Zadanie 1. Używając `R` obliczamy wartość krytyczną, oznaczaną przez `tc` dla dwustronnego testu na poziomie istności 0.05 i 10 stopniach swobody.

```
tc <- qt(1 - .05/2, 10)
```

Wartość `tc` jest równa:

```
## [1] 2.228
```

Następnie obliczamy wartość krytyczną, oznaczaną przez `Fc` dla testu F na poziomie istności 0.05 o 1 i 10 stopniach swobody.

```
Fc <- qf(1 - .05, 1, 10)
```

Wartość `Tc` jest równa:

```
## [1] 4.965
```

Zauważmy, że wartość `tc` podniesiona do kwadratu jest równa

```
## [1] 4.965
```

a więc kwadrat `tc` jest równy `Fc`.

Zadanie 2. W tym zadaniu analizować będziemy następującą tabelę ANOVA:

	df	*SS*
Model	1	100
Error	20	400

Na początek wyznaczymy n, tj. liczbę obserwacji. Przypomnijmy, że $SSE = n - 2$, a więc $n = 22$. W dalszej części podamy wartość estymatora σ. Przypomnjmy, że obciążony estymator σ^2, ozn. $\hat{\sigma^2}$ dany jest wzorem $\frac{SSE}{n}$, a zatem $\hat{\sigma} \approx 4.264$. Nieobciążony estymator σ^2, ozn. s^2 dany jest wzorem $\frac{SSE}{n-2}$, a zatem $s \approx 4.472$.
Następnie, przeprowadzając test F na poziomie istotności 0.05, sprawdzimy czy β_1 jest różna od 0:

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

Obliczamy statystykę testową F:

$$F = \frac{MSM}{MSE} = \frac{SSM}{SSE} \cdot dfE = \frac{20}{4} = 5.$$

Korzystając z `R`, otrzymujemy, że `Tc <- qf(1-.05,1,20)` wynosi 4.351. Widzimy więc, że wartość F jest znacząco większa od `Tc`. Możemy więc odrzucić $H_0 : \beta_1 = 0$.
W kolejnym kroku, korzystając ze wzoru $R^2 = \frac{SSM}{SST}$, otrzymujemy, że współczynnik determinacji jest równy 0.2. Oznacza, to że jedynie 20% zmienności w wektorze odpowiedzi stanowi zmienność wyjaśniana przez model. Dzięki R^2 możemy wyznaczyć również, z dokładnością do znaku, wyznaczyć wartość współczynnika korelacji próbkowej pomiędzy zmiennymi zależną i niezależną. W tym przypadku jest ona równa $\pm\sqrt{0.2}$, zatem otrzymana zależność jest umiarkowana.

Przejdziemy teraz do analizy zbioru danych `table1_6.txt`, który zawiera średnią ocen (GPA), wynik testu IQ (IQ), płeć (gender) oraz wynik testu Piersa-Harrisa (PHCSCS) dla 78 uczniów klasy siódmej.

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

Zadanie 3. W tym zadaniu zajmować będziemy się badaniem zależności liniowej pomiędzy GPA, a IQ. Zaczniemy od zastosowania regresji liniowej prostej zakładając, że regresorem jest IQ, a zmienną objaśnianą GPA.

```r
model <- lm(GPA ~ IQ, data)
```

Korzystając z funkcji `summary` otrzymujemy następujące dane dla zbudowanego modelu:

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

Widzmy zatem, że dopasowana prosta regresji dana jest wzorem $Y = -3.557 + 0.101X$. Współczynnik determinacji wynosi 0.402, a zatem około 40% wariancji GPA jest wyjaśniona poprzez IQ. Następnie zauważmy, że dla testu

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

otrzymujemy statystykę testową F równą 51.01 o 1 i 76 stopniach swobody. Obliczając Tc:

```r
qf(1 - .05, 1, 76)
```

```
## [1] 3.96676
```

widzimy, że wartość F jest znacznie większa od Tc, co świadczy o istnieniu zależności liniowej pomiędzy GPA, a IQ. Oczywiście wynika to również natychmiastowo z faktu, że p-wartość powyższego testu jest bardzo mała - wynosi zaledwie $4.737 \cdot 10^{-10}$.

W kolejnym kroku, przy użyciu zbudowanego modelu, dokonamy predykcji GPA dla ucznia o IQ równym 100 na poziomie ufności 0.9.

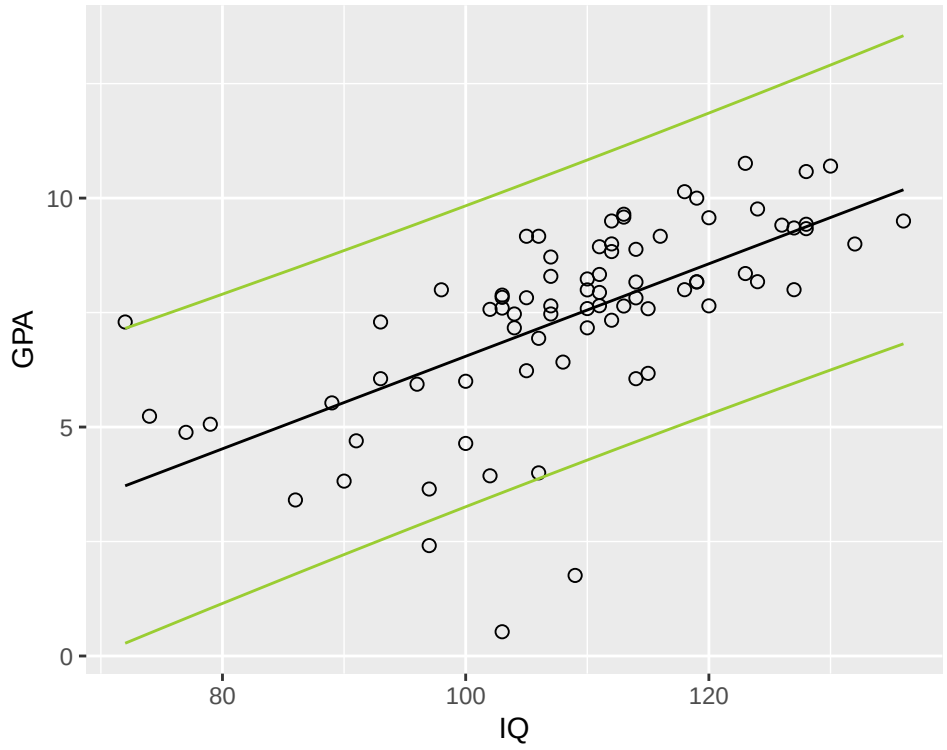
```r
new.data <- data.frame(IQ = 100)
predict(model, new.data, interval = "prediction", level = .9)
```

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

Otrzymujemy więc, że dla zmiennej niezależnej 100, zbudowany model dopasował wartość 6.545, natomiast z prawdopodobieństwem 0.9 prawdziwa wartość GPA dla IQ równego 100 znajduje się w przedziale [3.798, 9.293].

Zaznaczając 95-procentowe przedziały predykcyjne, dopasowaną prostą regresji oraz zbiór obserwacji na wykresie otrzymujemy:

```
plot1
```



Widzimy zatem, że dane pochodzące ze zbioru układają się w sposób "zbity" wzdłuż prostej. Ponadto jedynie 4 obserwacje (na 78) wypada poza uzyskane przez nas przedziały predykcyjne, co świadczy o dosyć dobrym dopasowaniu modelu.

