Modele liniowe raport nr 3

Dominik Mika

7 grudnia 2020

1 Zadanie 1

a)
Chcemy znaleźć kwantyl dla dwustronnego testu istotności t z $\alpha = 0.05$ oraz 10 stopniami swobody. W tym celu skorzystamy z funkcji w R-**qt**. Stąd nasz kwantyl jest równy 2.228

b)
W tym podpunkcie chcemy wyznaczyć wartość krytyczną dla testu F, czyli chcemy znaleźć kwantyl z rozkładu Fishera–Snedecora dla $\alpha = 0.05$, $dfM = 1$ i $dfE = 10$. Użyjemy do tego funkcji **qf**. Dstajemy, że $Fc = 4.965$

c)
Teraz chcemy sprawdzić, czy $tc^2 = Fc$.

$$tc^2 = (2.228)^2 = 4.965 = Fc$$

Widzimy, że zachodzi równość.

2 Zadanie 2

W tym zadaniu mamy daną niepełną tabelę analizy wariancji i za jej pomocą chcemy wyznaczyć odpowiednie wielkości.

Tabela 1: Tabela analizy wariancji

	df	SS
$Model$	1	100
$Error$	20	400

a)
Chcemy znaleźć liczbę obserwacji jaka znajduje się w naszym zbiorze. Wiemy, że $dfE = n - 2$, a w naszym przypadku $dfE = 20$. Stąd $n = 22$.

b)
Żeby znaleźć estymator odchylenia standardowego oznaczmy s jako estymator σ. Z wykładu wiemy, że $s^2 = MSE = \frac{SSE}{dfE} = \frac{400}{20} = 20$, czyli $s = 2\sqrt{5}$.

c)
Chcemy przestestować, czy nasz $\beta_1 = 0$, więc wykonujemy test F.

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

Szukamy teraz statystyki testowej F. Ma ona następującą postać $F = \frac{MSM}{MSE}$. Wiemy, że $MSE = s^2 = 20$ oraz $MSM = \frac{SSM}{dfM} = 100$, czyli $F = 5$. Odrzucamy H_0, gdy $F > F_c$, gdzie $F_c \sim F(1 - \alpha, dfM, dfE)$, czyli dla $\alpha = 0.05$, $F_c = 4.351$. Widzimy, że $F = 5 > 4.351 = F_c$, czyli odrzucamy hipotezę zerową.

d)
Chcemy znaleźć wpływ zmiennej objaśniającej na zmienność naszej zmiennej objaśnianej, czyli miarę jakości dopasowania modelu. Z wykładu wiemy, ze jest to określone przez współczynnik determinacji R^2. Jest on dany wzorem $R^2 = \frac{SSM}{SST}$. Znamy już SSM, więc został nam do policzenia SST, ale wiemy, że $SST = SSM + SSE = 400 + 100 = 500$. Stąd mamy $R^2 = 0.2$.

e)
Na koniec chcemy jeszcze znaleźć ile wynosi współczynnik próbkowej korelacji między naszymi zmiennymi. Jest dość proste, ponieważ współczynnik ten jest równy $R = 0.447$.

3 Zadanie 3

W tym i następnym zadaniu będziemy korzystać z jednej ramki danych **table1_6.txt**, która zawiera następujące informacje: GPA, wynik z testu IQ, płeć i wynik w teście psychologicznym.

a)
Za pomocą modelu liniowego chcemy wyznaczyć zależność GPA od wyników w teście IQ. Nasza estymowana prosta regresji ma postać:

$$Y_i = -3.557 + 0.101 \cdot X_i$$

Natomiast $R^2 = 0.402$. Następnie chcemy przetestować, czy GPA jest skorelowana z IQ.

$$H_0 : \beta_1 = 0 \quad \vee \quad H_1 : \beta_1 \neq 0.$$

Kwantyl $F_c = F^*(1 - \alpha, 1, n - 2)$, gdzie $\alpha = 0.05$ wynosi 3.967, statystyka testowa $F = 51.008$, a p-wartość jest równa 4.74×10^{-10}. Widzimy, że statystyka $F > F_c$, czyli na poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę zerową. Również p-wartość jest bardzo mała, czyli prawie zawsze będziemy odrzucać hipotezę zerową. Oznacz to, że GPA w jakimś stopniu zależy od IQ.

b)
W tym podpukcie chcemy przewidzieć jakie jest GPA dla studenta, którego iq wynosi 100, następnie wyznaczymy 90% przedział predykcyjny dla tej wartości. Do wyznaczenia predykcyjnej wartości dla $X = 100$ skorzystamy ze znanego już nam wzoru:

