Raport 3

Magdalena Potok

2023-12-11

Zadanie 1

Do zadania zostały użyte dane z pliku tabela1_6.txt, które zawierają średnią ocen (GPA), wynik w standardowym teście IQ, płeć oraz punktacje na teście psychologicznym Piers-Harris Children's Self -Concept Scale.

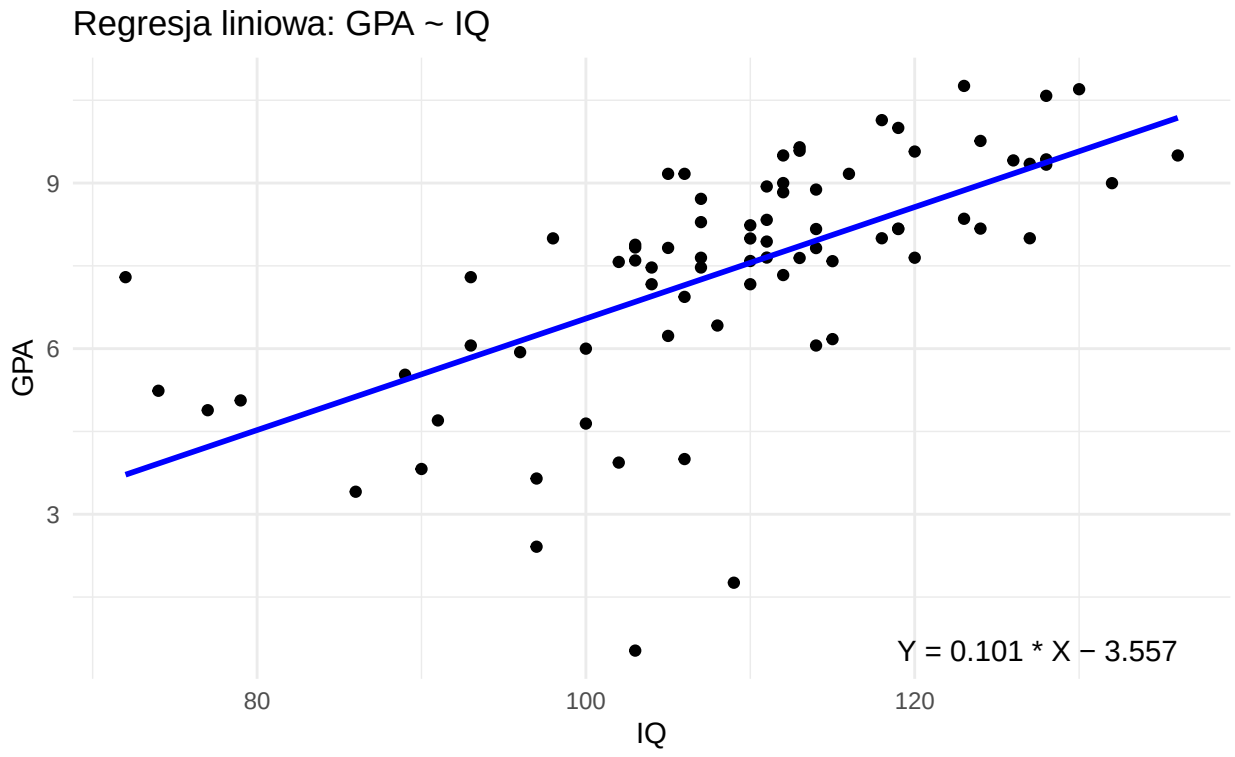
(a)
Użyję prostego modelu regresji do opisania zależności GPA od wyników testu IQ.

Z podsumowania modelu można odczytać, że:

- $\hat{\beta}_1 = 0.101$
- $\hat{\beta}_0$ = -3.557

Zatem równanie regresji wygląda tak: $Y = 0.101 * X - 3.557$.

Dane oraz prosta regresji na wykresie wyglądają tak:



Wykres przedstawia zależność GPA od wyników testu IQ. Na dane została nałożona prosta, która jest dopasowaną do danych prostą regresji.

Policzę teraz współczynnik determinacji R^2, aby zrobić to za pomocą wzorów teoretycznych należy policzyć:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

```
(R_squared <- 1 - sum((t$GPA - Beta0 - Beta1 * t$IQ)^2) / sum((t$GPA - mean(t$GPA))^2))
```

```
## [1] 0.4016146
```

Można również policzyć tę statystykę za pomocą poleceń wbudowanych w R.

```
summary(model1)$r.squared
```

```
## [1] 0.4016146
```

Na podstawie R^2 można powiedzieć, że 40% GPA jest wyjaśniane przez IQ.

(b)
Przetestuję hiipotezę, że GPA nie jest skorelowane z IQ na podstawie testu F.

$$H_0 : \beta_1 = 0 \; H_1 : \beta_1 \neq 0$$

Statystyka testowa dla testu F wyliczamy za pomocą wzorów teoretycznych w następujący sposób:

$$F = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \cdot dfE}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \cdot dfM}$$

Jest to statystyka z rozkładu Fishera dla 1 i $n-2$ stopni swobody, czyli w naszym przypadku 1 i 76. Odrzucamy hipotezę zerową wtedy, gdy $F > F_c = F^*(1-\alpha, 1, n-2)$. Przeprowadzę teraz ten test.

```
Fc <- qf(1-0.05, 1, n-2)
F <- sum((Beta0 + Beta1 * t$IQ - mean(t$GPA))^2)*(n-2)/sum((t$GPA - Beta0 - Beta1 * t$IQ)^2)
F > Fc
```

```
## [1] TRUE
```

W związku z czym na poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę zerową, co oznacza, że występuje zależność między GPA i IQ.

Ten sam problem możeby zbadać za pomocą funkcji wbudowanej w R.

```
summary(model1)$fstatistic[1] > Fc
```

```
## value
##  TRUE
```

I otrzymujemy ten sam wynik.

Innym sposobem na zbadanie, czy GPA jest skorelowane z IQ jest użycie p-wartości. W przypadku testu F p-wartość jest prawdopodobieństwm uzyskania wartości bardziej ekstremalnych od statystyki F, gdyby hipoteza zerowa była prawdziwa.

$$p - wartość = P(F > F_{statystyka})$$

```
(p_value <- 1 - pf(F, 1, 76))
```

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

Odrzucamy H_0 gdy $p - value < \alpha$, u nas p-wartość wyszła bardzo mała, bliska zera, zatem jest mniejsza od poziomu istotności $\alpha = 0.05$, zatem odrzucamy H_0.
P-wartość można również wyliczyć z polecen wbudowanych w R.

```
summary(model1)$coefficients["IQ", "Pr(>|t|)"]
```

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

Wynik wychodzi taki sam, jak dla obliczeń teoretycznych.

(c)
Przewidzę GPA dla uczniów, których IQ wynosi 75, 100, 140. Podam 90% przedziały predykcyjne.

```
iq_n <- c(75, 100, 140)
(gpa_s <- Beta0 + Beta1 * iq_n)
```

```
## [1]  4.019572  6.545114 10.585982
```

