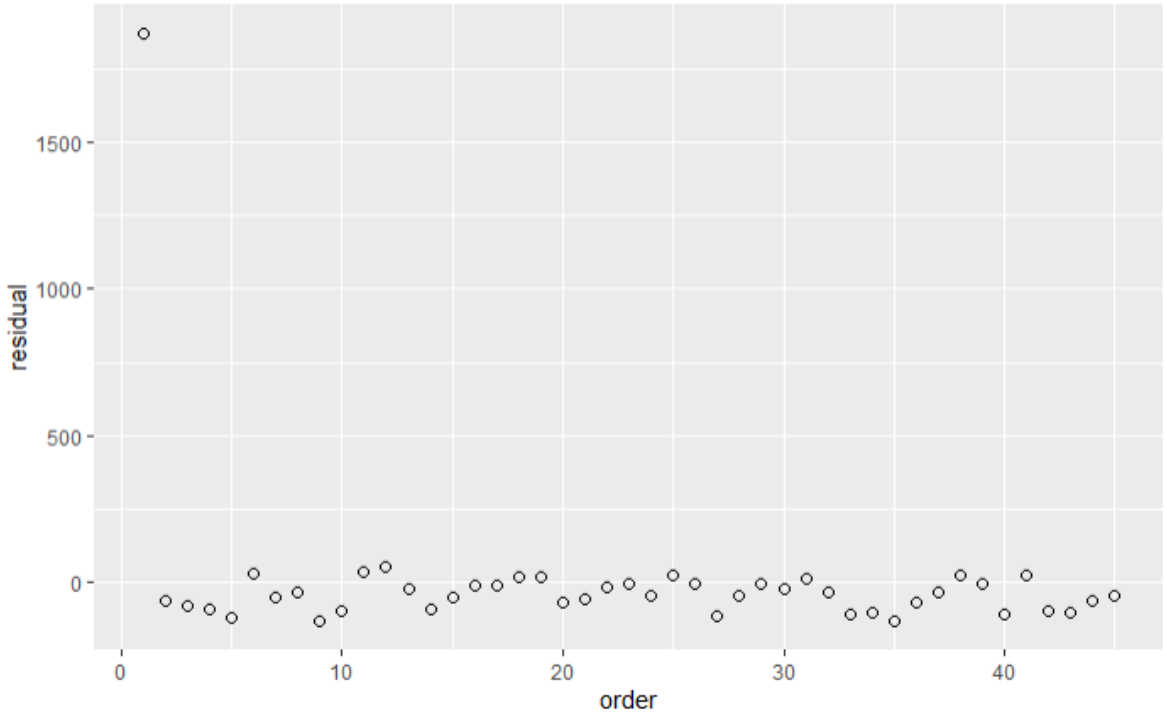
R^2 dla oryginalnego modelu wynosi 0.9575, co oznacza, że ponad 95% wariancji czasu jest wyjaśniane przez liczbę kopiarek. Dla modelu ze zmienionymi danymi, R^2 wynosi jedyne 0.000863.

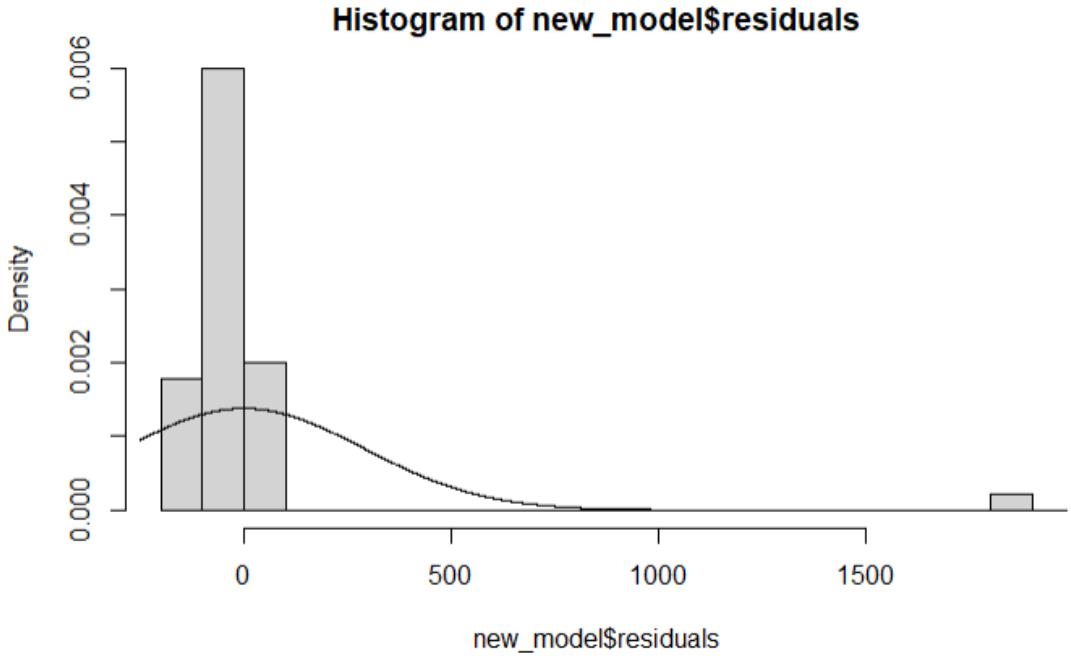
Standarowe odchylenie oryginalnego modelu jest znacznie niższe niż dla nowego modelu - 8.914 vs 292.8.