$$\hat{\mu}_{100} = \hat{\beta}_0 + \hat{\beta}_1 X_{100}$$

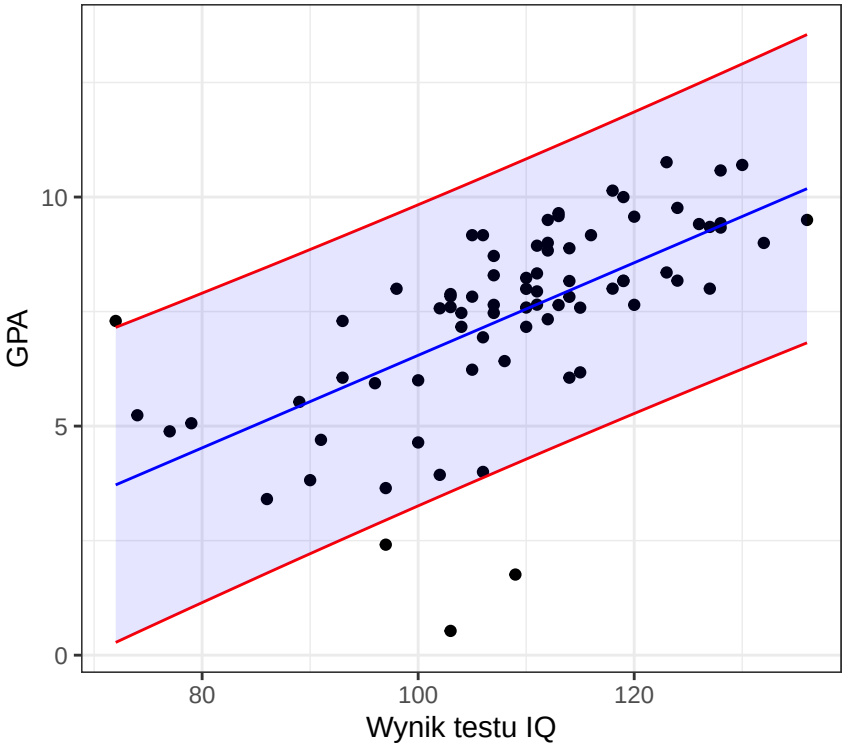
Stąd $\hat{\mu}_{100} = 6.545$
Natomiast nasz 90% przedział predykcyjny jest postaci:

$$\hat{\mu_{100}} \pm t_c s(pred), \quad \text{gdzie} \quad t_c = t^*(1 - \frac{\alpha}{2}, n - 2) \quad \text{i} \quad s^2(pred) = s^2(1 + \frac{1}{n} + \frac{(X_{100} - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2})$$

oraz wynosi:

$$[3.798, 9.293].$$

c)
Tym razem chcemy narysować pasmo predykcyjne dla 95% przedziałów predykcyjnych.



Rysunek 1: Pasmo predykcyjne dla 95% przedziałów

Widzimy, że 4 obserwacje na 78 znajdują się poza tym pasmem.

4 Zadanie 4

W tym zadaniu będziemy postępować bardzo podobnie jak w poprzednim, tylko tym razem naszym regresorem w modelu liniowym będzie wynik z testu psychologicznego.

a)
Nasza estymowana prosta regresji ma postać:

$$Y_i = 2.226 + 0.092 \cdot X_i$$

Natomiast $R^2 = 0.294$. Jest on nieco mniejszy, niż w zadaniu 3, więc możemy powiedzieć, że GPA zależy bardziej od IQ, niż od wyniku w teście psychologicznym.

b)
Następnie chcemy przetestować, czy GPA jest skorelowana z wynikiem testu psychologicznego.

$$H_0 : \beta_1 = 0 \quad \vee \quad H_1 : \beta_1 \neq 0.$$

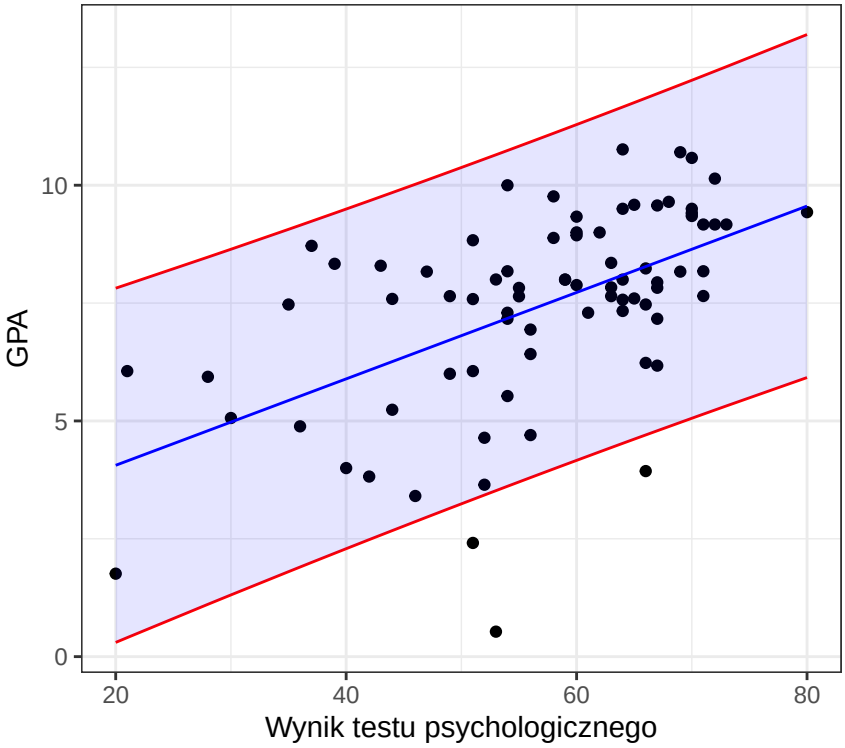
Kwantyl $F_c = F^*(1-\alpha, 1, n-2) = 3.967$, gdzie $\alpha = 0.05$, statystyka testowa $F = 31.585$, a p-wartość wynosi 3.01×10^{-7}.

