Modele liniowe Lista 3

Dominika Ochalik

2023-12-06

SPIS TREŚCI

Wstęp

Ten raport dotyczy analizy, jak dobrze dopasowana do danych jest prosta regresji. Analiza skupia się wokół obliczania współczynnika determinacji i korelacji, testu F, analizy wykresów, badania zachowania residuów oraz prób ulepszania modelu tak, aby uzyskać liniową zależność między danymi.

Podstawowe elementy teorii

Współczynnik R^2

Współczynnik determinacji R^2 określa, jaka część zmienności danych jest wyjaśniona przez model. Im większa wartość współczynnika, tym prosta regresji jest lepiej dopasowana do danych.

R^2 wyraża się wzorem:

$$R^2 = \frac{SSM}{SST}$$

gdzie:

- SSM - zmienność wyjaśniona przez model opisana wzorem:

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

- SST - zmienność całkowita opisana wzorem:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Test F

Test F sprawdza, czy istnieje liniowa zależność między zmiennymi X i Y. Problem sprowadza się do sprawdzenia, czy b_1=0.

$$H_0 : b_1 = 0$$

Statystyka testowa F jest postaci:

$$F = \frac{MSM}{MSE}$$

Odrzucamy H_0, gdy F>F_c, gdzie F_c jest kwantylem rzędu $1 - \alpha$ z rozkładu Fishera-Snedecora z dfM = 1 i dfE = n-2 stopniami swobody.

Transformacja Box'a - Cox'a.

Jest to procedura używana w sytuacjach, gdy nie mamy liniowej zależności pomiędzy zmiennymi X i Y. Chcemy nałożyć pewną funkcję na Y tak, aby otrzymać liniową zależność.

Otrzymujemy model postaci:

$$\tilde{Y} = b_0 + b_1 X + \epsilon$$

gdzie $\tilde{Y} = Y^{\lambda}$ lub $\tilde{Y} = \frac{Y^{\lambda}-1}{\lambda}$, gdy $\lambda \approx 0$.

Uwaga: granicznym przekształceniem wyrażenia $\tilde{Y} = \frac{Y^{\lambda}-1}{\lambda}$ *jest log(Y).*

Wartość parametru λ estymujemy metodą największej wiarogodności.

Zadanie 1: Badanie zależności GPA od IQ.

a) Równanie prostej regresji i współczynnik R^2.

W tym podpunkcie chcemy opisać zależność między GPA oraz wynikiem testu IQ. W tym celu znajdziemy równanie prostej regresji.

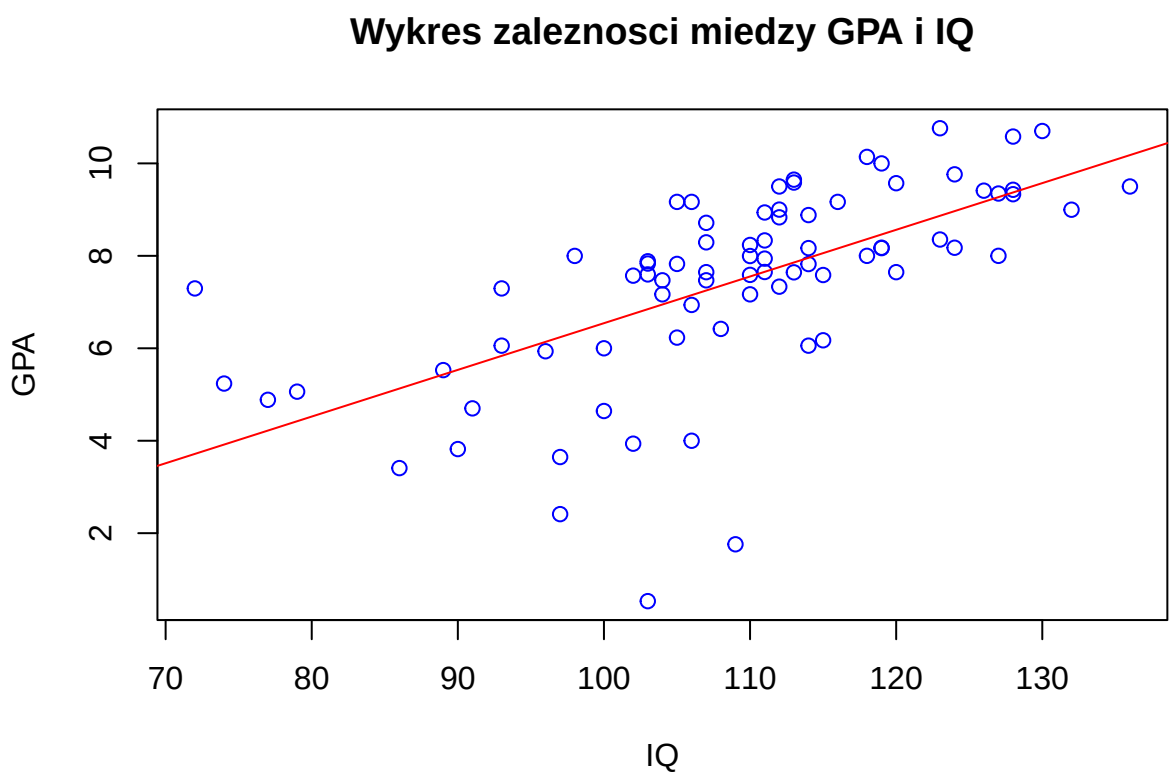
Równanie prostej regresji jest postaci:

$$Y = b_0 + b_1 X$$

gdzie:

- X to wartość wyniku testu IQ,
- Y to wartość GPA,
- b_0 wynosi -3.5571,
- b_1 wynosi 0.101.

Zobaczmy dane i prostą regresji na wykresie:



Następnie obliczymy wartość współczynnika determinacji R^2 za pomocą wzorów teoretycznych i poleceń wbudowanych w R.

Wartość obliczona za pomocą wzorów teoretycznych wynosi 0.4014, natomiast wartość obliczona korzystając z funkcji wbudowanych w R wynosi 0.4016.

Na podstawie współczynnika R^2 możemy wywnioskować, jak dobrym przybliżeniem wartości i zachowania danych jest prosta regresji. Im większa wartość tego współczynnika, tym dopasowanie jest lepsze. W tym

przypadku wartość R^2 wynosi 0.4014. Biorąc pod uwagę analizę wykresu zależności między GPA i IQ możemy wywnioskować, że dane są dosyć szeroko oddalone od prostej regresji (mają dużą wariancję), co wpływa na to, że prosta regresji nie jest ich dobrym przybliżeniem.

b) Test F.

Skoro ustalono, że prosta regresji nie jest dobrym przybliżeniem zależności między GPA i IQ, możemy przetestować hipotezę mówiącą, że GPA nie jest skorelowane z IQ. Zrobimy to za pomocą testu F.

Testujemy hipotezę zerową:

$$H_0 : b_1 = 0$$

W tym celu najpierw liczymy wartość statystyki testowej F:

$$F = \frac{MSM}{MSE}$$

Przyjmuje ona wartość F = 50.9862. Kwantyl $F^*(1-\alpha, 1, n-2)$ z rozkładu Fishera-Snedecora wynosi w przybliżeniu 4. Widzimy, że F $> F^*$, zatem odrzucamy hipotezę zerową. Ten sam wniosek możemy uzyskać, analizując p-wartość, czyli prawdopodobieństwo uzyskania takich wartości, jak F, i tych jeszcze mniej prawdopodobnych.

p-wartość możemy obliczyć za pomocą wyrażenia: 1- P(z>F), gdzie z jest zmienną losową z rozkładu Fishera-Snedecora z 1 i n-2 stopniami swobody.

Wynosi ona: 5×10^{-10}.

Wartość statystyki F, jak i p-wartość, możemy obliczyć, korzystając z poleceń wbudowanych w R.

Otrzymujemy:

- F = 51.0085,
- p-wartość = 5×10^{-10}.

Widzimy, że p-wartość jest bardzo mała, a w szczególności mniejsza od α=0.05, co pozwala nam odrzucić H_0.

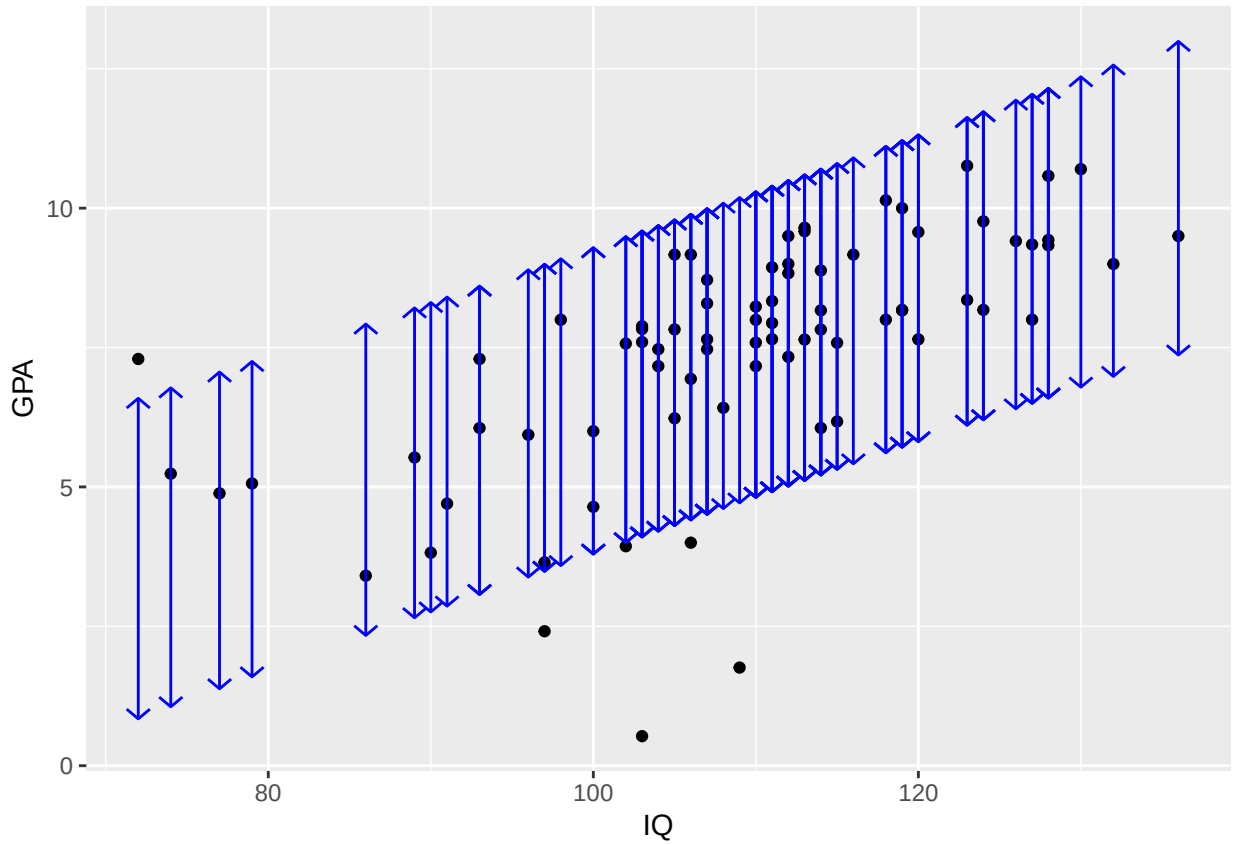
c) Przedziały predykcyjne dla IQ $\in$ {75, 100, 140}.

