Modele liniowe - raport 3

Łukasz Rębisz

23.12.2022

Zadanie 1

a)

Naszym zadaniem jest obliczenie t_c - krytycznej wartości dla niekierunkowego testu istototności dla poziomu istotności $\alpha = 0.05$.

Badana statystyka testowa $T \sim t(10)$.

```
alpha <- 0.05
df <- 10
t_c <- qt(1-alpha/2, df)
```

Otrzymana wartość $t_c = 2.2281389$.

b)

Naszym zadaniem jest obliczenie F_c - krytycznej wartości dla niekierunkowego testu istototności dla poziomu istotności $\alpha = 0.05$.

Badana statystyka testowa $F \sim F(1, 10)$.

```
F_c <- qf(1-alpha, 1, 10)
```

Otrzymana wartość $F_c = 4.9646027$.

c)

Sprawdźmy, czy $t_c^2 = F_c$:

```
(abs(t_c^2-F_c))
```

```
## [1] 8.881784e-16
```

Wartości F_c i t_c^2 są równe do szesnastego miejsca po przecinku, możemy więc przyjąć, że $t_c^2 = F_c$.

Zadanie 2

a)

Wiemy, że liczba stopni swobody błędu $dfE = n - 2$, więc rozmiar próby $n = dfE + 2 = 22$

b)

Estymacja σ:

$s^2 = MSE = \frac{SSE}{dfE} = \frac{400}{20} = 20$, zatem estymator $\hat{\sigma} = s = \sqrt{20} \approx 4.4721$.

c)

Test istotności dla β_1 (na poziomie istotności $\alpha = 0.05$):

$$H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0.$$

Wiemy, że gdy prawdziwa jest hipoteza zerowa, to statystyka $F = \frac{MSM}{MSE} = \frac{SSM/dfM}{SSE/dfE}$ pochodzi z rozkładu Fishera-Snedecora z dfM i dfE stopniami swobody ($F \sim F(dfM, dfE)$).

Jeżeli zaś zachodzi hipoteza alternatywna, to MSM jest średnio większa od MSE i w konsekwencji statystyka F przyjmuje duże wartości.

Wykonajmy test F sprawdzający, czy β_1 jest różna od zera:

odrzucamy hipotezę zerową, gdy $F = \frac{MSM}{MSE} > F_c$, gdzie $F_C = F^*(1-\alpha, 1, n-2)$

```
F_test <- function(SSM, SSE, dfM, dfE, alpha){
  F_c <- qf(1-alpha, dfM, dfE)
  F_stat <- (SSM/dfM)/(SSE/dfE)
  if(F_stat > F_c) return(c(F_stat, dfM, dfE, 1, F_c))
  return(c(F_stat, dfM, dfE, 0, F_c))
}

test <- F_test(SSM=100, SSE=400, dfM=1, dfE=20, alpha=0.05)
```

Otrzymana statystyka testowa $F = 5$ ~F(1, 20).

$F = 5 > 4.3512435 = F_C$, zatem odrzucamy hipotezę zerową mówiącą, że $\beta_0 = 0$ na poziomie istotności $\alpha = 0.05$.

d)

Proporcję tego, jak wariancja zmiennej objaśnianej jest wyjaśniana przez model określa współczynnik determinacji R^2 :

$R^2 = \frac{SSM}{SST} = \frac{SSM}{SSM+SSE} = \frac{100}{100+400} = 0.2.$

Współczynnik przyjmuje wartości pomiędzy 0 a 1. Wartość 0.2 świadczy o niskim poziomie dopasowania modelu do danych.

e)

Próbkowa korelacja

W regresji liniowej prostej współczynnik R^2 jest tożsamy z kwadratem próbkowej korelacji pomiędzy zmiennymi wyjaśniającą a wyjaśnianą, zatem próbkowa korelacja Pearsona wynosi (co do wartości bezwzględnej) $|cor(X, Y)| = \sqrt{R^2} \approx 0.447$.

Zadania 3, 4

```
data_3 <- read.table(file="tabela1_6.txt",header = F, sep="")
colnames(data_3) <- c("nr", "GPA", "IQ", "Gender", "PHtest")
```

Baza danych zawiera dane 78 uczniów siódmej klasy ze Stanów Zjednoczonych. Dla każdego ucznia opisywane dane to:

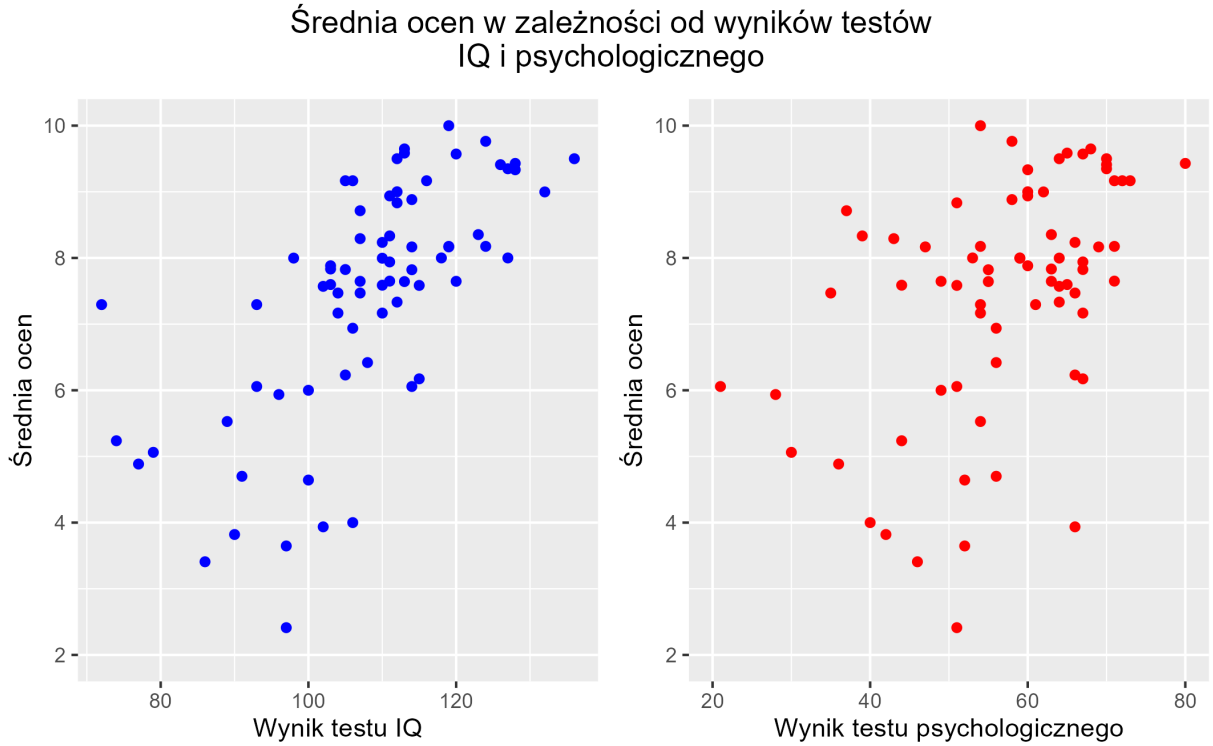
- GPA: średnia ocen (ocenom F, D-, D, D+, C-, C, C+, B-, B, B+, A-, A, odpowiadają liczby od 0 do 11);