Widzimy, że podobnie jak w zadaniu 3 statystyka $F > F_c$ czyli na poziomie istotności $\alpha = 0.05$ odrzucamy hipotezę zerową.

c) Chcemy znaleźć przewidywaną wartość GPA dla studentów, którzy otrzymali 60 punktów w teście psychologicznym. Postępujemy tak samo jak w zadaniu 3, tylko dla $X = 60$.

Przewidywana wartośc dla $X = 60$ wynosi 7.725, a przedział ma postać: $[4.747, 10.703]$.

d) Rysujemy 95% pasmo predykcyjne.



Rysunek 2: Pasmo predykcyjne dla 95% przedziałów

Tym razem 3 obserwacje na 78 znajdują się poza tym pasmem.

e)
Podsumowując dwa poprzednie zadania i ich rezultaty jasno możemy stwierdzić, że lepszym predyktorem GPA jest IQ.

5 Zadanie 5

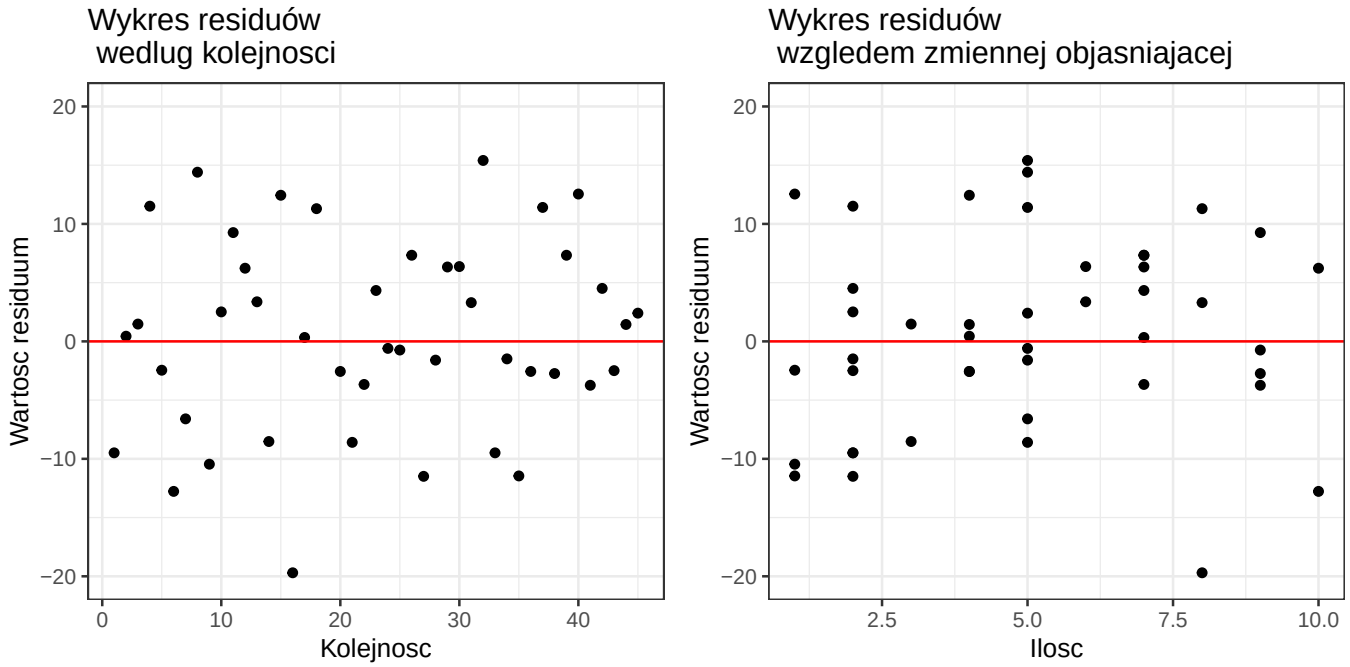
Do następnych dwóch problemów będziemy korzystać z tabeli danych **ch01pr20.txt**, w której znajdują się informacje o czasie serwisowania drukarek.

a)

Chcemy przetestować, czy suma residuów wynosi 0. Korzystamy ze wzoru i dostajemy: $\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = -1.176\,836 \times 10^{-14}$.