W tym podpunkcie należy przewidzieć wartość GPA dla uczniów, których IQ wynosi 75, 100, 140. W tym celu trzeba podać 90% przedziały predykcyjne.

IQ	wartość oczekiwana GPA	lewy kraniec przedziału	prawy kraniec przedziału	Długość przedziału
75	4.0179	1.1642	6.8716	5.7074
100	6.5429	3.7953	9.2905	5.4952
140	10.5829	7.7473	13.4185	5.6712

d) Przedziały predykcyjne dla wszystkich obserwacji.

W tym podpunkcie chcemy dodać do wykresu z danymi 90% przedziały predykcyjne.



Widzimy, że tylko 6 obserwacji nie należy do przedziałów predykcyjnych. Istnienie takich obserwacji wynika z faktu, że przedziały predykcyjne są konsruowane na poziomie ufności 90%, zatem spodziewamy się, że ok. 90% obserwacji wpadnie do przedziałów. W naszym przypadku 0.0769% obserwacji nie należy do przedziałów predykcyjnych, czyli ponad 90% obserwacji należy.

Zadanie 2: Badanie zależności GPA od PH.

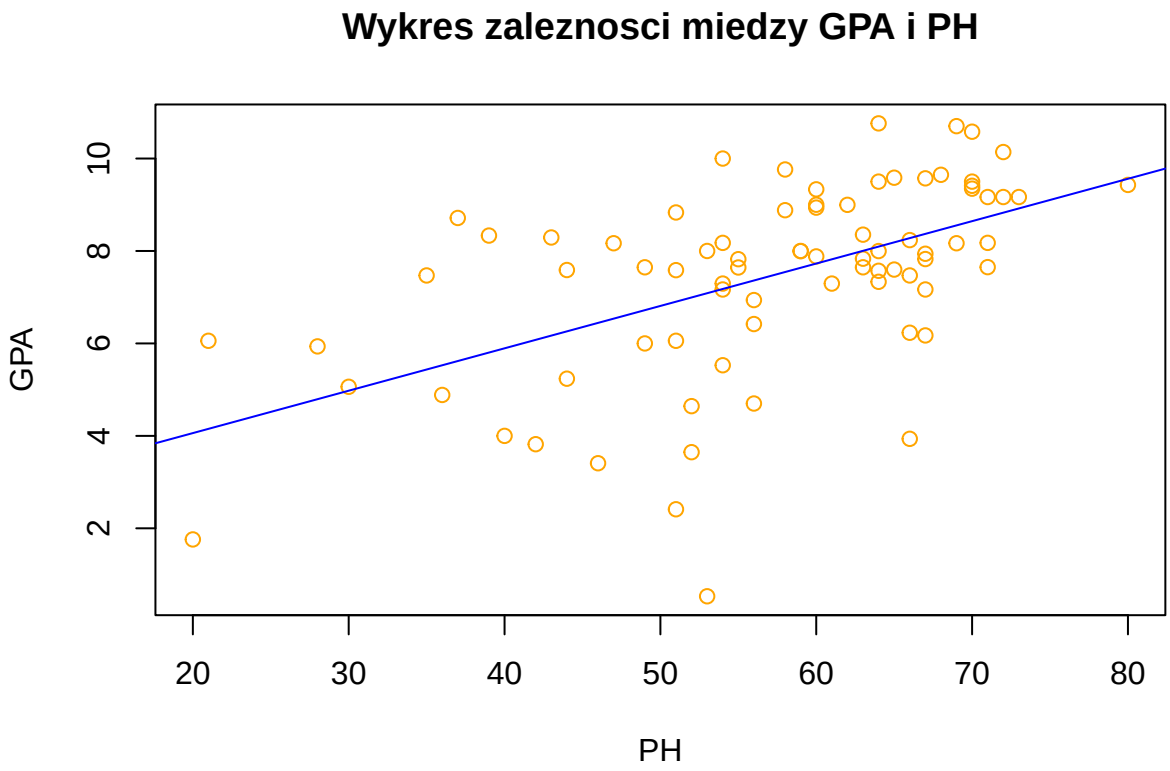
a) Prosta regresji i współczynnik R^2.

W tym zadaniu będziemy badać zależność między GPA oraz wyników testu HP. W celu znalezienia równania prostej regresji, najpierw znajdziemy wartości współczynników b_0 i b_1.

Wynoszą one:

- $b_0 = 2.2259$,
- $b_1 = 0.0917$.

Zobaczmy dane na wykresie wraz z dorysowaną prostą regresji:



Następnie obliczymy wartość współczynnika determinacji R^2 za pomocą wzorów teoretycznych i poleceń wbudowanych w R.

Wartość obliczona za pomocą wzorów teoretycznych wynosi 0.2939, natomiast wartość obliczona korzystając z funkcji wbudowanych w R wynosi 0.2936. Widzimy, że jest ona jeszcze mniejsza niż w przypadku analizy zależności między GPA i IQ. Ponadto, z wykresu możemy wywnioskować, że punkty są szeroko oddalone od prostej regresji (mają dużą wariancję).

b) Test F.

Skoro ustalono, że prosta regresji nie jest dobrym przybliżeniem zależności między GPA i PH, możemy przetestować hipotezę mówiącą, że GPA nie jest skorelowane z PH. Zrobimy to za pomocą testu F.

Testujemy hipotezę zerową:

$$H_0 : b_1 = 0$$

W tym celu najpierw liczymy wartość statystyki testowej F: $F = \frac{MSM}{MSE}$

Przyjmuje ona wartość F = 31.6178. Kwantyl $F^*(1-\alpha, 1, n-2)$ z rozkładu Fishera-Snedecora wynosi w przybliżeniu 4. Widzimy, że F > F^*, zatem odrzucamy hipotezę zerową.

Ten sam wniosek możemy uzyskać, analizując p-wartość, czyli prawdopodobieństwo uzyskania takich wartości, jak F, i tych jeszcze mniej prawdopodobnych.

p-wartość możemy obliczyć za pomocą wyrażenia: 1- P(z>F), gdzie z jest zmienną losową z rozkładu Fishera-Snedecora z 1 i n-2 stopniami swobody.

Wynosi ona: 2.971×10^{-7}.

Wartość statystyki F, jak i p-wartość, możemy obliczyć, korzystając z poleceń wbudowanych w R.

Otrzymujemy:

- F = 31.5852,
- p-wartość = 3.006×10^{-7}.

Widzimy, że p-wartość jest bardzo mała, a w szczególności mniejsza od α=0.05, co pozwala nam odrzucić H_0.

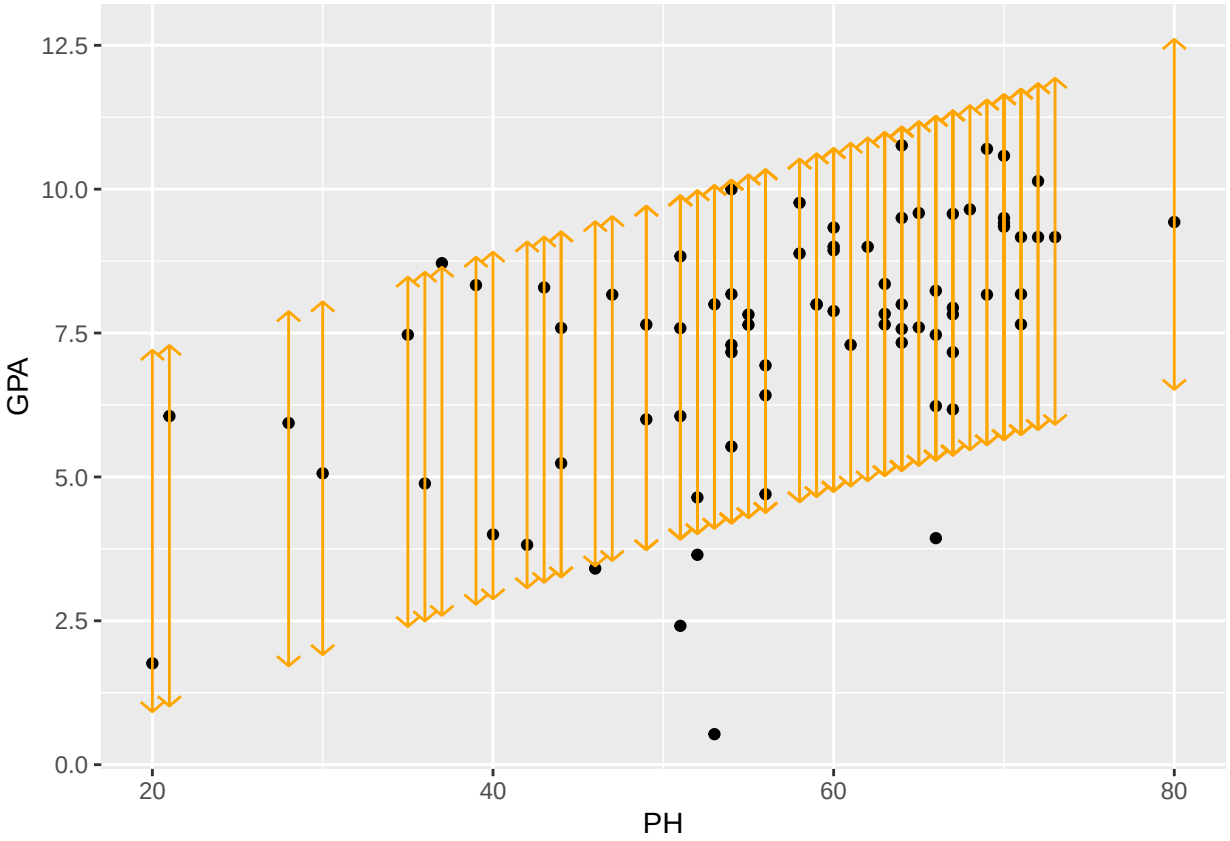
c) Przedziały predykcyjne dla PH ∈ {25, 55, 85}.

W tym zadaniu chcemy przewidzieć wynik GPA dla uczniów, których wyniki testu PH wynoszą 25, 55 lub 85. W tym celu skonstruujemy 90% przedziały predykcyjne.

PH	wartość oczekiwana GPA	lewy kraniec przedziału	prawy kraniec przedziału	Długość przedziału
25	4.5184	1.4179	7.6189	6.201
55	7.2694	4.2923	10.2465	5.9542
85	10.0204	6.948	13.0928	6.1448

d) Przedziały predykcyjne dla wszystkich obserwacji.

W tym podpunkcie dodamy do wykresu z danymi 90% przedziały predykcyjne, a następnie ustalimy, ile obserwacji znajduje się poza granicami przedziałów.



Widzimy, że tylko 6 obserwacji nie należy do przedziałów predykcyjnych. Istnienie takich obserwacji wynika z faktu, że przedziały predykcyjne są konsruowane na poziomie ufności 90%, zatem spodziewamy się, że ok. 90% obserwacji wpadnie do przedziałów. W naszym przypadku 0.0769% obserwacji nie należy do przedziałów predykcyjnych, czyli ponad 90% obserwacji należy.

Która ze zmiennych: IQ czy PH jest lepszy predyktorem GPA?

Podsumowując wszystkie uzyskane do tej pory wnioski możemy stwierdzić, że wynik testu IQ jest lepszym predyktorem wartości GPA. Postawiona teza wynika z faktu, że dla IQ i GPA mamy większą wartość współczynnika R^2 oraz przedziały predykcyjne mają trochę mniejszą długość.

Zadanie 3

W zadaniu 3 korzystamy z danych z pliku ch01pr20.txt, który zawiera informacje na temat liczby kopiarek (druga kolumna) i czasu potrzebnego na ich utrzymanie (pierwsza kolumna).

a) Sprawdzenie, czy suma residuów jest równa 0.