- IQ: wynik testu IQ;
- Gender: M-mężczyzna, F-kobieta;
- PHtest: wynik testu psychologicznego (Piers-Harris Children's Self-Concept Scale).

Celem zadania jest stworzenie **prostego modelu liniowego** na podstawie powyższych danych.

Regresja liniowa prosta

Zbadajmy regresję liniową prostą dla **średniej ocen** w zależności od wyników **testu IQ** oraz **testu psychologicznego**.



Powyższe wykresy nie wskazują na liniową zależność pomiędzy zmiennymi. Zwłaszcza w przypadku wyników testu psychologicznego trudno stwierdzić zależność liniową. Mimo tego, stwórzmy modele liniowe dla powyższych danych.

Zaimplemtujmy funkcję przeprowadzającą **test Snedecora** dla danej próby (wykorzystującą powyższą funkcję *F_test*):

```
F_test_data <- function(X, Y, alpha){
  dfM <- 1
  dfE <- length(X) - 2
  linMod <- lm(Y~X)
  beta_0 <- as.numeric(linMod$coefficients[1])
  beta_1 <- as.numeric(linMod$coefficients[2])
  Y_hat <- beta_0 + beta_1*X
  SSM <- sum((Y_hat - mean(Y))^2)
  SSE <- sum((Y-Y_hat)^2)
  return(F_test(SSM, SSE, dfM, dfE, alpha))
}
```

	beta_0	beta_1	R^2	Statystyka F	p-wartość	Statystyka F - własna
Średnia ocen~test IQ	-3.557	0.101	0.402	51.00845	0e+00	51.00845
Średnia ocen~test psychologiczny	2.226	0.092	0.294	31.58516	3e-07	31.58516

Analiza powyższych danych pokazuje, że:

- Otrzymaliśmy identyczne wartości statystyki T, wokrzystując funkcję wbudowaną i własną funkcję.
- Współczynnik determinacji R^2 jest większy dla modelu badającego wpływ wyników testu IQ na średnią ocen. Zatem wynik testu IQ ma zależność bliższą liniowej niż wynik testu psychologicznego - model liniowy jest dokładniejszy w przypadku testu IQ.
- Również statystyka T osiągająca większą wartość dla testu IQ świadczy o tym, że w tym przypadku z większą pewnością odrzucamy hipotezę zerową mówiącą, że zmienna wyjaśniającą (tu: wynik testu IQ) nie ma wpływu na zmienną wyjaśnianą (średnią ocen).

Podsumowując, wyniki testu IQ mają większy wpływ na średnią ocen niż wyniki testu psychologicznego.

Przedziały predykcyjne

Wyznaczmy 90% przedziały predykcyjne dla średniej ocen ucznia, którego:

i) wynik testu IQ jest równy 100,

ii) wynik testu psychologicznego jest równy 60.

```
x_pred_IQ <- data.frame(IQ=c(100))
conf_int_IQ <- predict(linMod3, x_pred_IQ, interval = "prediction", alpha=0.9)
x_pred_PH <- data.frame(PHtest=c(60))
conf_int_PH <- predict(linMod4, x_pred_PH, interval = "prediction", alpha=0.9)
```

Otrzymujemy następujące 90% przedziały predykcyjne:

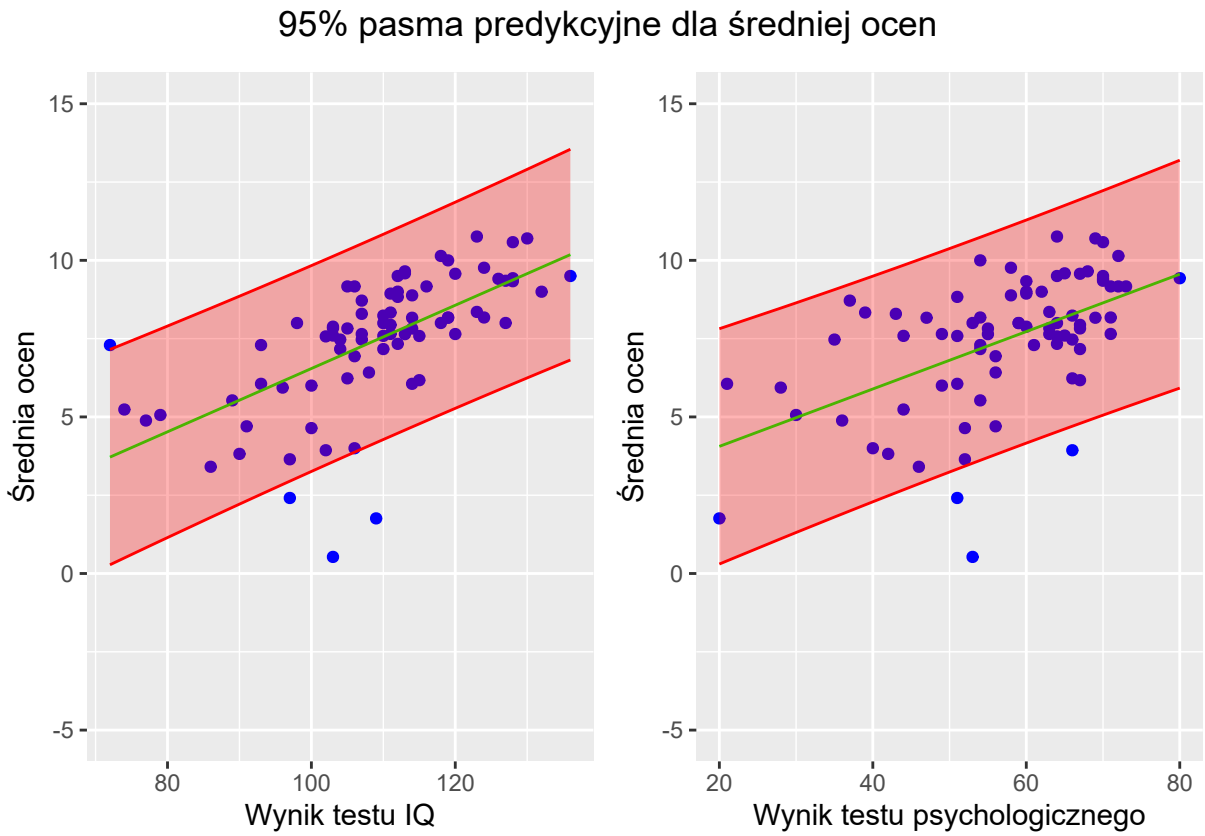
i) [3.259, 9.831], długość przedziału równa 6.572

ii) [4.163, 11.287], długość przedziału równa 7.124.

Przyjęte wyniki testów odpowiadają mniej więcej średniej wyników danego testu. Zauważmy, że dla danego wyniku testu IQ otrzymaliśmy węższy przedział predykcyjny dla średniej ocen.

Pasmo predykcyjne

Wyznaczmy 95% przedziały predykcyjne dla wszystkich zmienncyh objaśniających, tworząc tym samym 95% **pasmo predykcyjne dla średniej ocen**:



Poza pasmem ufności znajduje się ok.

5.13% obserwacji dla testu IQ,

3.85% obserwacji dla testu psychologicznego.

Otrzymane wyniki są zgodne z założonym poziomem istotności 0,05.

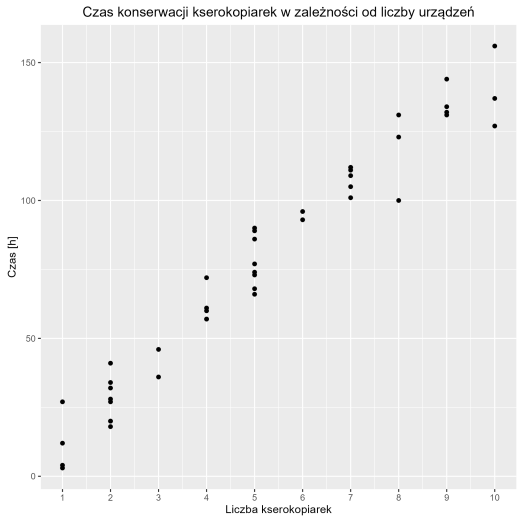
Podsumowanie

Tak jak pokazaliśmy powyżej, średnia ocen w większym stopniu zależy od wyników testu IQ niż wyników testu psychologicznego. Zatem wyniki testu IQ są lepszym predyktorem dla średniej ocen.

Zadania 5, 6

W zadaniu wykorzystamy dane obrazujące czas potrzebny do konserwacji danej liczby kserokopiarek.

```
data_5 <- read.table(file="CH01PR20.txt",header = F, sep="")
colnames(data_5) <- c("Time", "Number")
```



Powyższy wykres pozwala stwierdzić, że zależność pomiędzy czasem konserwacji a liczbą kserokopiarek **jest w przybliżeniu liniowa**.

Wyznaczmy model prostej regresji liniowej dla danych:

i) niezmienionych,

ii) po zmianie pierwszej obserwacji z wartości 20 na 2000.

Ciąg residuów będących estymatorami błędów losowych wyraża się następującym wzorem:

$$e_i = Y_i - \hat{Y}_i, \quad \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

Sprawdźmy, czy suma residuów jest równa zero:

Suma residuów jest równa $7.9580786 \times 10^{-13} \approx 0$ (dla niezmienionych danych) oraz $-2.5579538 \times 10^{-13} \approx 0$ (dla zmienionych danych).