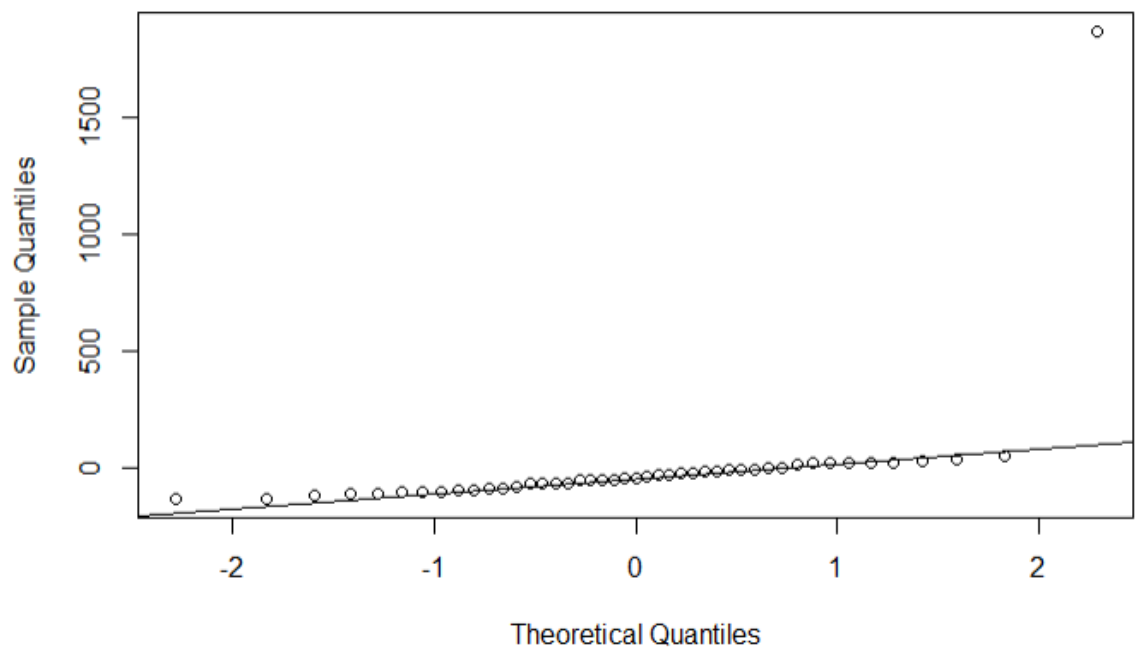
Powyższe obserwacje wskazują, że zmieniona wartość pierwszej obserwacji skrajnie pogorszyła wyniki modelu regresji liniowej.

b) Repeat points (b), (c) and (d) from problem 5 above on the modified data set and show the unusual observation on each of these plots.




residual
1500
1000
500
0
0
10
20
30
40
order


Histogram of new_model$residuals
Density
0.000
0.002
0.004
0.006
0
500
1000
1500
new_model$residuals



Powyższe wykresy pokazują, że resiuda nie mają rozkładu normalnego. Zmiana wartości pojedynczej obserwacji sprawiła, że założenia modelu liniowego zostały złamane.