Najpierw należy znaleźć równanie regresji liniowej. W tym celu potrzebne nam będą wartości estymatorów b_0 i b_1, które wynoszą odpowiednio:

- b_0 = -0.5802
- b_1 = 15.0352.

W równaniu regresji występują też zmienne X i Y, które w tym przypadku oznaczają odpowiednio liczbę kopiarek i czas.

Estymator Y, czyli $\hat{Y}$ liczymy ze wzoru

$$\hat{Y} = b_0 + b_1 \cdot X$$

dla obliczonych wyżej wartości b_0 i b_1.

Suma residuów to suma błędów, czyli wyrażenie postaci:

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

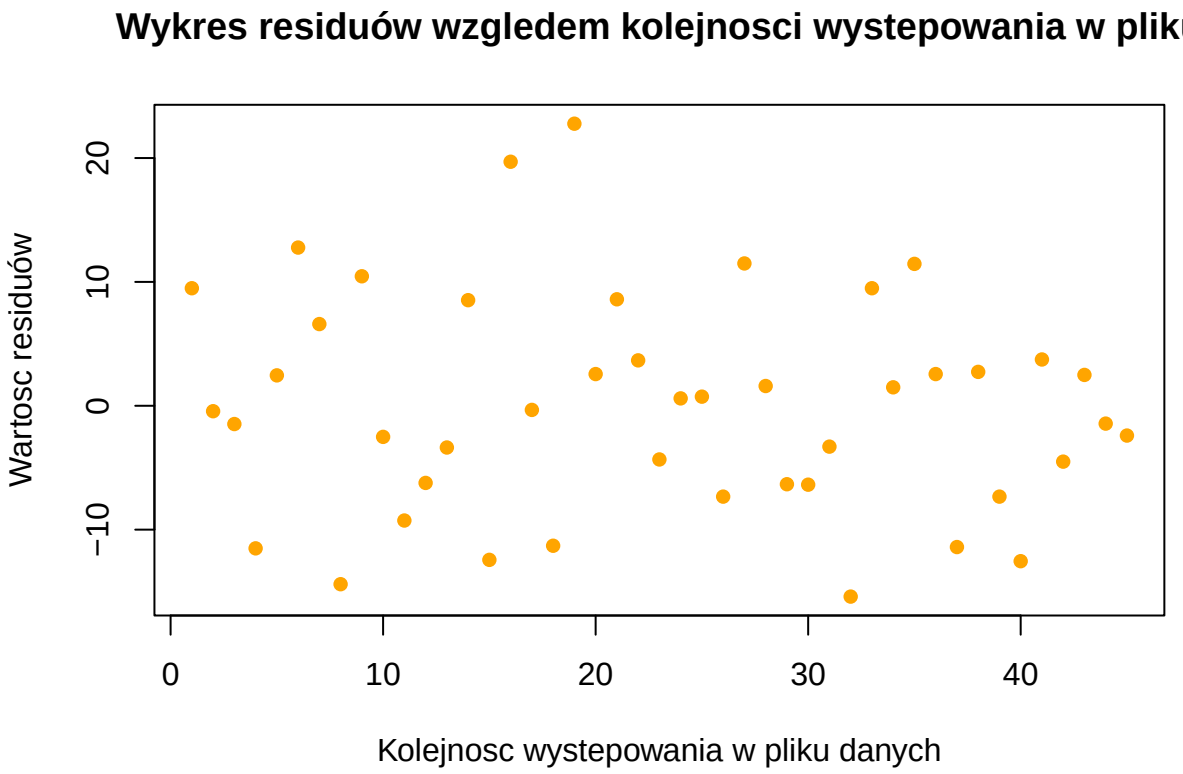
które wynosi -0.013. Widzimy, że jest różne od zera.

b) Wykres residuów względem zmiennej X.

W tym podpunkcie należy przedstawić wykres residuów względem zmiennej objaśniającej.

Większość wartości residuów mieści się w przedziale od -10 do 10 i są losowo rozrzucone w tym przedziale. Nie widać wyraźnej zależności pomiędzy kolejnymi punktami. Możemy zauważyć dwie obserwacje odstające, jedną dla X=8, a drugą dla X=10.

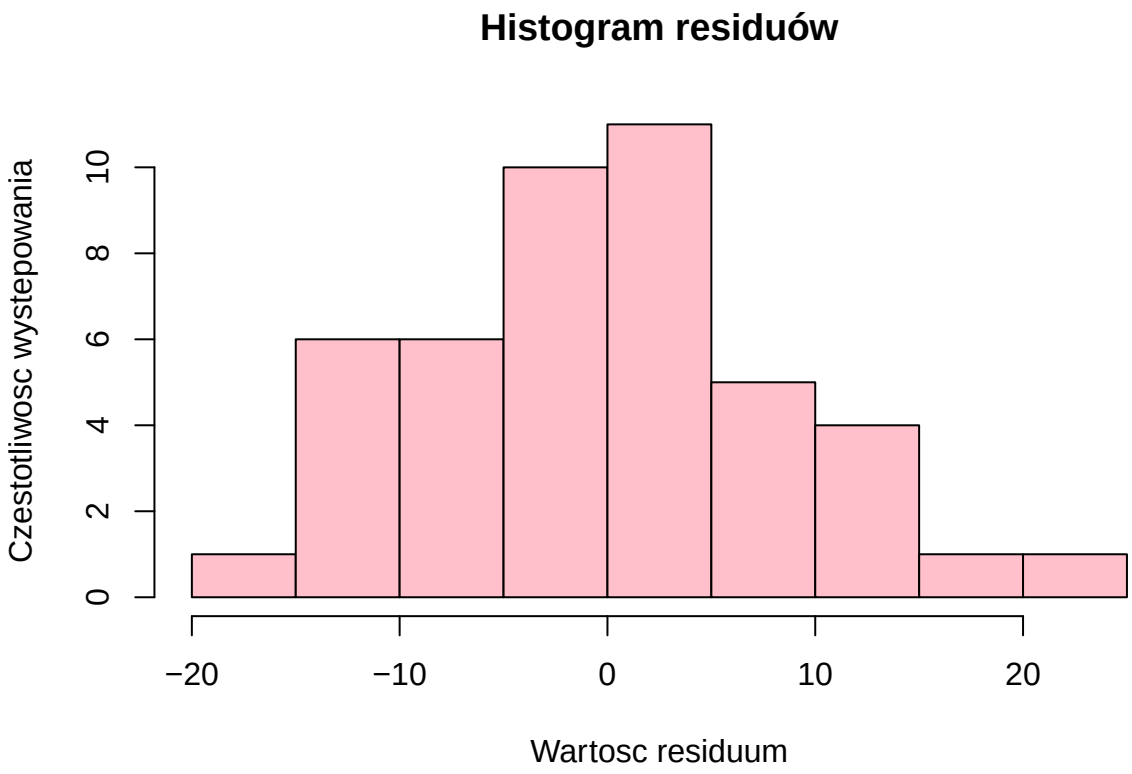
c) Wykres residuów względem kolejności występowania.

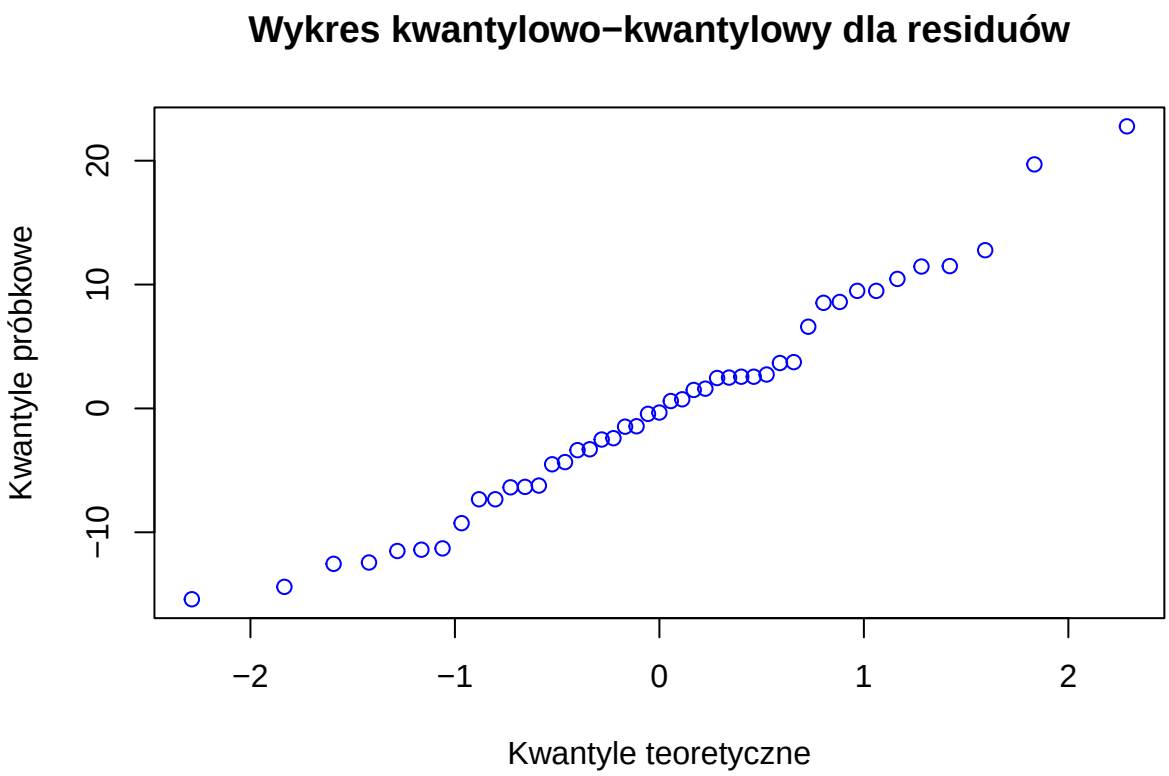


Nie widać zależności między kolejnymi wartościami błędów; wszystkie punkty zdają się być losowo rozrzucone między wartościami -10 i 10 (w większości przypadków). Możemy zauważyć, że są dwie obserwacje odstające, które osiagają wartość residuum blisko 20, a w pliku występują między 15 a 20 pozycją.

d) Badanie normalności rozkładu residuów.

W tym podpunkcie będziemy badać normalność rozkładu residuów za pomocą histogramu oraz wykresu kwantylowo-kwantylowego.


Histogram residuów
Czestotliwosc wystepowania
0
2
4
6
8
10
−20
−10
0
10
20
Wartosc residuum



Analizując histogram oraz wykres kwantylowo-kwantylowy możemy wywnioskować, że rozkład residuów jest w przybliżeniu normalny.

Zadanie 4: Dane z dodaną obserwacją (1000; 2).

a) Tabela porównująca.