Zadanie 4. W tym zadaniu powtórzymy rozumowanie z Zadania 3. przyjmując za regresor wynik testu Piersa-Harrisa (PHCSCS). Zaczniemy od zbudowania odpowiedniego modelu:

```
model <- lm(GPA ~ PHCSCS, data)
```

Korzystając z funkcji `summary` otrzymujemy następujące dane dla zbudowanego modelu:

```
##
## 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
```

Widzmy zatem, że dopasowana prosta regresji dana jest wzorem $Y = 2.226 + 0.917X$. Współczynnik determinacji wynosi 0.294, a zatem tylko około 30% zmienności GPA jest wyjaśniona poprzez PHCSCS. Zauważmy, że jest to słabszy rezultat, niż poprzednio, gdy za regresor braliśmy wartość IQ. Następnie zauważmy, że dla testu

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

otrzymujemy statystykę testową F równą 31.59 o 1 i 76 stopniach swobody. Obliczając Tc:

```
qf(1 - .05, 1, 76)

## [1] 3.96676
```

widzimy, że wartość F znów jest znacznie większa od Tc, co świadczy o istnieniu zależności liniowej pomiędzy GPA, a PHCSCS. Oczywiście wynika to również natychmiastowo z faktu, że p-wartość powyższego testu jest bardzo mała - wynosi zaledwie $3.006 \cdot 10^{-7}$.

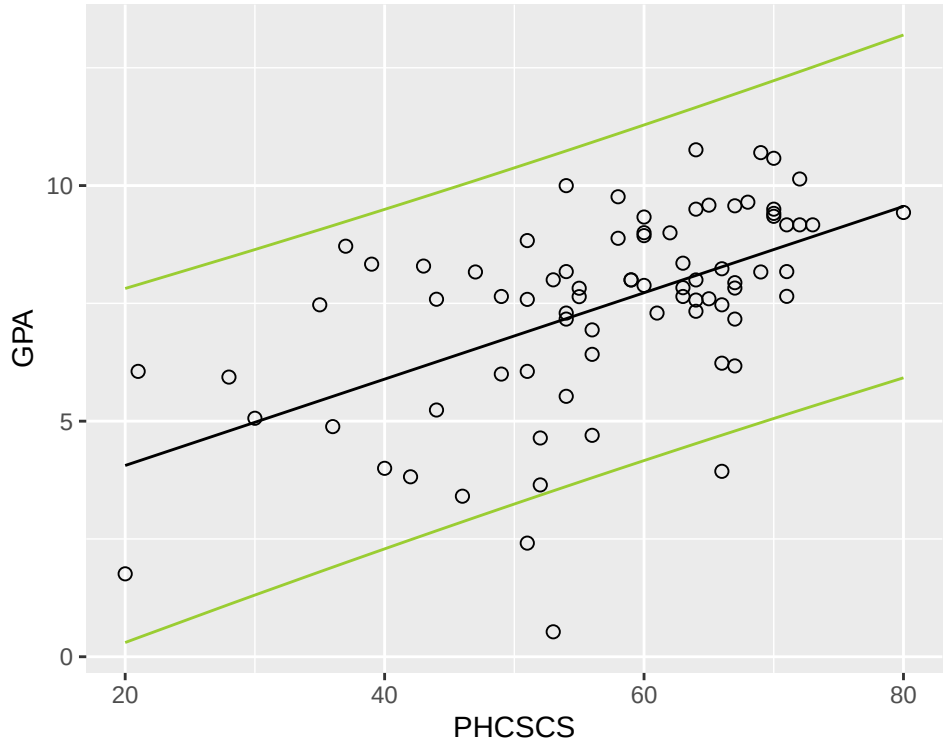
W kolejnym kroku, przy użyciu zbudowanego modelu, dokonamy predykcji GPA dla ucznia o PHCSCS równym 60 na poziomie ufności 0.9.

```
new.data <- data.frame(PHCSCS = 60)
predict(model, new.data, interval = "prediction", level = .9)

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

Otrzymujemy więc, że dla zmiennej niezależnej 60, zbudowany model dopasował wartość 7.725, natomiast z prawdopodobieństwem 0.9 prawdziwa wartość GPA dla PHCSCS równego 60 znajduje się w przedziale $[4.747,\ 10.703]$.

Zaznaczając 95-procentowe przedziały predykcyjne, dopasowaną prostą regresji oraz zbiór obserwacji na wykresie otrzymujemy:

```
plot46
```



Widzimy zatem, że w tym przypadku jedynie 3 obserwacje wypadają poza uzyskane przedziały predykcyjne. Nie świadczy to jednak o tym, że PHCSCS jest lepszym predyktorem dla GPA, niż IQ. Lepszym predyktorem dla GPA jest IQ ponieważ, po pierwsze, dla IQ wartość R^2 była większa niż dla PHCSCS. Ponadto wiemy, że im większa jest wartość statystyki F, tym silniejsza jest zależność liniowa między regresorem, a zmienną odpowiedzi. Zauważmy, że dla IQ wartość statystyki F była większa niż dla PHCSCS, czyli zależność liniowa między IQ, a GPA, jest silniejsza od zależności liniowej między PHCSCS, a GPA. Oczywiście wynika to również z faktu, że dla modelu $GPA \sim IQ$ p-wartość jest znacznie mniejsza od p-wartości dla modelu $GPA \sim PHCSCS$.
Na koniec zauważmy, że powyższe wnioski znajdują oparcie również w rozmieszczeniu obserwacji ze zbioru danych. Istotnie, zależność między PHCSCS, a GPA jest bardziej rozrzucona, niż w przypadku IQ, a GPA, które układały się w sposób zbliżony do liniowego.

Powrócimy teraz do analiz danych pochodzących z pliku `CH01PR20.txt` złożonych z dwóch kolumn: liczby kopiarek oraz czasu ich serwisowania (w godzinach).

```
data <- read.table("CH01PR20.txt", header = FALSE,
                   col.names = c("time", "copiers"))
```

Zakładając, że zmienną odpowiedzi jest czas serwisu, zaś regresorem - liczba kopiarek, budujemy następujący model regresji liniowej prostej