For next six problems you will use the solution concentration data ch03pr15.txt. The first column gives the values of the solution concentration and the second column gives the time.

```
data = read.table("CH03PR15.txt", header=FALSE, col.names=c("concentration", "time"))
```

Zadanie 7

Run the linear regression with time as the explanatory variable and the solution concentration as the response variable. Summarize the regression results by giving the fitted regression equation, the value of R2 and the results of the significance test for the null hypothesis that the solution concentration does not depend on time (formulate the statistical model, give null and alternative hypotheses in terms of the model parameters, test statistic with degrees of freedom, P-value, and brief conclusion in words).

Tworzę model regresji liniowej dla zmiennej objaśnianej będącej stężeniem, gdzie predyktor to czas.

```
model = lm(concentration ~ time, data)
summary(model)
```

```
Call:
lm(formula = concentration ~ time, data = data)

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 ***
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
```

Równanie regresji jest dane wzorem: $Y = -0.324X + 2.5753$.

R^2 wynosi 0.8116. 81% wariancji w stężeniu jest wyjaśniane przez czas.

Wykonam test: $H_0 : \beta_1 = 0 \; vs \; H_1 : \beta_1 \neq 0$

Statystyka testowa F wyniosła 55.99, co jest znacząco większe od:

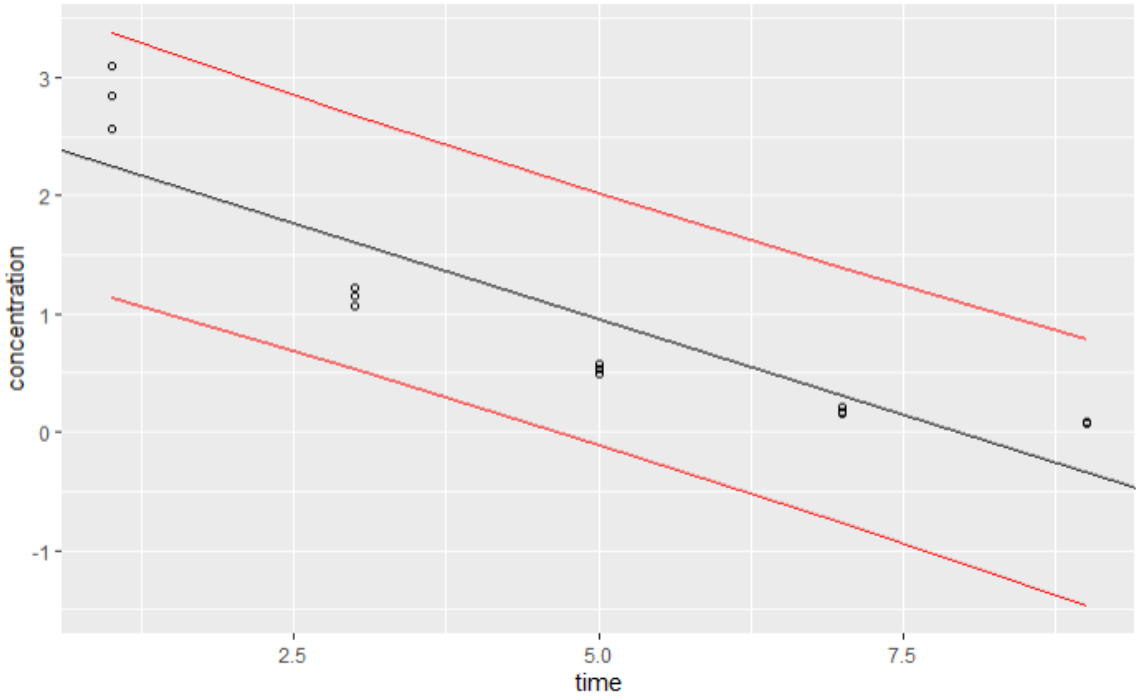
```
qt(1-0.05, 1, 13)
```

```
[1] 4.667193
```

możemy więc odrzucić hipotezę zerową o braku zależności liniowej pomiędzy czasem a stężeniem. Wynika to również z faktu, że wartość-p wyniosła jedynie $4.611 * 10^{-6}$

Zadanie 8.

Plot the solution concentration versus time. Add a fitted regression line and a band for 95% prediction intervals. What do you conclude ? Calculate the correlation coefficient between the observed and predicted value of the solution concentration.