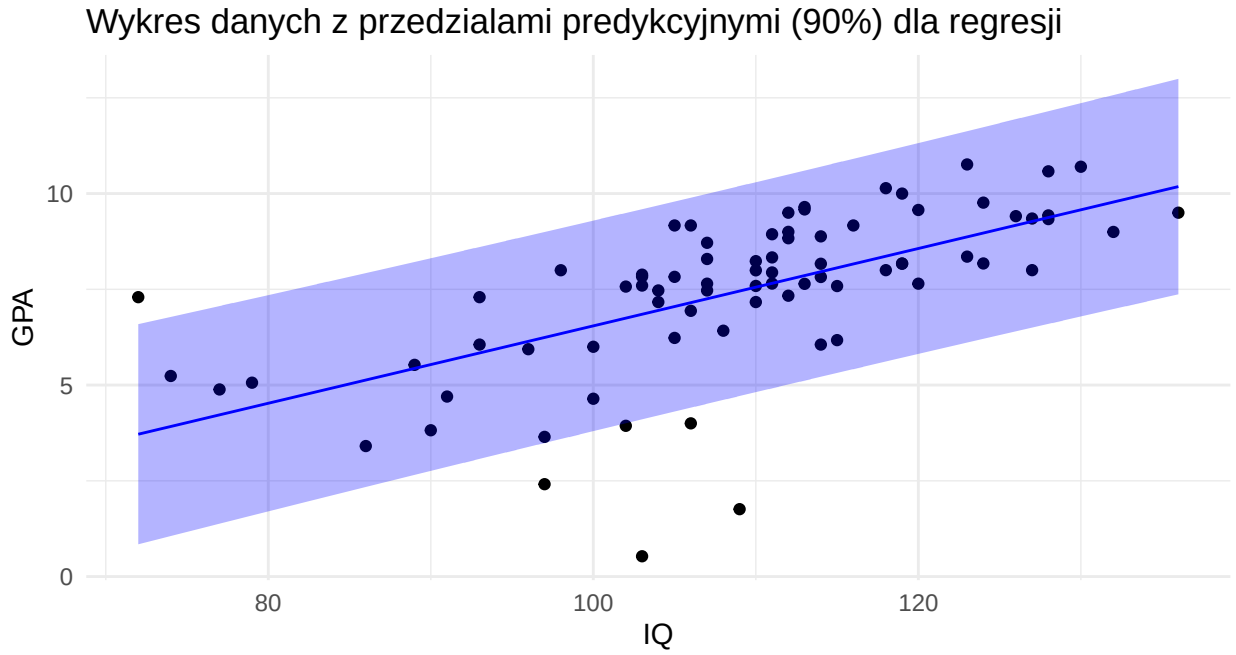
Tyle wynoszą przewidywane wartości średnich ocen, następnie wyznaczę przedziały predykcyjne za pomocą wzoru

$$[\hat{\mu}_h - t_c s(pred),\ \hat{\mu}_h + t_c s(pred)]$$

gdzie $s^2(pred) = s^2\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$.

IQ	Przewidywane GPA	Lewy Przedział	Prawy Przedział
75	4.020	0.606	7.433
100	6.545	3.259	9.831
140	10.586	7.194	13.978

(d)
Do wykresu z podpunktu **(a)** dodam 90% przedziały predykcyjne.



Z wykresu można odczytać, że z 78 obserwacji jedynie 6 obserwacji wychodzi poza obszar 90% pasma predykcji. Przedziały predykcyjne są dosyć szerokie, wynika to z dużego rozproszenia punktów na wykresie.

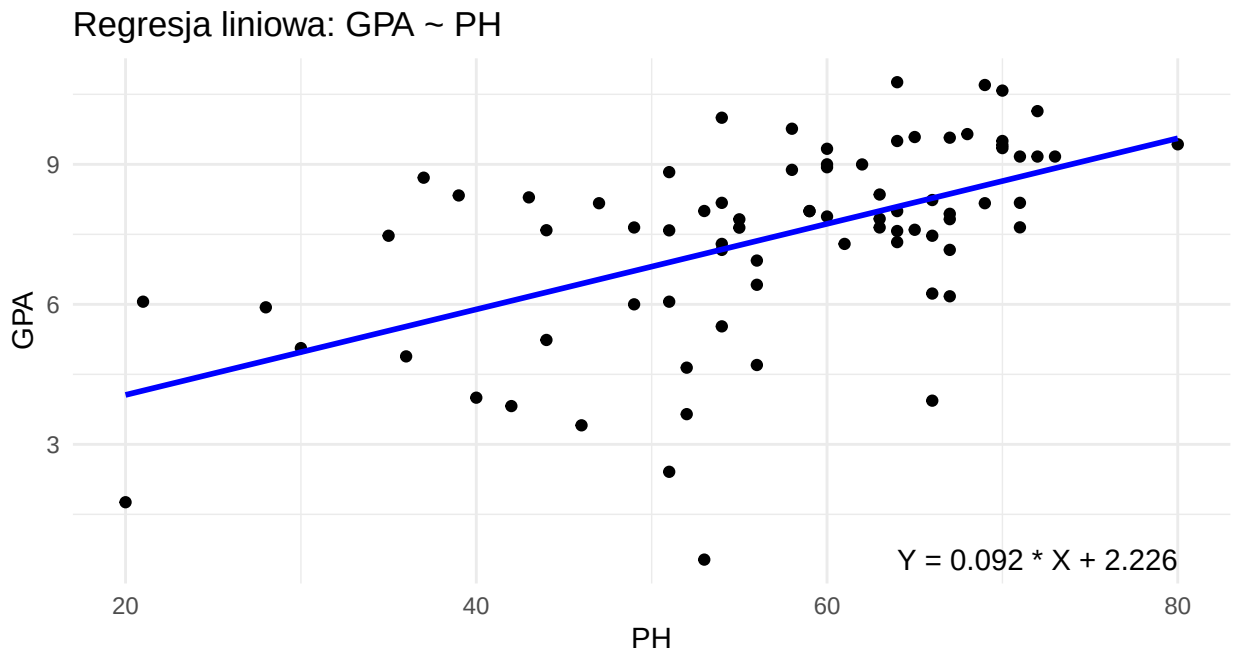
Zadanie 2

W zadaniu 2 zostały użyte te same dane, co w zadaniu 1.

(a) Użyję prostego modelu regresji do opisania zależności GPA od wyników testu PH.

Z podsumowania modelu można odczytać, że:

- $\hat{\beta}_1 = 0.092$
- $\hat{\beta}_0 = 2.226$ Zatem równanie regresji wygląda tak: $Y = 0.092 * X + 2.226$.
  Prosta regresji oraz dane można przedstawić na wykresie



Współczynnik determinacji wynosi

```
(R_squared <- 1 - sum((t$GPA - round(summ2$coefficients[1,1], 3) - round(summ2$coefficients[2,1], 3) *
```

```
## [1] 0.2934875
```

Można również policzyć tę statystykę za pomocą poleceń wbudowanych w R.

```
summary(model2)$r.squared
```

```
## [1] 0.2935829
```

Na podstawie R^2 można powiedzieć, że 29% GPA jest wyjaśniane przez PH.

(b) Przetestuję hipotezę, że GPA nie jest skorelowane z PH na podstawie testu F.

$$H_0 : \beta_1 = 0 \; H_1 : \beta_1 \neq 0$$

Do przetestowania hipotezy, tak jak w zadaniu 1., potrzebujemy statystyki F z rozkładu Fishera dla 1 i $n-2$ stopni swobody, czyli w naszym przypadku 1 i 76. Odrzucamy hipotezę zerową wtedy, gdy $F > F_c = F^*(1-\alpha, 1, n-2)$. Przeprowadzę teraz ten test.

```
summary(model2)$fstatistic[1] > Fc
```

```
## value
##  TRUE
```

Statystyka F wyszła około 31.59, a wartość krytyczna 3.97, zatem na poziomie istotnosci $\alpha = 0.05$ odrzucamy hipotezę zerową i są skorelowane.
Policzę teraz p-wartość.

```
summary(model2)$coefficients["PH", "Pr(>|t|)"]
```

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

Ta wartość jest bardzo mała, zatem na pewno mniejsza od 0.05, ten test również uzasadnił, że GPA i PH są skorelowane.