```
model <- lm(time ~ copiers, data)
```

Zadanie 5. W tym zadaniu, analizując odpowiednie wykresy, będziemy sprawdzać założenia regresji liniowej dla w.w. danych. Zaczniemy, jednak od sprawdzenia, że suma residuów jest równa zero.

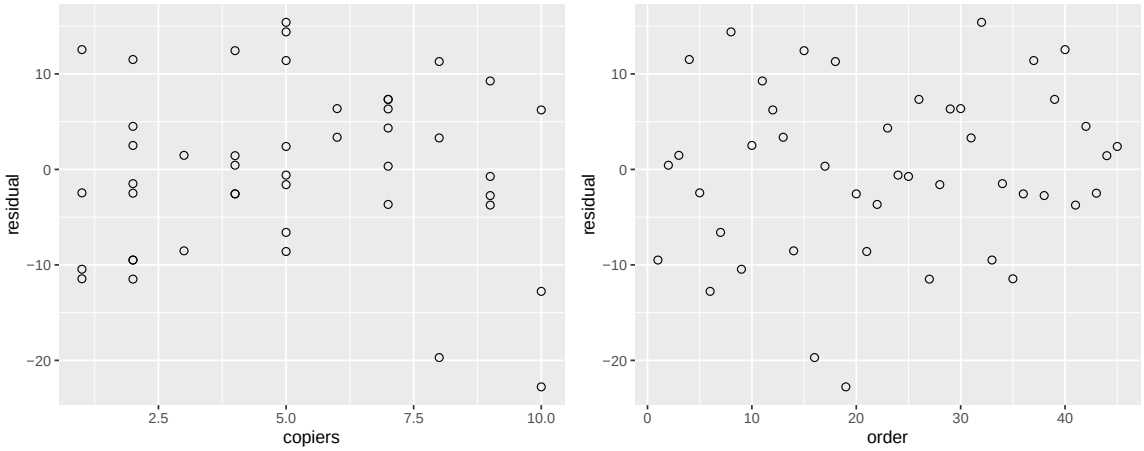
```
sum(model$residuals)

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

Zauważmy, że dla dowolnego modelu regresji liniowj prostej suma residuów wynosi 0, ponieważ, korzystając z faktu że $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ otrzymujemy:

$$\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)) = \sum_{i=1}^{n}(Y_i - \bar{Y} + \hat{\beta}_1(\bar{X} - X_i)) = \sum_{i=1}^{n} Y_i - n\bar{Y} + \hat{\beta}_1(n\bar{X} - \sum_{i=1}^{n} X_i) = 0.$$

Przejdziemy teraz do analizy wykresów residua vs. zmienne objaśniające oraz residua vs. kolenność pojawiania się obserwacji w zbiorze danych.

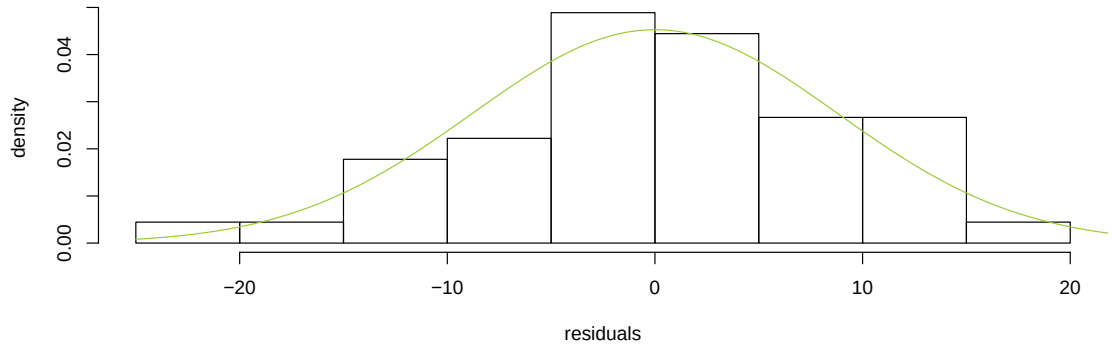
```
grid.arrange(plot3, plot4, ncol = 2)
```



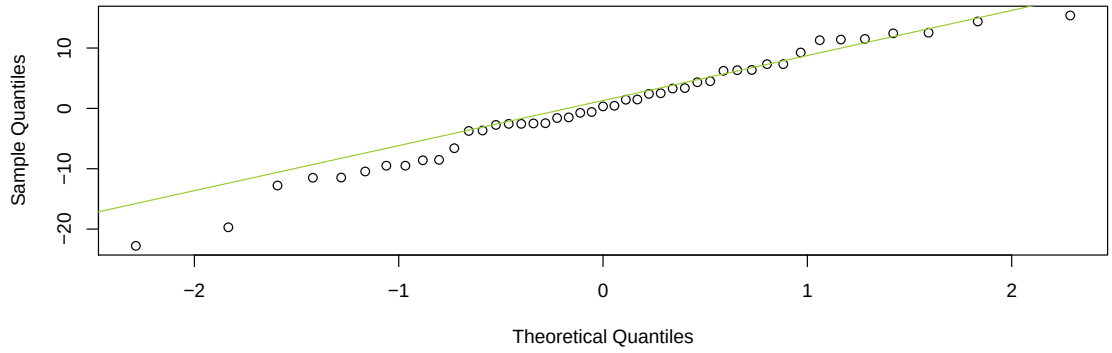
Zauważmy, że na wykresie residua vs. zmienne objaśniające pojawiają się obserwacje odstające. Pozbywając się ich ze zbioru danych moglibyśmy stwierdzić, że wahanie przy każdym regresorze jest w przybliżeniu takie samo. Otrzymujemy zatem, że, w przybliżeniu, wariancja składnika losowego jest taka sama dla wszystkich obserwacji, a zatem reszty mają jednorodną wariancję.
Na wykresie residua vs. kolejność pojawiania się obserwacji w zbiorze danych możemy zauważyć, że rozrzut residuów jest dosyć losowy, a zatem reszty mają strukturę losową.

Pozostaje sprawdzić, że residua mają rozkład normalny. W tym celu rozważmy histogram reszt wraz z krzywą rozkładu normalnego oraz wykres kwantylowo-kwantylowy.

```
hist(model$residuals, probability = TRUE, nclass = 7, xlab = "residuals", ylab = "density", main = "
lines(t, dnorm(t, 0, sd(model$residuals)), col = "yellowgreen")
```




```
qqnorm(model$residuals, main = "")
qqline(model$residuals, main = "", col = "yellowgreen")
```



Na podstawie powyższych wykresów możemy stwierdzić, że residua mają rozkład normalny. Podsumowując, otrzymaliśmy, że rozważane dane spełniają założenia regresji liniowej.

Zadanie 6. W tym zadaniu zmienimy wartość czasu serwisu dla pierwszej obserwacji z 20 na 2000. Na podstawie otrzymanych danych budujemy model regresji liniowej:

```
data[1, 1] <- 2000
model <- lm(time ~ copiers, data)
```

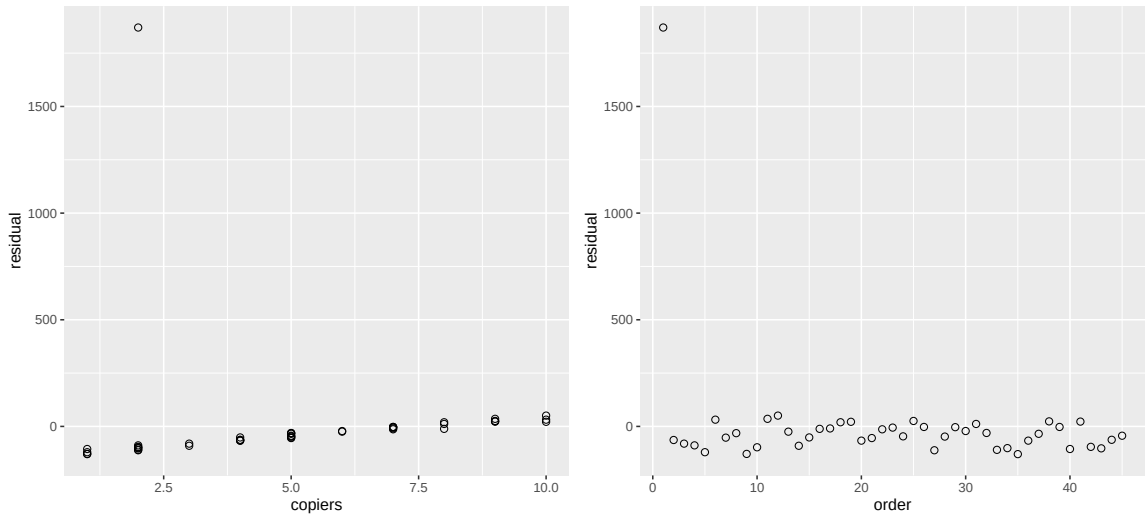
W poniższej tabeli dokonamy porównania modelu z Zadania 5. (ozn. 1. model) z powyżej otrzymanym modelem (ozn. 2. model).

