Lista 3

Helena Sękowska-Słoka, nr indeksu 321531

2023-12-20

SPIS TREŚCI

Zadanie 1

Importujemy dane z pliku, a następnie tworzymy prosty model regresji, aby otrzymać zależność GPA od wyników testu IQ.

Równanie regresji

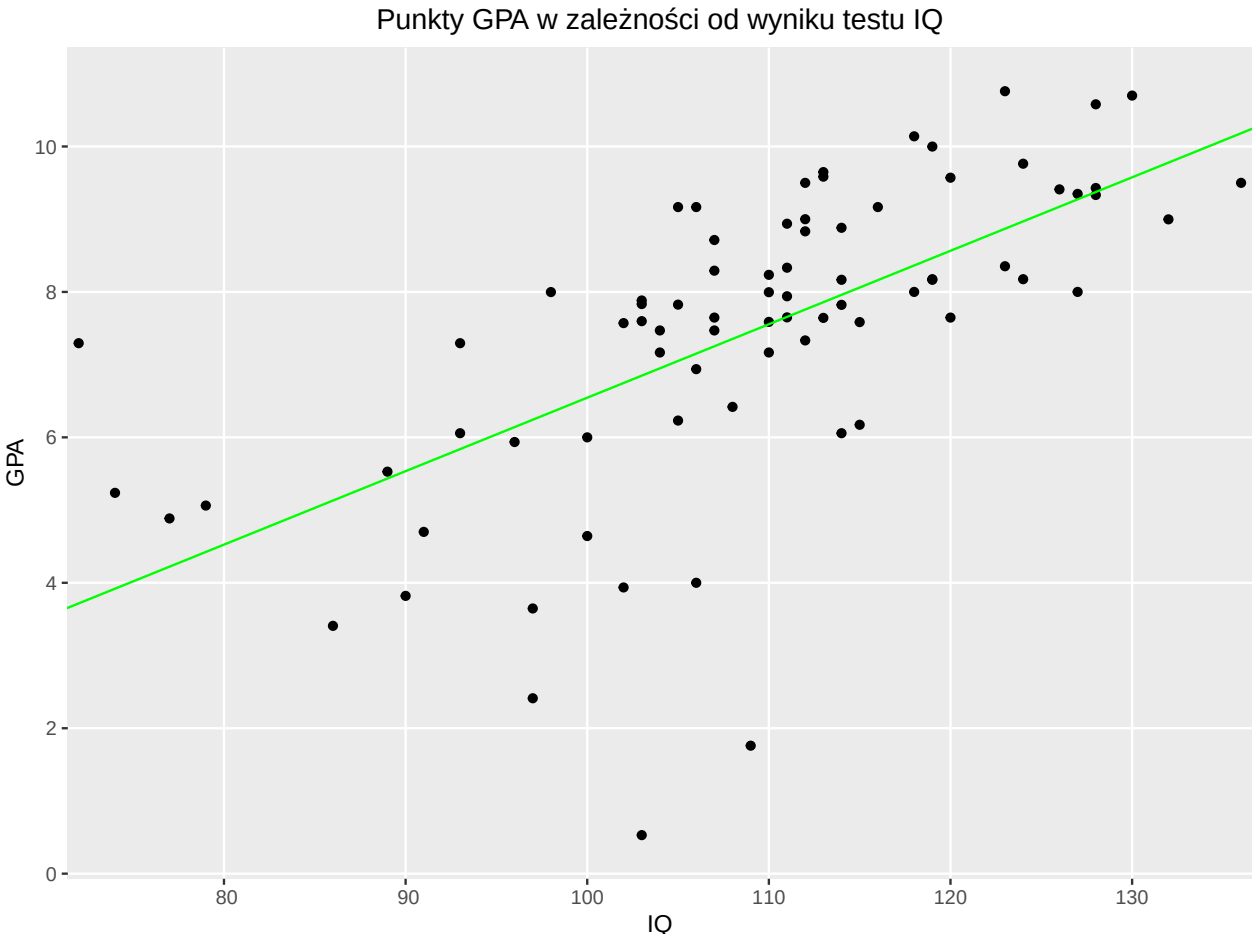
```
## [1] "GPA = -3.55706 + 0.101022 * IQ"
```

Przy czym:

```
## [1] "intercept = -3.55706"
```

```
## [1] "slope = 0.101022"
```

Wykres z naniesioną prostą regresji



Obserwacje na wykresie są dość rozstrzelone, nie leżą szczególnie blisko prostej regresji. Jednakże poza kilkoma wyjątkowo niskimi średnimi dla środkowych wartosci IQ można przyjąć, że układają się w przybliżeniu w jednym paśmie.

Współczynnik determinancji R^2 wyznaczony za pomocą R

```
## [1] 0.401615
```

Współczynnik determinancji R^2 wyznaczony teoretycznie

Skorzystamy ze wzorów:

- $R^2 = \frac{SSM}{SST}$
- $SSM = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$
- $SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Ostateczny wynik:

```
## [1] 0.401617
```

Jak widać, obliczone różnymi metodami R^2 różnią się dopiero na 6. miejscu po przecinku, zatem możemy uznać, że metody są zgodne.

Wnioski na podstawie wartości R^2

Współczynnik determinacji określa, jaką część całkowitej zmienności w wektorze stanowi zmienność wyjaśniona przez model. Zatem im większa wartość R^2, tym lepszy jest model (tym więcej jesteśmy w stanie na jego podstawie przewidzieć). Biorąc pod uwagę, że wartości tego współczynnika pochodzą z przedziału $[0, 1]$, dla podanych danych model liniowy jest słabo dopasowany. Jednym z powodów takiego stanu może być brak zależności liniowej pomiędzy IQ i GPA, a nawet brak korelacji pomiędzy nimi w ogóle. Zbadajmy tę hipotezę.

Testowanie hipotezy, że GPA nie jest skorelowane z IQ na podstawie testu F

Hipotezy:

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

Licząc wartość statystyki testowej F teoretycznie, skorzystamy ze wzoru $F = \frac{MSM}{MSE}$, gdzie $MSM = SSM$, zaś $MSE = \frac{SSE}{dfE}$.
Teoretyczna wartość F:

```
## [1] 51.009
```

Wartość F według R:

```
##  value
## 51.008
```

Wyniki różnią się dopiero na 3. miejscu po przecinku, zatem możemy przyjąć, że są zgodne.
Otrzymane wartości porównamy z $F_c = F^*(1 - \alpha, 1, n - 2)$, czyli kwantylem rzędu $1 - \alpha$ z rozkładu Fishera-Snedecora z dfM i dfE stopniami swobody. Niech $\alpha = 0.05$. W naszym przypadku n to:

```
## [1] 78
```

Wobec czego szukany rozkład to $F(1, 76)$.

Zatem F_c wynosi:

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

Teoretyczna p-wartość:

Skorzystamy ze wzoru $p = P(z > F) = 1 - P(z \leq F)$, gdzie $z \sim F(1, 76)$, zatem możemy skorzystać z dystrybunanty rozkładu $F(1, 76)$.

```
## [1] 4.736551e-10
```

p-wartość według R:

```
##           IQ
## 4.737341e-10
```

Wyliczone p-wartości różnią się na trzecim miejscu po przecinku, co, ze względu na ich rząd, jest różnicą mało znaczącą.
Ponieważ w obu przypadkach $F > F_c$ (i to znacznie), a także p-wartości są bardzo małe, zatem możemy stanowczo odrzucić hipotezę zerową. To oznacza, że zachodzi pewna korelacja pomiędzy GPA i IQ.

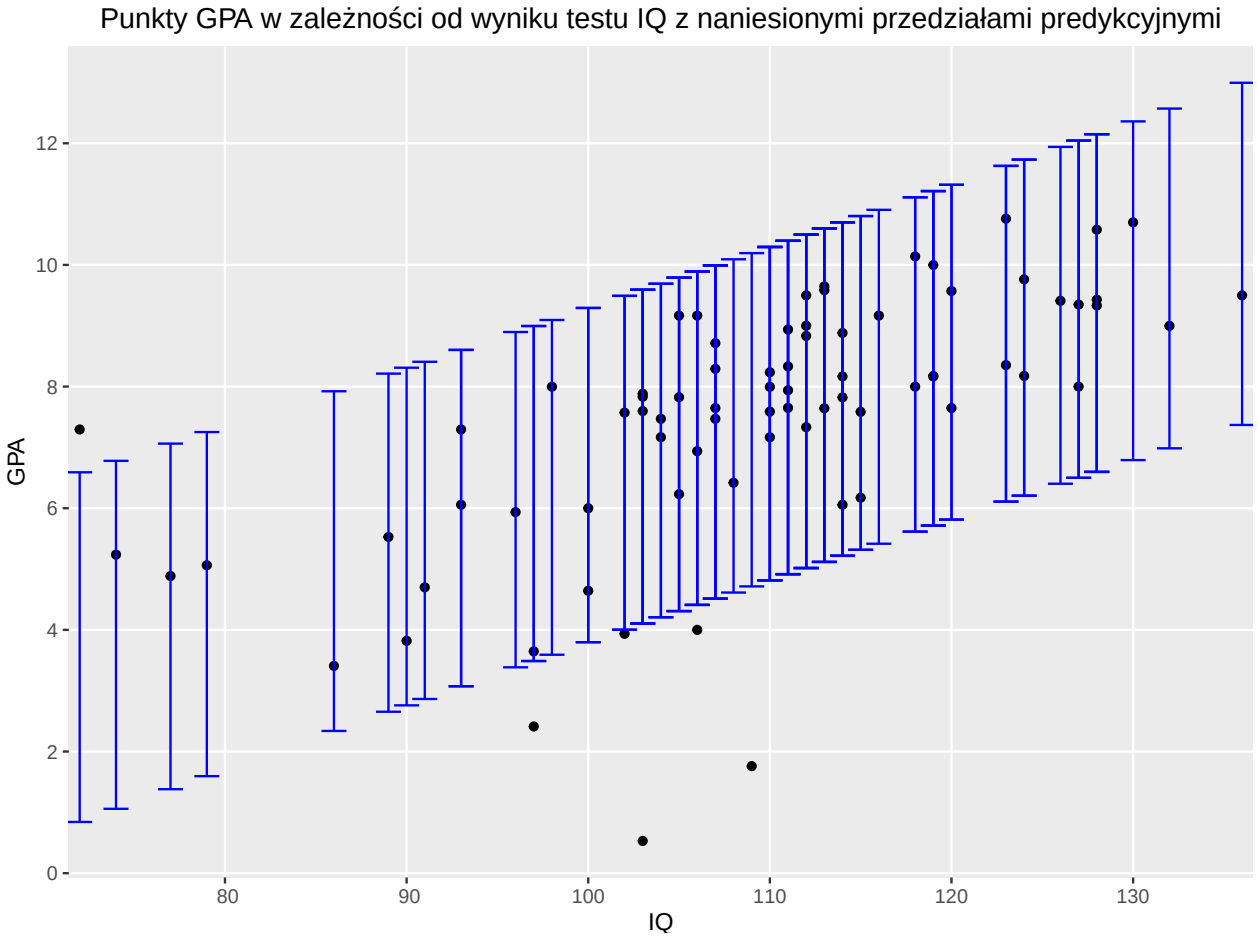
Przewidywane wartości GPA dla IQ $\in$ {75, 100, 140}

Table 1: 90% przedziały predykcyjne

IQ	estymator	lewy koniec	prawy koniec	długość przedziału
75	4.020	1.166	6.873	5.707
100	6.545	3.798	9.293	5.495
140	10.586	7.750	13.422	5.672

Przedziały są tym dłuższe, im bliżej skrajnych danych się znajdujemy, ale jednak są to stosunkowo bardzo małe różnice, patrząc na to, jak bardzo oddalamy się od środkowych wartości w zadanych danych.
Ponadto wszystkie przedziały są szerokie (ich długość to około połowa różnicy pomiędzy maksymalnym a minimalnym GPA). Nie dziwi nas to ze względu na fakt, że punkty na wykresie z danymi były wyraźnie rozstrzelone, a nie skupione, a także ze względu na wyliczoną wartość R^2, która była stosunkowo niewielka, przez co nie jesteśmy w stanie przewidzieć bardzo dokładnie spodziewanych wartości dla nowych obserwacji.

Wykres z naniesionymi przedziałami predykcyjnymi



Po raz kolejny obserwujemy, że ze względu na rozłożenie danych przedziały predykcyjne są bardzo szerokie - zajmują one powierzchniowo około połowy wykresu.
Sześć obserwacji się poza granicami wyznaczonych przedziałów. Obliczmy zatem, jaka część obserwacji do nich należy:

```
## [1] 0.93878
```

Jak widać, zdecydowana większość obserwacji znajduje się w tych przedziałach, więcej niż 90% będace poziomem ufności wyznaczonych przedziałów.
Podsumowując, wyznaczone przedziały predykcyjne wskazują na to, że potencjalne nowe obserwacje również będą podobnie rozstrzelone jak pierwotne obserwacje, a ich zakres może być bardzo szeroki.