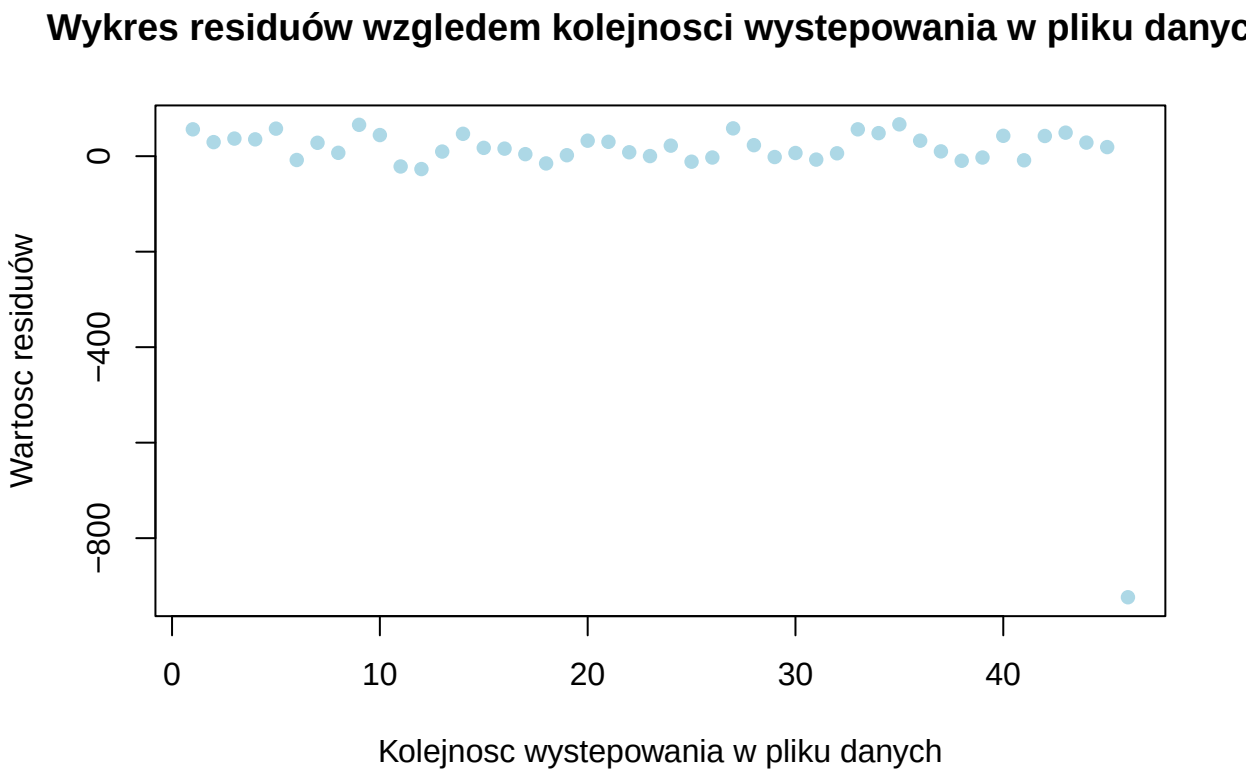
Parametry	Oryginalne dane	Dane z dodaną wartością
Równanie regresji	Y = -0.5802 + 15.0352X	Y = 63.0915 + 6.5939X
Wartość statystyki t	31.1233	0.8625
p-wartość	0	0.3931
R^2	0.9574955	0.0166256
Estymator wariancji	77.6449324	1.9997694×10^4

Widzimy, że dla danych z dodaną wartością współczynnik R^2 osiąga bardzo małą wartość, co oznacza, że prosta regresji nie przybliża dobrze wartości i zachowania danych.

b)

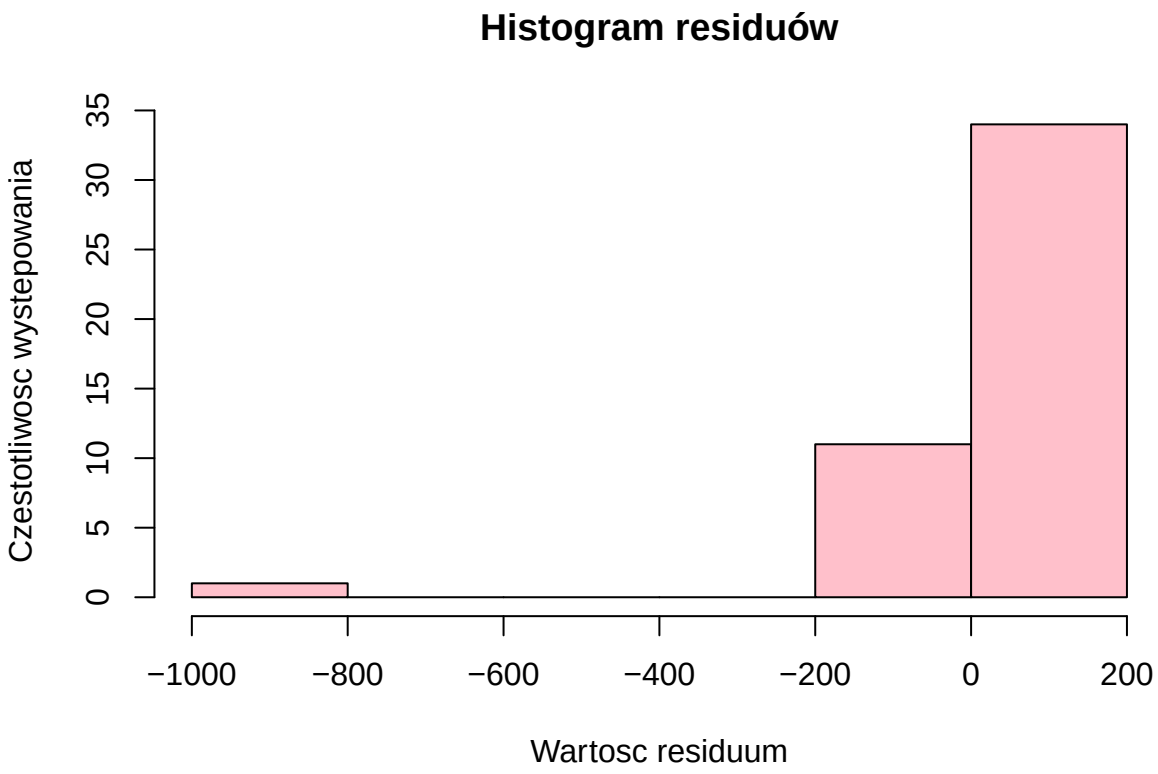
Teraz będziemy powtarzać podpunkty b), c) i d) z zadania 3 na zmodyfikowanym w tym zadaniu zbiorze danych.

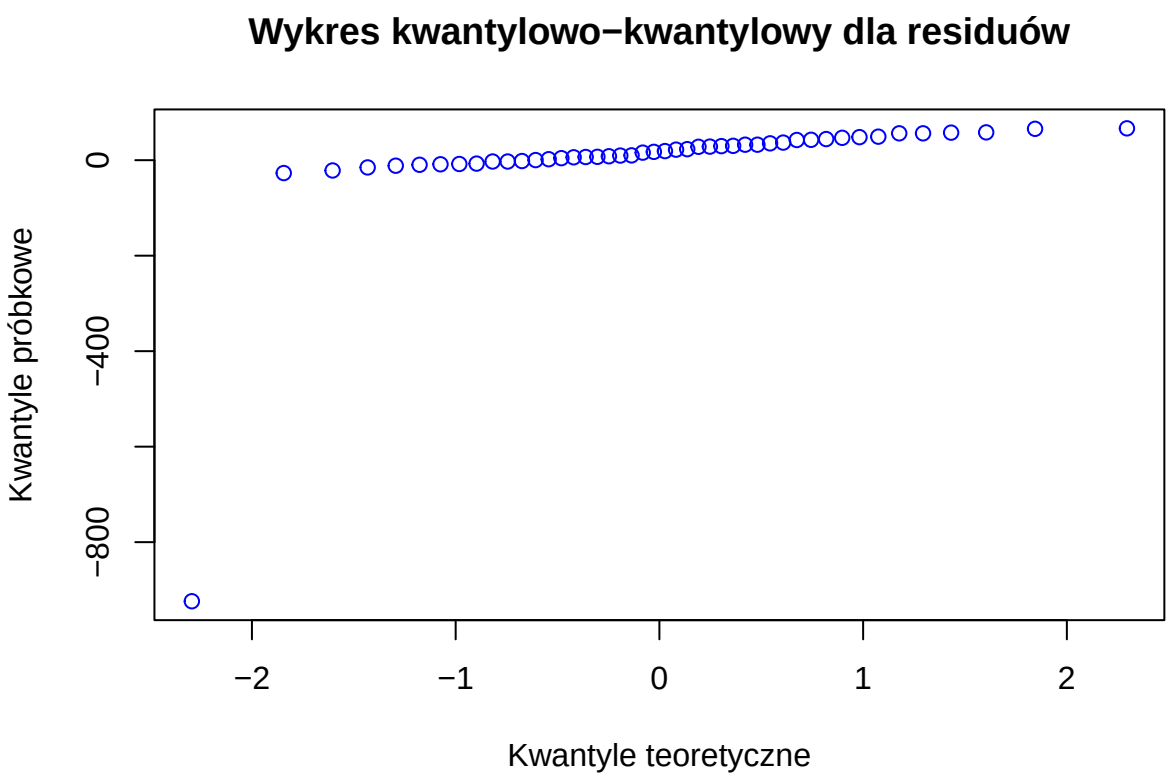
Widzimy, że dodana obserwacja bardzo wyróżnia się na powyższym wykresie ze względu na wartość odstającą zmiennej X (czasu). Implikuje to wyraźnie większą wartość residuum. Pozostałe obserwacje osiągają podobną wartość residuum.



Ponownie możemy dostrzec odstającą wartość residuum dla dodanej zmiennej. Co ciekawe, wartości błędów pozostałych obserwacji zdają się być od siebie zależne, ponieważ nie są rozmieszczone losowo, tylko układają się w kształt fali (językiem matematyka: w kształt sinusoidy).

Zbadajmy normalność wartości residuów:


Histogram residuów
Czestotliwosc wystepowania
0
5
10
15
20
25
30
35
−1000
−800
−600
−400
−200
0
200
Wartosc residuum

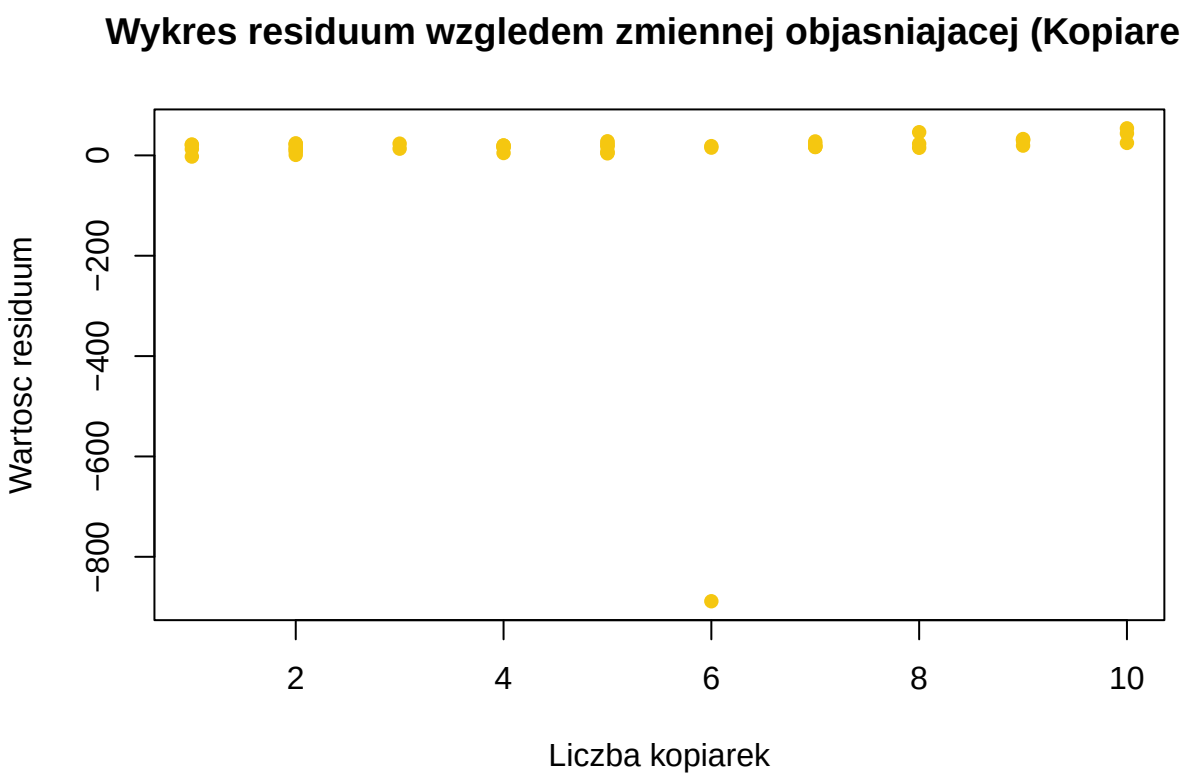


Widzimy, że histogram nie przypomina swoim kształtem krzywej Gaussa. Należy też zauważyć, że ze względu na obserwację odstającą nie da się wiele z niego odczytać.