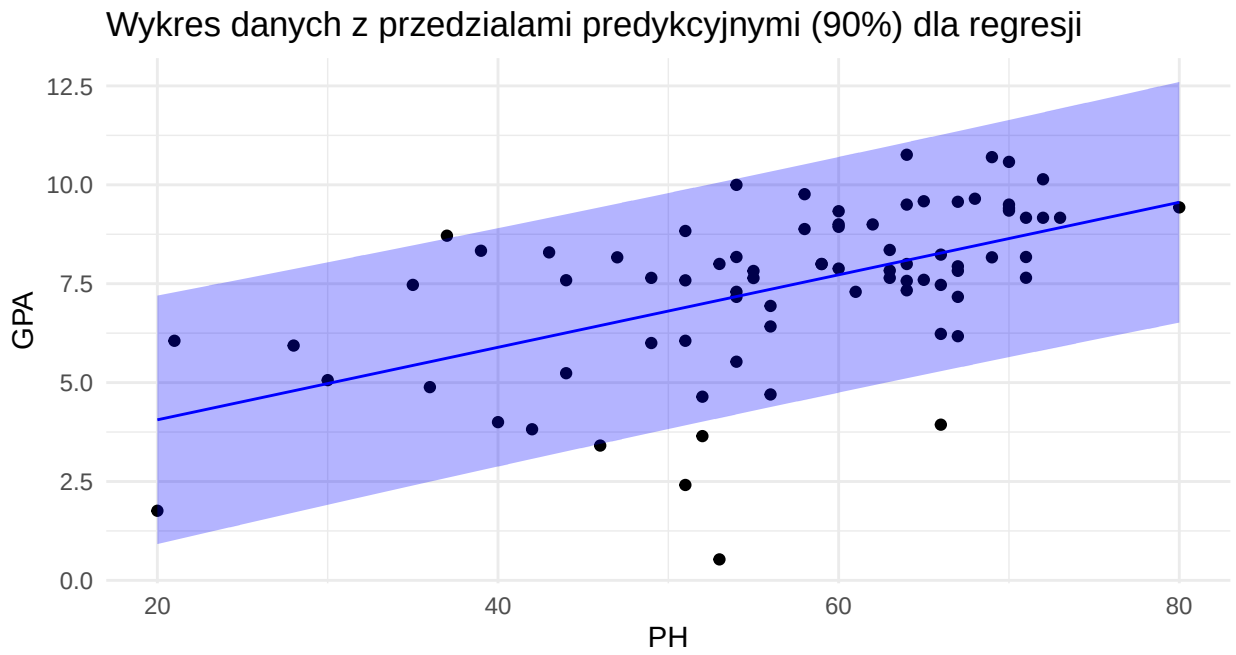
(c) Przewidzę GPA dla uczniów, których PH wynosi 25, 55, 85. Podam 90% przedziały predykcyjne.

```
## [1]  4.526  7.286 10.046
```

Tyle wynoszą przewidywane wartości średnich ocen, następnie wyznaczę przedziały predykcyjne.

PH	Przewidywane GPA	Lewy Przedział	Prawy Przedział
25	4.526	0.817	8.235
55	7.286	3.725	10.847
85	10.046	6.371	13.721

(d)
Do wykresu z podpunktu **(a)** dodam 90% przedziały predykcyjne.



Poza przedziałami predykcyjnymi znajduje się 6 obserwacji, tak samo, jak w przypadku zadania 1. Również przedział predykcyjne są całkiem szerokie, z tego samego powodu, co w zadaniu 1.

(e)
Współczynnik determinacji R^2 mierzy stopień, w jakim zmienność zmiennej zależnej (w naszym przypadku GPA) jest wyjaśniana przez zmienne niezależne (IQ lub PH) w modelu regresji. Im bliżej R^2 do 1, tym lepiej model pasuje do danych i lepiej wyjaśnia zmienność zmiennej objaśnianej. W naszym przypadku dla IQ $R^2 = 0.4$, natomiast dla PH $R^2 = 0.29$. Zatem IQ ma większy wpływ na predykcję wartości GPA i jest lepszym predyktorem.

Zadanie 3

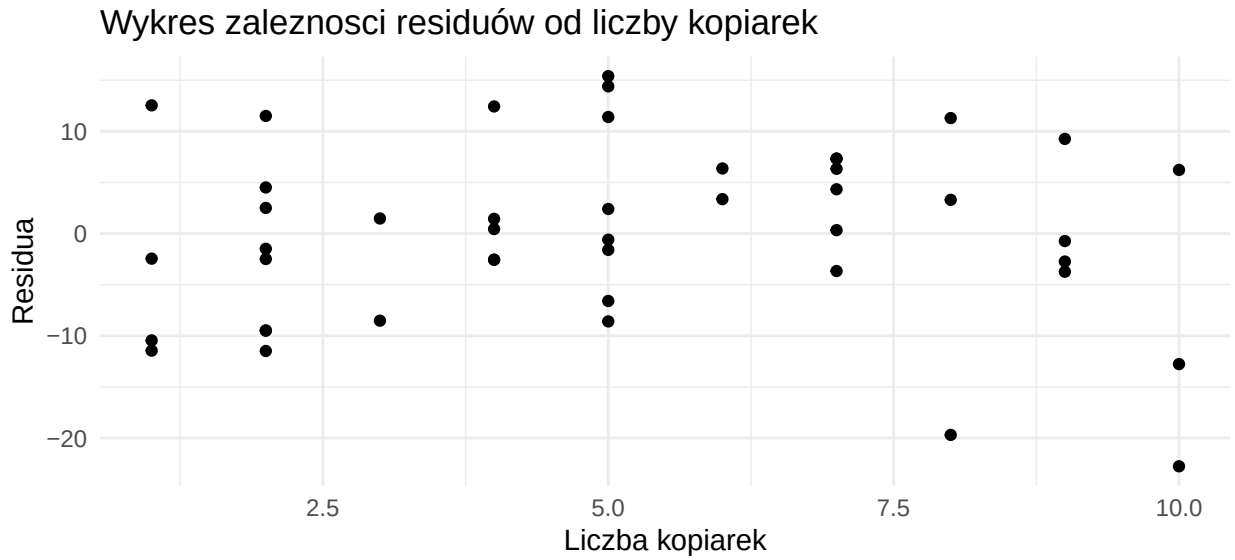
W tym zadaniu zostaną wykorzystane dane z pliku **ch01pr20.txt**, który zawiera liczbę kopiarek oraz czas (w godzinach) potrzebny na utrzymanie tych kopiarek.

(a) Sprawdzę ile wynosi suma residuów.

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

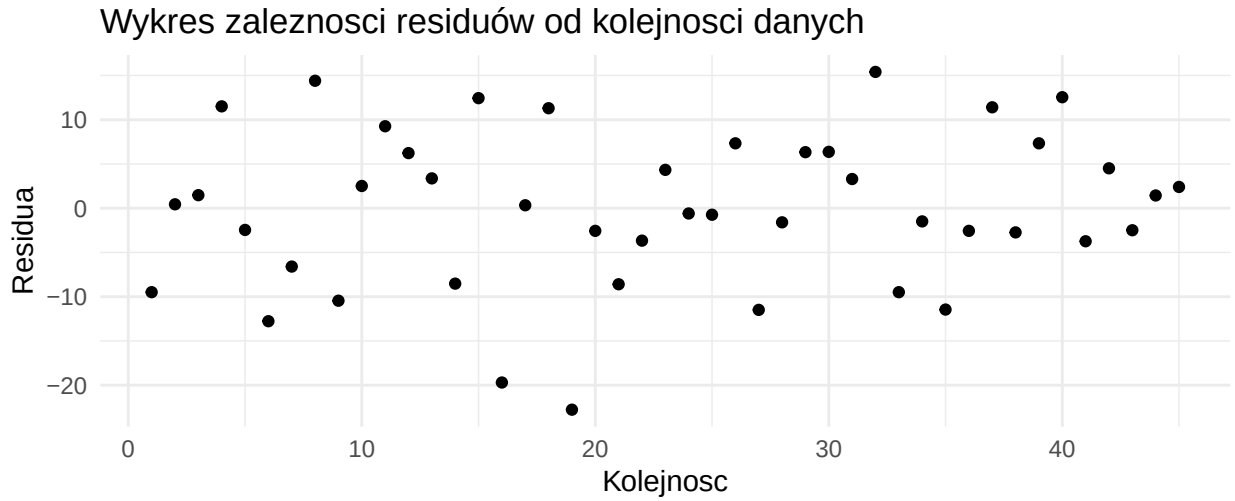
Suma reszt w modelu liniowym wynosi $-1.1768364 \times 10^{-14} \approx 0$, co jest wynikiem bardzo bliskim i przybliżanym do 0.

(b) Przeddstawię wykres residuów względem zmiennej objaśniającej.