	1. model	2. model
równanie regresji	$-0.58 + 15.035X$	$135.9 - 3.059X$
statystyka F	968.7	0.037
p-wartość	$< 2.2 \cdot 10^{-16}$	0.848
R^2	0.958	0.001
$\hat{\sigma}$	8.914	292.8

Widzimy więc, że w 1. modelu p-wartość jest bardzo mała, a zatem istnieje duża zależność liniowa między regresorem, a zmienną odpowiedzi. W 2. modelu p-wartość wynosi aż 0.848, co oznacza

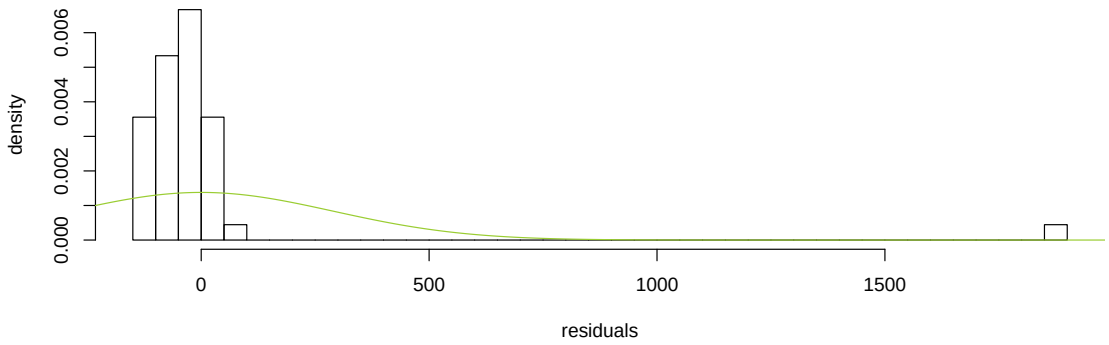
słabą zależność liniową między liczbą kopiarek, a czasem serwisu, równoważnie możemy przyjąć, że współczynnik β_1 wynosi 0. Dodatkowo w 1. modelu aż 96% zmienności w wektorze odpowiedzi jest wyjaśniona przez regresor, natomiast w 2. modelu R^2 wynosi zaledwie 0.001. Ponadto w modelu 1. odchylenie standardowe jest dosyć małe, zaś w modelu 2. osiąga aż 292. Na podstawie powyższych obserwacji możemy zatem wnioskować, że wprowadzenie obserwacji odstającej zaburza strukturę liniową w omawianych danych.
W dalszej części sprawdzimy, że w przypadku modelu 2. łamane są założenia regresji liniowej. Zaczniemy od analizy wykresów residua vs. zmienne objaśniające oraz residua vs. kolejność pojawiania się obserwacji w zbiorze danych.

```
grid.arrange(plot5, plot6, ncol = 2)
```

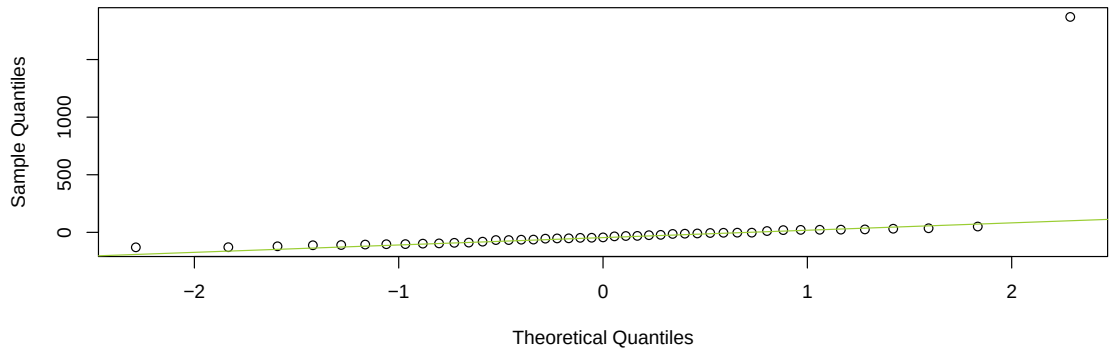


Możemy zauważyć, że obserwacja odstająca zniekształca 1. wykres. Dodatkowo na podstawie 1. wykresu widzimy, że wariancja nie będzie jednorodna. Na podstawie 2. wykresu widzimy, że reszty nie mają struktury losowej, możemy zauważyć pewien okresowy trend. Na koniec rozważymy histogram reszt wraz z krzywą rozkładu normalnego oraz wykres kwantylowo-kwantylowy.

```
hist(model$residuals, probability = TRUE, nclass = 30, xlab = "residuals", ylab = "density", main =
lines(t, dnorm(t, 0, sd(model$residuals)), col = "yellowgreen")
```



```
qqnorm(model$residuals, main = "")
qqline(model$residuals, main = "", col = "yellowgreen")
```



Jak widać reszty nie mają rozkładu normalnego, a więc łamane są założenia modelu liniowego. Podsumowując, obserwacja odstająca zaburza strukturę liniową omawianych danych.

W dalszej części omawiać będziemy zbiór danych `CH03PR15.txt`, który zawiera stężenie pewnego roztworu i czas.

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

Zadanie 7. Zaczniemy od zastosowania regresji liniowej prostej zakładając, że regresorem jest czas, a zmienną objaśnianą jest stężenie.

```
model <- lm(concentration ~ time, data)
```

Korzystając z funkcji `summary` otrzymujemy następujące dane dla powstałego modelu:

równanie regresji	$2.575 - 0.324X$
statystyka F	55.99
p-wartość	$4.611 \cdot 10^{-6}$
R^2	0.812

Widzimy więc, że aż 81% zmienności w stężeniu jest wyjaśniane poprzez czas. Dodatkowo wykonując następujący test:

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

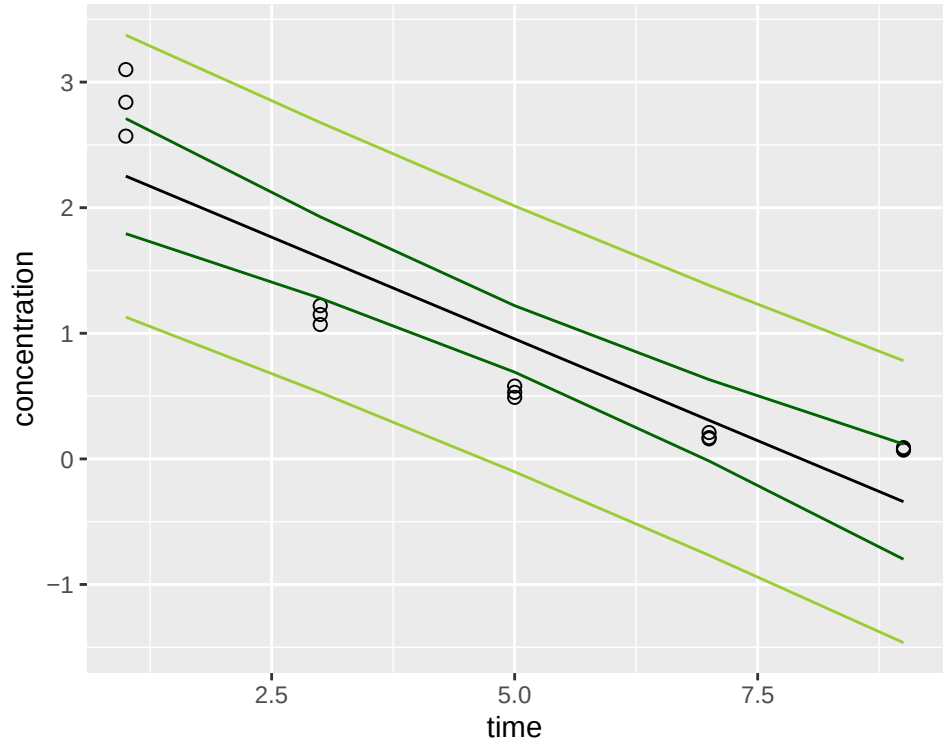
przy użyciu statystyki testowej F, która ma wartość 55.99, o stopniach swobody 1 i 13, otrzymujemy, że F jest znacząco większa od `Tc <- qf(1-.05,1,13)`, który wynosi

```
## [1] 4.667
```

a zatem możemy odrzucić H_0, otrzymując tym samym, że stężenie jest liniowo zależne od czasu. Oczywiście wynika to również z faktu, że p-wartość, tj. $4.611 \cdot 10^{-6}$ jest bardzo mała.

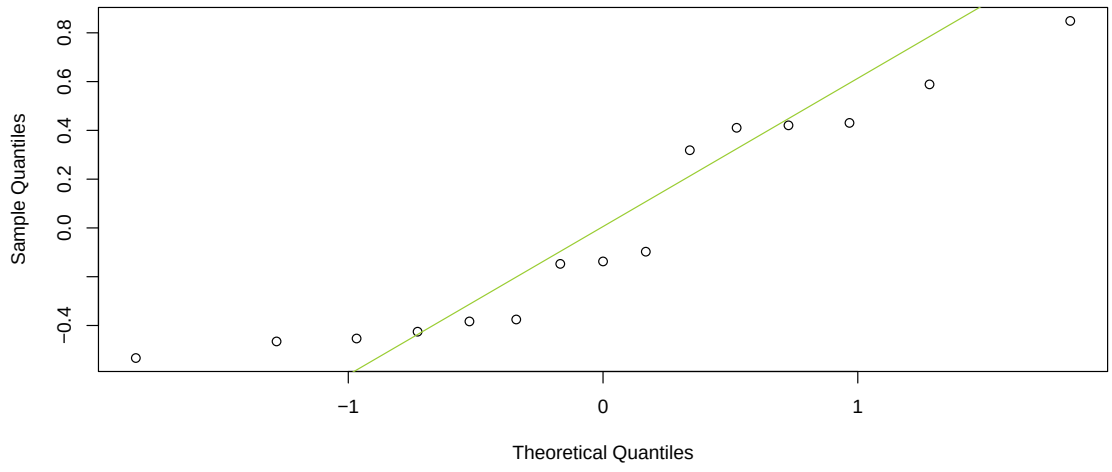
Zadanie 8. Korzystając z danych z poprzedniego zadania, przedstawimy na wykresie 95-procentowy przedziały predykcyjne i przedziały ufności, dopasowaną prostą regresji oraz dane ze zbioru obserwacji.

```
plot7
```



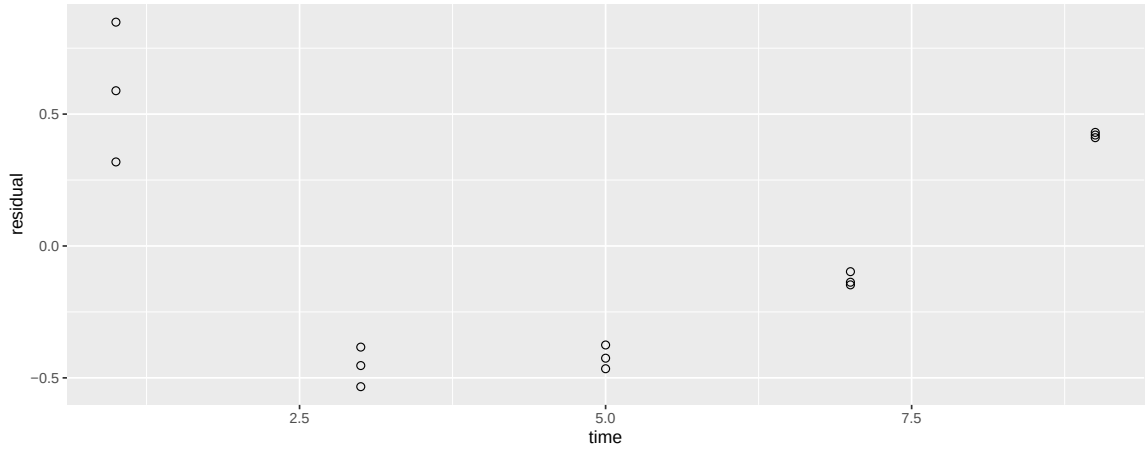
Widzimy więc, że w przedziale predykcji mieszczą się wszystkie obserwacje, natomiast większość obserwacji wypada z przedziału ufności, a zatem możemy podejrzewać, że łamane są założenia modelu liniowego. Istotnie, na podstawie poniższego wykresu kwantylowo-kwantylowego dla residuów widzimy, że nie mają one rozkładu normalnego.

```
qqnorm(model$residuals, main = "")
qqline(model$residuals, main = "", col = "yellowgreen")
```



W szczególności na podstawie wykresu residua vs. zmienna objaśniająca możemy zaobserwować, że w tym przypadku mamy do czynienia z niejednorodnością wariancji.

```
plot20
```