Regresja liniowa

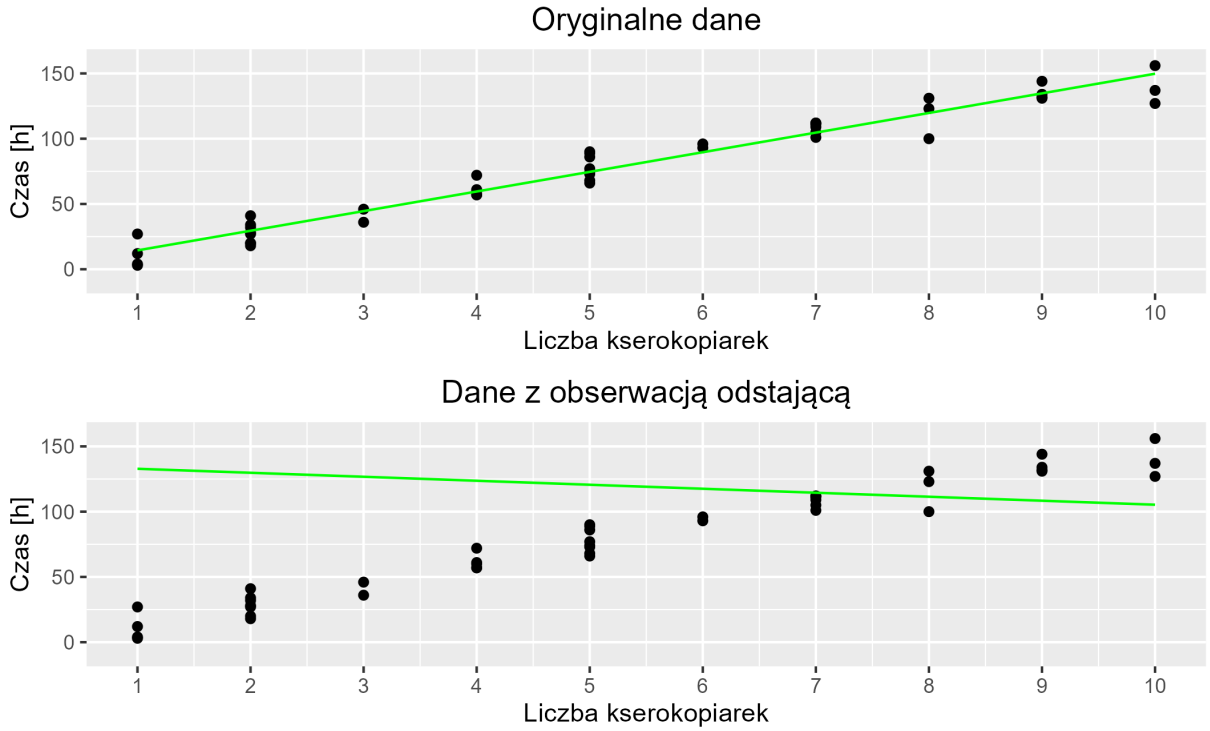
Dla powyższych danych otrzymujemy modele regresji liniowej o następujących parametrach:

	beta_0	beta_1	R^2	Statystyka F	p-wartość	s^2
Oryginalne dane	-0.58	15.035	0.957	968.6571960	0.000000	79.45063
Zmiana 1-szej obserwacji	135.90	-3.059	0.001	0.0371408	0.848086	85759.43314

Analiza powyższych danych pokazuje, że dodanie obserwacji odstającej:

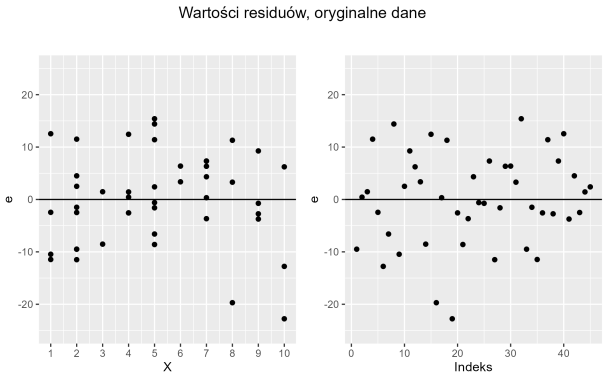
- Zdecydowanie zwiększyła wartość wyrazu wolnego - parametru β_0.
- Współczynnik kierunkowy zmienił nie tylko wartość, ale również znak, aby uwzględnić dużą, odstającą wartość.
- Współczynnik determinacji R^2 z wartości bliskiej 1 spadł praktycznie do zera - model prawie idealnie liniowy stał się zupełnie nieliniowy.
- P-wartość (równa 0) świadcząca o tym, że nie ma żadnych podstaw do twierdzenia, że zmienna objaśniająca nie zależy od objaśnianej, przyjęła dużą wartość - w tym przypadku nie mamy podstaw do odrzucenia hipotezy, że $\beta_1 = 0$.

- Wyestymowana wartość wariancji znaczącą wzrosła - zmieniona obserwacja znaczącą odstaje od pozostałych wyników.

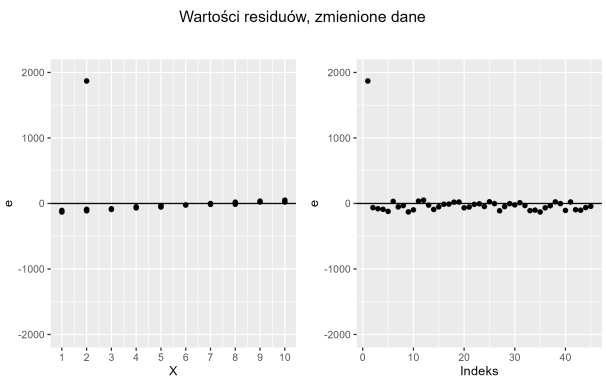


Powyższe wykresy z zaznaczonymi prostymi regresji liniowej jasno pokazują, jak odstająca obserwacja zmieniła cały model. Prosta na dolnym wykresie w żadnym stopniu nie oddaje danych, nie pokrywa się z obserwacjami w przeciwieństwie dla prostej wyznaczonej dla oryginalnych danych.