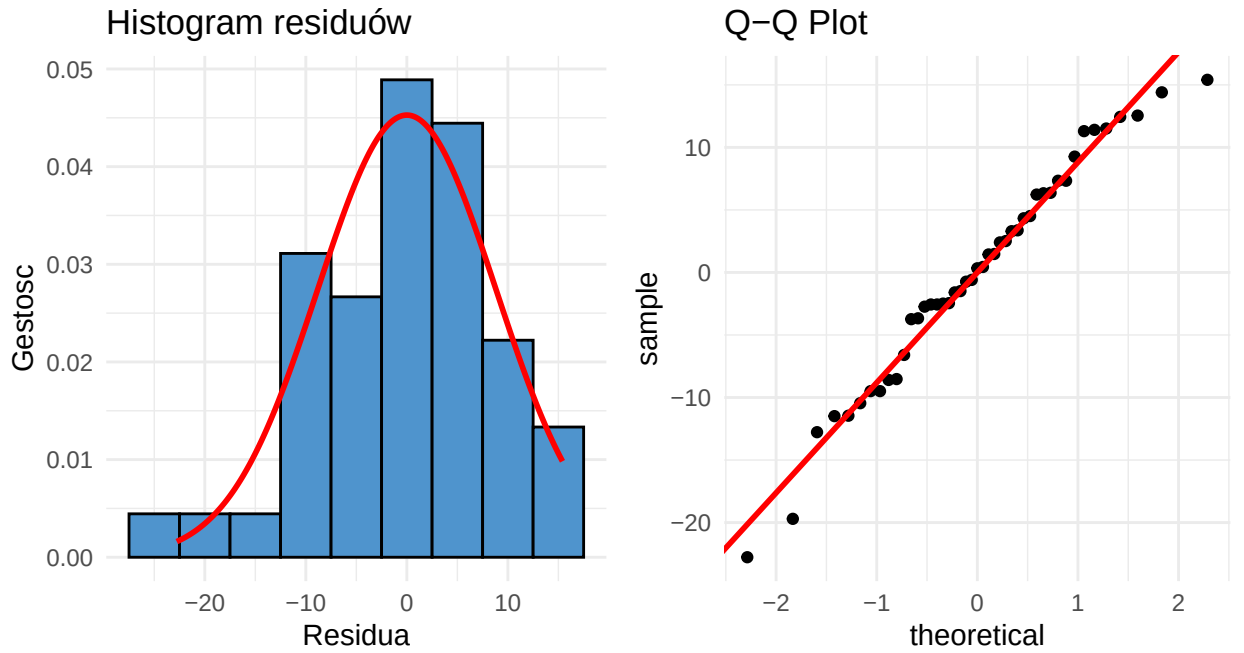
Każdy punkt na wykresie to osobna obserwacja, a odległość punktu od poziomej linii (y = 0) pokazuje jak zła byłą predykcja dla konkretnej obserwacji. Można zauważyć, że najwięcej punktów znajduje się w okolicy 0 oraz, że są one w miarę symetrycznie rozłożone wokół prostej $y = 0$. W związku z tymi cechami suma residuów jest bliska 0. Nietypowymi punktami są punkty oddalone o 20 (na poziomie y = -20) od 0, ale jest to zrównoważone ilością punktów, które są oddalone o 10, jest ich znacznie więcej.

(c) Przedstawię wykres residuów względem kolejności, w jakiej dane pojawiają się w pliku danych.



Wykres pokazuje wielkość reszt w zależności od kolejności dokonywanych obserwacji. Punkty są takie same, jak w poprzednim podpunkcie, tylko teraz obserwujemy, czy wielkość reszt w modelu liniowym zależała od czasu wykonywania pomiaru. Ciężko na tym wykresie wychwycić jakieś wyraźne wzorce, punkty są poukładane przypadkowo, więc można stwierdzić, że residua mają strukturę losową, co oznaczałoby, że błędy losowe są względem siebie niezależne.

(d) Sprawdzę rozkład residuów za pomocą histogramu i wykresu kwantylo-kwantylowego.



Na histogram została nałożona krzywa wyznaczająca teoretyczną gęstość rozkładu normalnego. Większość histogramu znajduje się pod krzywą, co sugeruje nam, że reszt mają zbliżony rozkład do rozkładu normalnego. Wykres prawdopodobieństwa jest prosty i nie ma punktów, które mocna odchylały by się od teoretycznej prostej, co również wskazuje na rozkład normalny residuów.

Zadanie 4

Zmodyfikuję dane z pliku **ch01pr20.txt** dodając dodatkową obserwację (1000;2).

```
nowa_obserwacja <- data.frame(czas = 1000, kopiarki = 2)
dane2_n <- rbind(dane2, nowa_obserwacja)
```

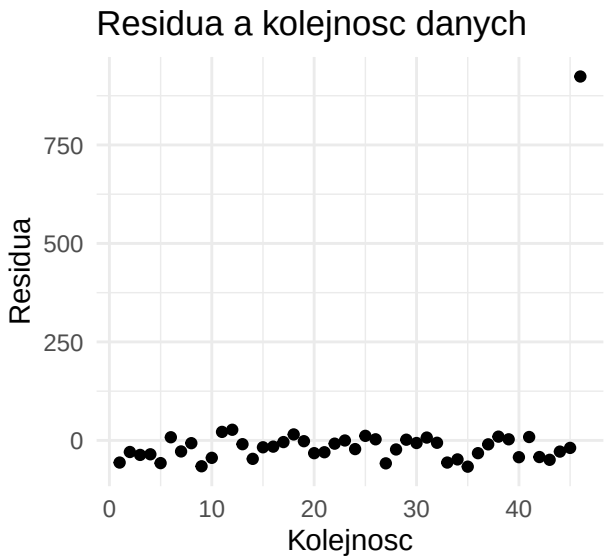
(a) Przeprowadzę regresję ze zmienionymi danymi i utworzę tabelę porównującą wyniki tej analizy z wynikami analizy oryginalnych danych.

Model	Dopasowane.równanie.regresji	p.wartość	R.kwadrat	Estymator.sigma.2
Zmodyfikowany model	Y = 63.091 + X * 6.594	0.393	0.017	143.011
Stary model	Y = -0.58 + X * 15.035	0.000	0.957	8.914

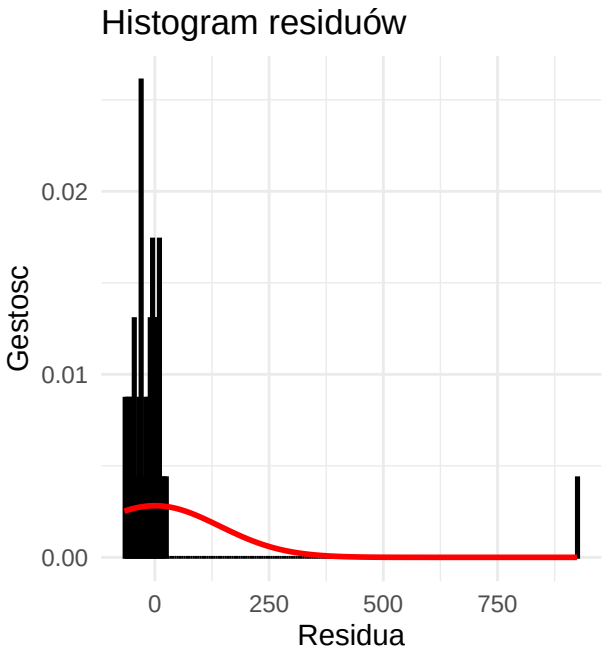
Można zaobserwować, że dodanie jednej zmiennej znacząco wpłynęło na dopasowanie całego modelu - wszystkie parametry z tabeli uległy zmianie. Przede wszystkim, p-wartość w starym modelu jest przybliżana do 0, co za tym idzie $p - wartość < \alpha = 0.05$, a w nowym modelu $p - wartość = 0.393 > 0.05$, co oznaczałoby, że nie możemy odrzucić hipotezy zerowej, gdzie slope = 0, czyli X i Y są nieskorelowane. W starym modelu $R^2 = 0.96$, co oznacza, że 96% zmiennych X wyjaśniało zmienną Y, w nowym moodelu $R^2 = 0.017$, co wskazuje na bardzo słabe dopasowanie modelu. Bardzo dużą różnicę można zauważyć również w estymatorze σ^2, w modelu zmodyfikowanym wynosi ona aż 143.011, co wskazuje na duże rozrzucenie zmiennych, natomiast

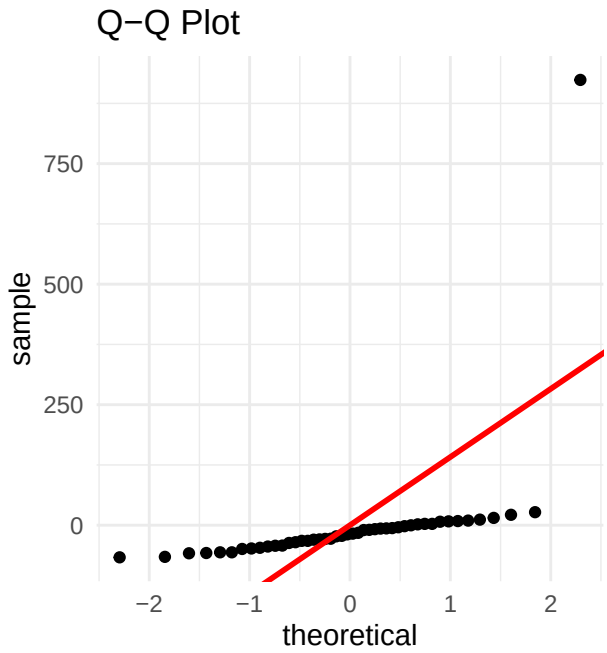
dla starego modelu wynosi jedynie 8.914. Obserwacja, która tak znacząco wpływa na statystyki rozkładu zwana jest **obserwacją wpływową**, jeśli zauważamy taką obserwację w naszych danych powinniśmy się zastanowić skąd ona pochodzi i czy aby na pewno nie jest to błąd, ale żeby móc to stwierdzić musimy sprawdzić tę teorię u źródła.