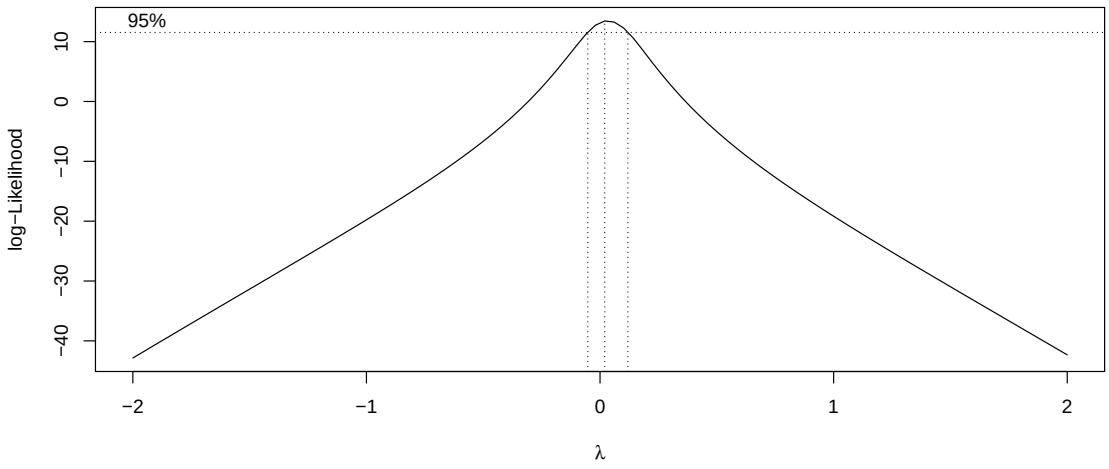


Widzimy więc, że regresja liniowa nie jest dobrze dopasowana. Następnie obliczmy korelację pomiędzy daną objaśnianą i jej predykcją, która naturalnie wskazuje na mocną zależność między nimi.

```
## [1] 0.901
```

Zadanie 9. Przy użyciu metody Boxa-Coxa znajdziemy odpowiednie przekształcenie zmiennej objaśnianej, tak, aby otrzymać zależność liniową z regresorem. Otrzymujemy następującą krzywą wiarygodności:

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



Łatwo zauważyć, że jest ona maksymalizowana dla $\lambda = 0$, zatem odpowiednie przekształcenie Y będzie dane przez $\log(Y)$, to znaczy musimy przekształcić rozważane dane logarytmicznie.

Zadanie 10. Przekształcając rozważane dane logarytmicznie otrzymujemy nowy model

```
new.data <- data
new.data$concentration <- log(new.data$concentration)
model.log <- lm(concentration ~ time, new.data)
```

Korzystając z funkcji `summary` otrzymujemy następujące dane dla powstałego modelu:

równanie regresji	$1.508 - 0.45X$
statystyka F	1838
p-wartość	$2.118 \cdot 10^{-15}$
R^2	0.993

Widzimy więc, że prawie 100% zmienności w stężeniu jest wyjaśniane poprzez czas, co świadczy o idealnym wręcz dopasowaniu modelu do danych. W szczególności R^2 jest tutaj większe, niż w przypadku poprzednio rozważanego modelu. Dodatkowo wykonując test:

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

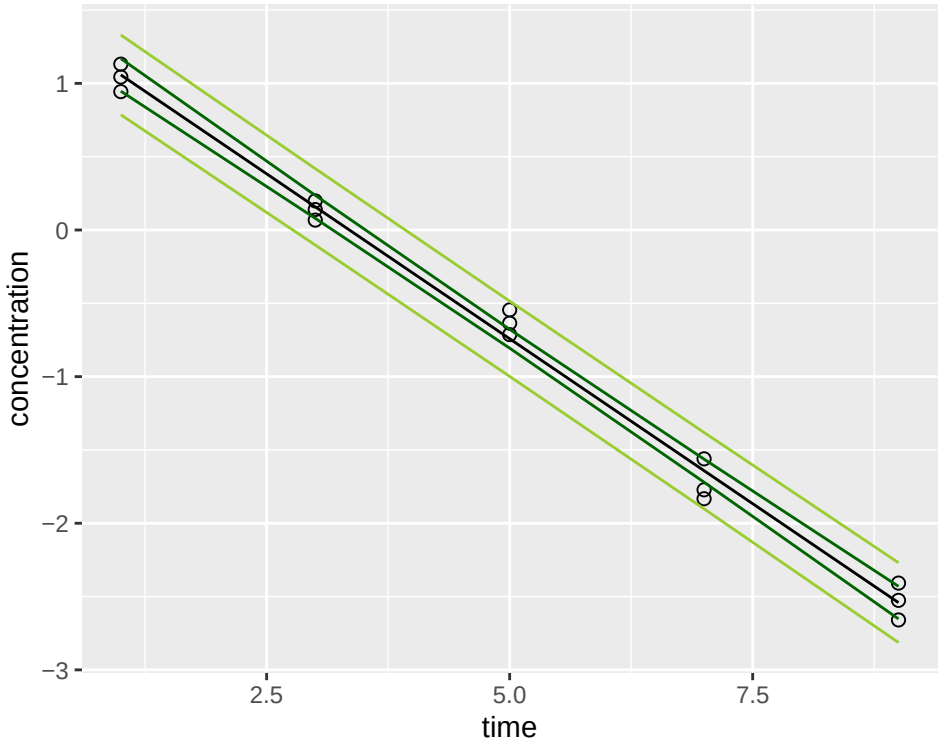
przy użyciu statystyki testowej F, która ma wartość 1838, o stopniach swobody 1 i 13, otrzymujemy, że F jest znacząco większa od `Tc <- qf(1-.05,1,13)`, który wynosi

```
## [1] 4.667
```

a zatem możemy odrzucić H_0, otrzymując tym samym, że stężenie jest liniowo zależne od czasu. Zauważmy, że otrzymana wartość F jest tutaj znacznie większa od wartości F dla poprzedniego modelu, a zatem mamy tutaj do czynienia z silniejszą zależnością liniową. Oczywiście wynika to również z faktu, że p-wartość, tj. $4.611 \cdot 10^{-6}$ jest w tym przypadku znacznie mniejsza.

Zbadamy teraz dopasowanie modelu do obserwacji na podstawie przedziałów ufności i predykcji, co przedstawione jest na poniższym wykresie:

```
plot8
```



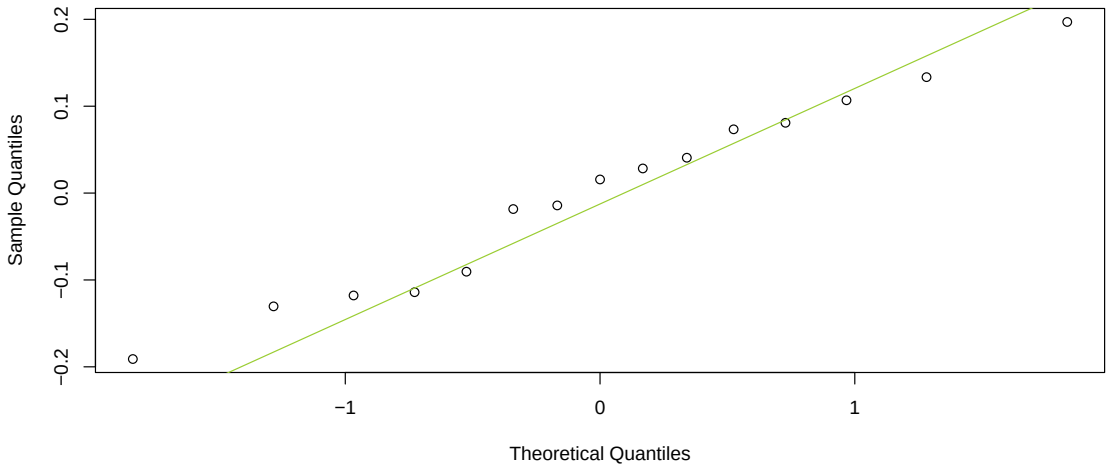
Widzimy, że przedziały ufności i predykcji są w tym przypadku są dużo węższe, a zatem dzięki transformacji Boxa-Coxa zyskaliśmy większą dokładność predykcji. Dodatkowo widzimy, że wzrosła liczba obserwacji wpadających do przedziału ufności. Podsumowując, powstały model jest znacznie lepiej dopasowany do omawianych danych, niż oryginalny model.