Residua

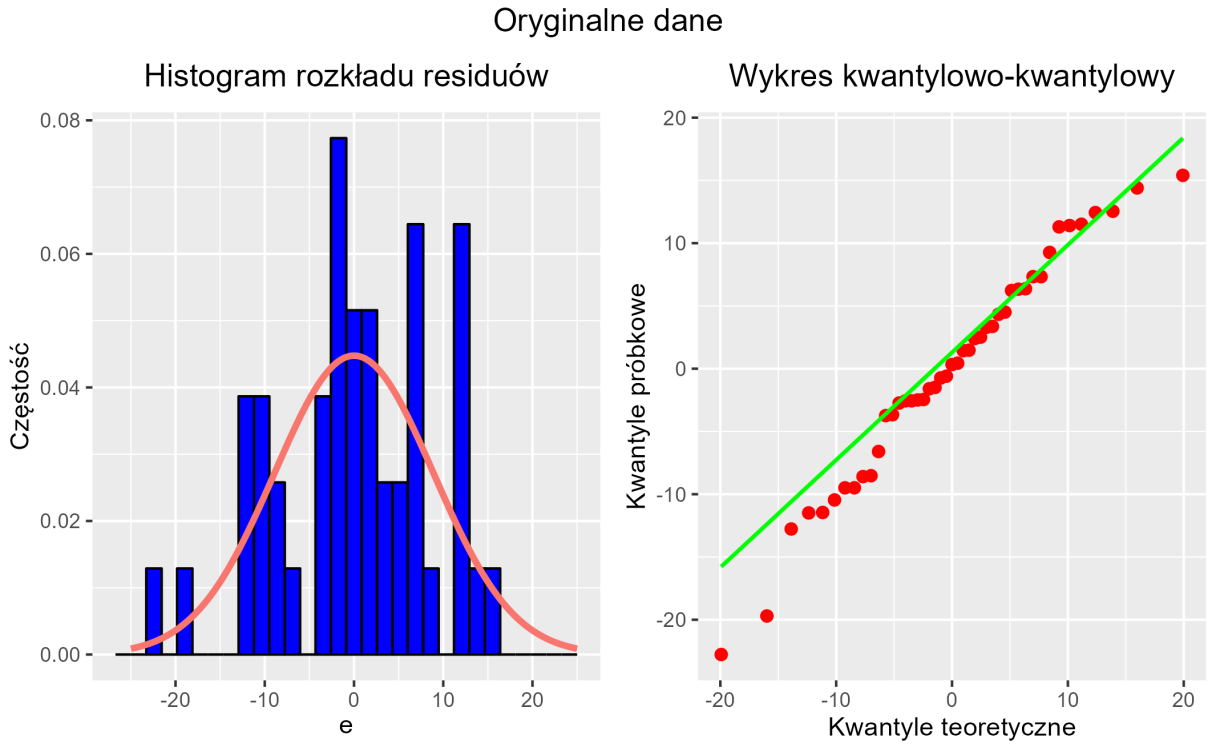


Rozkład residuów dla oryginalnych danych nie wykazuje większych odstępstw od rozkładu normalnego. Oba powyższe wykresy są symetryczne względem zera, brak wyraźnych zależności, wzorów, które nie są chaotyczne. Świadczy to o normalności rozkładu residuów.



Powyższe wykresy wyraźnie obrazują, jak duży wpływ na rozkład residuów ma obserwacja odstająca. Zupełnie zaburza rozkład. Trudno mówić o normalności rozkładu residuów w tym przypadku.

Testy normalności



Powyższy histogram nie potwierdza tego, że rozkład residuów jest normalny. Szczególnie obserwacje dla wartości równych ok. 10 znaczącą przewyższają krzywą rozkładu normalnego. Natomiast wykres kwantylowo-kwantylowy pokazuje, że poza wartościami residuów poniżej -5 rozkład residuów jest w przybliżeniu normalny - kwantyle teoretyczne i próbkowe praktycznie się pokrywają.

Wykonajmy dokładniejsze testy normalności dla powyższych danych:

Testy sprawdzające normalność rozkładu:

- **Test Shapiro-Wilka** (dobrze wykrywa skośność oraz ciężkie ogony rozkładu). "Jest najbardziej zalecanym testem normalności rozkładu, jednakże może dawać mylne wyniki dla liczebności próbek powyżej 2000. Wymaga, żeby cecha miała rozkład ciągły."(Billewicz K. 2011, s. 78). Był pierwszym testem, który był w stanie wykryć odstępstwa od normalności z powodu skośności i/lub kurtozy. Stał

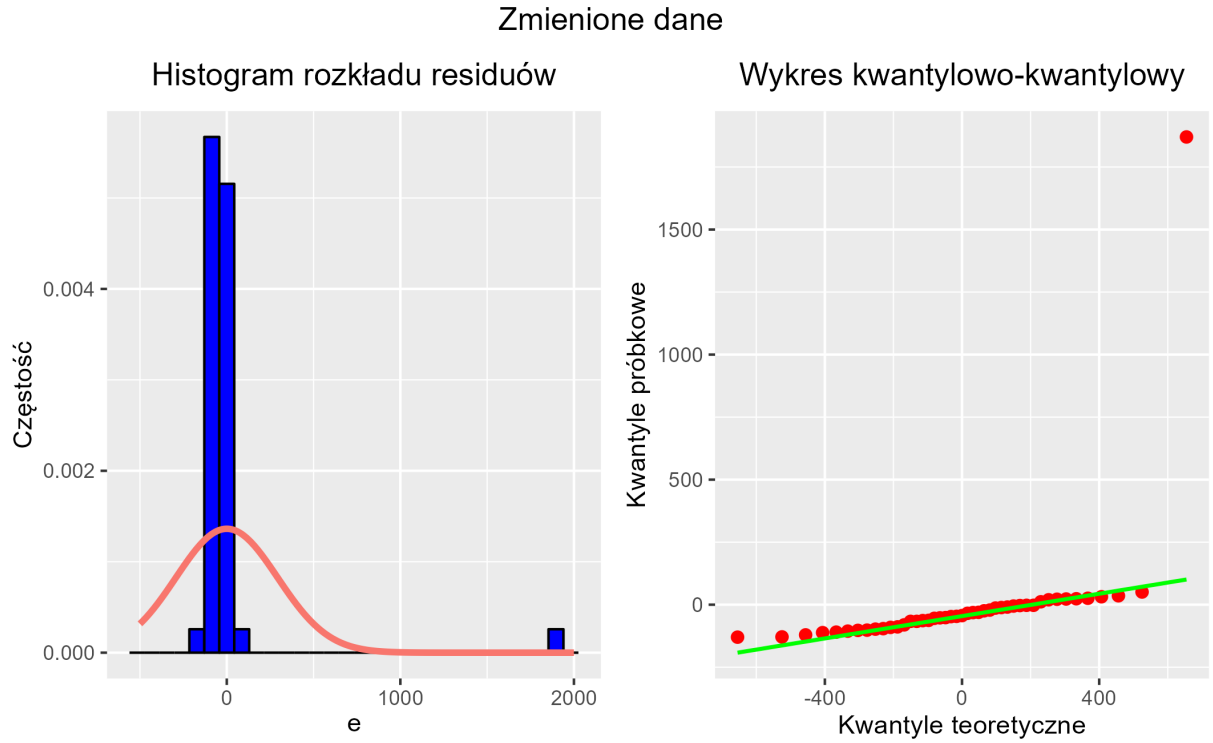
się preferowanym testem ze względu na swoją silną moc. Test jest oparty na szacowaniu średniej odległości wykresu kwantyl-kwantyl od prostej. Może być stosowany dla małych prób. Jest mało wrażliwy np. na autokorelację.