(b) Powtórzę podpunkty (b), (c) i (d) z zadania 3. na powyżej zmodyfikowanym zbiorze danych.



Z obu wykresów łatwo zauważyć obserwacje odstającą (wpływową), która jest oddalona od prostej y = 0 o ponad 800. Tak wyglądające wykresy residuów wskazywałyby na to, że model regresji nie radzi sobie dobrze z wyjaśnieniem lub przewidywaniem tej konkretnej obserwacji. Moglibyśmy się zastanawiać, czy pomiar został poprawnie przeprowadzony, takiej obserwacji nie możemy usunąć z danych, a jedynie zweryfikować jej poprawność u źródła pytając, czy były jakieś nietypowe okoliczności, które mogłyby wyjaśnić tę wartość odstającą.





Na histogramie można zaobserwować obserwcję obstającą na samym brzegu wykresu, na wyrkesie QQ również łatwo ją dostrzec - jest zdecydowanie oddalona od pozostałych danych. Wpływająca obserwacja

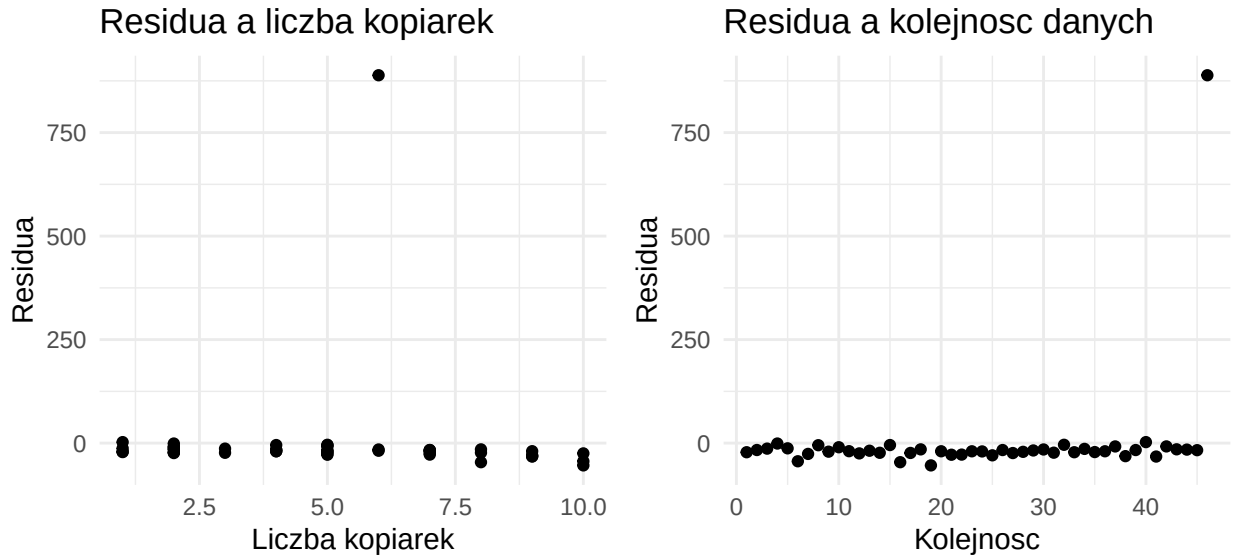
wpłwa na teoretyczną prostą wyliczoną dla rozkładu normalnego, wobec czego dane nie pokrywają się z tą prostą. To samo stało się z histogramem - krzywa gęstości znaczącą część histogramu danych nie obejmuje. Z histogramu oraz wykresu kwantylo-kwantylowego można dojść do wniosku, że dane nie podchodzą z rozkładu normalnego.

(c) Tym razem do początkowego pliku **ch01pr20.txt** dodam obserwację $(1000, 6)$ i sprawdzę podpunkty (a) i (b) z tego zadania.

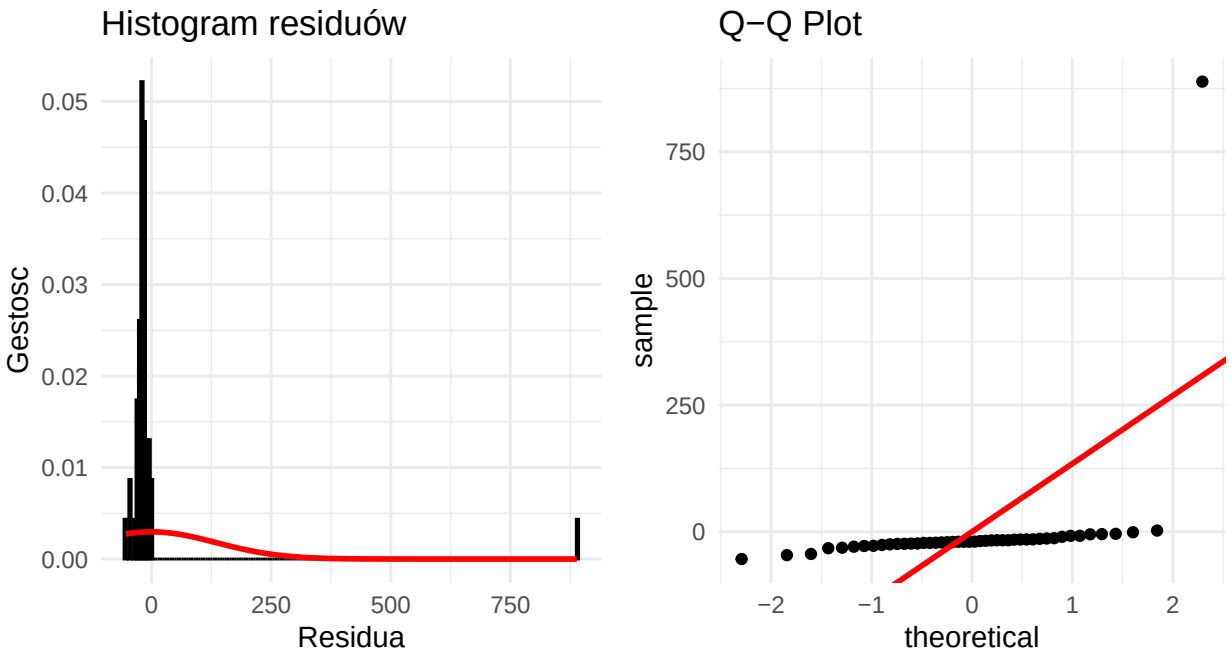
```
nowa_obserwacja_1 <- data.frame(czas = 1000, kopiarki = 6)
dane2_n <- rbind(dane2, nowa_obserwacja_1)
```

Model	Dopasowane.równanie.regresji	p.wartość	R.kwadrat	Estymator.sigma.2
Zmodyfikowany model	Y = 7.308 + X * 17.355	0.023	0.112	135.875
Stary model	Y = -0.58 + X * 15.035	0.000	0.957	8.914

Ponownie wszystkie paramtery zostały zmienione przed dodanie jednej obserwacji, co by oznaczało, że jest to **obserwacja wpływowa**. Wartość R^2 dla zmodyfikowanego modelu wynosi zaledwie 0.112, co wskazuje na słabe dopasowanie mdoelu liniowego do danych. Wartość estymatora σ^2 również uległa dużej zmianie, wynika to z bardzo oddalonej jednej obserwacji od reszty danych. Ciekawie w tym zestawieniu wychodzi p-wartość, ponieważ jak i dla starego i dla zmodydikowanego modelu jest ona mniejsza od $\alpha = 0.05$, czyli w obu przypadkach na tym poziomie istotności możemy odrzucić hipotezę, że zmienne są nieskorelowane.