Zadanie 2

Równanie regresji

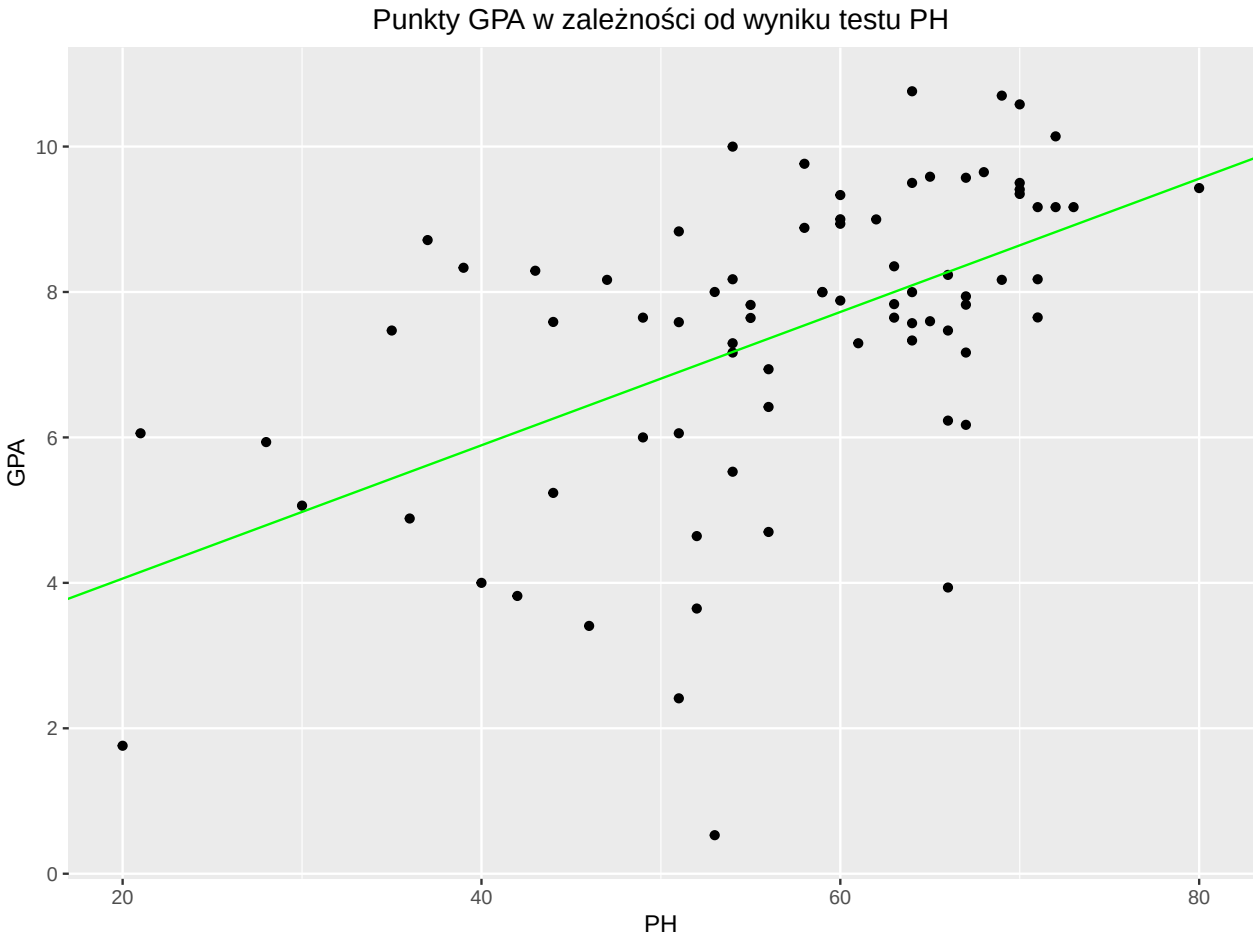
```
## [1] "GPA = 2.22588 + 0.091652 * PH"
```

Przy czym:

```
## [1] "intercept = 2.22588"
```

```
## [1] "slope = 0.091652"
```

Wykres z naniesioną prostą regresji



Na pierwszy rzut oka obserwacje są znacznie bardziej rozstrzelone niż w przypadku IQ. Przeanalizujmy jednak wartość współczynnika determinancji, żeby móc powiedzieć więcej.

Współczynnik determinancji R^2 wyznaczony za pomocą R

```
## [1] 0.293583
```

Współczynnik determinancji R^2 wyznaczony teoretycznie

Korzystamy ze wzorów jak uprzednio.
Wynik:

```
## [1] 0.293581
```

Wartości różnią się dopiero na 6. miejscu po przecinku, zatem możemy powiedzieć, że są zgodne.
Jak widać, w tym przypadku uzyskaliśmy jeszcze mniejsze R^2. Oznacza to, że model jest jeszcze słabiej dopasowany. Przyjrzyjmy się zatem, jak wyglądają wyniki testu F.

Testowanie hipotezy, że GPA nie jest skorelowane z PH na podstawie testu F

Hipotezy:

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

Korzystamy ze wzorów jak uprzednio.
Teoretyczna wartość F:

```
## [1] 31.5849
```

Wartość F według R:

```
##   value
## 31.5852
```

Wyniki różnią się dopiero na 3. miejscu po przecinku, zatem możemy przyjąć, że są zgodne.
Otrzymane wartości porównamy jak w poprzednim zadaniu z F_c, które pozostało bez zmian, ponieważ nie zmieniła się ani liczba stopni swobody, ani α. Przypomnijmy zatem wartość F_c:

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

Teoretyczna p-wartość:

Skorzystamy ze wzoru jak poprzednio. Otrzymany wynik:

```
## [1] 3.006706e-07
```

p-wartość według R:

```
##           PH
## 3.006416e-07
```

Wyliczone p-wartości różnią się na czwartym miejscu po przecinku, co, ze względu na ich rząd, jest różnicą mało znaczącą.
Po raz kolejny w obu przypadkach F jest znacznie większe od F_c, a także p-wartości są bardzo małe, zatem możemy stanowczo odrzucić hipotezę zerową. To oznacza, że zachodzi pewna korelacja pomiędzy GPA i PH.

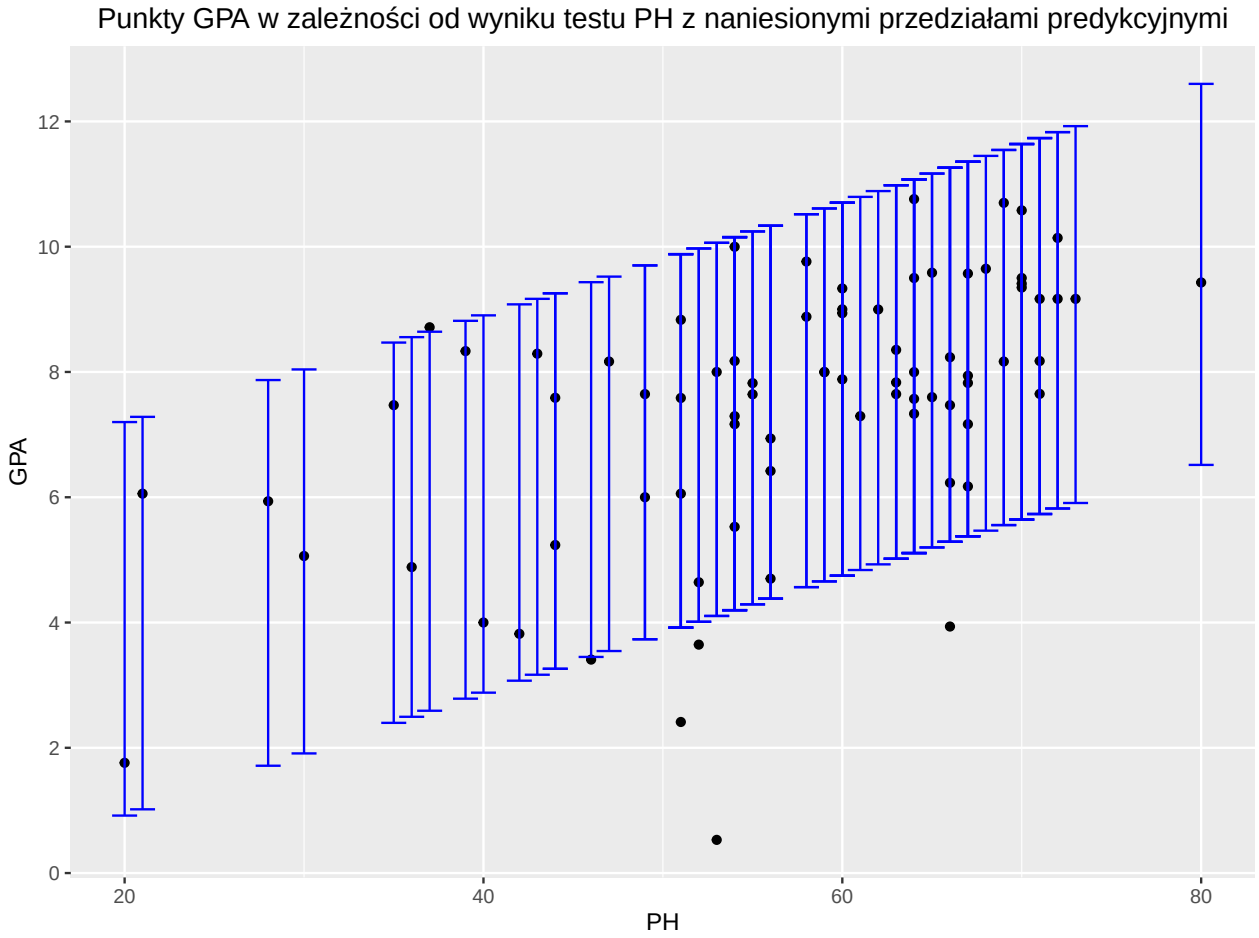
Przewidywane wartości GPA dla PH $\in$ {25, 55, 85}

Table 2: 90% przedziały predykcyjne

PH	estymator	lewy koniec	prawy koniec	długość przedziału
25	4.517	1.417	7.618	6.201
55	7.267	4.290	10.244	5.954
85	10.016	6.944	13.089	6.145

Po raz kolejny przedziały są tym dłuższe, im bliżej skrajnych danych się znajdujemy, ale jednak są to stosunkowo bardzo małe różnice, patrząc na to, jak bardzo oddalamy się od środkowych wartości w zadanych danych.
Ponadto wszystkie przedziały są szerokie, jeszcze szersze niż w przypadku IQ. Można to uzasadnić jeszcze mniej liniowym ułożeniem punktów na wykresie, a także mniejszą wartością współczynnika R^2.

Wykres z naniesionymi przedziałami predykcyjnymi



Wykres prezentuje się podobnie jak dla IQ, z tą różnicą, że przedziały są jeszcze dłuższe i zajmują powierzchniowo jeszcze więcej miejsca.Poza granicami przedziałów znowu znajduje się 6 obserwacji.

Co jest lepszym predyktorem GPA - wynik testu IQ czy wynik testu PH?

Biorąc pod uwagę zarówno wszystkie wykresy, jak i współczynnik determinancji R^2, a wreszcie długość przedziałów predykcyjnych, można ocenić, że wynik testu IQ okazał się predyktorem lepszym od wyniku testu PH. Przy czym warto dodać, że nie oznacza to, iż jest to dobry predyktor. Otrzymane w jego przypadku wyniki są jedynie nieco lepsze od wyników dla PH, ale żadna z wartości nie jest zadowalająca, w szczególności $R^2 < 0.5$ oraz stosunkowo długie przedziały predykcyjne 90%.

Zadanie 3

Sprawdzenie, czy suma residuów jest równa 0

Importujemy dane z pliku, a następnie tworzymy prosty model regresji, aby otrzymać zależność czasu od liczby kopiarek. Na tej podstawie zbadamy residua. Żeby były one zgodne z modelem teoretycznym, muszą spełniać następujące warunki:

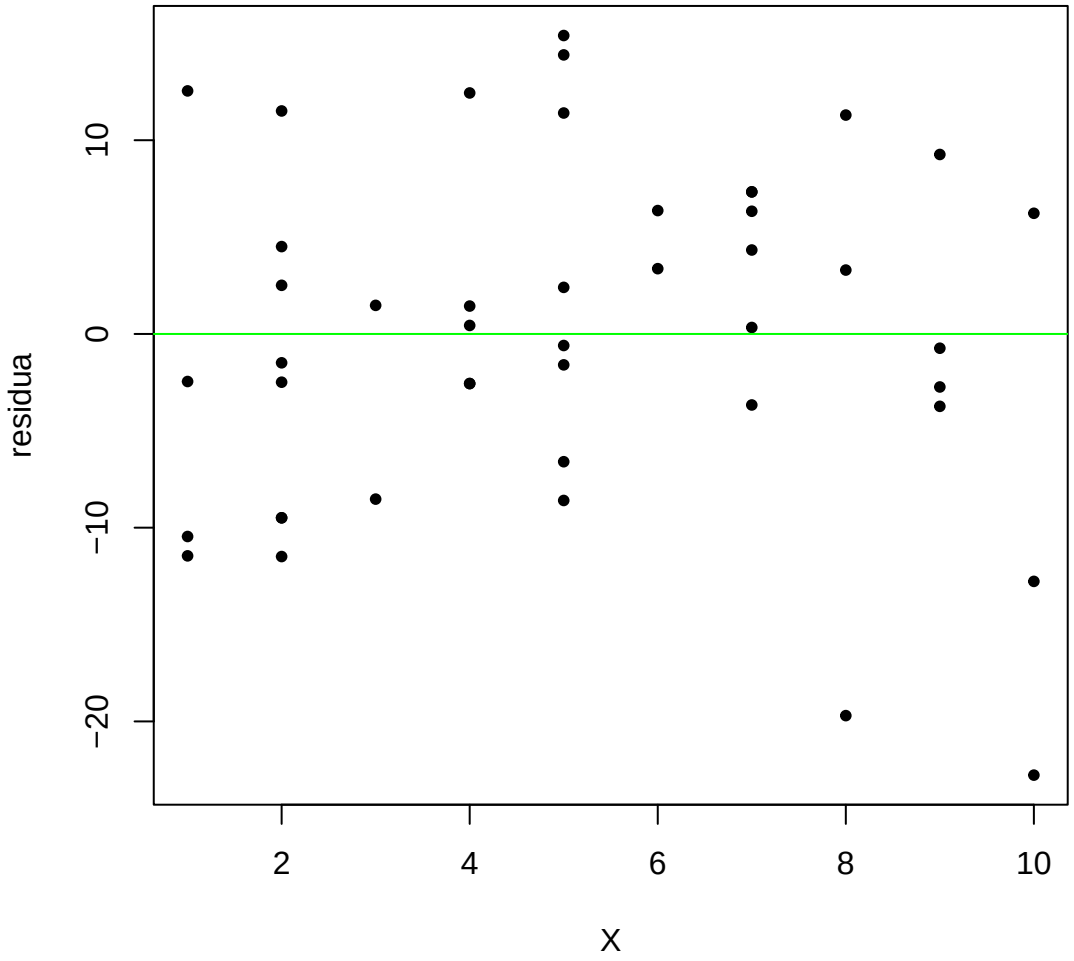
- niezależność
- wartość oczekiwana równa 0 i stała wariancja σ^2
- pochodzenie z rozkładu normalnego