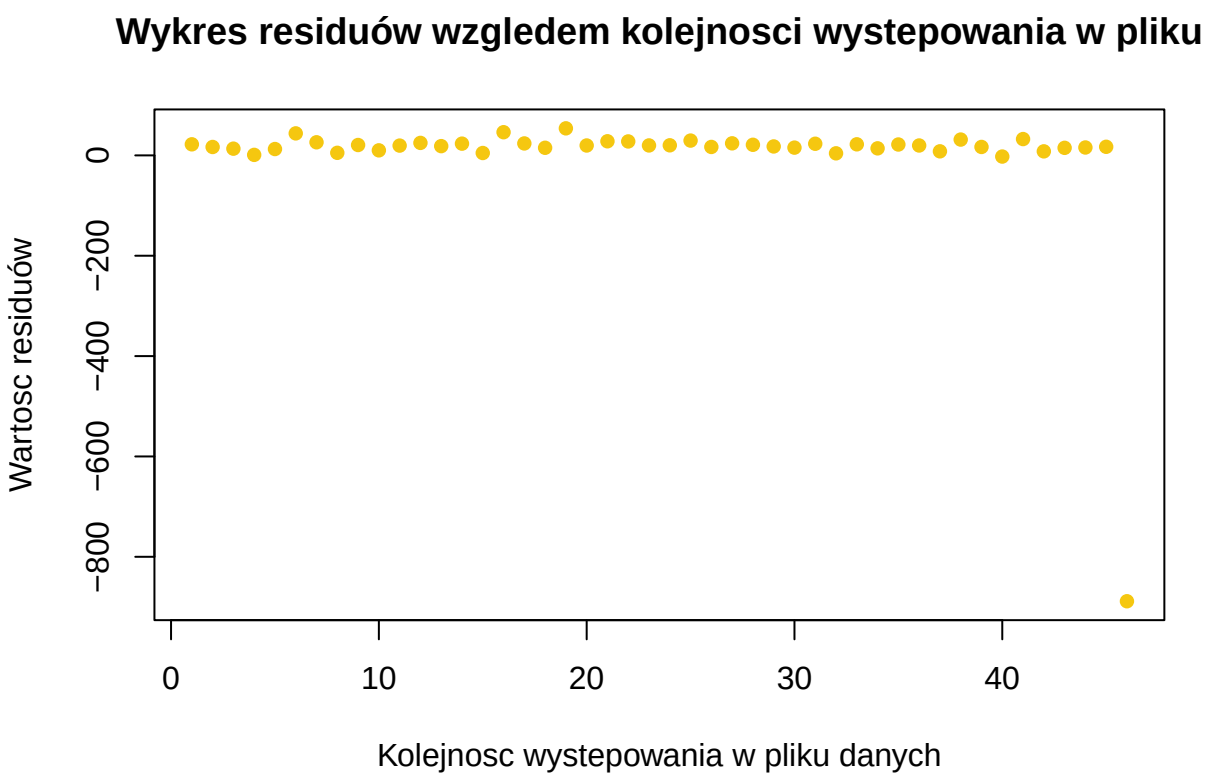
Punkty na wykresie kwantylowo-kwantylowym ze względu na obserwację odstającą sprawiają wrażenie, jakby bardziej układały się w linię prostą. Obserwacja odstająca jest bardzo charakterystyczna na wykresie (podobnie, jak na histogramie).

c) Dodanie obserwacji (1000; 6) do początkowego pliku.

Parametry	Oryginalne dane	Dane z dodaną wartością
Równanie regresji	Y = -0.5802 + 15.0352X	Y = 7.3078 + 17.3552X
Wartość statystyki t	31.1233	2.3594
p-wartość	0	0.3931
R^2	0.9574955	0.1123102
Estymator wariancji	77.6449324	1.805187×10^4

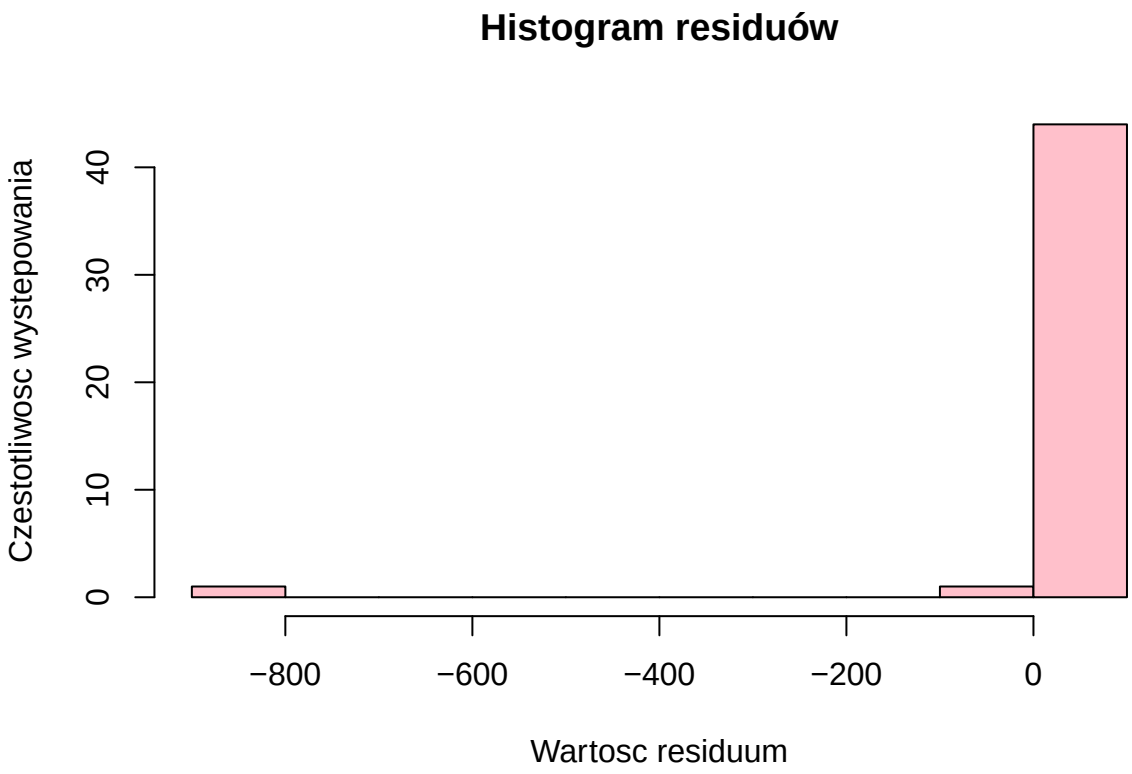


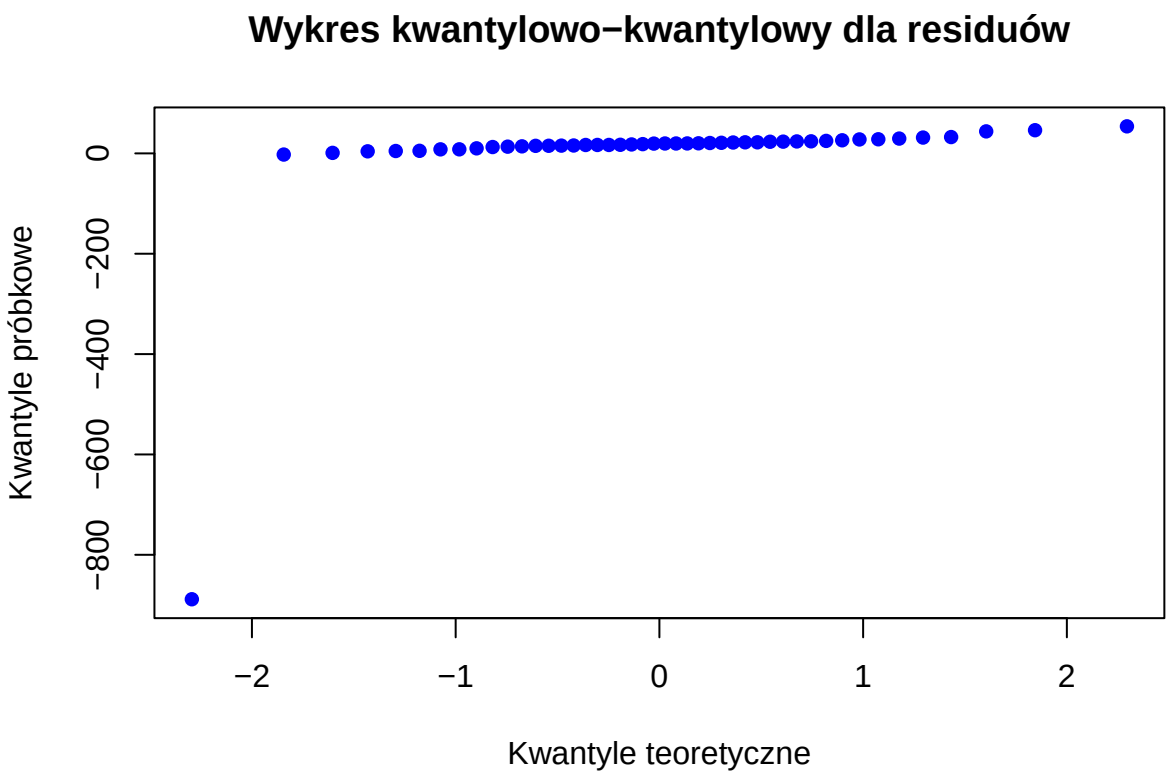
Dodana obserwacja wyróżnia się na wykresie ze względu na bardzo dużą wartość residuum. Pozostałe obserwacje mają zbliżone do siebie wartości residuów.



Dodana obserwacja znowu wyróżnia się na wykresie. Pozostałe punkty na wykresie mają zbliżone do siebie wartości residuów.

Zbadajmy normalność wartości resiuów:


Histogram residuów
Czestotliwosc wystepowania
0
10
20
30
40
−800
−600
−400
−200
0
Wartosc residuum



Histogram, ze względu na wartość odstającą, jest bardzo nieczytelny i niewiele można z niego odczytać. Na wykresie kwantylowo-kwantylowym widzimy, że wszystkie obserwacje (poza jedną odstającą) układają się w linię prostą.

Zadanie 5

W tym zadaniu będziemy pracować z danymi z pliku ch03pr15.txt. W pliku znajdują się dane na temat stężenia roztworu (pierwsza kolumna) i czasu (druga kolumna).

a) Równanie regresji i 95% przedziały predykcyjne.