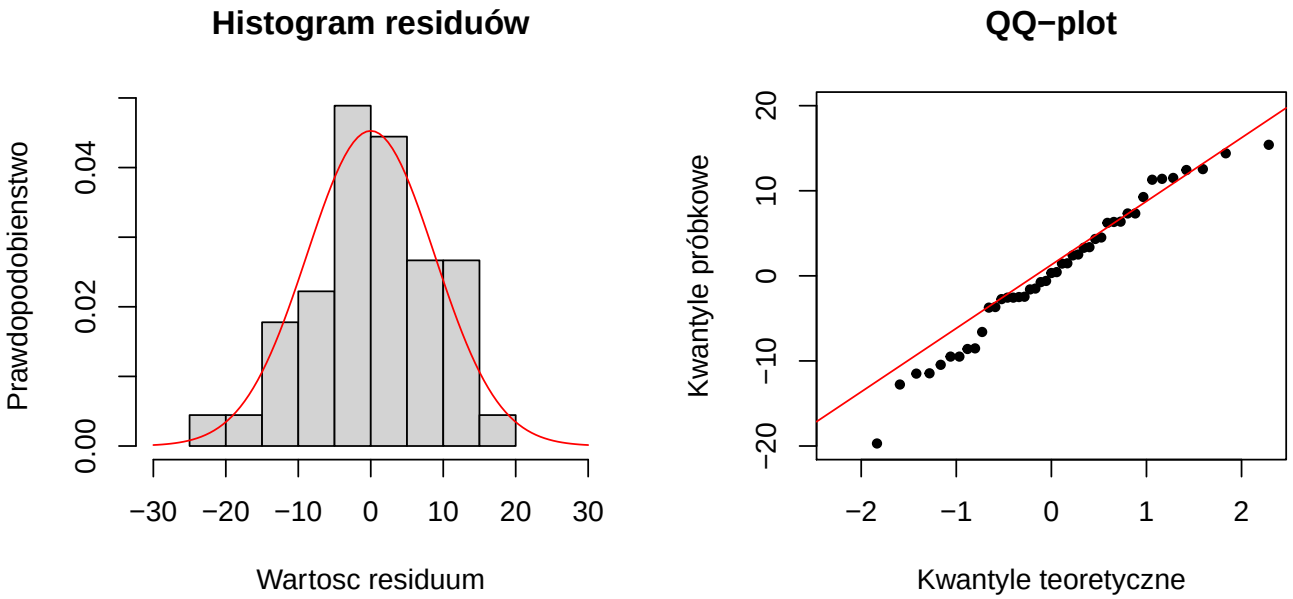
Jest to bardzo mała liczba, więc możemy ją traktować jako 0.

b), c)



Rysunek 3: Wykres residuów względem zmiennej objaśniającej oraz resiuduów w kolejności

Na wykresie nie widzimy żadnych trendów. Punkty wydają się być rozrzucone niezależnie względem siebie. Wariancja błędów wydaje się być stała. Punkt o wartości residuum mniejszej niż -20 można uznać za wartość wpływową.



Rysunek 4: Histogram i QQ-plot residuów

Widzimy, że rozkład residuów, czyli zmiennej ϵ (błędów losowych) wydaje się być normalny, czyli dane spełaniają założenia modelu liniowego.

6 Zadanie 6

W tym zadaniu chcemy zmodyfikować nasze dane. Zamienimy czas serwisowania dla pierwszej obserwacji na 2000 i zobaczymy jak zmieni to nasze wyniki.

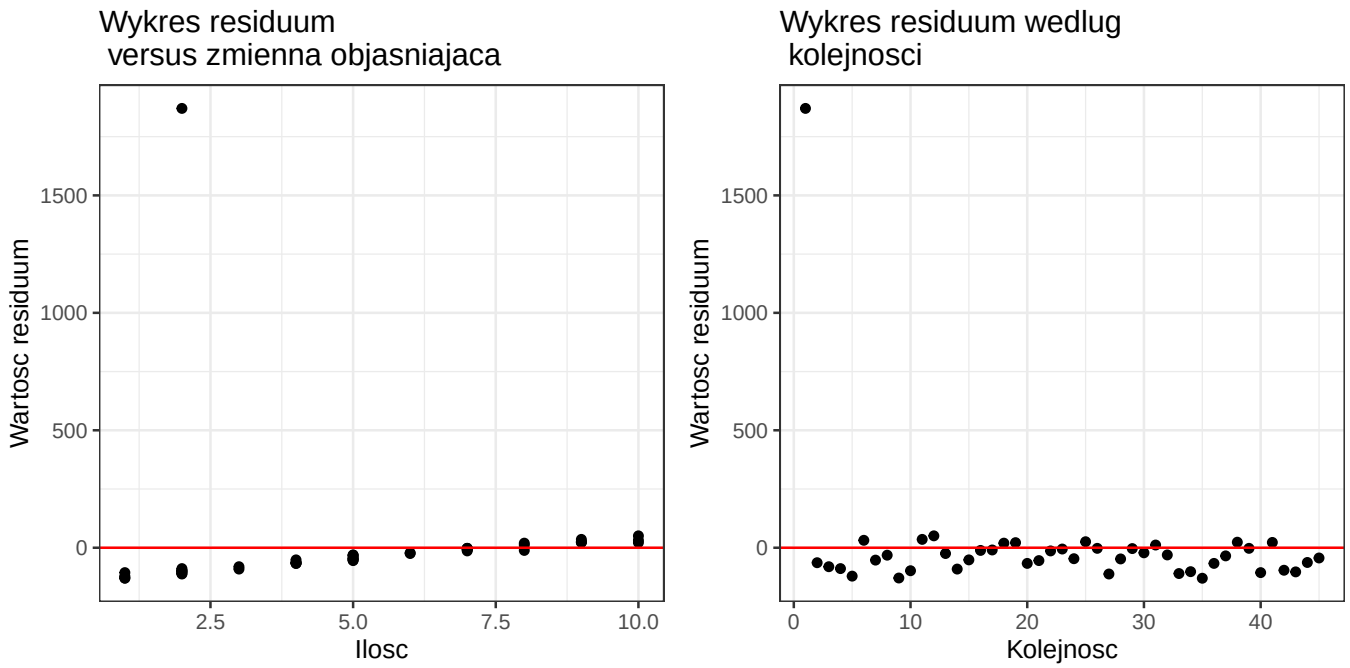
a)

Chcemy wyznaczyć równanie prostej regresji i inne parametry dla naszego modelu. Wszystkie poniższe parametry wyznaczamy znanymi nam już metodami.