Zacznijmy od sumy residuów:

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

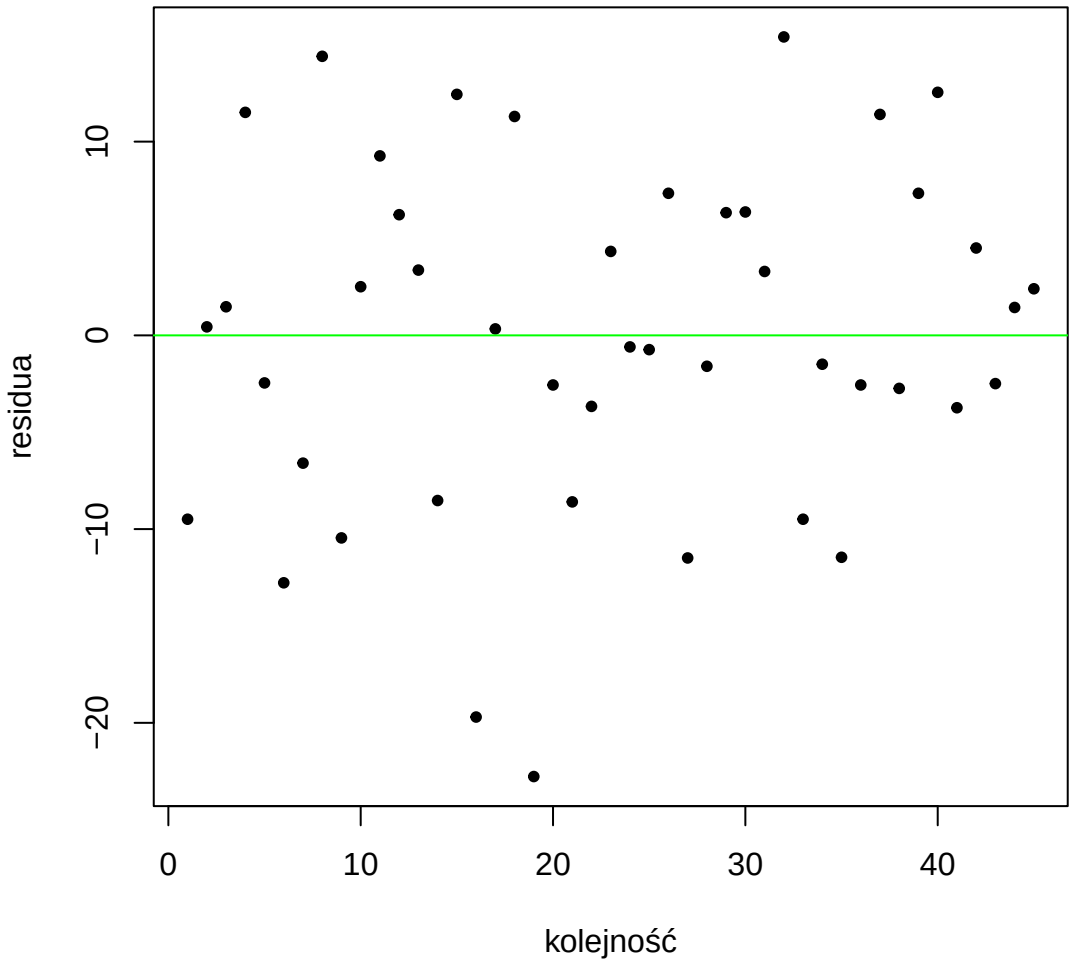
Nie wynosi ona dokładnie 0, ale jest tej wartości bardzo bliska, zatem możemy uznać ten wynik za zgodny z założeniami.

Wykres residuów względem zmiennej objaśniającej



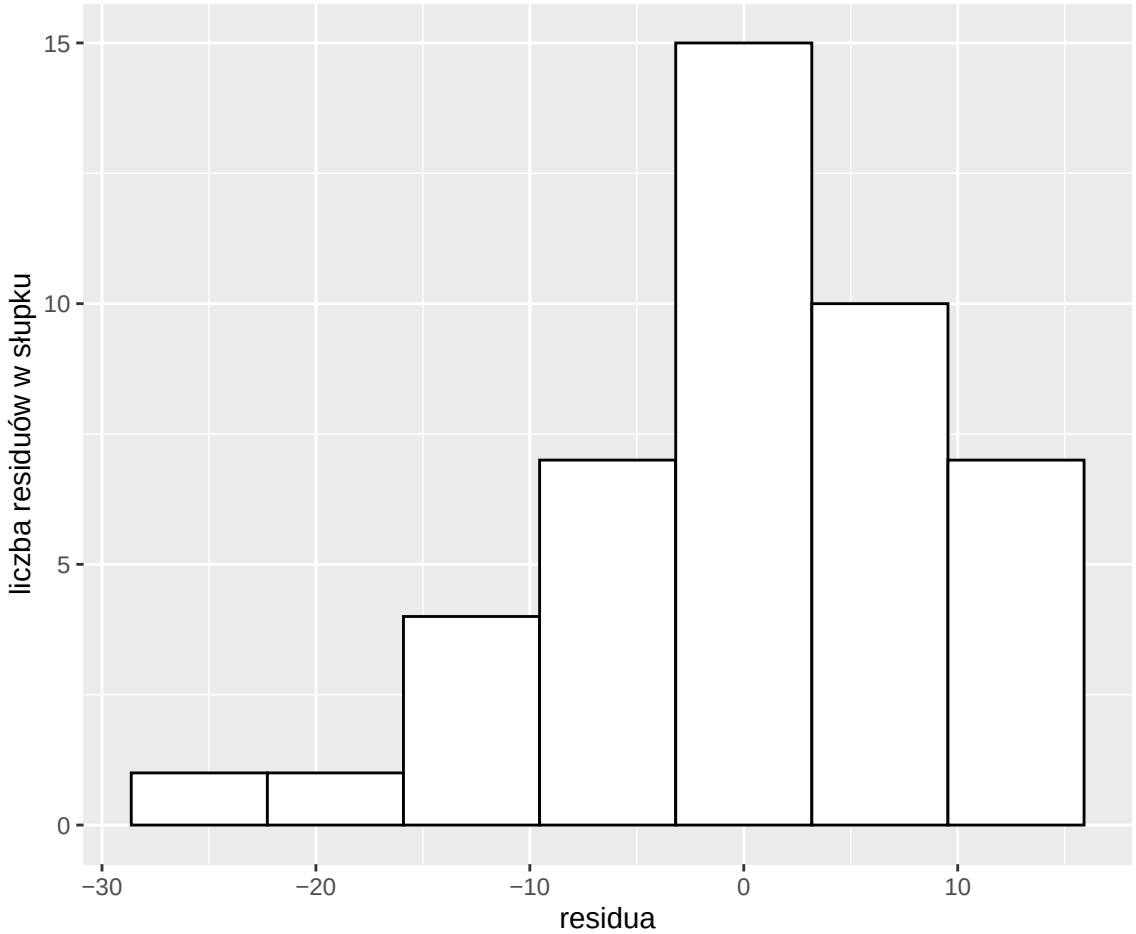
W lewej części wykresu residua rozmieszczone są w przybliżeniu równomiernie po obu stronach zera, czyli tak, jak oczekiwaliśmy. W prawej części wykresu widać więcej wartości dodatnich, zaś równoważą dwie wartości ujemne dla 8 i 10, które są co do modułu sporo większe od pozostałych. Na wykresie nie widać cykliczności ani innych nietypowych struktur.

Wykres residuów względem kolejności, w jakiej dane pojawiają się w pliku



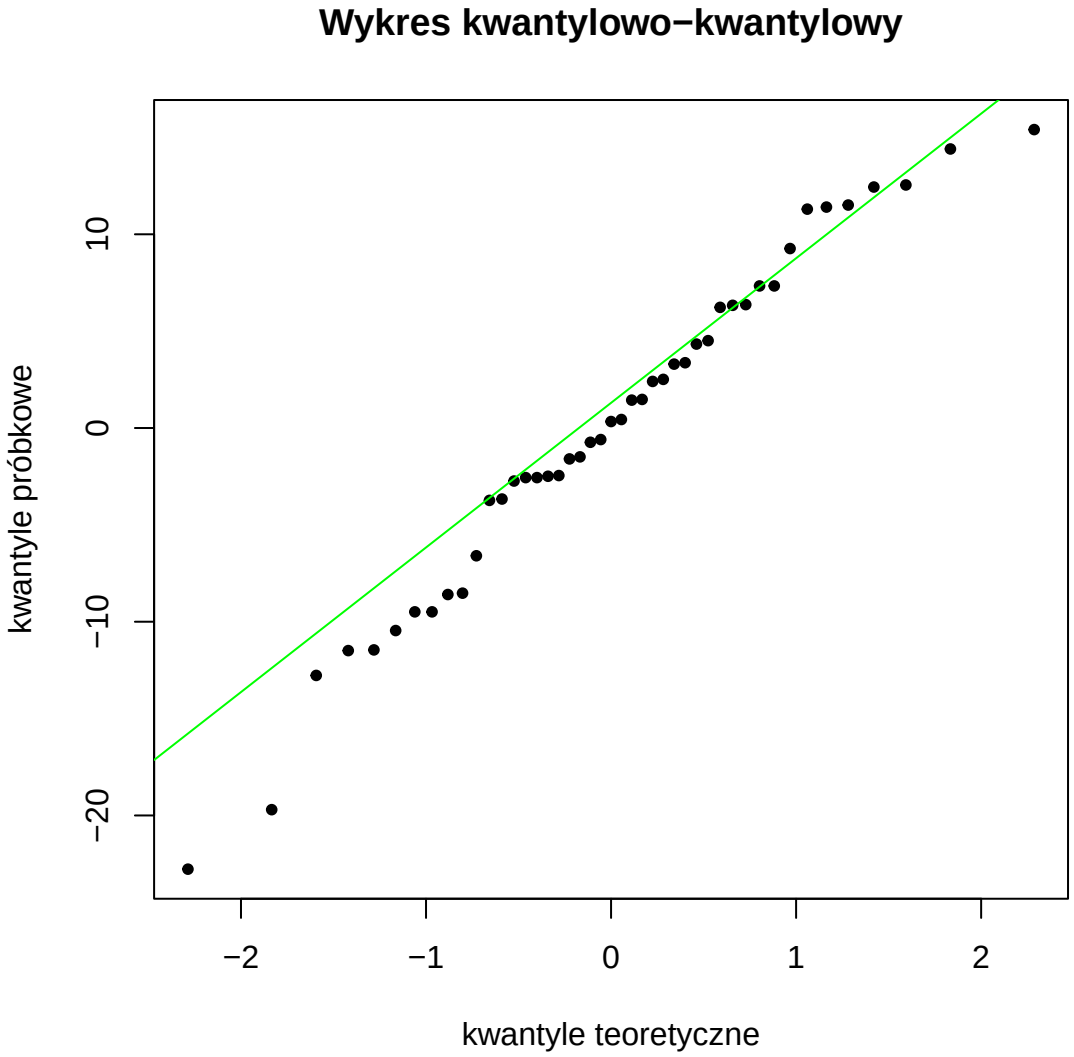
Na wykresie nie widać zgrupowań punktów świadczących o zależności residuów, ani formacji typowych dla niestałej wariancji. Punkty są rozłożone stosunkowo równomiernie. Jedyną cechą stanowiącą pewne odstępstwo od normy jest niewielka przewaga liczebna wartości dodatnich nad ujemnymi, z kolei residua ujemne są bardziej różnorodne co do wartości.

Histogram residuów



Histogram przypomina wizualnie rozkład normalny. Najwięcej błędów znajduje się w okolicach zera, słupki robią się coraz niższe w miarę oddalania się od środka. Po raz kolejny widać, że więcej jest obserwacji dodatnich, jednak obserwacje ujemne osiągają większe co do modułu wartości.

Wykres kwantylowo-kwantylowy residuów



Większość punktów na wykresie leży blisko prostej. Odstępstwa widać jedynie dla wartości krańcowych, w szczególności ujemnych.

Wnioski

Na podstawie dwóch pierwszych wykresów możemy uznać, że najprawdopodobniej wariancja jest stała, a błędy niezależne. Uwzględniając te wnioski, ze względu na sumę residuów oraz wygląd histogramu i wykresu kwantylowo-kwantylowego możemy postawić tezę, że residua pochodzą z rozkładu $N(0, \sigma^2)$ dla pewnego σ.

Zadanie 4

Dodanie obserwacji (1000;2)

Dodajemy nową obserwację (1000;2) do pliku z zadania 3, a następnie przeprowadzamy regresję ze zmienionymi danymi.