P-wartość = 0.4614257. Wartość świadczy o tym, że nie mamy podstaw do odrzucenia hipotezy zerowej mówiącej, że rozkład danych jest normalny.

- **Gładki test Neymana**: dobrze radzi sobie z rozpoznaniem zaburzenia masy w środku rozkładu, ma szerokie spektrum wykrywania możliwych odstępstw.

P-wartość = 0.3494297 również nie daje nam podstaw do odrzucenia normalności danych.

Podsumowując, powyższe analizy i wykonane testy świadczą o tym, że rozkład residuów dla oryginalnych danych jest w przybliżeniu normalny, czyli zgodny z założeniami modelu regresji liniowej.



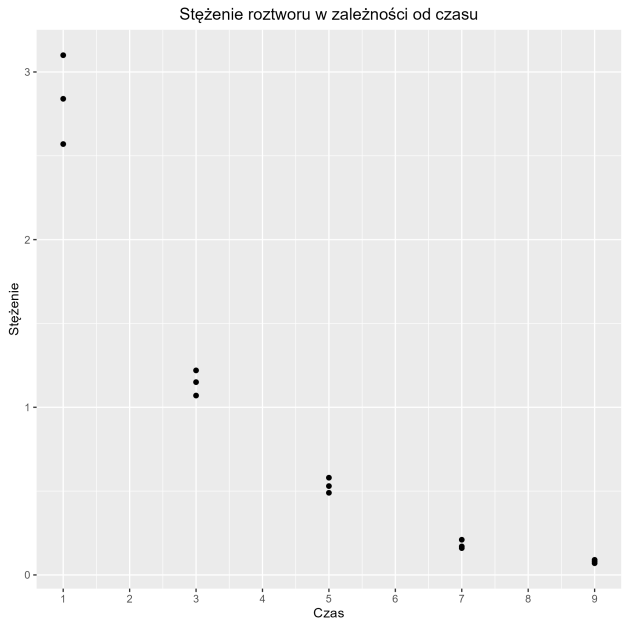
Histogram wartości residuów dla zmienionych danych jasno pokazuje, że w tym przypadku nie ma zgodności z krzywą rozkładu normalnego. Również wykres kwantylowo-kwantylowo obrazuje istotność obserwacji odstającej na decyzję o odrzuceniu normalności rozkładu residuów.

Dane z obserwacją odstającą nie mogą być uznane za dane o rozkładzie nawet w przybliżeniu normalnym.

Zadania 7 - 12

W kolejnych zadaniach będziemy badać różne modele regresji liniowej dla danych obrazujących stężenie procentowe roztworu w zależności od czasu, starając się dopasować model jak najlepiej dopasowany do danych.

```
data_7 <- read.table(file="CH03PR15.txt",header = F, sep="")
colnames(data_7) <- c("Concentration", "Time")
```



Powyższy wykres obrazujący badane dane świadczy o tym, że dane nie są w przybliżeniu liniowe. Krzywymi, które lepiej oddają charakter danych wydają się krzywe typu $\exp(-x)$ lub $\frac{1}{\sqrt{x}}$.

Metoda Boxa - Coxa

Naszym zadaniem jest zalezienie odpowiedniego przekształcenia zmiennej Y, korzystając z transformacji Boxa-Coxa w celu poprawienia liniowości modelu.

Transformacja dopasowuje do danych model postaci

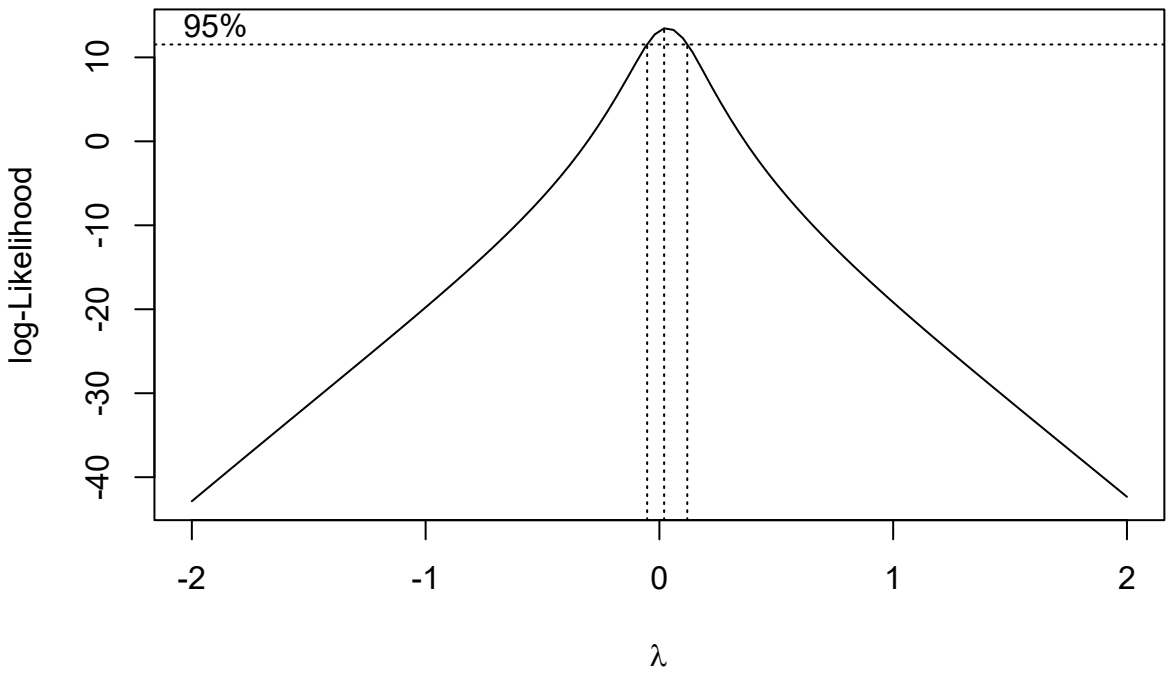
$$f_\lambda(Y) = \tilde{Y} = \beta_0 + \beta_1 X_i + \epsilon_i,$$