Powyższy wykres przedstawia obserwacje, na podstawie których powstał model regresji liniowej, którego prosta jest zaznaczona czarnym kolorem. Czerwone proste są 95-procentowymi przedziałami predykcyjnymi. Żadne obserwacje nie znalazły się poza nimi, jednak same przedziały są dość szerokie.

Policzę korelację pomiędzy prawdziwymi wartościami stężenia, a tymi danymi przez model regresji liniowej:

```
preds = predict(model, data, interval = "prediction")[, 1]
cor(preds, data$concentration)
```

```
[1] 0.9008759
```

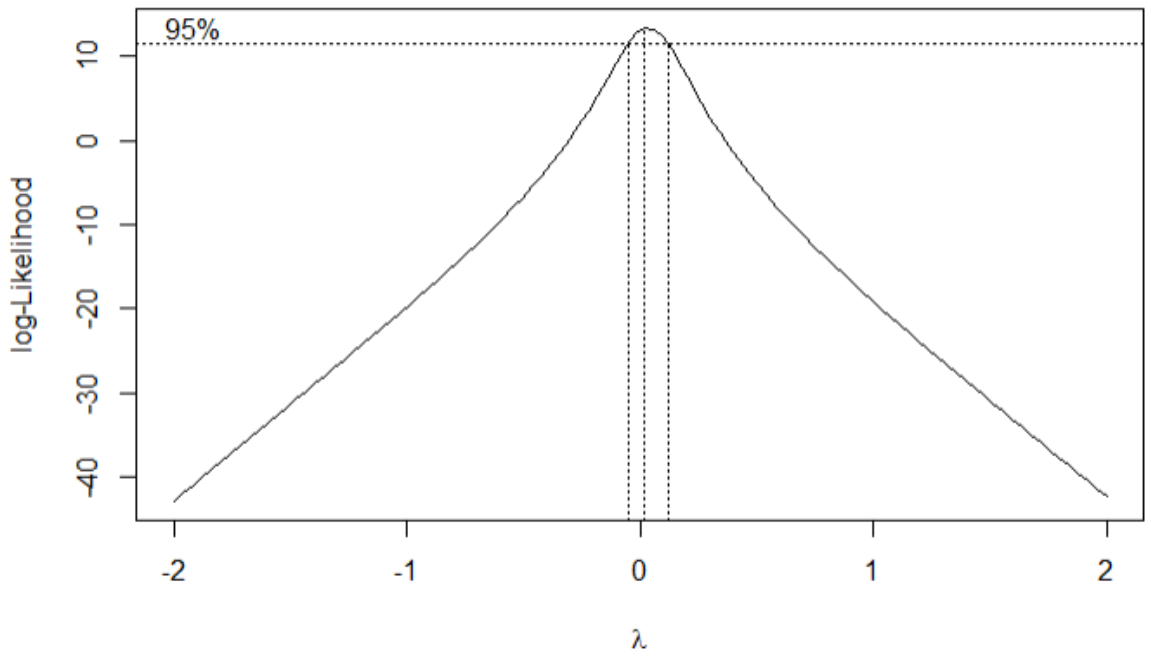
Możemy zauważyć, że wartości są ze sobą mocno skorelowane.

Zadanie 9

Use the Box-Cox procedure to find an appropriate transformation for the solution concentration.

Skorzystam z funkcji boxcox, aby znaleźć odpowiednie przekształcenie dla zmiennej opisującej stężenie.

```
library(MASS)
boxcox(data$concentration ~ data$time)
```



Funkcja osiąga swoje maksimum dla $\lambda = 0$, więc najlepszym przekształceniem będzie $log(Y)$.

Zadanie 10

Construct a new response variable by taking the log of the solution concentration (define logy = log(Y)). Repeat points 7 and 8 of this homework with logy as the response variable (and time as the explanatory variable). Summarize your results.

Przekształcam zmienną objaśnianą i ponownie wyznaczam model regresji liniowej.

```
data$concentration = log(data$concentration)
```

```
model = lm(concentration ~ time, data)
summary(model)
```

```
Call:
lm(formula = concentration ~ time, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19102 -0.10228  0.01569  0.07716  0.19699

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.50792    0.06028   25.01 2.22e-12 ***
time        -0.44993    0.01049  -42.88 2.19e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.115 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: 2.188e-15
```

Równanie regresji jest dane wzorem: $Y = -0.44993X + 1.50792$.