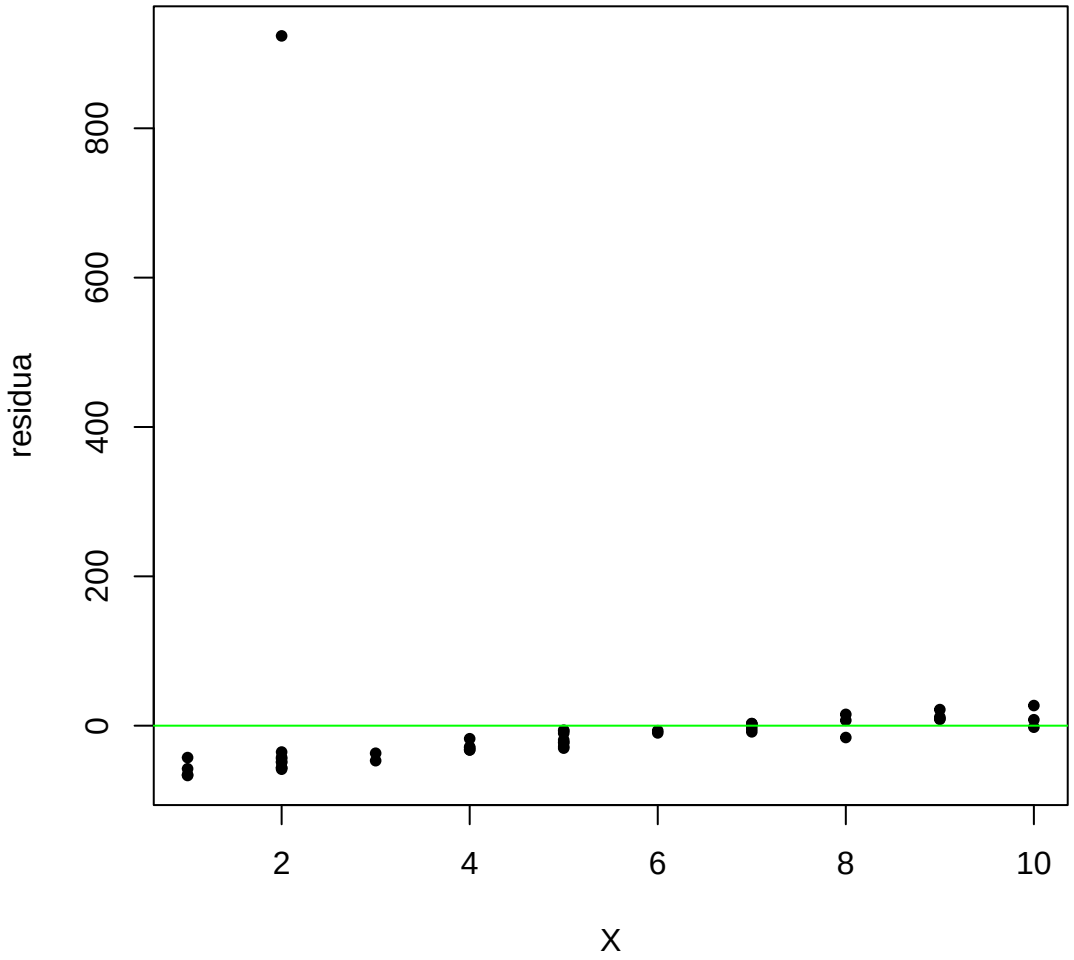
Tabela porównawcza po dodaniu obserwacji (1000;2)

Table 3: Tabela porównawcza

	oryginalne dane	dane z dodaną obserwacją (1000;2)
równanie regresji	Y = -0.58016 + 15.035 * X	Y = 63.091 + 6.5939 * X
wartość statystyki t	31.123	0.86249
p-wartość	4.009e-31	0.39309
współczynnik determinancji	0.9575	0.0166
estymator wariancji	79.451	20452

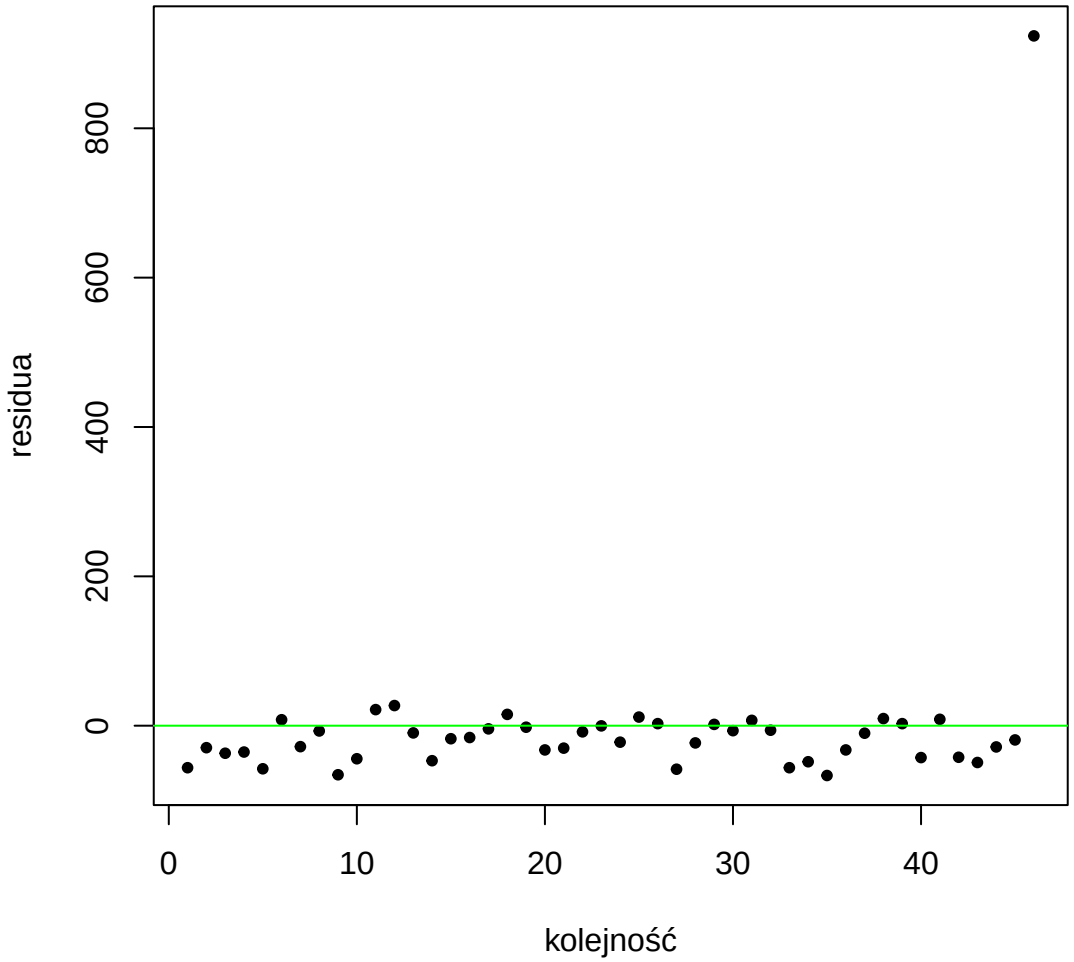
W oryginalnych danych regresja liniowa była możliwa do zastosowania, natomiast w danych z dodatkową obserwacją odstającą takiej możliwości nie ma. Współczynnik determinancji jest tak mały, że dopasowanie modelu liniowego traci sens, dodatkowo wyjątkowo duża p-wartość pozwala odrzucić hipotezę mówiącą, że między zmiennymi X(kopiarki) i Y(czas) zachodzi zależność liniowa. Warto też zwrócić uwagę na bardzo dużą różnicę w wariancji residuów - w przypadku tabeli z dodaną obserwacją odstającą nawet odchylenie standardowe (pierwiastek z wariancji) jest większe niż wartość większości obserwacji.
Podsumowując, przy stosunkowo niewielkiej próbce (tu o liczności 45) już jedna obserwacja odstająca może znacząco wpłynąć na wyniki regresji, a nawet na sens jej przeprowadzania.

Wykres residuów względem zmiennej objaśniającej dla zbioru z obserwacją (1000;2)



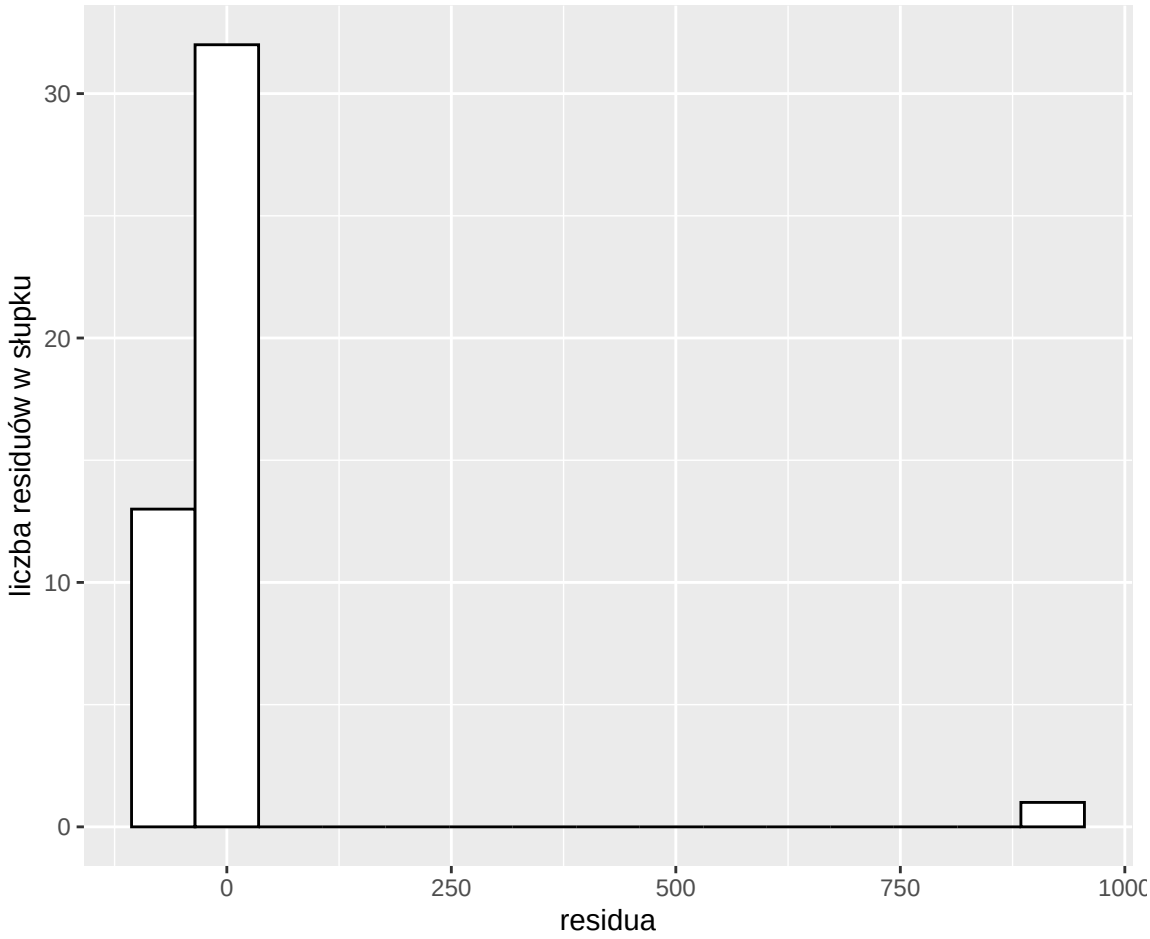
Zauważmy, że na wykresie residuów dla zbioru z dodatkową obserwacją widać zupełnie inne stuktury niż na wykresie residuów oryginalnego zbioru. Ze względu na bardzo duże różnice w wartości między błędem dodatkowej obserwacji a błędami obserwacji oryginalnych, wszystkie punkty poza nowym błędem dla $X = 2$ wydają się skupione wokół zera. Zmieniła się jednak ich struktura - układają się one w przybliżeniu liniowo, a ich wartości rosną wraz ze wzrostem X. Są to sygnały świadczące o tym, że zbiór może nie spełniać założeń regresji liniowej, w szczególności błąd może być zależny od wartości X.