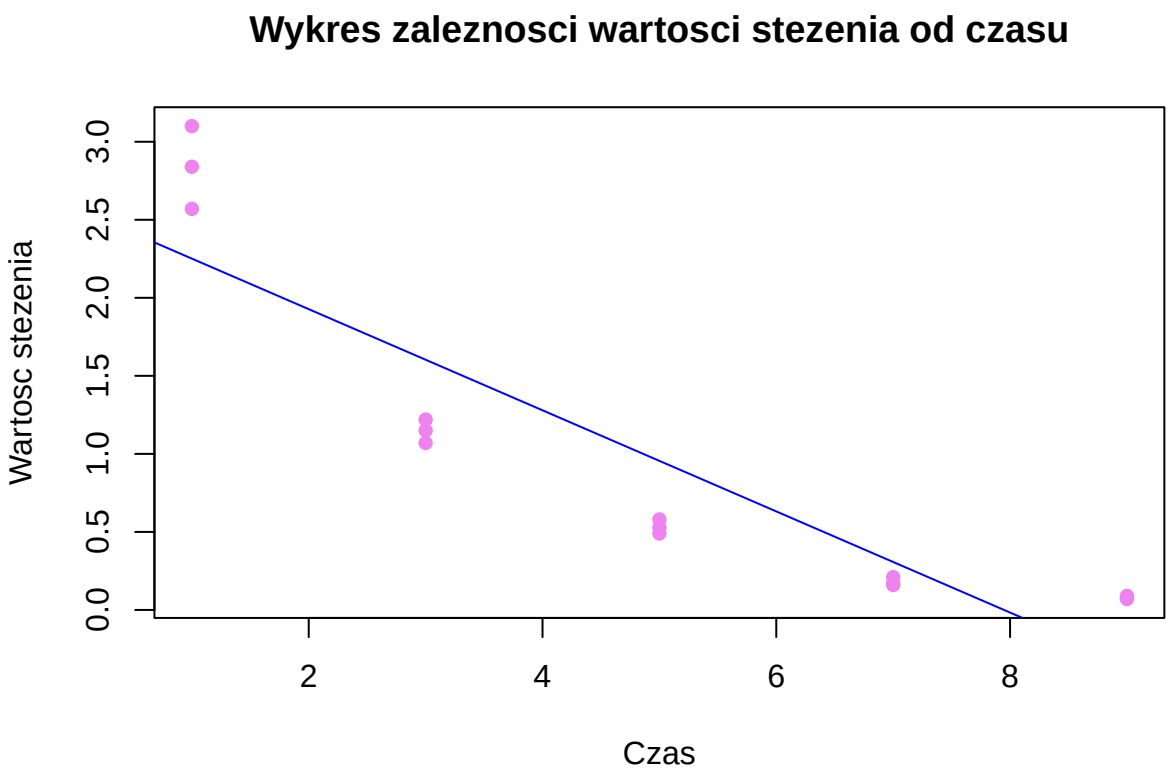
Aby znaleźć równanie regresji, potrzebujemy znaleźć wartości współczynników b_0 i b_1. Ponadto, stężenie roztworu będzie zmienną odpowiedzi (Y w równaniu), a czas zmienną objaśniającą (X).

Wartości współczynników b_0 i b_1 obliczone przy użyciu poleceń wbudowanych w R:

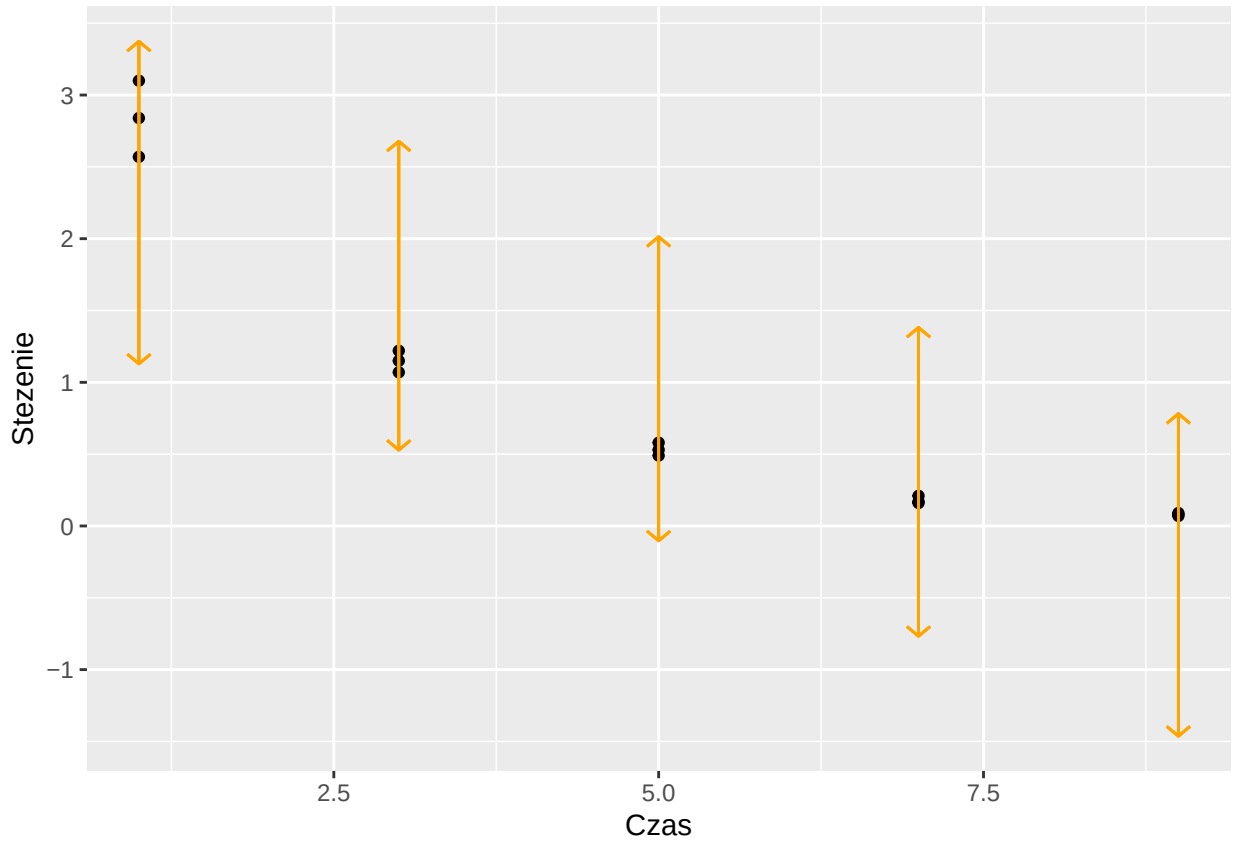
- b_0 = 2.5753,
- b_1 = -0.324.

Zatem równanie regresji wyraża się wzorem: Y = 2.5753 + -0.324X.

Zobaczmy dane wraz z prostą regresji na wykresie:



Dodajmy do wykresu 95% przedziały predykcyjne:



Widzimy, że wszystkie obserwacje mieszczą się w wyznaczonych przedziałach predykcyjnych, które są dosyć szerokie.

b) Analiza regresji: współczynnik R^2 i test istotności dla b_1.

W celu zbadania dokładności dopasowania prostej regresji do danych obliczymy współczynnik R^2.

Wynosi on: 0.8116. Widzimy, że jest dosyć duży, co sugeruje, że prosta regresji jest dobrze dopasowana do danych.

Zbadajmy, czy istnieje zależność między zmiennymi. W tym celu testujemy nastepującą hipotezę:

$$H_0 : b_1 = 0$$

Hipoteza alternatywna zaprzecza hipotezie zerowej, czyli jest postaci:

$$H_a : b_1 \neq 0$$

W celu przetestowania hipotez możemy użyć testu t-studenta albo testu Fishera. Z racji, że są równoważne, nie ma znaczenia, który test wybierzemy. Zbadajmy zatem hipotezę zerową przy pomocy testu F.

Tradycyjnie przyjmujemy α=0.05.

Statystyka testowa jest postaci:

$$F = \frac{MSM}{MSE}$$

Podstawiając, otrzymujemy: F = 55.9938.

Teraz sprawdzimy, czy wartość statystyki F jest większa od kwantyla $F_c = F^*(0.95, 1, 13)$, który wynosi 4.6672.

Widzimy, że F $> F_c$, zatem odrzucamy H_0 mówiącą, że b_1=0.

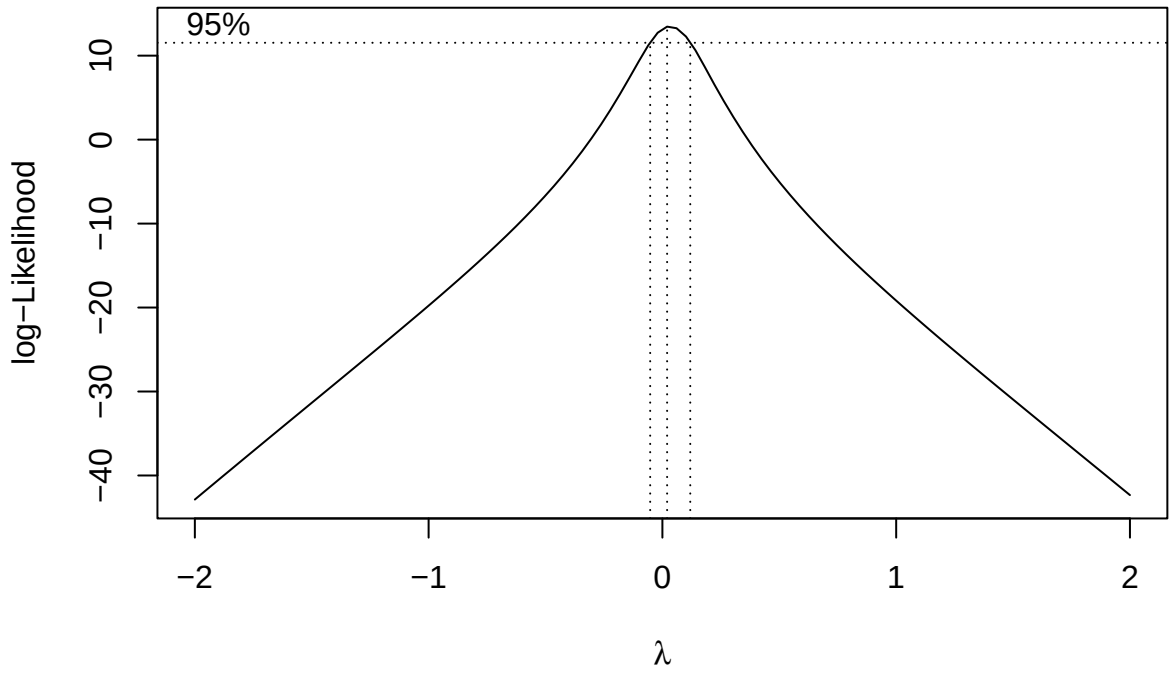
Ten sam wniosek możemy uzyskać analizując p-wartość, która wynosi 4.6112155×10^{-6}. Widzimy, że jest zdecydowanie mniejsza niż α=0.05, zatem odrzucamy H_0.

Wniosek: Z prawdopodobieństwem 95% wartość $b_1 \neq 0$, co oznacza, że istnieje zależność między stężeniem roztworu a czasem.

c) Współczynnik korelacji między przewidywaną a obserwowaną wartością roztworu.