R^2 wynosi 0.993. Aż 99% wariancji w stężeniu jest wyjaśniane przez czas.

Wykonam test: $H_0 : \beta_1 = 0 \; vs \; H_1 : \beta_1 \neq 0$

Statystyka testowa F wyniosła 1838, co jest znacząco większe od:

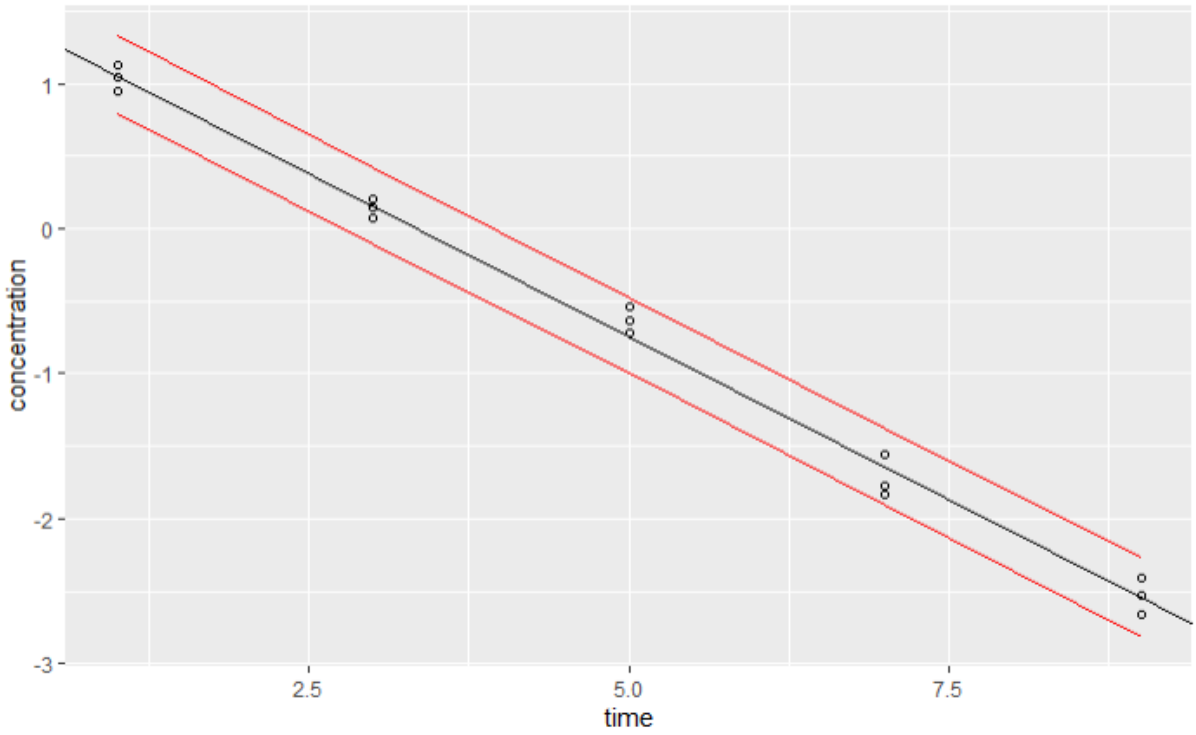
```
qt(1-0.05, 1, 13)
```

```
[1] 4.667193
```

możemy więc odrzucić hipotezę zerową o braku zależności liniowej pomiędzy czasem a logarytmem stężenia.

Zarówno wyższa wartość F jak i R^2 sugeruje, że pomiędzy czasem a logarytmem stężenia występuje bardzo silna zależność liniowa, silniejsza niż w przypadku zależności pomiędzy czasem a stężeniem.