Wykres residuów względem kolejności, w jakiej dane pojawiają się w pliku dla zbioru z obserwacją (1000;2)



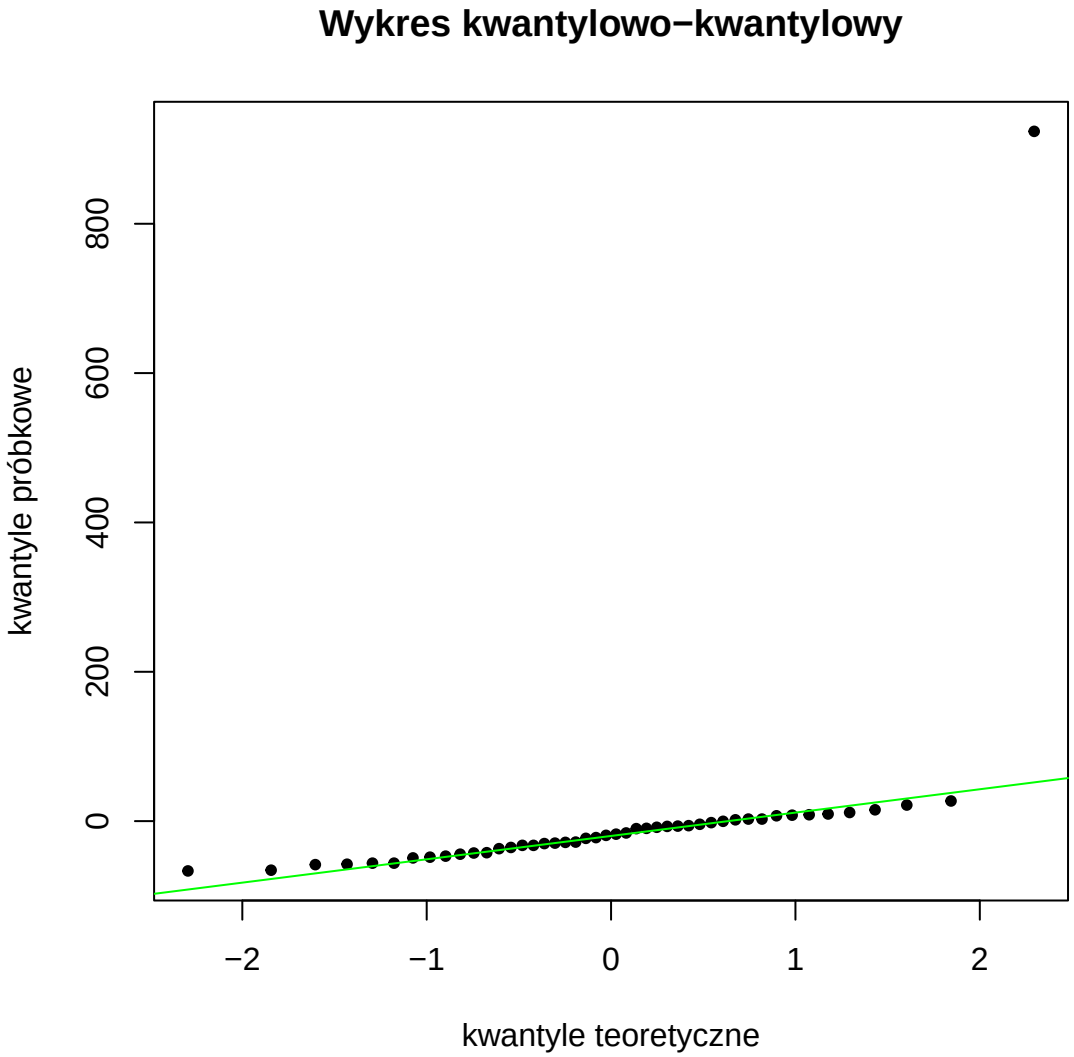
W przeciwieństwie do pierwotnego wykresu, tu widać nie tylko dodatkowy punkt w prawym górnym rogu, ale również charakterystyczne ułożenie pozostałych punktów, sugerujące, że kolejne błędy nie są niezależne - punkty układają się niemal cyklicznie, kolejne residua leżą blisko siebie (za wyjątkiem odstającego). Widać również przewagę liczebną wartości ujemnych nad dodatnimi.

Histogram residuów dla zbioru z obserwacją (1000;2)



Ze względu na odstającą obserwację histogram jest nieczytelny. Nie widać na nim struktury danych, w szczególności punktów poza dodanym. Nie przypomina on kształtem rozkładu normalnego.

Wykres kwantylowo-kwantylowy residuów dla zbioru z obserwacją (1000;2)



Ze względu na dodatkową obserwację pozostałe punkty na wykresie kwantylowo-kwantylowym wydają się leżeć znacznie bliżej prostej niż na oryginalnym wykresie. Samo ułożenie prostej na wykresie sprawia, że od razu zauważamy odstającą obserwację w prawym górnym rogu.

Wnioski

Dodanie obserwacji (1000;2) wystarczyło, żeby residua przestały spełniać założenia regresji liniowej, a współczynnik determinancji R^2 osiągnął wartość tak niską, że dopasowany do danych model liniowy nie ma w praktyce zastosowania.

Dodanie obserwacji (1000;6)

Dodajemy nową obserwację (1000;6) do początkowego pliku z zadania 3, a następnie przeprowadzamy regresję ze zmienionymi danymi.

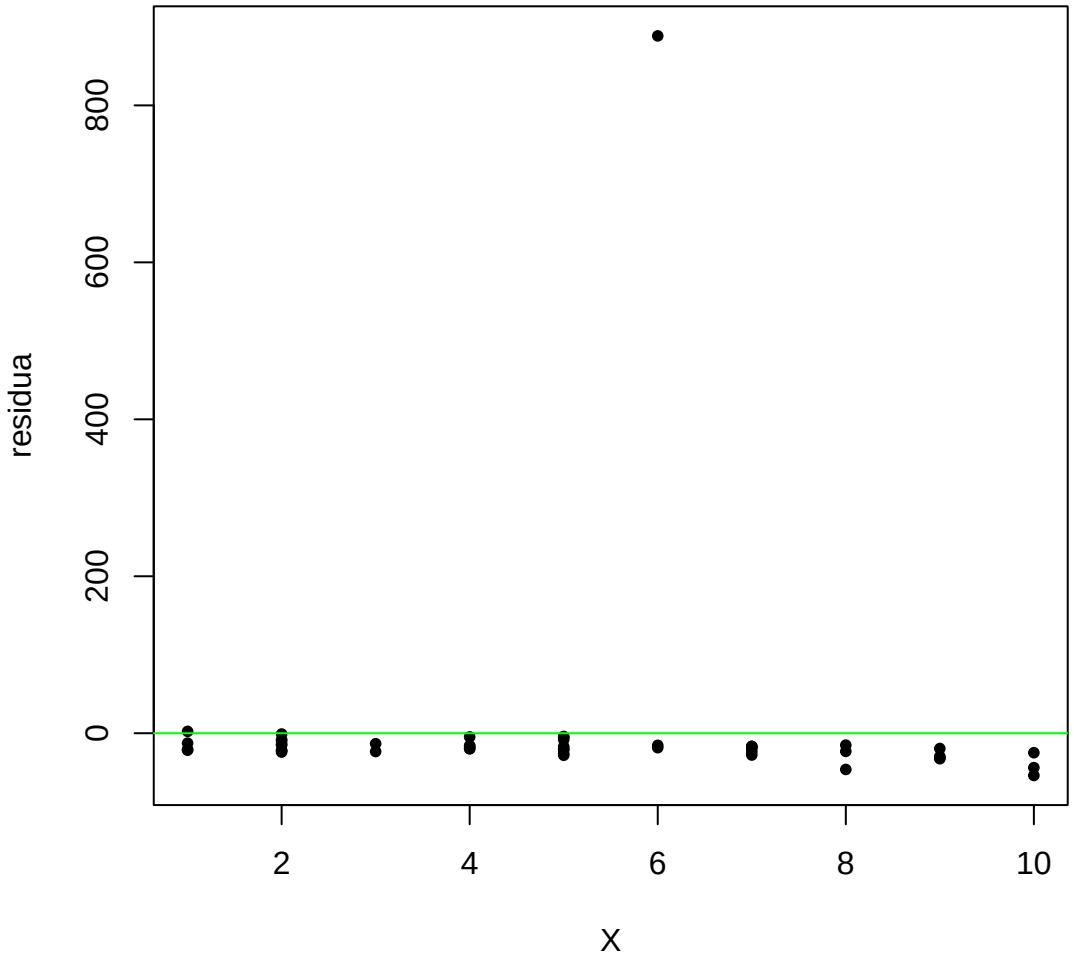
Tabela porównawcza po dodaniu obserwacji (1000;6)

Table 4: Tabela porównawcza

	oryginalne dane	dane z dodaną obserwacją (1000;6)
równanie regresji	Y = -0.58016 + 15.035 * X	Y = 7.308 + 17.3552 * X
wartość statystyki t	31.123	2.35942
p-wartość	4.009e-31	0.02281
współczynnik determinancji	0.9575	0.1123
estymator odchylenia standardowego	79.451	135.88

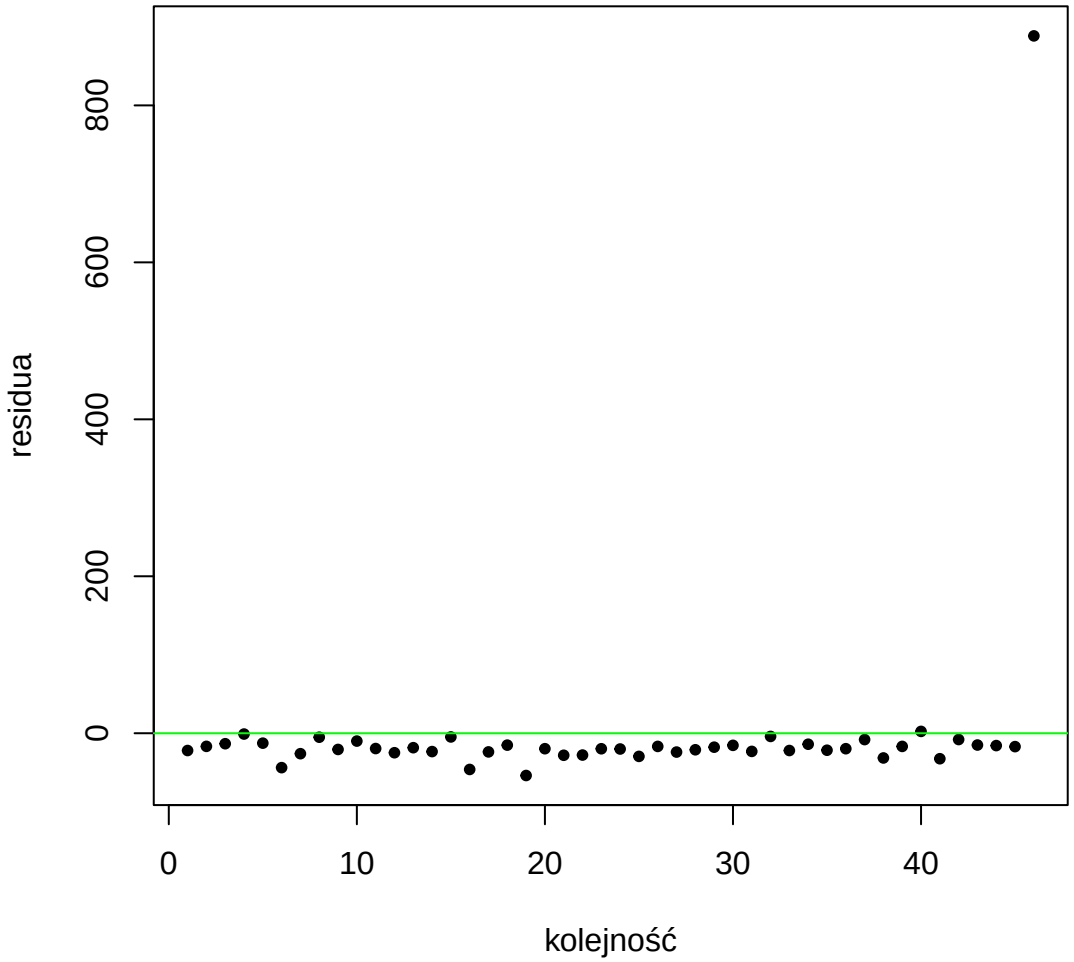
Dla obserwacji (1000;6) statystyka t jest zdecydowanie większa, zaś p-wartość: zdecydowanie mniejsza. Wzrósł również współczynnik determinancji, choć nadal jego wartość jest niska i wskazuje na bardzo słabe, nieużyteczne dopasowanie modelu. Estymator odchylenia standardowego jest niższy dla najnowszej wersji danych, ale wciąż jego wartość jest bardzo wysoka i tego samego rzędu, znacznie większa od wartości dla oryginalnego zbioru danych.

Wykres residuów względem zmiennej objaśniającej dla zbioru z obserwacją (1000;6)



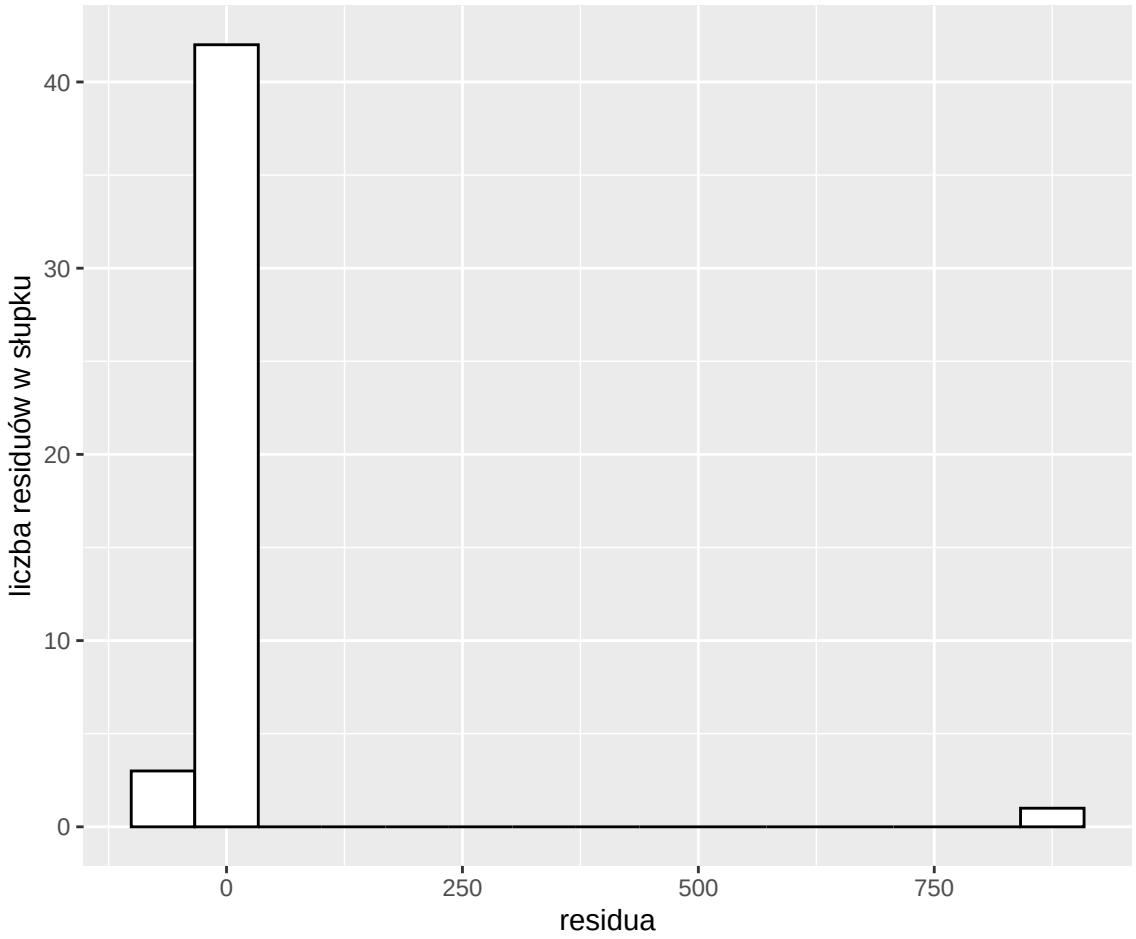
Wykres wygląda podobnie jak w przypadku dodania obserwacji (1000;2), jednak tutaj punkty zdają się układać w przybliżeniu liniowo w drugą stronę - maleją wraz ze wzrostem X, co po raz kolejny wskazuje na zależność między błędem a X.
Zdecydowanie więcej jest residuów o wartości ujemnej niż tych o wartości dodatniej.