Na obu wykresach łatwo zauważyć obserwacje odstającą, jest ona wyraźnie pokazana u góry wykresów. Możemy zauważyć, że dla wykresu residuów dla liczby kopiarek obserwacja odstająca jest mniejwięcej w połowie wykresu, co sugeruje pewną regularność w występowaniu tego odstępstwa. Ten wykres oraz niska p-wartość tych danych oznacza,że obserwacja odstająca nie różni się istotnie od reszty danych.



Na histogramie oraz wykresie QQ łatwo dostrzec obserwację odstającą, w obu przypadkach znajduje się ona na brzegu wykresu. Tak jak w przypadku podpunktu (b) wpływa ona na teoretyczną krzywą gęstości rozkładu normalnego oraz teoretyczną prostą, wobec czego ani histogram, ani wykres kwantylo-kwantylowy się z nimi nie pokrywają. W związku z czym dochodzimy do wniosku, że dane nie pochodzą z rozkładu normalnego.

Zadanie 5

W tym zadaniu zostaną wykorzystane dane z pliku **ch03pr15.txt** dotyczące stężenia roztworu. Zawierają wartości stężenia roztworu oraz czas.

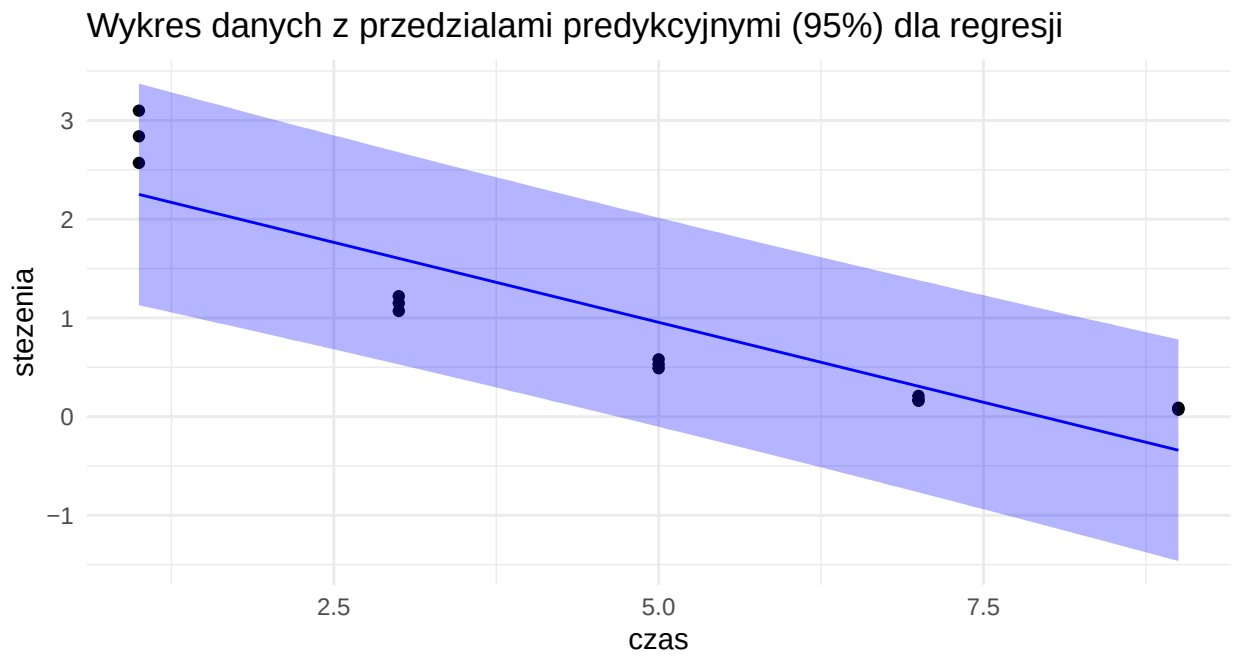
(a) Przeprowadzę regresję liniową z czasem jako zmienną objaśniającą i stężeniem roztworu jako zmienną odpowiedzi. Podam odpowiednie równanie regresji, przedstawię dane i prostą regresji na wykresie. Do wykresu dodam 95% przedział prdeykcyjny dla poszczególnych obserwacji.

Z podsumowania modelu można odczytać, że:

- $\hat{\beta}_1$ = -0.324
- $\hat{\beta}_0$ = 2.575

Zatem równanie regresji wygląda tak: $Y = -0.324 * X - 2.575$.

Dane oraz prosta regresji na wykresie wraz z 95% przedziałem predykcyjnym wyglądają tak:



Na wykresie można zauważyć, że obserwacje (punkty) mniejwięcej układają się tak, jak prosta regresji, niestety widać, że nie jest to najlepsze dopasowanie, bo żaden z punktów nawet nie pokrywa się z tą linią. Można więc szukać lepszego modelu do przewidywania na podstawie tych danych. Prasmo predykcyjne obejmuje wszystkie obserwacje, ale zarazem jest bardzo szerokie, co wskazuje na znaczną niepewność prognozowanego modelu w odniesieniu do rzeczywistych wartości. Powoduje to konieczność ostrożnego interpretowania wyników oraz dalszej analizy w celu poprawy skuteczności modelu.

(b) Podam wartość R^2 i wyniki testu istotności dla hipotezy zerowej, że stężenie roztworu nie zależy od czasu.

```
summ3$r.squared
```

```
## [1] 0.8115774
```

Współczynnik determinacji R^2 wynosi około 0.812, co oznacza, że około 81.2 zmienności zmiennej zależnej można wyjaśnić za pomocą modelu regresji liniowej. Jest to dość wysoki wynik, co oznacza, że ten model regresji jest stosunkowo skuteczny w wyjaśnianiu zmienności danej zmiennej.

$$H_0: \ \beta_1 = 0 \quad vs \quad H_1: \beta_1 \neq 0$$

Odrzucamy hipotezę zerową wtedy, gdy $F > F_c = F^*(1-\alpha, 1, n-2)$, gdzie $n = 15$, a α przyjmę 0.05. Przeprowadzę teraz ten test.

```
Fc <- qf(1-0.05, 1, 15-2)
summ3$fstatistic[1] > Fc
```

```
## value
##  TRUE
```

Statystyka testowa F wyszła 56 ilosc stopni swobody wynosi 1 i 13, p-wartość 4.6111995×10^{-6}. Z przeprowadzonego wyżej testu można wywnioskować, że hipoteza zerowa zostałą odrzucona, co oznacza, że z 90% pewnością możemy stwierdzić, że zmienna objaśniana i objaśniająca są ze sobą skorelowane. Niska p-wartość również o tym świadczy, ponieważ jest zdecydowanie mniejsza od ustalonego poziomu istotności $\alpha = 0.05$.

(c) Obliczę współczynnik korelacji między obserwowaną i przewidywaną wartością stężenia roztworu.

```
cor(dane3$stezenia,data.frame(predictions)$fit)
```

```
## [1] 0.9008759
```

Współczynnik korelacji między obserwowaną i przewidywaną wartością stężenia roztworu wynosi 0.9008759. Oznacza, to, że zależność między zmiennymi jest silna, czyli jeśli jedna zmienna rośnie, to bardzo prawdopodobne jest, że druga zmienna również wzrośnie.