Poniższy wykres pokazuje, że 95-procentowe przedziały predykcyjne są tym razem o wiele węższe:



Korelacja jest większa niż dla poprzedniego modelu:

```
preds = predict(model, data, interval = "prediction")[, 1]
cor(preds, data$concentration)
```

```
[1] 0.9964826
```

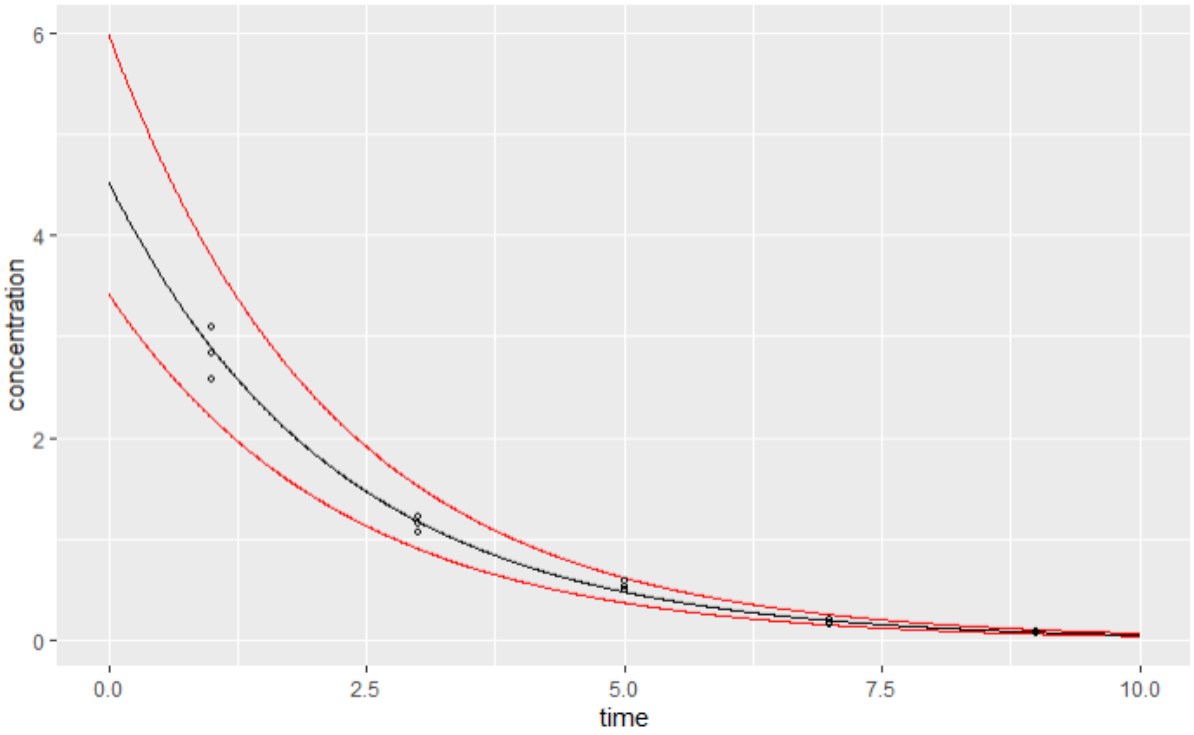
Transformacja Boxa-Coxa pozwoliła uzyskać o wiele lepiej dopasowany model.

Zadanie 11

Plot the solution concentration versus time. Add a regression curve and a band for 95% prediction intervals based on the results obtained in point 10. Compare to the graph obtained in point 8. Calculate the correlation coefficient between the observed solution concentration and the predictions based on the model from point 10.

Po wykonaniu transformacji uzyskałem model dany wzorem $log(Y) = -0.44993X + 1.50792$, więc $Y = e^{-0.44993X+1.50792}$

Po odpowiednich przekształceniach, wyznaczam wykres z oryginalnymi obserwacjami, krzywą regresji liniowej oraz 95-procentowymi przedziałami predykcyjnymi:



Wszystkie obserwacje znajdują się w obrębie przedziałów predykcyjnych. Krzywa wydaje się być idealnie do nich dopasowana.

Korelacja pomiędzy prawdziwymi wartościami, a przekształconymi predykcjami modelu wynosi:

```
cor(exp(predict(model, data, interval="prediction")[, 1]), old_data$concentration)
```

```
[1] 0.9945587
```

Co jest wyższą wartością niż w przypadku modelu z zadania 8., gdzie wyniosła ona 0.9008

Zadanie 12

Construct a new explanatory variable t1 = time−1/2 . Repeat points 10 and 11 of this exercise using the regression model with the solution concentration as the response variable and t1 as the explanatory variable. Summarize your results. Which model seems to be the best?