Wykres residuów względem kolejności, w jakiej dane pojawiają się w pliku dla zbioru z obserwacją (1000;6)



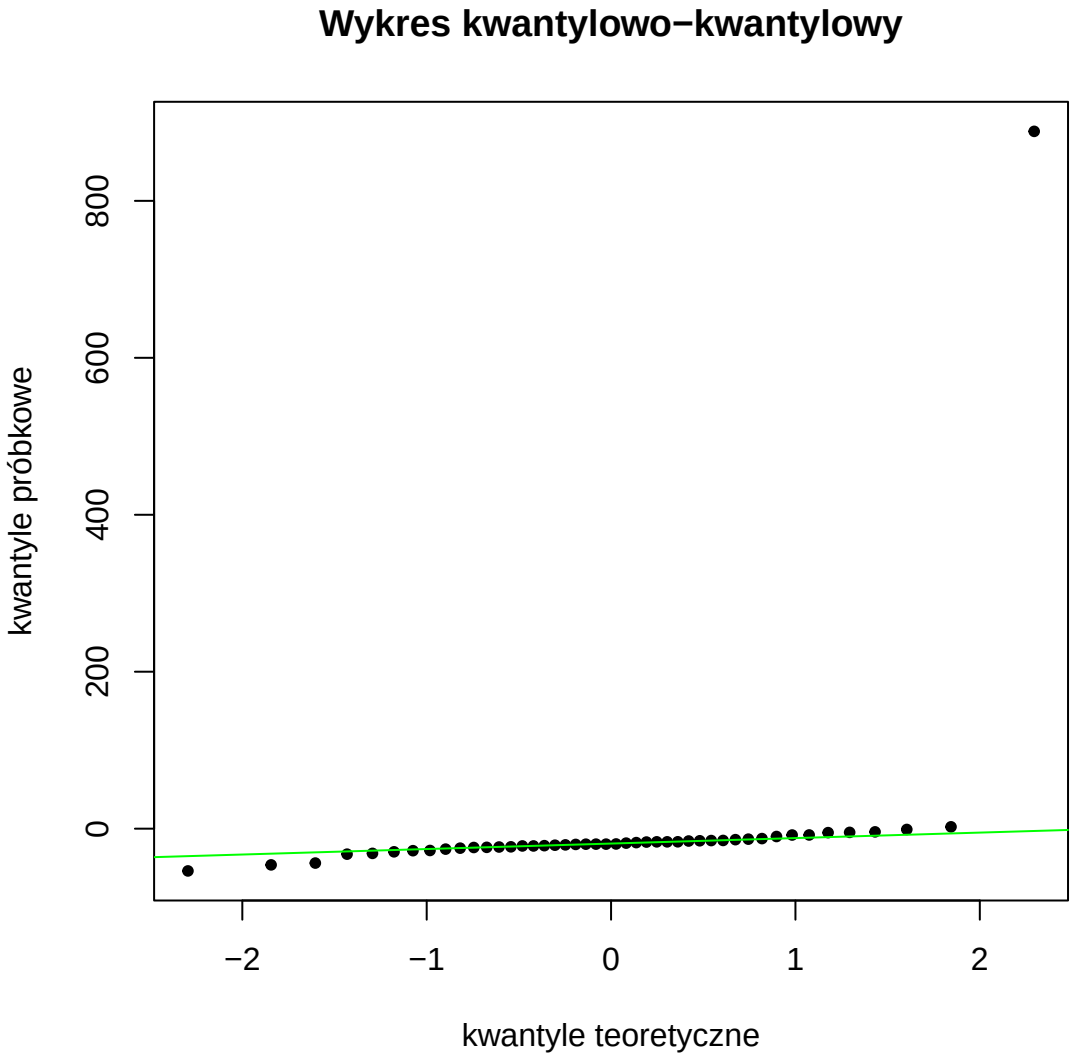
Na wykresie nie widać cykliczności jak w przypadku wykresu dla zbioru z obserwacją (1000;2), jednak wciąż w większości kolejne residua leżą blisko siebie (za wyjątkiem tego dla obserwacji odstającej), nie są one rozłożone równomiernie po obu stronach zera. Wskazuje to na brak niezależności błędów.

Histogram residuów dla zbioru z obserwacją (1000;6)



Podobnie jak we wcześniejszym przypadku, histogram jest nieczytelny i nie przypomina wizualnie rozkładu normalnego.

Wykres kwantylowo-kwantylowy residuów dla zbioru z obserwacją (1000;6)



W przypadku tego zbioru punkty na wykresie kwantylowo-kwantylowym zdają się układać jeszcze bliżej prostej (za wyjątkiem odstającego). Po raz kolejny widać tu pewne odbicie w stosunku do zbioru z obserwacją (1000;2): choć wartości również rosną wraz ze wzrostem kwantyli teoretycznych, to w tamtym przypadku zaczynały się one nad prostą, a kończyły poniżej niej, zaś tu jest odwrotnie - punkty początkowo znajdują się poniżej prostej, a wraz ze wzrostem przechodzą nad nią.

Wnioski

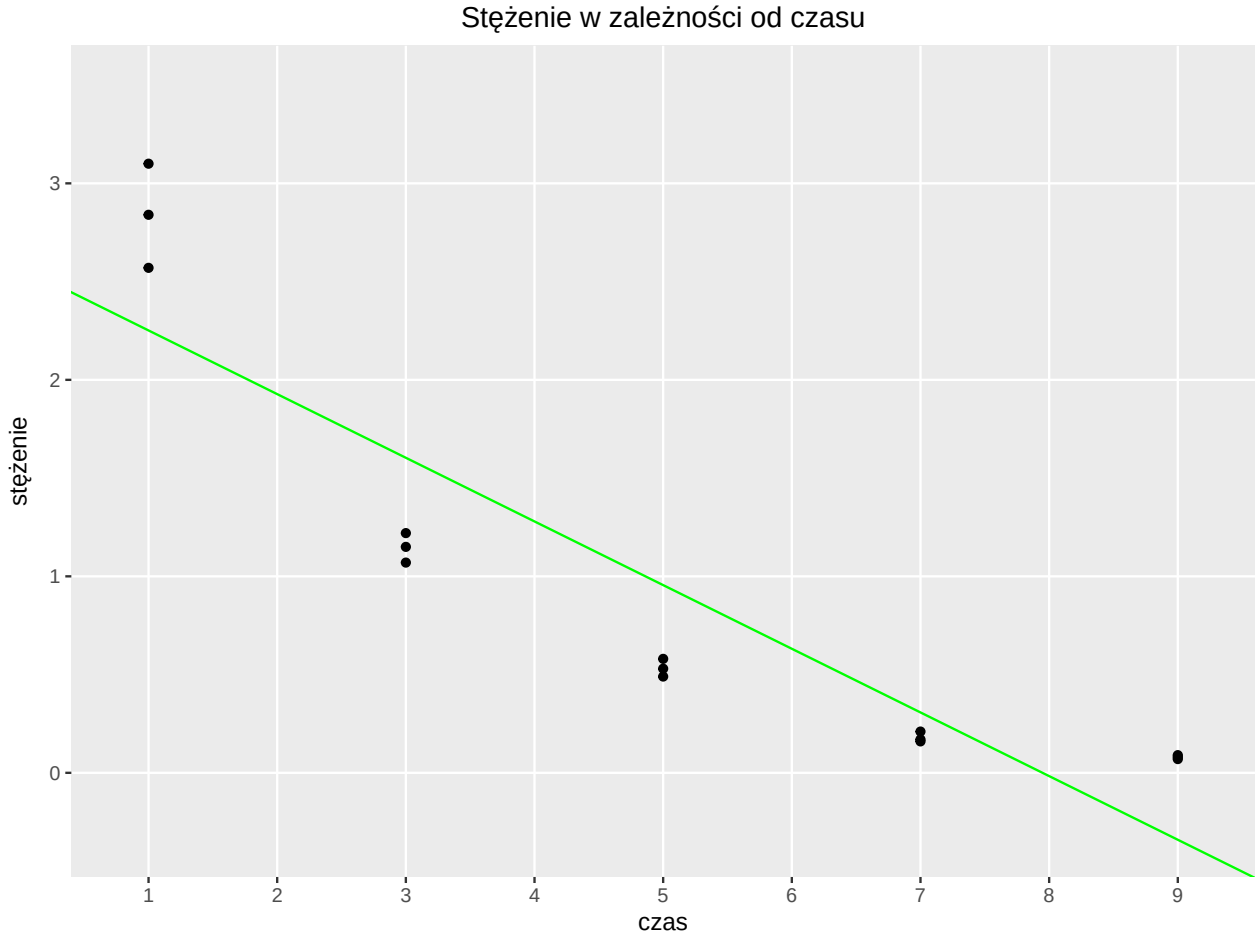
Ostatecznie wyniki prezentują się nieco lepiej niż w przypadku zbioru z obserwacją (1000;2). Dodanie obserwacji odstającej dla "środkowej" wartości $X = 6$ zaburzyło zbiór mniej niż dodanie takiej obserwacji dla wartości skrajnej $X = 2$, jednak wciąż przeprowadzanie regresji liniowej na takim zbiorze mija się z celem, ponieważ nie spełnia on założeń regresji liniowej.

Zadanie 5

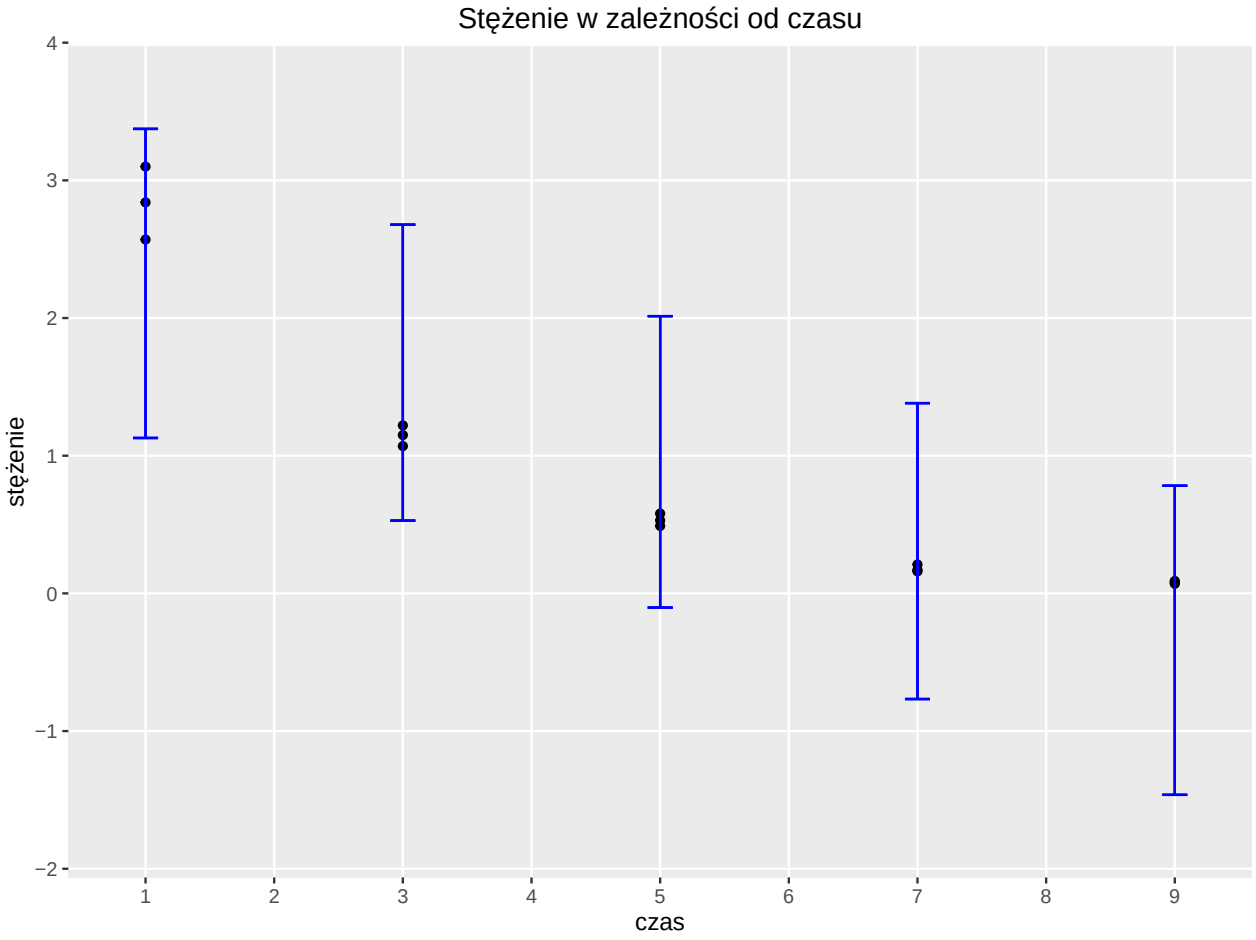
Równanie regresji

```
## [1] "stężenie = 2.5753 - 0.324 * czas"
```

Wykres z naniesioną prostą regresji



Wykres z naniesionymi przedziałami predykcyjnymi



Wnioski

Patrząc na to, jak blisko siebie znajdują się wartości stężeń dla danego czasu, a także na brak obserwacji odstających, przedziały predykcyjne są zaskakująco szerokie.
Choć początkowo prosta wydawała się całkiem dobrze dopasowana, bo bliższym przyjrzeniu sie ułożeniu punktów na wykresie bardziej przypomina ono funkcję wykładniczą o ujemnej potędze lub rozpad połowiczny, aniżeli zależność liniową. Przyjrzyjmy się temu bliżej.