gdzie $\tilde{Y} = Y^\lambda$ lub $\tilde{Y} = (Y^\lambda - 1)/\lambda$.

Następnie przy użyciu metody największej wiarogodności estymuje optymalną wartość parametru λ.

Wykorzystajmy wbudowaną funkcję *boxcox* z pakietu *MASS*:

```
LinMod7 <- lm(Concentration~Time, data_7)
b7 <- boxcox(LinMod7)
```




```
lambda <- b7$x[which.max(b7$y)]
```

Wyznaczona powyższą metodą optymalna wartość dla λ wynosi ok. $0.0202 \approx 0$, zatem optymalne przekształcenie to $\tilde{Y} = (Y^\lambda - 1)/\lambda$, co w przypadku granicznym $\lambda \to 0$ zadaje przekształcenie $\log Y$.

Zbadajmy modele regresji liniowej następujących zmiennych:

- $Y \sim X$,
- $\log(Y) \sim X$ (zgodnie z wynikiem metody Boxa-Coxa),
- $\log(Y) \sim X^{-1/2}$,
- $Y \sim X^{1/2}$,

gdzie X - czas, Y - stężenie roztworu.

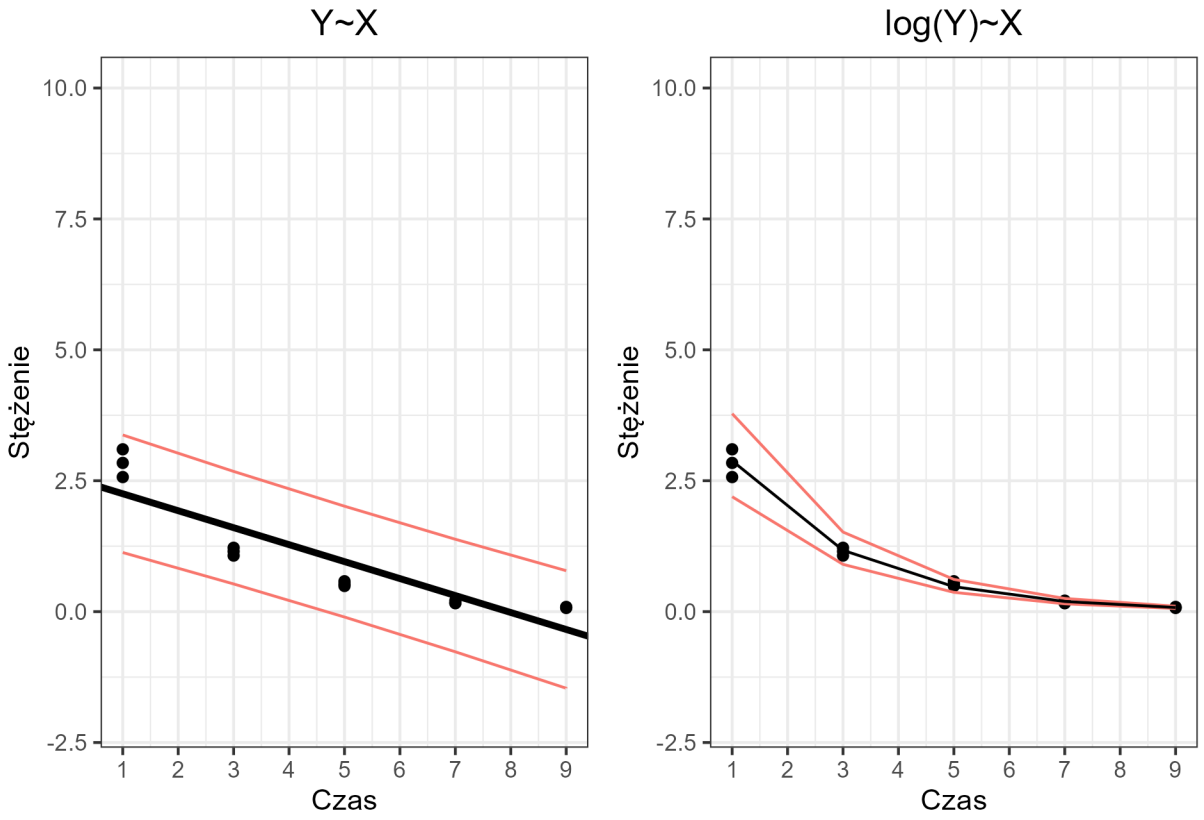
Modele regresji liniowej

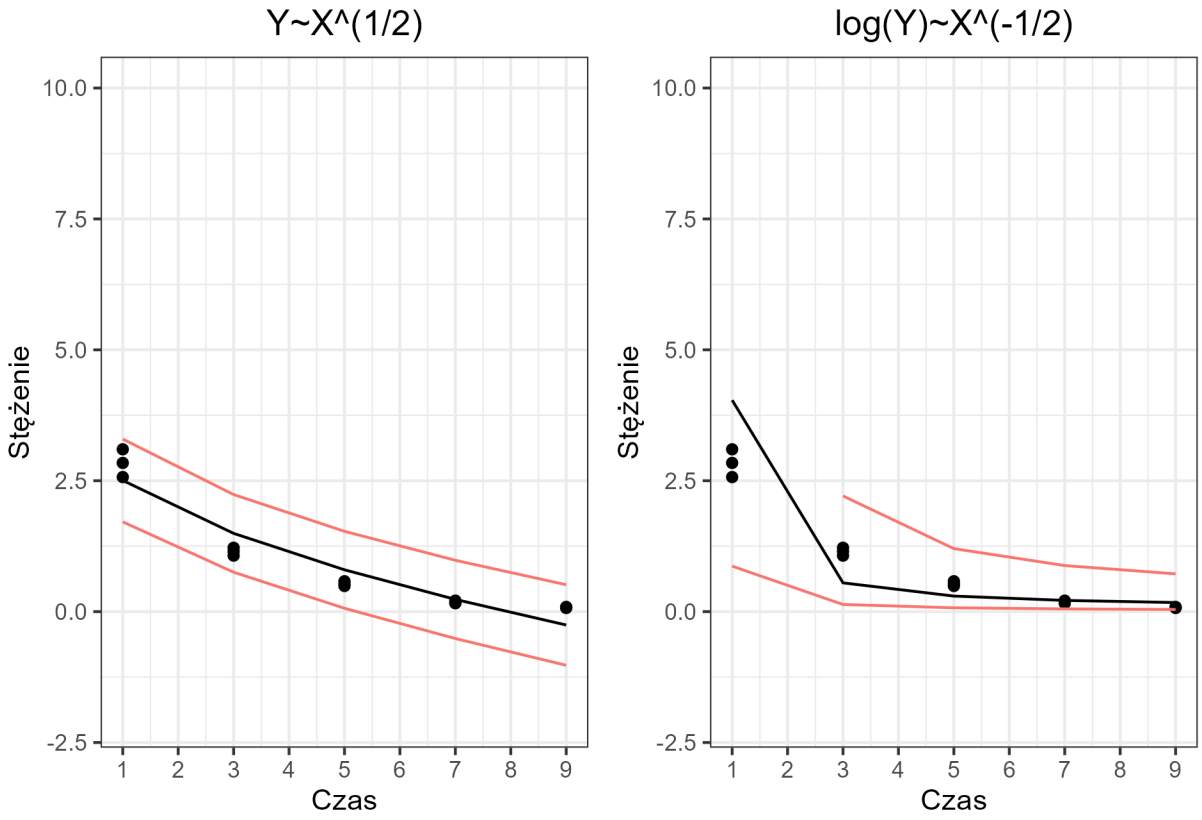
	beta_0	beta_1	Statystyka t	p-wartość	R^2	wariancja	korelacja
Y~X	2.575333	-0.3240000	-7.482903	4.6e-06	0.8115774	0.2249733	-0.9008759
log(Y)~X	1.507916	-0.4499258	-42.874527	0.0e+00	0.9929776	0.0132149	-0.9964826
log(Y)~X^(-1/2)	-3.324813	4.7208160	7.065941	8.5e-06	0.7934131	0.3887613	0.8907374
Y~X^(1/2)	3.885447	-1.3803229	-11.413593	0.0e+00	0.9092623	0.1083392	-0.9535525

Analiza powyższych danych pokazuje, że:

- Współczynnik determinacji osiągą najwyższą wartość (bliską 1) dla modelu $\log(Y) \sim X$. Dla takiego modelu zależność pomiędzy zmiennymi jest więc prawie idealnie liniowa - zgodna z modelem regresji liniowej.
- Korelacja pomiędzy zmienną objaśniającą a objaśnianą również wskazuje na model $\log(Y) \sim X$ jako ten, w którym zmienne są najbardziej skorelowane.