	Pierwotne dane	Zmodyfikowane dane
Wyestymowane równanie	$15.035X - 0.58$	$-3.059X + 135.9$
Statystyka testowa T dla slope'a	31.1	-0.193
p-wartość	4.01×10^{-31}	0.848
R^2	0.957	0.001
Estymator σ^2	79.451	85759

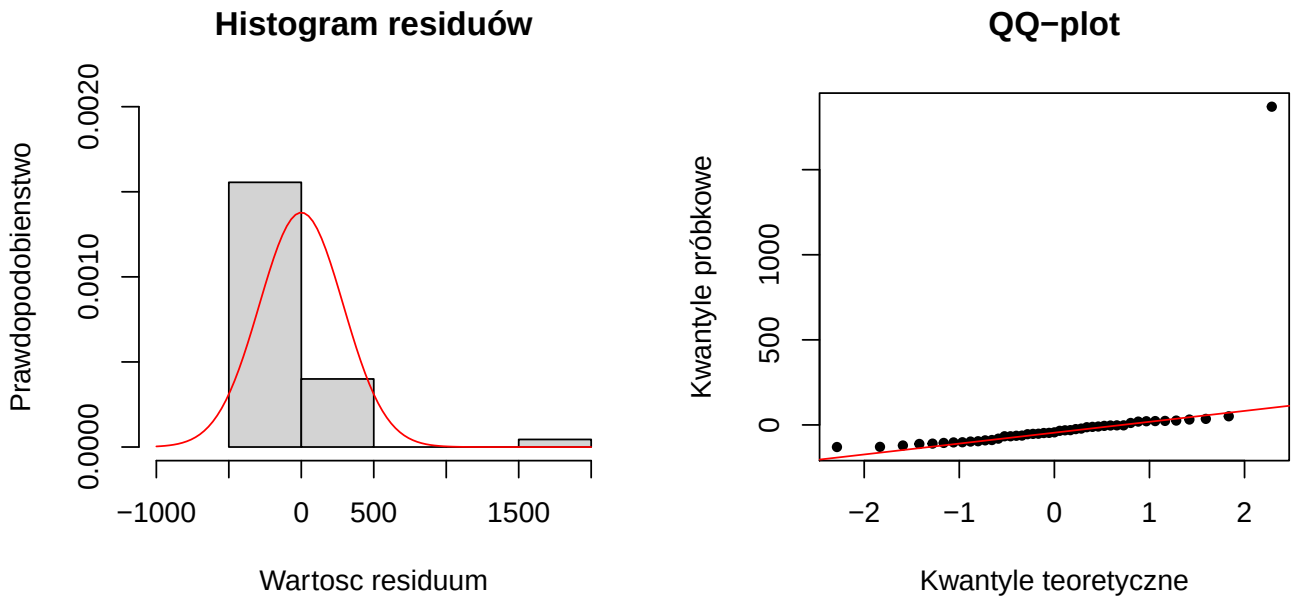
Widzimy, że wszystkie wielkości dla zmodyfikowanych danych drastycznie odstają od tych dla naszych wyjściowych danych. Wskazuje to na to, że rozważanie takiego modelu nie ma sensu. Gdybyśmy rozważyli test sprawdzający, czy $\beta_1 = 0$, to patrząc na p-wartość, dla każdego „wiarygodnego" poziomu istotności test nie odrzuciłby hipotezy zerowej.

b) W tym podpunckie dla naszych zmodyfikowanych danych powtórzymy rozważania z podpunktów (b), (c) i (d) z zadania nr 5.



Rysunek 5: Wykresy residuum według kolejności i dla zmiennej objaśniającej

W dwóch powyższych wykresach widzimy jakąś strukturę, czyli nasze residua są zależne w jakiś sposób od siebie.



Rysunek 6: Histogram i QQ-plot residuów

Wnioskując z histogramu i wykresu kwantylowo-kwantylowego możemy wnioskować, że residua nie mają rozkładu normalnego.

Podsumowując, tak zmodyfikowany zbiór nie pozwala nam rozważać modelu liniowego, ponieważ łamane są założenia o błędach losowych. Nie mają one rozkładu normalnego i nie są niezależne.

W następnych sześciu zadaniach będziemy pracować na tabeli danych **ch03pr15.txt**, która w pierwszej kolumnie zawiera informacje o stężeniu roztworu, a w drugiej o czasie.

7 Zadanie 7

Tworzymy model liniowy, gdzie zmienną objaśnianą jest stężenie roztworu, a zmienną objaśniającą czas. Chcemy podsumować nasz model w tym celu wyznaczymy estymowaną prostą regresji, obliczymy R^2 i wykonamy test statystyczny, aby sprawdzić, czy $\beta_1 = 0$.

Nasza prosta regresji wygląda następująco:

$$Y_i = 2.575 - 0.324 X_i.$$

Wartość R^2 wynosi 0.812.