Wartość współczynnika determinancji R^2

```
## [1] 0.811577
```

Współczynnik determinancji jest większy niż 0.8, co wskazuje na dość dobre dopasowanie prostej regresji do punktów na wykresie.

Test istotności dla hipotezy zerowej, że stężenie roztworu nie zależy od czasu

Hipotezy

- H_0: $\beta_1 = 0$ (stężenie roztworu nie zależy od czasu)
- H_A: $\beta_1 \neq 0$ (stężenie roztworu zależy od czasu)

Test będziemy przeprowadzać na poziomie istotności $\alpha = 0.05$.

Wartość statystyki testowej:

```
## [1] -7.483
```

Liczba stopni swobody $n - 2$:

```
## [1] 13
```

p-wartość:

```
## [1] 4.611199e-06
```

Jak widać, $p < \alpha$, zatem możemy odrzucić hipotezę zerową. Oznacza to, że istnieje zależność pomiędzy stężeniem a czasem.

Współczynnik korelacji między obserwowaną i przewidywaną wartością stężenia roztworu

```
## [1] 0.90088
```

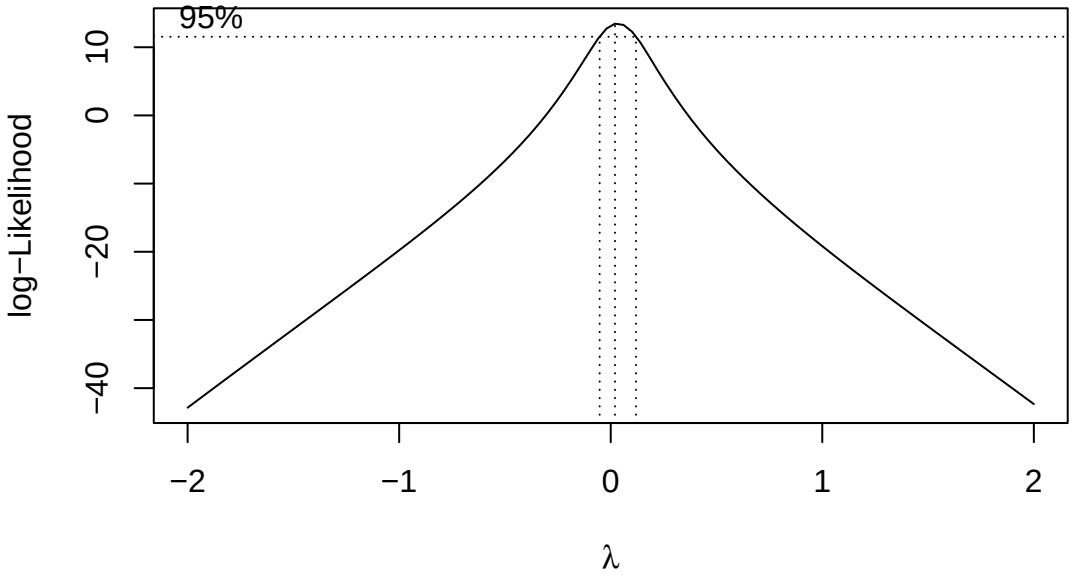
Wartość wyliczonego współczynnika korelacji jest dodatnia, bliska 0.9, zatem zachodzi korelacja pomiędzy obserwowaną i przewidywaną przez wyznaczony model liniowy wartością stężenia roztworu.

Podsumowując wszystkie obliczenia, wykresy i obserwacje - stężenie roztworu jest zależne od czasu, jednak podejrzewamy, że niekoniecznie jest to zależność liniowa (choć dla wybranego zbioru obserwacji może być ona bliska liniowej). Będziemy kontynuować badanie tego zbioru, zaczynając od poszukiwania odpowiedniej transformacji zbioru, aby uzyskać zależność liniową pomiędzy zmienną objaśniającą i objaśnianą.

Zadanie 6

Przeprowadzenie procedury Boxa-Coxa na zbiorze danych z zadania 5

Zobaczmy, jak wygląda wykres dla λ powstały w wyniku procedury:



Wyznaczamy wartość λ:

```
## [1] 0.020202
```

Ponieważ jest ona bliska zeru, przyjmujemy transformację $\tilde{Y} = \log(Y)$.

Zadanie 7

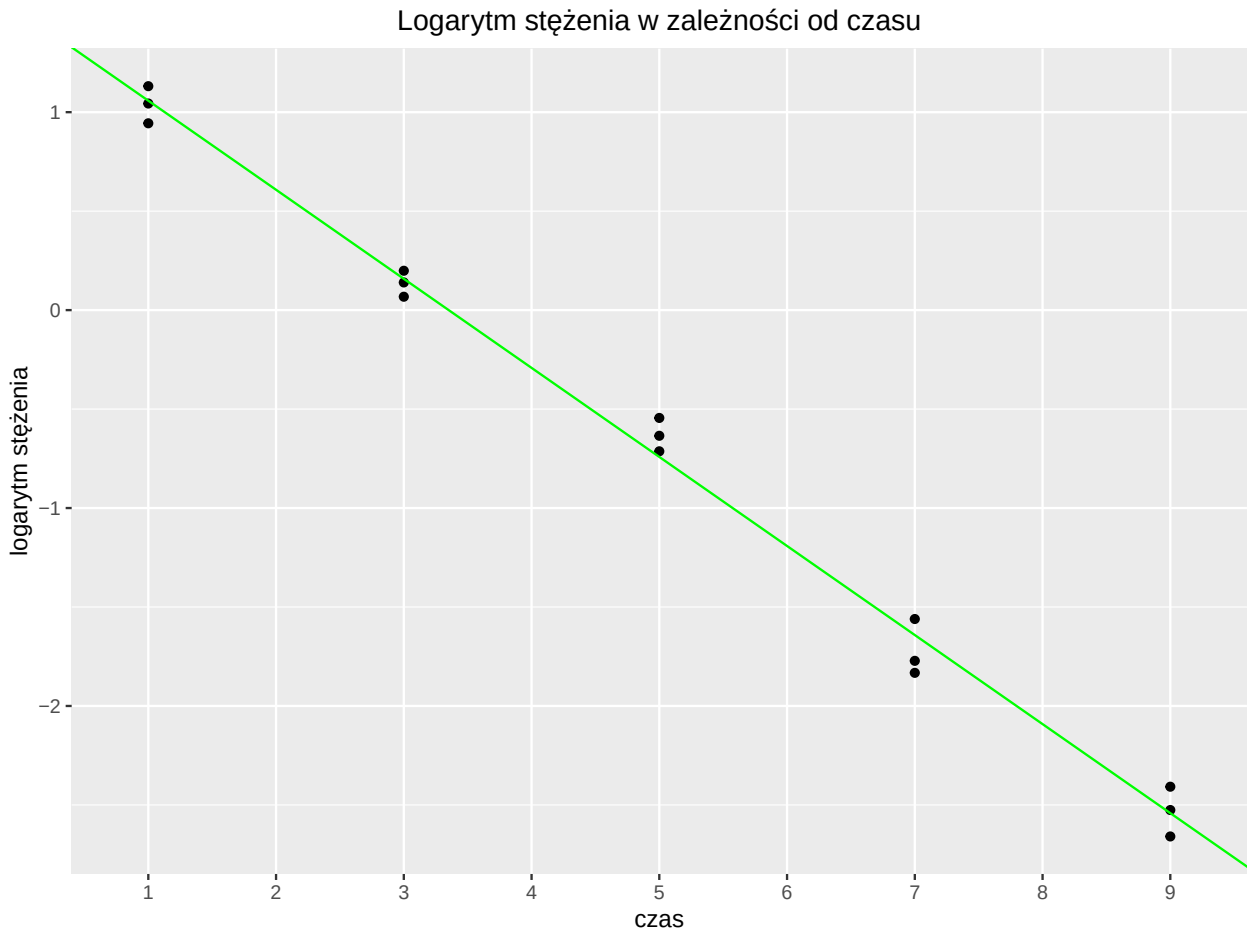
Powtarzamy zadanie 5, ale ze zmienioną zmienną $\tilde{Y} = \log(Y)$.

Równanie regresji

```
## [1] "stężenie = 1.5079 - 0.44993 * czas"
```

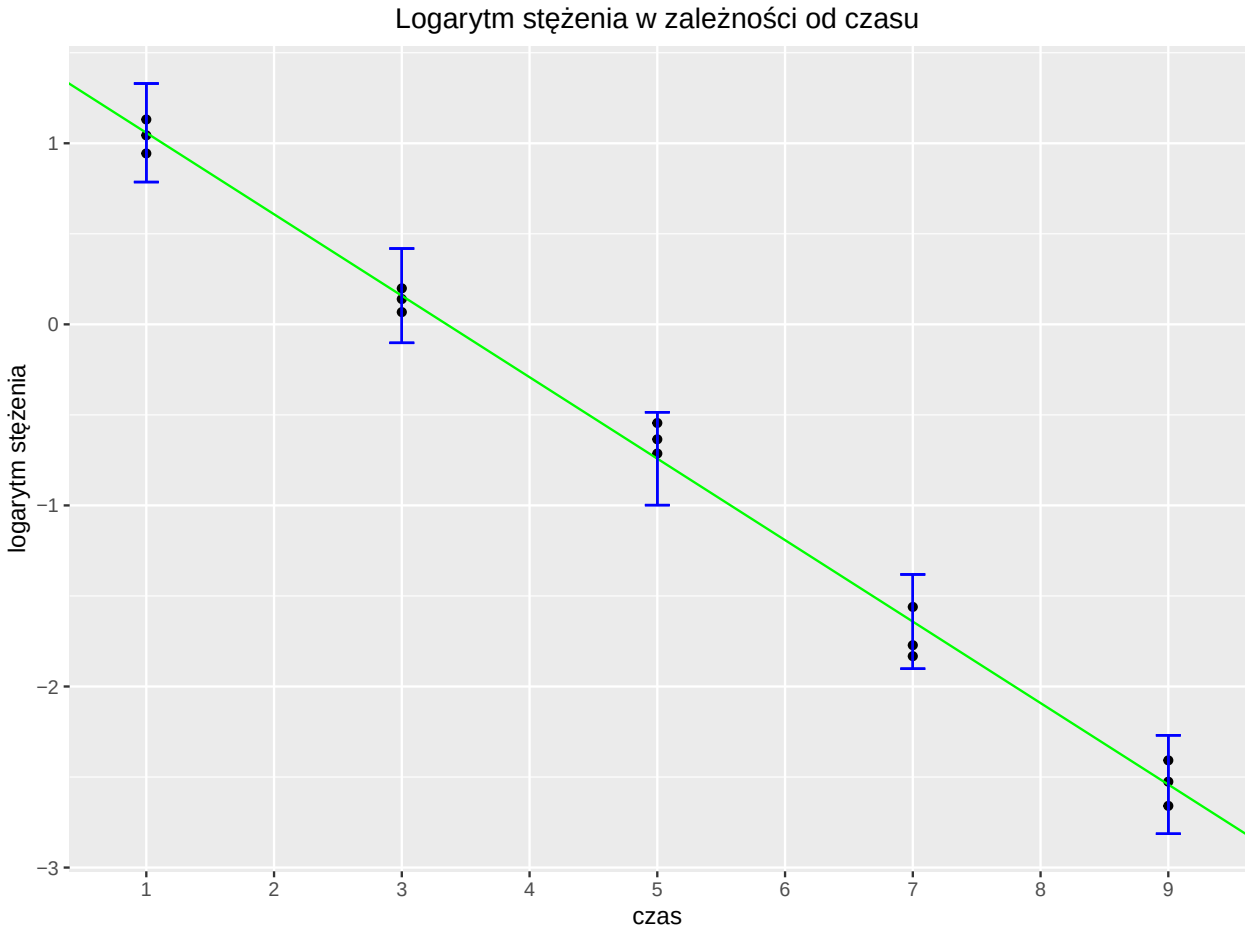
Oba współczynniki zmniejszyły się w stosunku do tych z pierwotnego równania. Zobaczmy, jak wygląda zmieniony wykres.