Współczynnik korelacji wynosi -0.9009. Jest ujemny i bliski -1, co oznacza, że istnieje silny związek między stężeniem roztworu a czasem. Ujemna wartość współczynnika oznacza, że wraz ze wzrostem czasu wartość stężenia maleje.

Zadanie 6: procedura Box'a-Cox'a.



Wartość parametru λ wynosi 0.0202. Widzimy, że jest prawie równa zero, co oznacza, że dokonujemy podstawienia:

$$\tilde{Y} = log(Y)$$

Zadanie 7

Utworzenie nowej zmiennej odpowiedzi.

Wykorzystując procedurę Box'a - Cox'a, tworzymy nową zmienną odpowiedzi $\tilde{Y}$, która wyraża się wzorem $\tilde{Y} = log(Y)$.

Powtórz zadanie 5 dla nowej zmiennej odpowiedzi.

Sprawdzimy, jak zmienią się uzyskane wnioski, gdy powtórzymy wszystkie kroki z zadania 5 dla nowej zmiennej odpowiedzi $\tilde{Y}$.

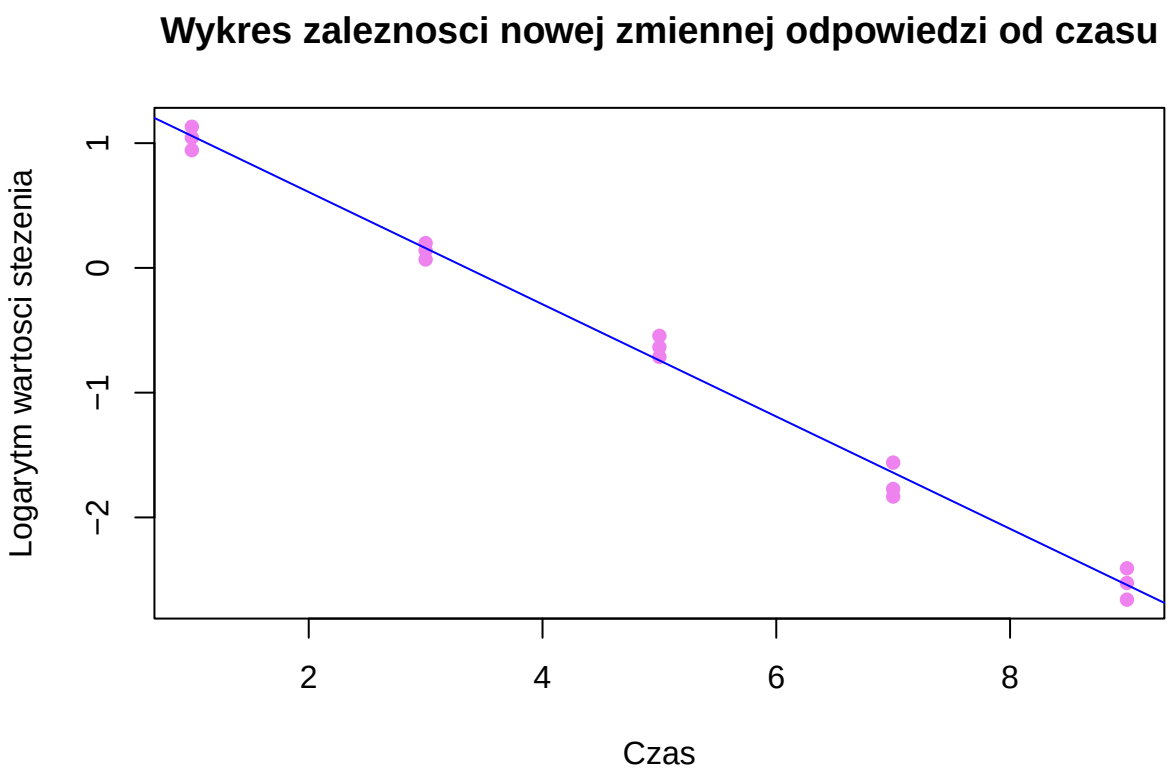
Regresja liniowa i przedziały predykcyjne.

Wartości współczynników wynoszą:

- b_0 = 1.5079,
- b_1 = -0.4499.

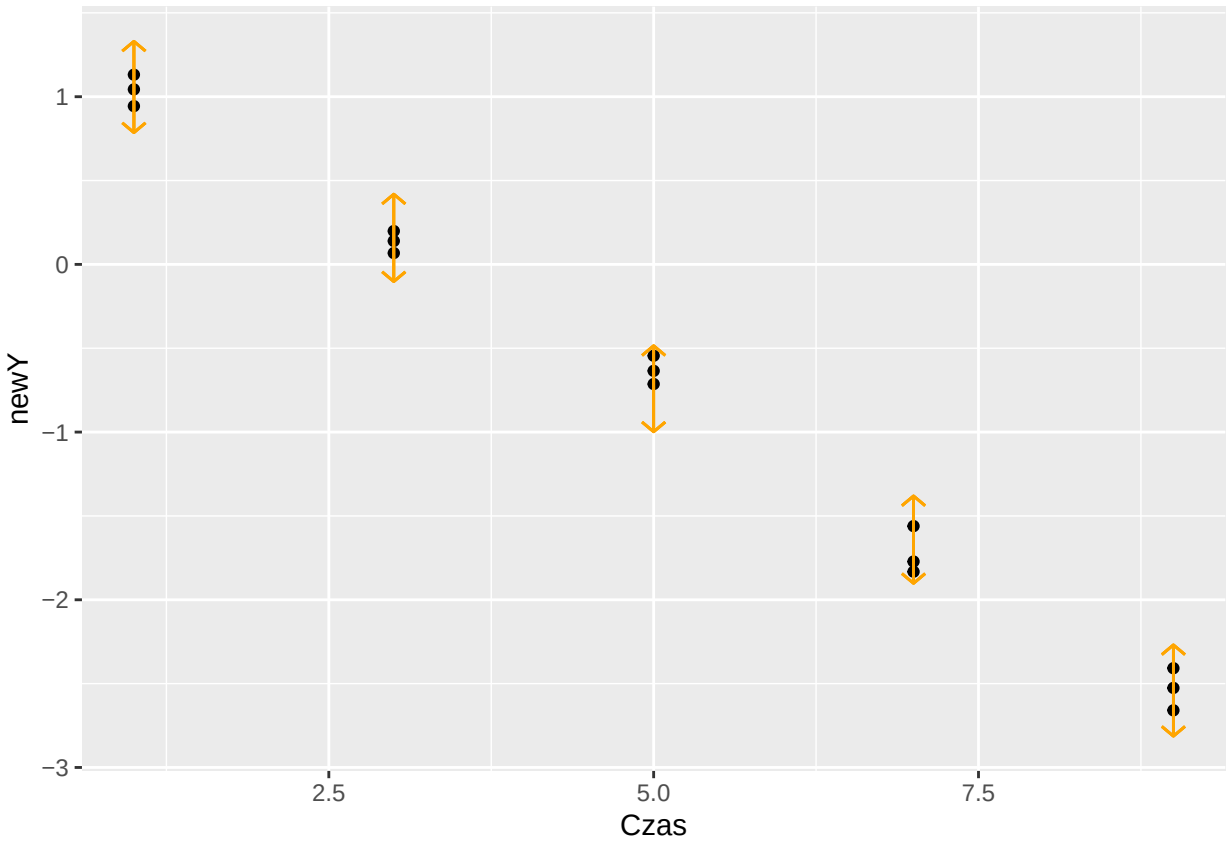
Zatem równanie regresji wyraża się wzorem: Y = 1.5079 + -0.4499X.

Zobaczmy dane wraz z prostą regresji na wykresie:



Patrząc na wykres widzimy, że dane bardziej układają się w linię prostą niż w poprzednim przypadku.

Dodajmy do wykresu 95% przedziały predykcyjne:



Widzimy, że wszystkie obserwcje mieszczą się w wyznaczonych przedziałach predykcyjnych, które mają dość małą długość. Oznacza to, że dobrze przybliżają wartości zmiennych odpowiedzi.

Analiza regresji: współczynnik R^2 i test istotności dla b_1.

W celu zbadania dokładności dopasowania prostej regresji do danych obliczymy współczynnik R^2.

Wynosi on: 0.993. Widzimy, że jest prawie równy 1, co oznacza, że prosta regresji jest bardzo dobrze dopasowana do danych.

Zbadajmy, czy istnieje zależność między zmiennymi. W tym celu testujemy nastepującą hipotezę:

$$H_0 : b_1 = 0$$

Hipoteza alternatywna zaprzecza hipotezie zerowej, czyli jest postaci:

$$H_a : b_1 \neq 0$$

W celu przetestowania hipotez możemy użyć testu t-studenta albo testu Fishera. Z racji, że są równoważne, nie ma znaczenia, który test wybierzemy. Zbadajmy zatem hipotezę zerową przy pomocy testu F.

Tradycyjnie przyjmujemy α=0.05.

Statystyka testowa jest postaci:

$$F = \frac{MSM}{MSE}$$

Podstawiając, otrzymujemy: F = 1838.225.

Teraz sprawdzimy, czy wartość statystyki F jest większa od kwantyla $F_c = F^*(0.95, 1, 13)$, który wynosi 4.6672.

Widzimy, że F $> F_c$, zatem odrzucamy H_0 mówiącą, że b_1=0.

Ten sam wniosek możemy uzyskać analizując p-wartość, która wynosi 2.220446×10^{-15}. Widzimy, że jest zdecydowanie mniejsza niż α=0.05, zatem odrzucamy H_0.

Wniosek: Z prawdopodobieństwem 95% wartość $b_1 \neq 0$, co oznacza, że istnieje zależność między stężeniem roztworu a czasem.

Współczynnik korelacji między przewidywaną a obserwowaną wartością roztworu.

Współczynnik korelacji wynosi -0.9965. Jest ujemny i bliski -1, co oznacza, że istnieje silny związek między stężeniem roztworu a czasem. Ujemna wartość współczynnika oznacza, że wraz ze wzrostem czasu wartość stężenia maleje.

Porównując uzyskaną wartość z tą uzyskaną w zadaniu 5, widzimy, że wartość współczynnika korelacji dla zmienionej zmiennej odpowiedzi jest co do modułu bliższa 1, co oznacza, że korelacja jest silniejsza.

Podsumowując, model ze zmienioną zmienną odpowiedzi $\tilde{Y}$ jest dużo lepszy od pierwotnego.

Przeanalizujmy jeszcze jeden model.

Zadanie 8: nowa zmienna objaśniająca.

W tym zadaniu przekształcamy zmienną objaśniającą X zgodnie ze wzorem:

$$\tilde{X} = X^{-0.5}$$

Powtarzamy kroki z zadania 7 dla $\tilde{X}$ jako zmiennej objaśniającej oraz Y jako zmiennej odpowiedzi.