Wyznaczam model regresji dla zmiennej objaśniającej (będącej czasem) przekształconej do postaci $\frac{1}{\sqrt{X}}$.

```
data$time = 1 / sqrt(data$time)
model = lm(concentration ~ time, data)
summary(model)
```

```
Call:
lm(formula = concentration ~ time, data = data)
```

```
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 1.99e-10 ***
time         4.19632    0.12792   32.80 6.90e-14 ***
---
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: 6.898e-14
```

Równanie regresji jest dane wzorem: $Y = 4.19632X - 1.34078$.

R^2 wynosi 0.9881. Aż 98% wariancji w stężeniu jest wyjaśniane przez czas.

Wykonam test: $H_0 : \beta_1 = 0 \; vs \; H_1 : \beta_1 \neq 0$

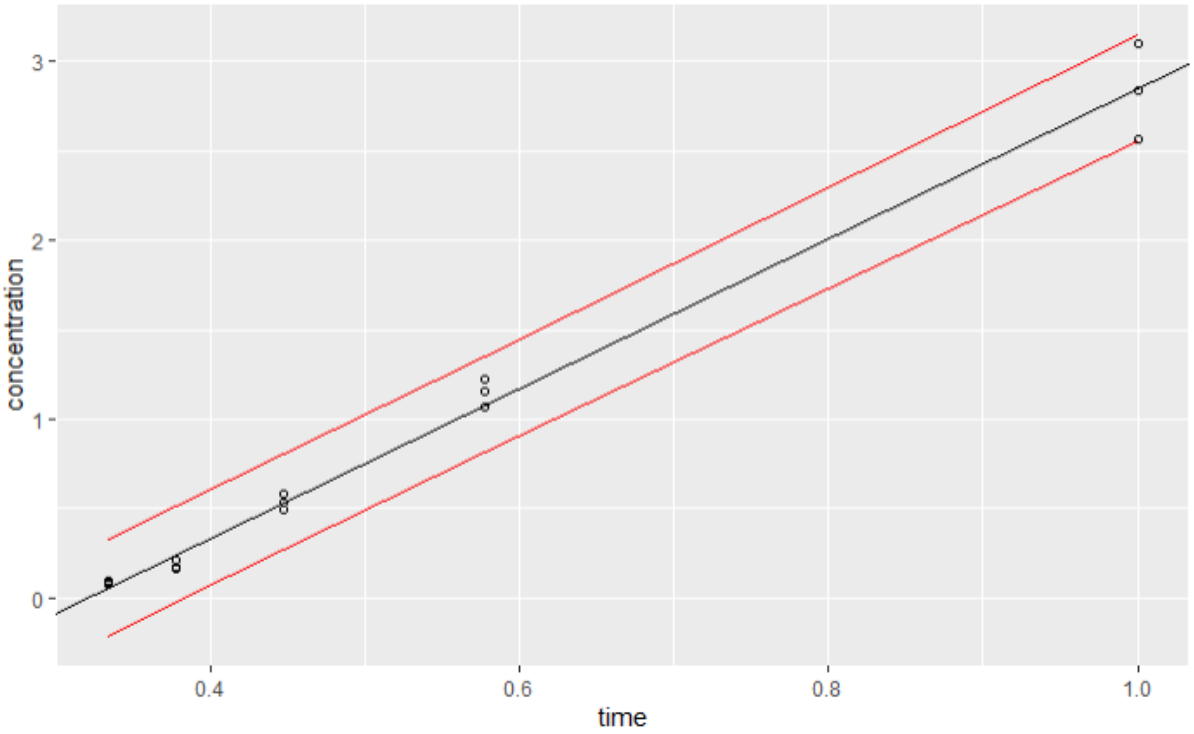
Statystyka testowa F wyniosła 1076, co jest znacząco większe od:

```
qt(1-0.05, 1, 13)
```

```
[1] 4.667193
```

możemy więc odrzucić hipotezę zerową o braku zależności liniowej pomiędzy czasem a logarytmem stężenia.

Poniższy wykres przedstawia 95-procentowe przedziały predykcyjne:



Zauważamy, że wszystkie obserwacje znajdują się w obrębie przedziałów.

Korelacja pomiędzy predykcjami, a prawdziwymi wartościami stężenia wyniosła:

```
[1] 0.9940136
```

Ostatecznie, najlepszym modelem okazał się być model z zadania 10. Wynika to z faktu, że zarówno korelacja pomiędzy predykcjami a prawdziwymi wartościami zmiennej objaśnianej, jak i wartość R^2 były dla tego modelu największe.