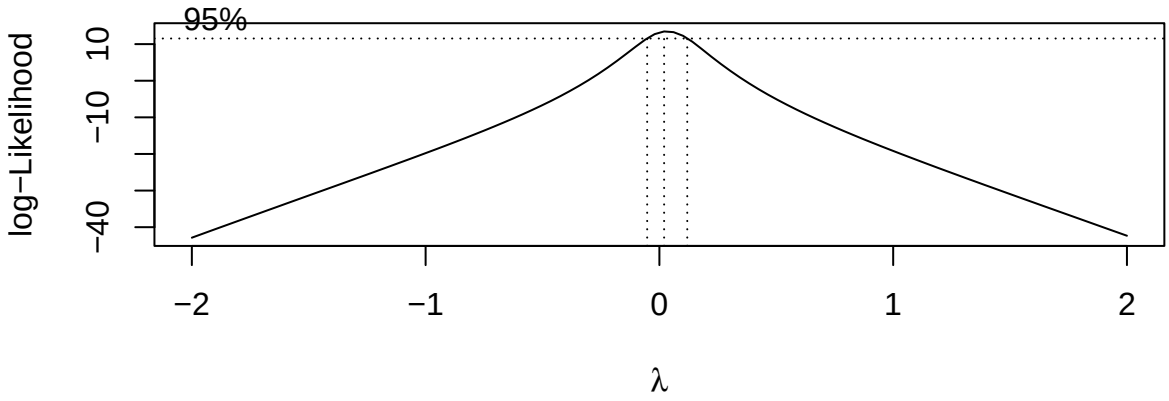
Zadanie 6

Użyję procedury Box'a-Cox'a, aby znaleźć odpwiednią transformację dla stężenia roztworu do danych użytych w poprzednim zadaniu.
Transormacja Boxa-Coxa umożliwia wybór optymalnego przekształcenia, dopasowuje ona do danych model postaci:

$$f_\lambda(Y) = \tilde{Y} = \beta_0 + \beta_1 X_i + \epsilon_i$$

gdzie $\tilde{Y} = Y^\lambda$ lub $\tilde{Y} = (Y^\lambda - 1)/\lambda$. Następnie przy użyciu metody największej wiarogodności estymuje optymalną wartość parametru λ.

```
bc <- boxcox(model3)
```




```
(lambda <- bc$x[which.max(bc$y)])
```

```
## [1] 0.02020202
```

Procedura Boxa-Coxa jest używana do znalezienia optymalnej transformacji danych, aby lepiej spełniać założenia regresji liniowej. Polecenie **boxcox()** zwraca wykres wartoścy kryterium Boxa-Coxa dla różnych potencjalnych transformacji parametru (lambda), który odpowiada za różne rodzaje transormacji danych. Optymalną wartością lambda jest taki x, dla którego wartość y na wykresie jest największa, w przypadku naszych danych ten wynik to 0.020202. Sugerowana wartość lambda przez tę procedurę może być stosowana jako potęga do której dane będą podnoszone, aby uzyskać liniową zależność danych. W tym przypadku sugerowana transformacja jest zbliżona do logarytmicznej.

Zadanie 7

W tym zadaniu będę używać danych tych samych, co w poprzednich dwóch zadaniach.

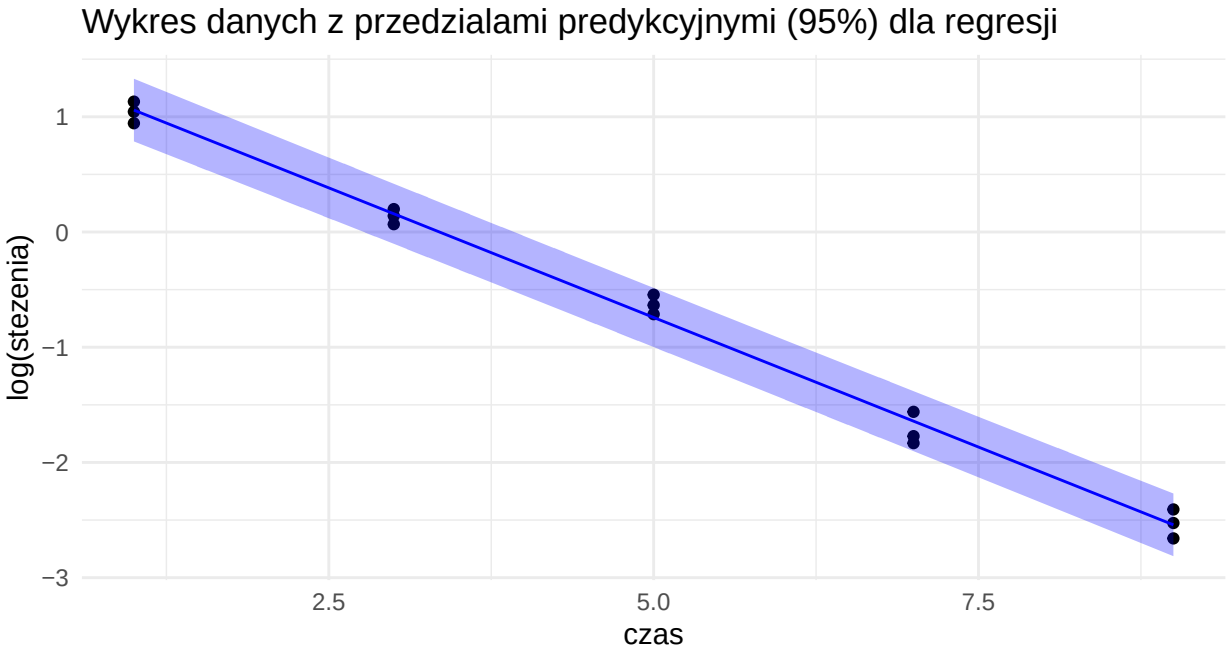
(a) Utworzę nową zmienną odpowiedzi, biorąc logarytm stężenia roztworu.

Z podsumowania modelu można odczytać, że:

- $\hat{\beta}_1$ = -0.45
- $\hat{\beta}_0$ = 1.508

Zatem równanie regresji wygląda tak: $\tilde{Y} = log(Y) = -0.45 * X - 1.508$.

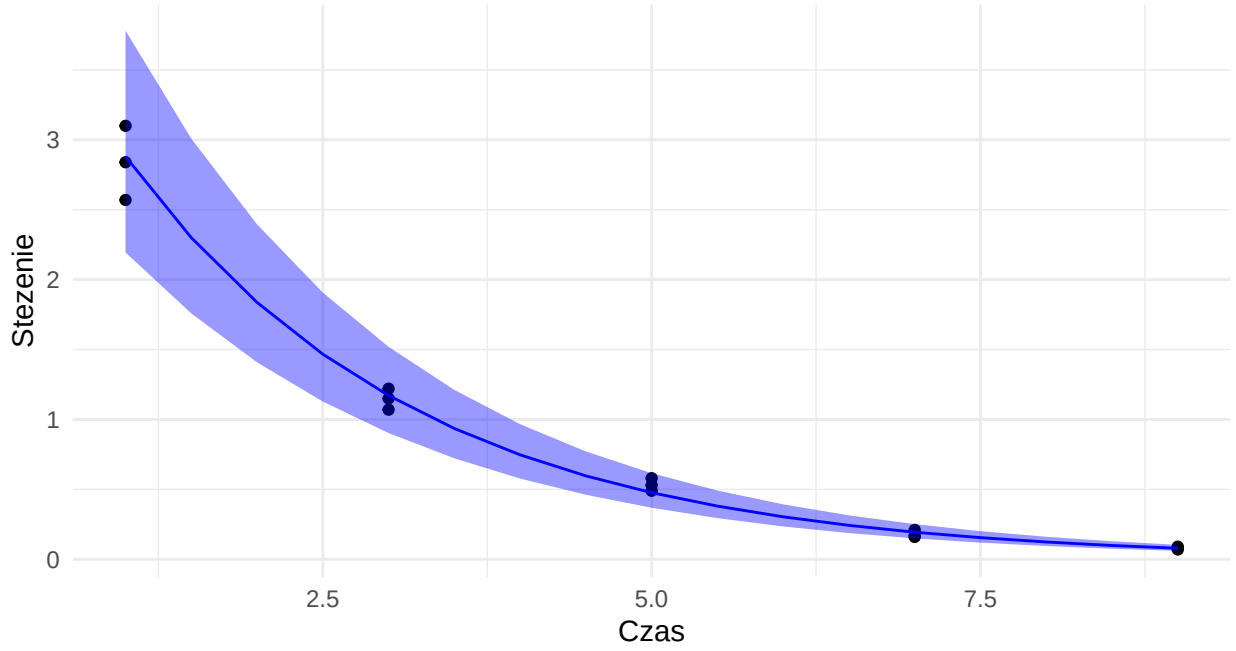
(b) Powtórzę zadanie 5. z $\tilde{Y}$ jako zmienną odpowiedzi i czasem jako zmienną objaśniającą ($\tilde{Y} \sim$ time)



	R_2	F_statistic	Fc	F > Fc	Korelacja
wartość	0.993	1838.225	4.667	TRUE	0.996

Z wykresu możemy odczytać, że wszystkie punkty zawierają się w 95% paśmie predykcyjnym oraz jest dużo węższy niż w zadaniu 5., czyli model ma dobrą zdolność do przewidywania i jest wiarygodny. Potwierdza to również bardzo wysoki współczynnik determinacji, ponieważ wynosi on $R^2 = 0.99$, czyli 99% danych jest wyjaśnianych przez model. Dla hipotezy, że zmienne są niezależne (hipoteza zerowa: $\beta_1 = 0$) została policzona statystyka F oraz wartość krytyczna Fc dla tego testu na poziomie ufności 95%, wynik tego testu jest w kolumnie `F > Fc`. Oznacza to, że z 95% pewnością możemy odrzucić tę hipotezę i uznać, że zmienne są zależne. Współczynnik korelacji między obserwowaną i przewidywaną wartością stężenia również wyszedł bardzo wysoki, bliski 1.