Wykonamy teraz test statystyczny na poziomie istotności $\alpha = 0.05$

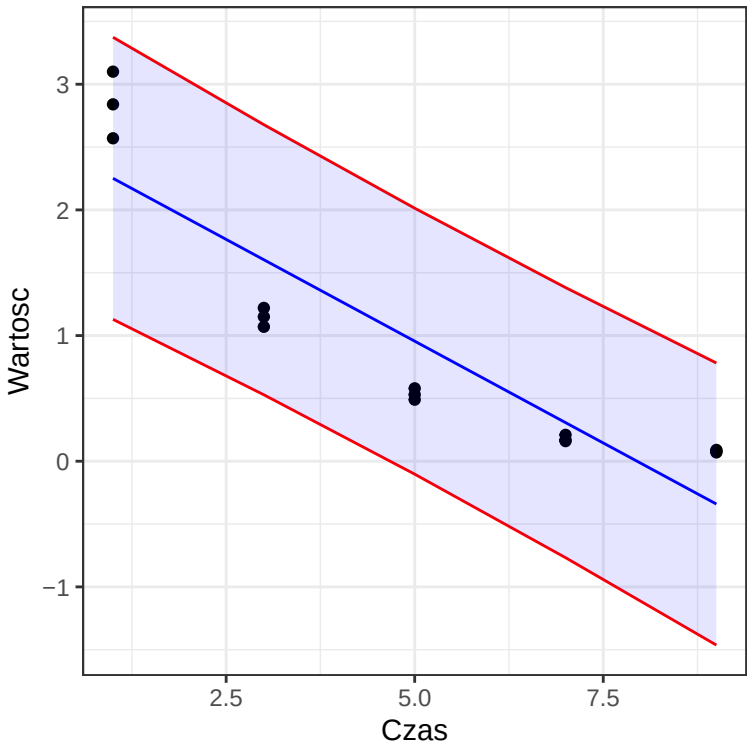
$$H_0 : \beta_1 = 0 \quad \vee \quad H_1 : \beta_1 \neq 0.$$

Kwantyl jest równy $F_c = F^*(1-\alpha, 1, n-2) = 4.667$, gdzie $\alpha = 0.05$. Nasza statystyka testowa wynosi $F = 55.994$. Stąd $F > F_c$, czyli możemy odrzucić hipotezę zerową. Natomiast p-wartość wynosi 4.61×10^{-6}.

Możemy, więc powiedzieć, że nasz model jest dobrze dopasowany, czyli stężenie roztworu zależy od czasu.

8 Zadanie 8

W tym zadaniu chcemy narysować wykres stężenia względem czasu. Następnie dopasowaną wcześniej prostą regresji i 95% pasmo predykcyjne. Na koniec chcemy wyznaczyć współczynnik korelacji między zaobserwowaną, a przewidzianą wartością stężenia roztworu.



Rysunek 7: Wykres stężenia od czasu wraz z 95% pasmem predykcyjnym

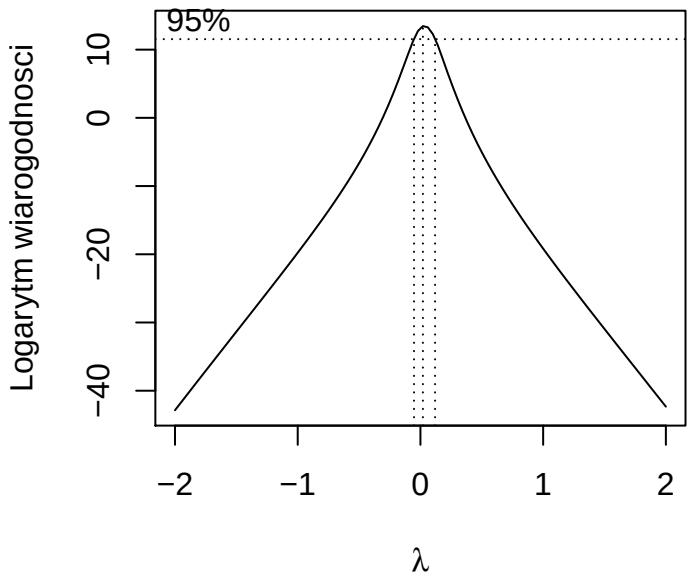
Widzimy, że nasze pasmo jest dość szerokie i wszystkie nasze obserwacje się w nim znajdują.

Natomiast współczynnik korelacji wynosi 0.901. Jest on dość blisko 1, czyli zmienne zależą od siebie w jakimś stopniu.

9 Zadanie 9

W tym zadaniu używając metody Boxa-Coxa chcemy znaleźć najlepsze przekształcenie stężenia roztworu.

W tym celu użyjemy funkcji w R **boxcox**.



Rysunek 8: Wykres funkcji boxa-coxa

Z wykresu widzimy, że $\lambda \approx 0$. W takim przypadku nasze przekształcenie jest postaci: $\tilde{Y} = \frac{Y^{\lambda}-1}{\lambda}$. Co w granicy daje nam logarytm.

10 Zadanie 10

W poprzednim zadaniu wywnioskowaliśmy, że najlepszą transformacją naszej zmiennej objaśnianej będzie logarytm. Dlatego teraz stworzymy nową zmienną $logy = \log Y$, gdzie Y to stężenie roztworu. Następnie powtórzymy rozważania z zadania nr 7 i 8.