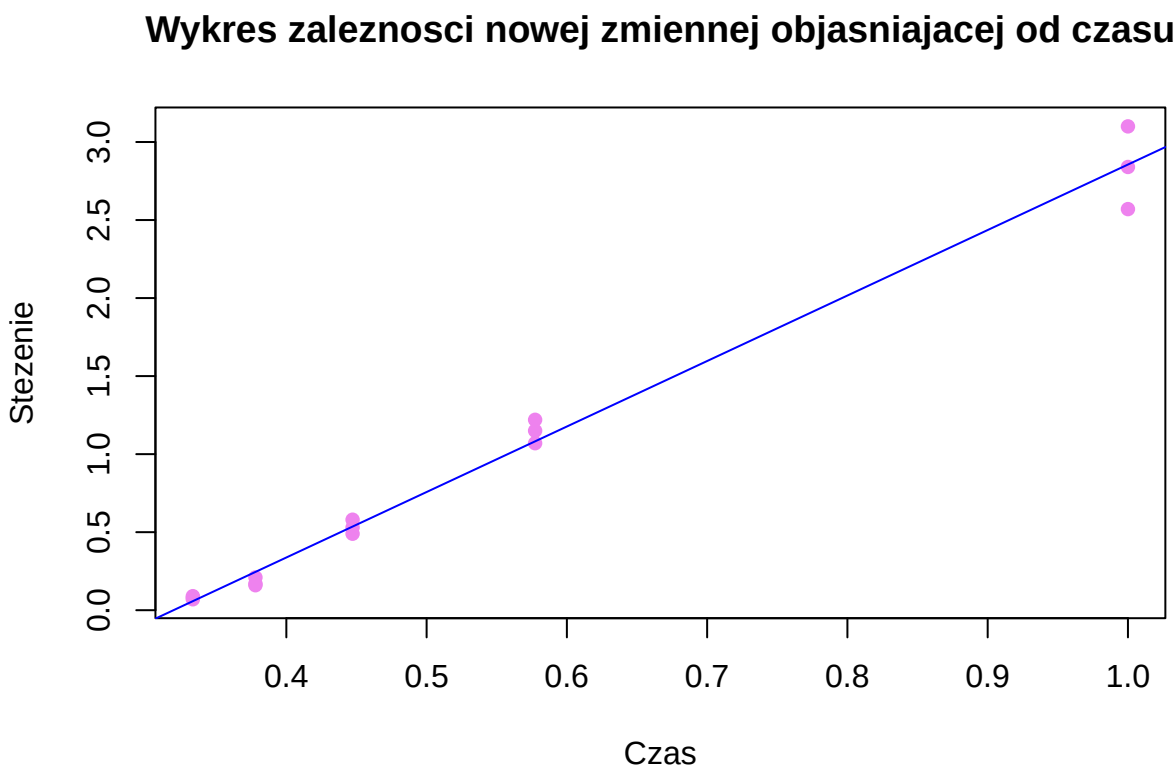
Regresja liniowa i przedziały predykcyjne.

Wartości współczynników wynoszą:

- b_0 = -1.3408,
- b_1 = 4.1963.

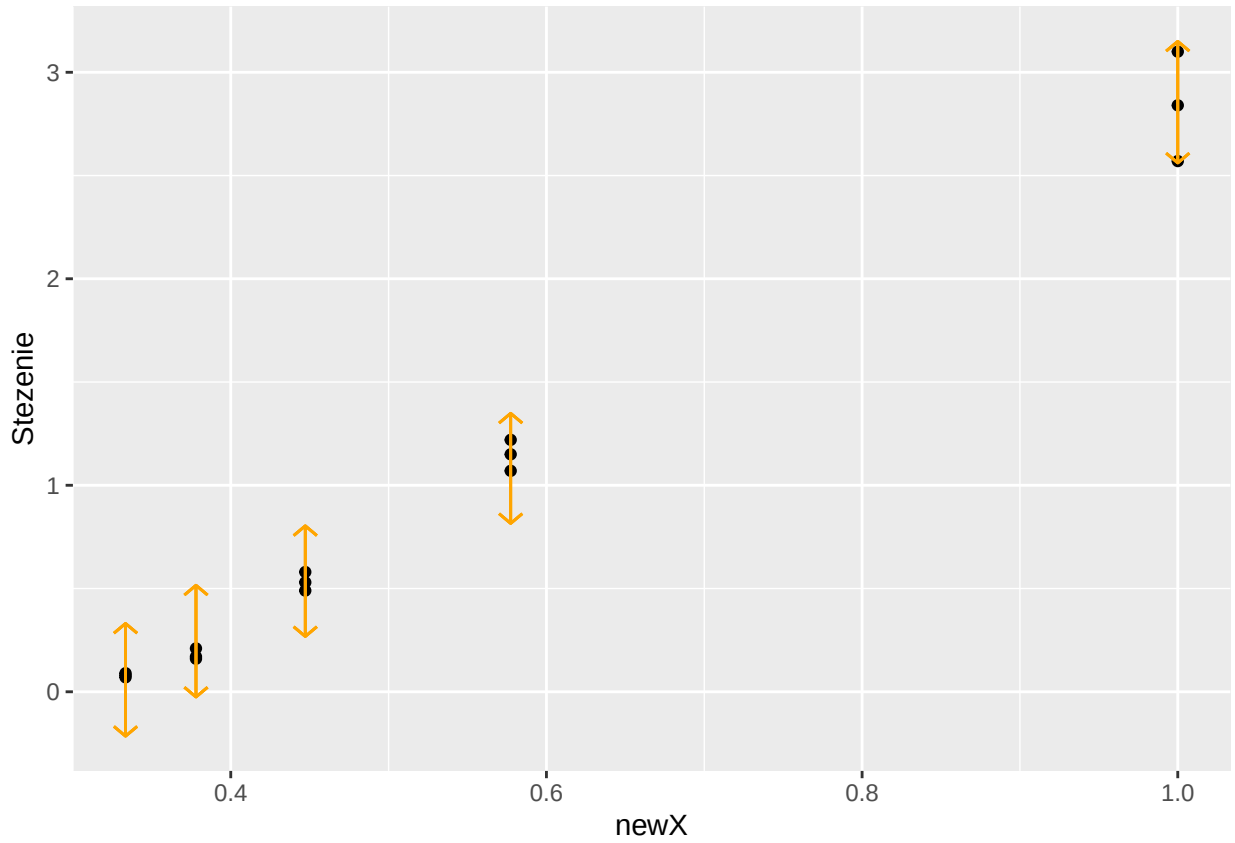
Zatem równanie regresji wyraża się wzorem: Y = -1.3408 + 4.1963X.

Zobaczmy dane wraz z prostą regresji na wykresie:



Widzimy, że prosta na wykresie zmieniła kierunek: jest funkcją liniową rosnącą. Jest mniej dopasowana do danych niż w przypadku poprzedniego modelu, ale lepiej niż w pierwotnym modelu.

Dodajmy do wykresu 95% przedziały predykcyjne:



Widzimy, że przedziały są nieco dłuższe niż w poprzednim modelu, ale węższe niż w pierwotnym. Wszystkie obserwacje wpadają do przedziałów predykcyjnych.

Analiza regresji: współczynnik R^2 i test istotności dla b_1.

W celu zbadania dokładności dopasowania prostej regresji do danych obliczymy współczynnik R^2.

Wynosi on: 0.9881. Widzimy, że jest prawie równy 1, co oznacza, że prosta regresji jest bardzo dobrze dopasowana do danych.

Zbadajmy, czy istnieje zależność między zmiennymi. W tym celu testujemy nastepującą hipotezę:

$$H_0 : b_1 = 0$$

Hipoteza alternatywna zaprzecza hipotezie zerowej, czyli jest postaci:

$$H_a : b_1 \neq 0$$

W celu przetestowania hipotez możemy użyć testu t-studenta albo testu Fishera. Z racji, że są równoważne, nie ma znaczenia, który test wybierzemy. Zbadajmy zatem hipotezę zerową przy pomocy testu F.

Tradycyjnie przyjmujemy α=0.05.

Statystyka testowa jest postaci:

$$F = \frac{MSM}{MSE}$$

Podstawiając, otrzymujemy: F = 1076.0501.

Teraz sprawdzimy, czy wartość statystyki F jest większa od kwantyla $F_c = F^*(0.95, 1, 13)$, który wynosi 4.6672.

Widzimy, że F $> F_c$, zatem odrzucamy H_0 mówiącą, że b_1=0.

Ten sam wniosek możemy uzyskać analizując p-wartość, która wynosi $6.9055872 \times 10^{-14}$. Widzimy, że jest zdecydowanie mniejsza niż α=0.05, zatem odrzucamy H_0.

Wniosek: Z prawdopodobieństwem 95% wartość $b_1 \neq 0$, co oznacza, że istnieje zależność między stężeniem roztworu a czasem.

Współczynnik korelacji między przewidywaną a obserwowaną wartością roztworu.

Współczynnik korelacji wynosi 0.994. Jest dodatni i bliski 1, co oznacza, że istnieje silny związek między stężeniem roztworu a czasem. Dodatnia wartość współczynnika oznacza, że wraz ze wzrostem czasu wartość stężenia również wzrasta.

Który model jest najlepszy?

Podsumowując, najlepszy jest model log(Y)~X. Trochę gorszy, aczkolwiek dalej dobry i lepszy od pierwitnego, okazał się model Y~$X^{-0.5}$. Wnioski zostały wyciągniete na podstawie:

- analizy wartoścy R^2 - dla modelu log(Y)~X jego wartość była największa,
- długości przedziałów predykcyjnych - dla modelu log(Y)~X przedziały osiągały najmniejszą długość.

Zadania teoretyczne

Zadanie 1

Obliczanie wartości krytycznej t-testu

Za pomocą R należy obliczyć wartość krytyczną(t_c) dla dwukierunkowego t-testu istotności z r stopniami swobody, gdzie:

- $r \in \{5, 10, 50\}$,
- $\alpha = 0.05$.

liczba stopni swobody	wartość t_c
5	2.5706
10	2.2281
50	2.0086

Wartość krytyczna testu F

Za pomocą R należy obliczyć wartość krytyczną (F_c) dla testu istotności F z 1 stopniem swobody w liczniku i r stopniami swobody w mianowniku, gdzie:

- $r \in \{5, 10, 50\}$,
- $\alpha = 0.05$.

liczba stopni swobody w mianowniku	wartość F_c
5	6.6079
10	4.9646
50	4.0343

Sprawdzenie, czy $t_c^2 = F_c$

Wartość r	Wartość t_c^2	Wartość F_c
5	6.6079	6.6079
10	4.9646	4.9646
50	4.0343	4.0343

Widzimy, że $t_c^2 = F_c$.

Zadanie 2

Ile obserwacji znajduje się w pliku?

Skoro dfE = n-2 = 20, to wiemy, że n=22. Czyli mamy 22 obserwacje w pliku.

Estymator σ.

Estymator σ^2, czyli s^2, możemy obliczyć ze wzoru:

$$s^2 = \frac{SSE}{dfE} = \frac{400}{20} = 20$$

$$s = \sqrt{s^2} = \sqrt{20} = 2\sqrt{5}$$

co wynosi w przybliżeniu 4.4721.

Sprawdź, czy slope jest równy 0.

$$H_0 : b_1 = 0$$

Statystyka testowa F jest postaci

$$F = \frac{MSM}{MSE}$$

$MSM = \frac{SSM}{dfM} = 100$ $MSE = \frac{SSE}{dfE} = 20$

Zatem statystyka testowa F ma wartość $F = \frac{100}{20} = 5$ oraz liczbę stopni swobody 1 i 20. Kwantyl F_c dla α=0.05 wynosi $F_c = 4.3512$. Widzimmy, że F>F_c, więc odrzucamy H_0.

Jaką część zmienności zmiennej odpowiedzi wyjaśnia model?

W tym celu trzeba policzyć wartość współczynnika R^2, który wyraża się wzorem:

$$R^2 = \frac{SSM}{SST}$$

gdzie SST = SSM + SSE = 500. Podstawiając, otrzymujemy $R^2 = 0.2$.

Próbkowy współczynnik korelacji między zmienną odpowiedzi a zmienną objaśniającą.

W tym podpunkcie interesuje nas wartość R. Znając R^2, możemy łatwo obliczyć R, który wynosi $0.4472.