(c) Na wykresie przedstawię stężenie roztworu względem czasu. Dodam krzywą regresji i pasmo dla 95% przedziałów predykcji na podstawie wyników uzyskanych w punkcie (b). Porównam z wykresem uzyskanym w zadaniu 5.



Porównując wykres z wykresem z zadania 5. można zauważyć, że krzywa dopasowania jest o wiele lepiej dopasowana do danych. To znaczy, że znajduje się bliżej puktów na wykresie i bardziej przypomina linie łączącą te punkty. Pasmo predykcyjne również jest węższe, co wskazuje na lepsze dopasowanie.

(d) Obliczę współczynnik korelacji między obserwowanym, a przewidywanym stężeniem roztworu opartym na modelu z punktu (b) i porównam z odpowiednim wynikiem z zadania 5.

```
cor(exp(data.frame(predictions)$fit), datlog6$stezenie)
```

```
## [1] 0.9945587
```

Korelacja wyszła wyższa niż w zadaniu 5., co wskazuje na lepsze dopasowanie modelu do danych.

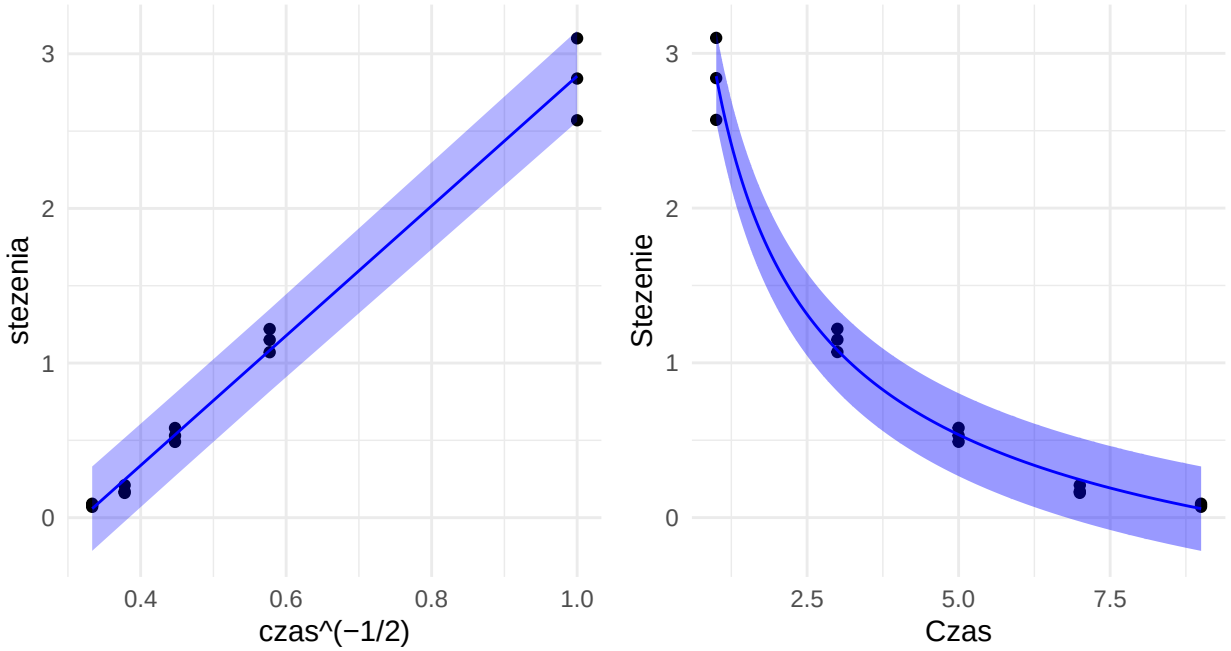
Zadanie 8

Używajac danych z zadania 5. skonstruuję nową zmienną objaśnijącą $\tilde{t} = time^{-1/2}$. Powtórzę zadanie 7., używając modelu regresji ze stężeniem roztworu jako zmienną odpowiedzi i $\tilde{t}$ jako zmienną objaśniającą ($Y \sim \tilde{t}$). Podsumuję wyniki.

Z podsumowania modelu można odczytać, że:

- $\hat{\beta}_1 = 4.196$
- $\hat{\beta}_0$ = -1.341

Zatem równanie regresji wygląda tak: $\tilde{Y} = 4.196 * X - 1.341$.



	R_2	F_statistic	Fc	F > Fc	Korelacja	Korelacja2
wartość	0.988	1076.05	4.667	TRUE	0.901	0.994

Przedziały predykcyjne, choć zawierają wszystkie obserwacje, są szersze niż w przypadku gdy $\tilde{Y} = log(Y)$. Proste regresji również przechodzącą przez większą ilość punktów, co wskazywałoby na odpowiednie dopasowanie modelu. Wartość $R^2 = 0.988$ oznacza, że 98% danych jest wyjaśnone przez model, co jest wysokim wynikiem. Test istotności dla slope'a utwierdza nas w tym, że zmienne są zależne. Z wysokiego poziomu korelacji w obu przypadkach możemy wnioskować i pozostałych wartości statystyk możemy wnioskować, że ten model jest dobrze dopasowany do naszych danych.

Porównując szerokości przedziałów predykcyjnych, odległość prostych liniowych od punktów oraz statystyki przedstawionych w tabelce można stwierdzić, że $\tilde{Y} = log(Y)$ jest lepszym wyborem przekształcenia zmiennej Y. Ponadto zastosowanie metody Boxa-Coxa również wskazało to przekształcenie.

Zadania teoretyczne

```
#1a
(tc <- qt(1 - 0.05/2,df=  c(5,10,15)))
```

```
## [1] 2.570582 2.228139 2.131450
```

```
#1b
(Fc <- qf(1 - 0.05,df1 = 1,df2=  c(5,10,15)))
```

```
## [1] 6.607891 4.964603 4.543077
```

```
#1c
tc^2
```

```
## [1] 6.607891 4.964603 4.543077
```

```
#jak widać tc^2 = Fc, wynika to z tego, że zmienna jeśli X ~ t-student(df = n), to X^2 ~ F(1, n)
```

```r
#2a
#dfe = n-2 = 20 -> n = 22

#2b
#dfE = 20
#SSE = 400
#MSE = SSE/dfE = s^2 -> s = sqrt(SSE/dfE)
(s <- sqrt(400/20))
```

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

```r
#2c
#odrzucamy H_0 gdy MSM/MSE > Fc
# MSM = SSM/dfM = 100/1
# MSE = SSE/dfE = 400/20
MSM <- 100
MSE <- 400/20
MSM/MSE > qf(1 - 0.05, 1, 20)
```

```
## [1] TRUE
```

```r
#zatem odrzucamy hipoteze zerowa o rownosci srednich grup na poziomie istotnosci 0.05, statystyka F = 5

#2d

#wsp determinacji R^2 mowi jaka czesc calkowitej zmiennosci w wektorze Y stanowi zmiennosc wyjasniona p
#R^2 = SSM/SST = 1 - SSE/SST = 1 - SSE/(SSM + SSE)
R2 <- 1 - 400/(100+400)
#model wyjasnia 20% zmiennych objasniajacych

#2e
(corr <- sqrt(R2))
```

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

```r
#nieski współczynnik wskazuje na słabą korelację
```