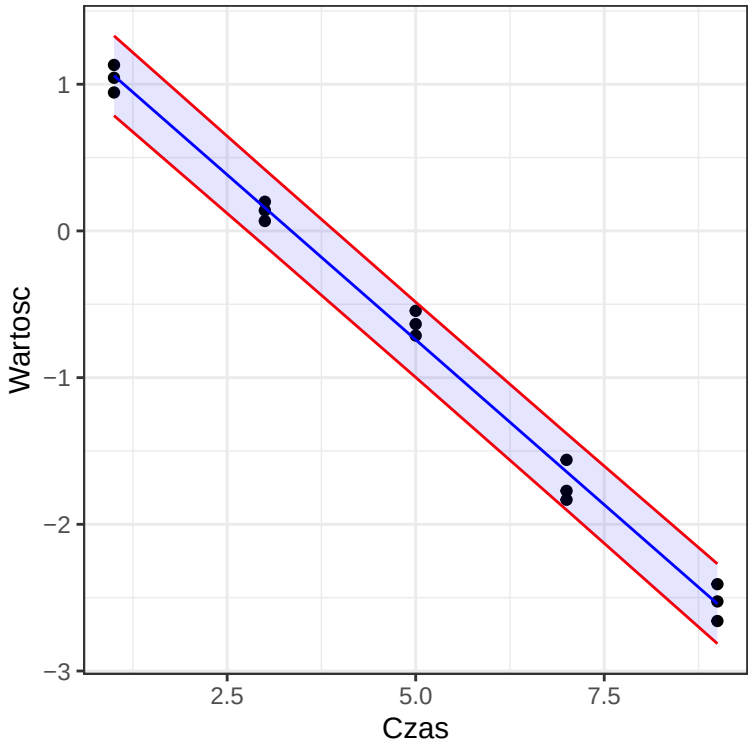
Wyniki przedstawię w tabeli:

Podsumowanie modelu	
Wyestymowane równanie	$-0.45X + 1.508$
Statystyka testowa F dla slope'a	1840
p-wartość	2.19×10^{-15}
R^2	0.993

Tabela 2: Tabela z analizą modelu

Obserwując powyższe wielkości widzimy, że nasz model jest teraz jeszcze lepiej dopasowany R^2 jest bliższe 1 i p-wartość jest bardzo mała. Patrząc na p-wartość test istotności β_1 prawie zawsze odrzuci hipotezę zerową.

Następnie narysujemy 95% procentowe pasmo predykcyjne



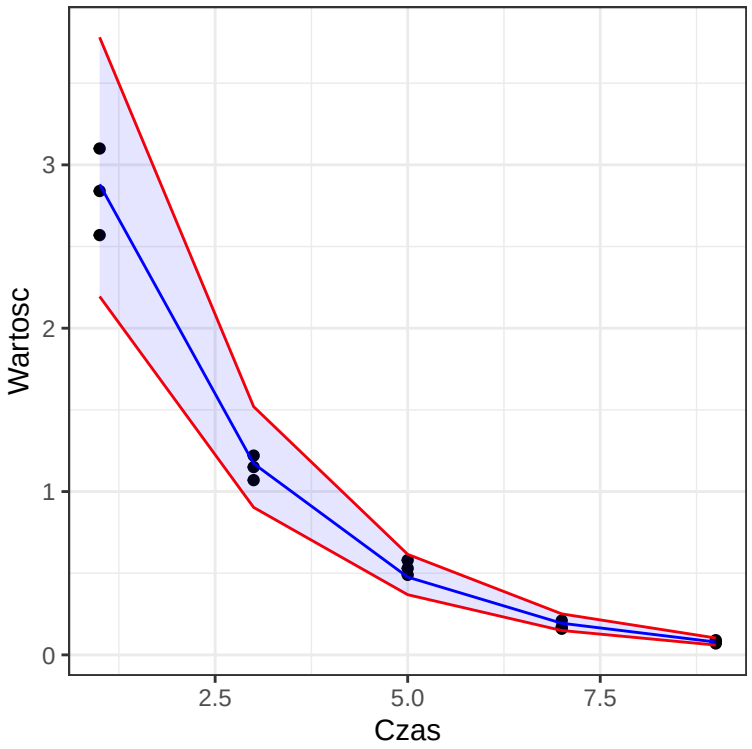
Rysunek 9: Wykres logarytmu stężenia od czasu wraz z 95% pasmem predykcyjnym

Widzimy, że po przekształceniu pasmo predykcyjne jest węższe, niż przed przakształceniem. Natomiast współczynnik korelacji między zaobserwowanymi danymi, a tymi przewidzianymi wynosi 0.996.

Po przeanalizowaniu tego zadania widzimy, że model liniowy po przekształceniu jest lepiej dopasowany.

11 Zadanie 11

W tym zadaniu chcemy narysować wykres stężenia roztworu względem czasu. Jako, że w poprzednim zadaniu nałożyliśmy logarytm na naszą zmienną, więc musimy „wrócić" przekształceniem odwrotnym do naszej pierwotnej zmiennej. Następnie sprawdzimy, czy nasze wyniki się zmieniły w porównaniu z zadaniem nr 8.



Rysunek 10: Pasmo predykcyjne na poziomie 95%

Widzimy, że pasmo znacząco się zmieniło, ponieważ już nie ograniczają go proste tylko łamane i jest ono coraz węższe wraz ze wzrostem czasu. Natomiast współczynnik korelacji wynosi 0.995. Jest ona również większa od tej, którą policzyliśmy w zadaniu nr 7.

12 Zadanie 12

W ostatnim zadaniu chcemy stworzyć nową zmienną $t1 = time^{-1/2}$. Następnie powtórzymy rozważania z zadania 10 i 11.