- P-wartości praktycznie równe zero uzyskaliśmy dla modeli $\log(Y) \sim X$ oraz $Y \sim X^{1/2}$. W tych przypadkach z prawdopdobieństwem 1 odrzucamy hipotezę mówiącą, że zmienna objaśniana nie zależy od zmiennej objaśniającej.
- Model $\log(Y) \sim X$ charakteryzuje się ponadto najmniejszą wariancją.

Podsumowując, model najbliższy modelowi prostej regresji liniowej uzyskujemy w przypadku $\log(Y) \sim X$, czyli zgodnym z wynikiem metody Boxa-Coxa.





Analiza powyższych wykresów pozwala stwierdzić, że:

- Model $Y \sim X$ nie oddaje dobrze danych, ponieważ prosta regresji liniowej nie przechodzi w tym przypadku przez prawie żaden z punktów, przedziały predykcyjne są szerokie.
- Model $Y \sim X^{1/2}$ lepiej oddaje dane - krzywa lepiej pokrywa się z danymi, przedziały predykcyjne są węższe niż w przypadku $Y \sim X$.
- Modelem, który najlepiej oddaje dane, znów okazuje się model $\log Y \sim X$. Krzywa przechodzi przez wszystkie punkty. Przedziały predykcyjne są wąskie i zmniejszają się wraz z przysrostem czasu.
- Model $\log Y \sim X^{-1/2}$ zachowuje się niejednoznacznie. Z jednej strony krzywa dość dobrze pokrywa się z danymi. Przedziały predykcyjne dla małych wartości X są natomiast bardzo szerokie - na tyle szerokie, że zostały przycięte na otrzymanym wykresie.

Wniosek

Przekształcenie danych w sposób zgodny z metodą Boxa-Coxa, tzn. zbadanie regresji liniowej $\log Y \sim X$, a następnie powrócenie do wyjściowej sytuacji (poprzez funkcję wykładniczą) pozwoliło otrzymać model dobrze oddający otrzymane dane - krzywą pokrywającą się z danymi oraz wąskie przedziały predykcyjne.