Następnie obliczamy korelację pomiędzy daną objaśnianą i jej predykcją, która naturalnie wskazuje na mocną zależność między nimi. W tym przypadku jest ona bardzo bliska 1, czyli jest mocniejsza niż w Zadaniu 7.

```
## [1] 0.996
```

Na koniec, dla porównania z poprzednim rezultatem, prezentujemy wykres kwantylowo-kwantylowy dla residuów z nowego modelu.

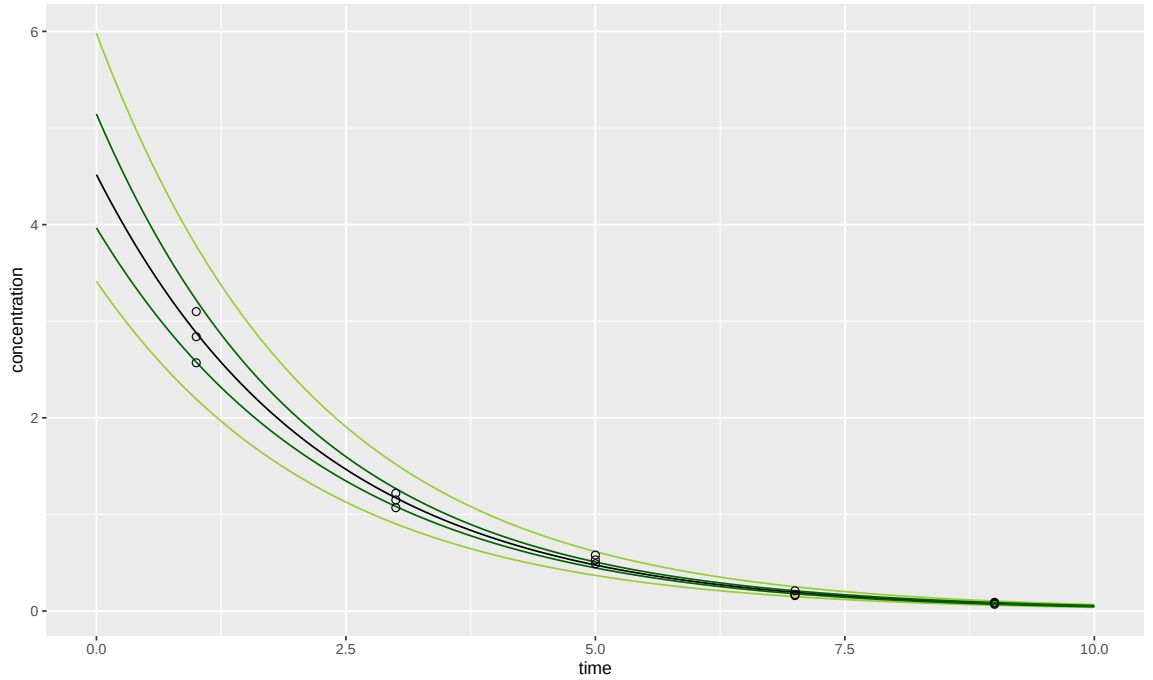
```
qqnorm(model.log$residuals, main = "")
qqline(model.log$residuals, main = "", col = "yellowgreen")
```



Widzimy tutaj, że dopasowanie prostej do reszt jest znacznie lepsze, a więc rozkład residuów jest bliski rozkładowi normalnemu. W szczególności możemy wnioskować, że podstawowe założenia modelu regresji liniowej są w tym przypadku spełnione.

Zadanie 11. W tym zadaniu będziemy dopasowywać uzyskaną w poprzednim zadaniu prostą do oryginanych danych. Przekształcając wykładniczo prostą regresji $\log(Y) = 1.508 - 0.45X$ otrzymujemy $Y = e^{1.508-0.45X}$. Dodatkowo, przekształcając w analogiczny sposób uzyskane w poprzednim zadaniu przedziały ufności i predykcji otrzymujemy następujący wykres:

```
plot12 + geom_point(data = data, aes(time, concentration), shape = 1, size = 2)
```



Możemy zaobserwować, że w przedziale ufności leżą niemal wszystkie obserwacje, co jest zgodne z

wartością poziomu ufności, którą przyjęliśmy, tj. 0.95.

Następnie obliczamy korelację pomiędzy daną objaśnianą i jej predykcją, która naturalnie wskazuje na mocną zależność między nimi. W tym przypadku jest ona bardzo bliska 1, czyli jest mocniejsza niż w Zadaniu 7.

```
## [1] 0.995
```

Zauważmy, że korelacja w Zadaniu 7. wynosiła 0.901, a więc zastosowana powyżej transformacja regresora wpływa pozytywnie na dopasowanie modelu do danych.

Zadanie 12. W tym zadaniu stworzymy nową zmienną objaśniającą $t1 = \text{czas}^{-1/2}$ i powtórzymy dla niej rozumowanie z dwóch poprzednich zadań.

```
new.data <- data
new.data$time <- new.data$time^-.5
model.t1 <- lm(concentration ~ time, new.data)
```

Korzystając z funkcji summary otrzymujemy następujące dane dla powstałego modelu:

równanie regresji	$-1.341 + 4.196X$
statystyka F	1076
p-wartość	$6.898 \cdot 10^{-14}$
R^2	0.988

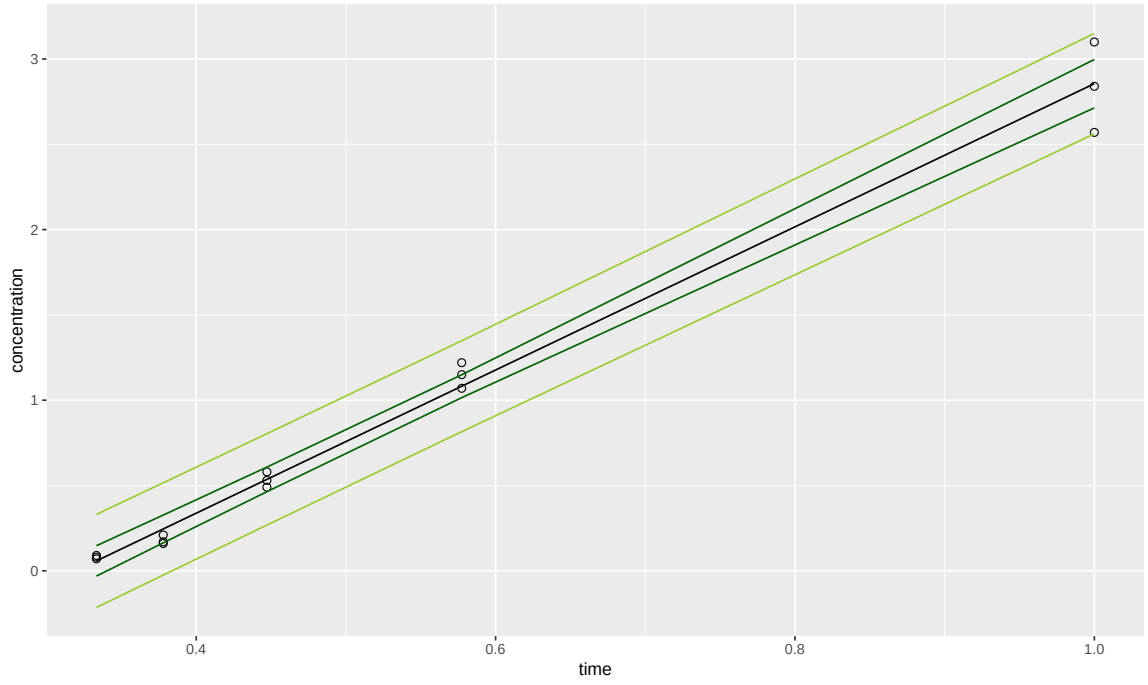
Na ich podstawie możemy zauważyć, że współczynnik determinacji jest większy niż w przypadku modelu z Zadania 7., ale mniejszy niż dla modelu z Zadania 10. Podobnie wartość statystyki testowej F jest większa od wartości F z Zadania 7. oraz mniejsza od wartości F z Zadania 10. W szczególności oznacza to, że wykonując test:

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

możemy odrzucić H_0. Dodatkowo widzimy, że zależność liniowa jest w tym przypadku słabsza od tej z Zadania 10., ale silniejsza od tej w Zadaniu 7.

Zbadamy teraz dopasowanie modelu do obserwacji na podstawie przedziałów ufności i predykcji co przedstawione jest na poniższym wykresie:

```
plot13
```



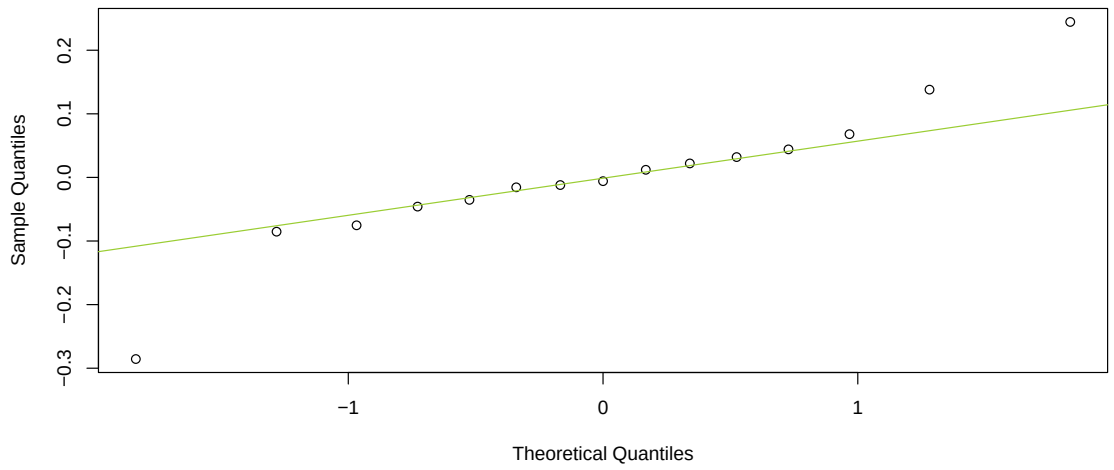
Możemy zauważyć, że jedynie trzy obserwacje wypadają z przedziału ufności. Oznacza to, że 0.8 wszystkich obserwacji leży w przedziale ufności. Jest to mniejsza proporcja niż teoretycznie zakładane przy konstrukcji przedziału 0.95. Przypomnijmy jednak, że dla modelu z Zadania 7. proporcja ta była znacznie mniejsza - wynosiła jedynie ok. 47% (poza przedziałem znajdowało się 8 obserwacji). Z drugiej strony w Zadaniu 10. otrzymana proporcja wynosiła 0.93, a zatem była większa od 0.8. W szczególności była bliższa teoretycznej wartości wynikającej z konstrukcji przedziału tj. 0.95.

Następnie, obliczając korelację pomiędzy daną objaśnianą i jej predykcją, otrzymujemy:

```
## [1] 0.994
```

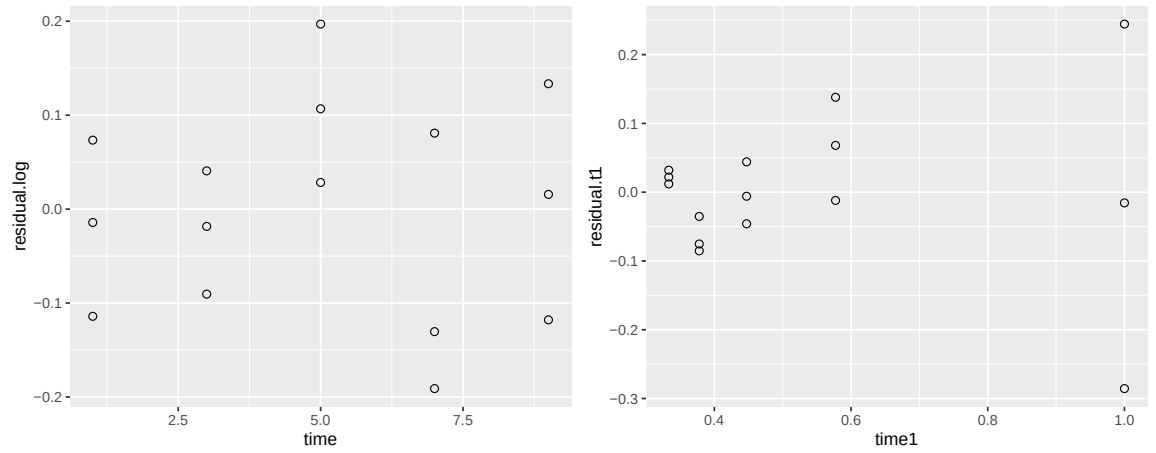
Zgodnie z przewidywaniami, obserwujemy, że obliczona korelacja jest większa niż w Zadaniu 7. i mniejsza niż w Zadaniu 10.
Na koniec, dla porównania z poprzednim rezultatem, prezentujemy wykres kwantylowo-kwantylowy dla residuów z nowego modelu.

```
qqnorm(model.t1$residuals, main = "")
qqline(model.t1$residuals, main = "", col = "yellowgreen")
```



Widzimy, że pewne obserwacje znacząco odchodzą od prostej dopasowania. dopasowanie prostej do reszt jest znacznie lepsze. W szczególności dopasowanie do prostej jest słabsze niż w przypadku Zadania 10. Istotnie, jeśli rozważymy wykresy residua vs. zmienne objaśniające kolejno dla modeli z Zadania 10. oraz 12.:

```
grid.arrange(plot3, plot4, ncol = 2)
```



możemy zauważyć, że w przypadku omawianego w tym zadaniu modelu wariancja nie jest jednorodna, co oznacza łamanie założeń modelu, natomiast dla modelu z Zadania 10. wariancja jest jednorodna.

Podsumowując: na podstawie uzyskanych obserwacji i wniosków możemy stwierdzić, że najlepiej dopasowany do danych jest model z Zadania 10.