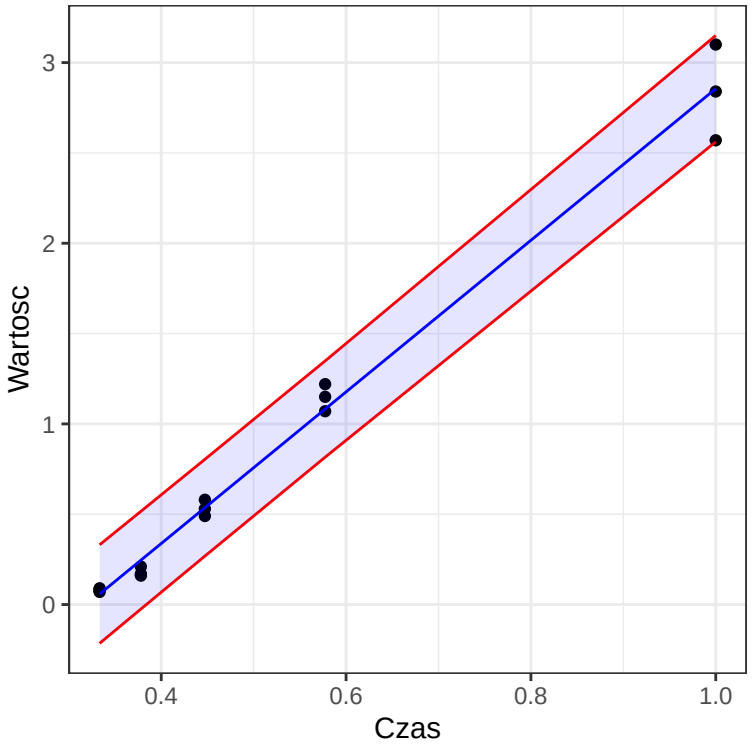
Nasze wyniki przedstawimy w tabeli:

Podsumowanie modelu	
Wyestymowane równanie	$4.196X - 1.341$
Statystyka testowa F dla slope'a	1080
p-wartość	6.9×10^{-14}
R^2	0.988

Tabela 3: Tabela z analizą modelu

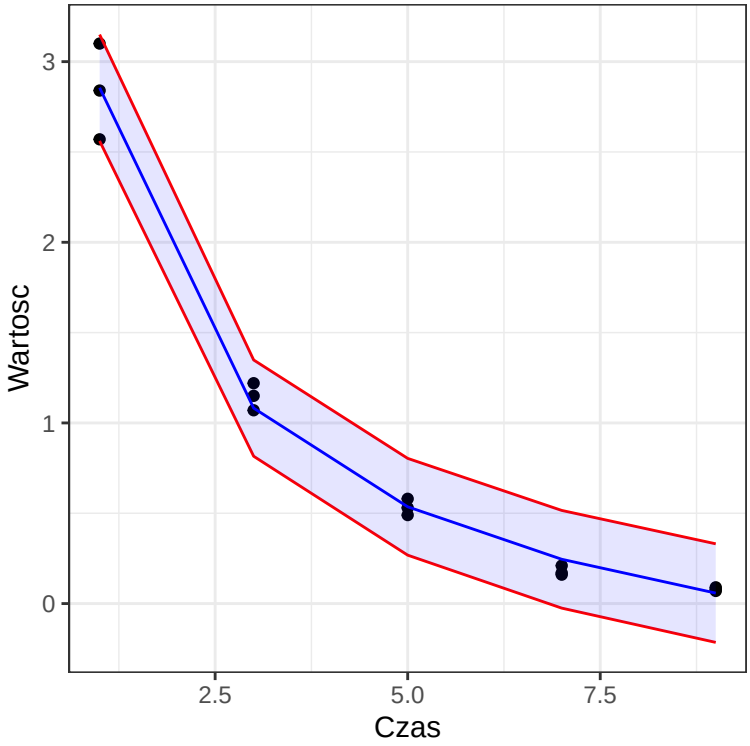
Wnioskując z p-wartości test istotności β_1 prawie zawsze odrzuci hipotezę zerową.

Rysujemy 95% pasmo predykcyjne dla $t1$ i stężenia.



Rysunek 11: Pasmo predykcyjne dla $t1$ i stężenia

Korelacja w tym przypadku wynosi 0.994. Następnie sprawdzamy jak zachowa się pasmo, gdy „wrócimy" do wyjściowego czasu.



Rysunek 12: Pasmo predykcyjne dla czasu i stężenia

Podobnie jak przy pierwszym przekształceniu widzimy, że pasma się zmieniają (stają się węższe od tego z zadania nr 7), czyli transformacja nieco ulepszyła nasz model. Natomiast współczynnik korelacji w tym przypadku wynosi 0.994.

Podsumowanie

Gdy spojrzymy na wszystkie rozważania nie możemy jasno stwierdzić, które przekształcenie jest najlepsze. W pyrzpadku przekształcenia naszej zmiennej Y na logarytm z niej, to wartość R^2 jest większa i częściej odrzucamy hipotezę zerową ($\beta_1 = 0$), ale przedziały predykcyjne zmieniają szerokość. Natomiast w przypadku drugiego przekształcenia mamy lepsze przedziały predykcyjne. Prawie wszędzie mają tą samą szerokość.