Wykres z naniesioną prostą regresji



Wykres wygląda znacznie lepiej niż ten w zadaniu 5. Punkty układają się niemal dokładnie wzdłuż linii.

Wykres z naniesionymi przedziałami predykcyjnymi



Przedziały predykcyjne są proporcjonalnie o wiele węższe niż przed transformacją, zajmują o wiele mniejszą część wykresu.

Wartość współczynnika determinancji R^2

```
## [1] 0.992978
```

Współczynnik determinancji jest bliski jedynki, co wskazuje na bardzo dobre, niemal idealne dopasowanie modelu do danych.

Test istotności dla hipotezy zerowej, że logarytm stężenia roztworu nie zależy od czasu

Hipotezy

- H_0: $\beta_1 = 0$ (logartym stężenia roztworu nie zależy od czasu)
- H_A: $\beta_1 \neq 0$ (logartym stężenia roztworu zależy od czasu)

Test będziemy przeprowadzać na poziomie istotności $\alpha = 0.05$.

Wartość statystyki testowej:

```
## [1] -42.875
```

Liczba stopni swobody $n-2$:

```
## [1] 13
```

p-wartość:

```
## [1] 2.188252e-15
```

Jak widać, $p < \alpha$ (i to znacznie), zatem możemy odrzucić hipotezę zerową. Oznacza to, że istnieje zależność pomiędzy logarytmem stężenia a czasem.

Współczynnik korelacji między obserwowaną i przewidywaną wartością stężenia roztworu

```
## [1] 0.99648
```

Współczynnik korelacji jest bliski 1, zatem zachodzi znacząca korelacja pomiędzy obserwowaną i przewidywaną przez wyznaczony model liniowy wartością logarytmu stężenia roztworu.
Podsumowując, model z $\tilde{Y} = \log(Y)$ jest zdecydowanie lepszy od pierwotnego. Porównajmy go jednak z jeszcze inną możliwością: $\tilde{t} = time^{-\frac{1}{2}}$.

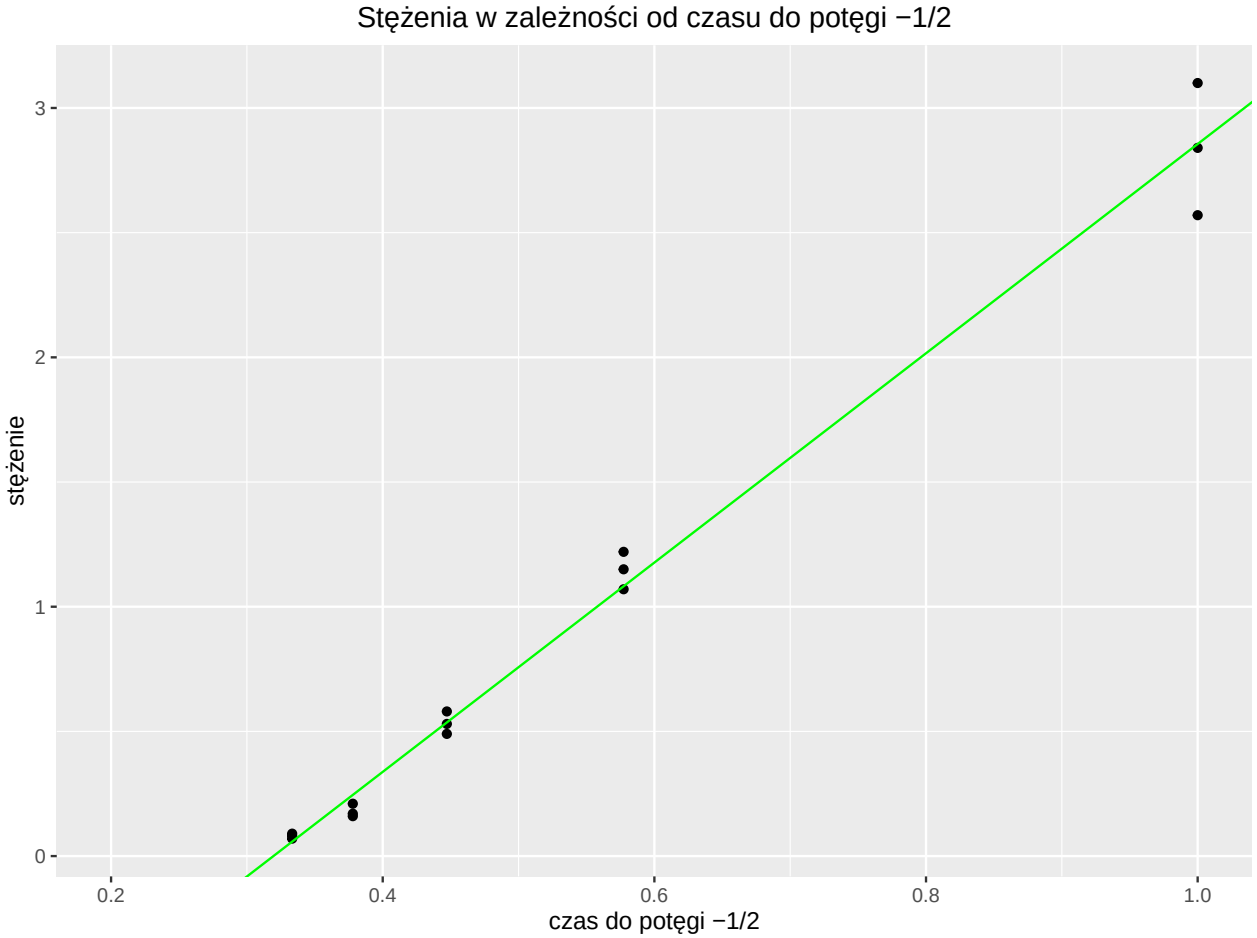
Zadanie 8

Konstruujemy nową zmienną objaśniającą $\tilde{t} = time^{-\frac{1}{2}}$, a następnie model $Y \sim \tilde{t}$.

Równanie regresji

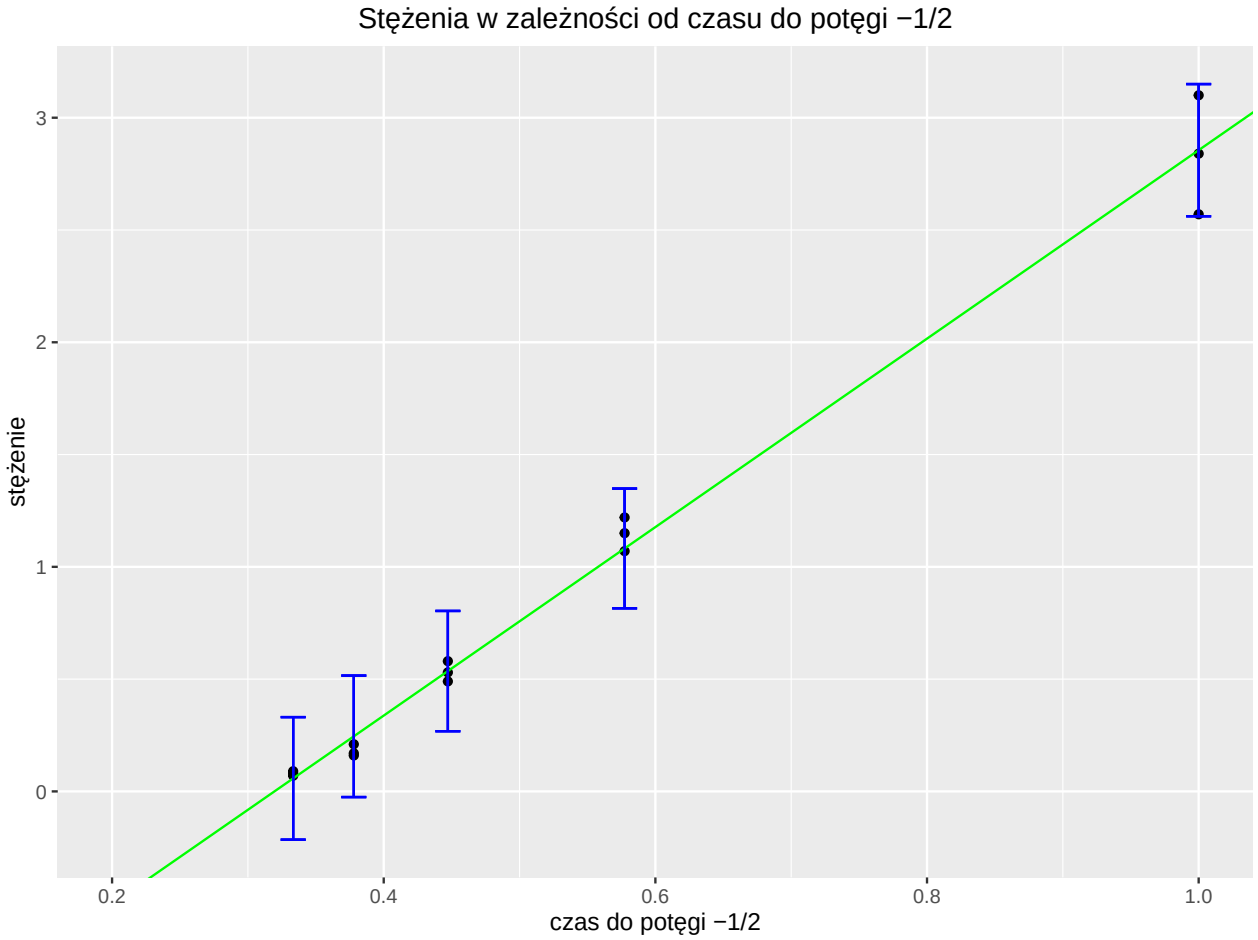
```
## [1] "stężenie = -1.3408 + 4.19632 * czas"
```

Wykres z naniesioną prostą regresji



Prosta na wykresie zmieniła kierunek. Wydaje się mniej dopasowana do punktów względem poprzedniej transformacji, ale lepiej niż w przypadku pierwotnych danych.

Wykres z naniesionymi przedziałami predykcyjnymi



Przedziały predykcyjne są stosunkowo nieco szersze niż w przypadku transformacji zmiennej objaśnianej, jednak węższe niż w przypadku oryginalnych danych.

Wartość współczynnika determinancji R^2

```
## [1] 0.988063
```

Współczynnik jest znacznie bliższy 1 niż pierwotny, jednak jego wartość jest niższa niż dla transformacji $\tilde{Y}$.

Test istotności dla hipotezy zerowej, że stężenie roztworu nie zależy od czasu do potęgi -1/2

Hipotezy

- H_0: $\beta_1 = 0$ (stężenie roztworu nie zależy od czasu do potęgi -1/2)
- H_A: $\beta_1 \neq 0$ (stężenie roztworu zależy od czasu do potęgi -1/2)

Test będziemy przeprowadzać na poziomie istotności $\alpha = 0.05$.

Wartość statystyki testowej:

```
## [1] 32.803
```

Liczba stopni swobody $n-2$:

```
## [1] 13
```

p-wartość:

```
## [1] 6.897696e-14
```

Jak widać, $p < \alpha$ (i to znacznie), zatem możemy odrzucić hipotezę zerową. Oznacza to, że istnieje zależność pomiędzy stężeniem a czasem do potęgi -1/2.

Współczynnik korelacji między obserwowaną i przewidywaną wartością stężenia roztworu

```
## [1] 0.99401
```

Współczynnik korelacji jest nieznacznie mniejszy niż dla transformacji $\tilde{Y}$, jednak wciąż bardzo bliski 1, zatem zachodzi znacząca korelacja pomiędzy obserwowaną i przewidywaną przez wyznaczony model liniowy wartością stężenia roztworu.

Wnioski

Model $\log(Y) \sim t$ okazał się lepszy niż model $Y \sim t^{-\frac{1}{2}}$, a oba te modele są lepsze niż dla pierwotny $Y \sim t$. Z praktycznego punktu widzenia model $\log(Y) \sim t$ jest lepszy przede wszystkim ze względu na węższe przedziały ufności, ale również pozostałe jego parametry prezentują się korzystniej (wyższa wartość R^2, a zatem lepsze dopasowanie, nieco większa korelacja). W zależności od kontekstu, może to mieć również uzasadnienie teoretyczne. Przykładowo w reakcjach pierwszego rzędu, czyli takich, w których równaniu kinetycznym (w postaci jednomianu potęgowego) suma wykładników potęg jest równa 1, wzór na stężenie c po czasie t (c_t), gdzie c_0 to stężenie początkowe, wyprowadza się następująco:

$$\frac{dc}{dt} = k \cdot c$$

$$\Updownarrow$$

$$-\frac{dc}{c} = k \cdot dt$$

$$\Updownarrow$$

$$\int_{c_0}^{c_t} \frac{1}{c}\, dc = \int_0^t -k\, dt$$

$$\Updownarrow$$

$$\ln(c_t) - \ln(c_0) = -kt$$

$$\Updownarrow$$

$$c_t = c_0 \cdot e^{-kt}$$

Zatem rzeczywiście w tym przypadku logarytm stężenia zależy liniowo od czasu.

Zadania teoretyczne

Zadanie 1

a)

Wartości t_c:

```
## [1] 2.570582 2.228139 2.008559
```

b)

Wartości F_c:

```
## [1] 6.607891 4.964603 4.034310
```

c)

Sprawdźmy, jak wygląda wektor różnic $t_c^2 - F_c$:

```
## [1] 0 0 0
```

Jak widać, wektor ten to same zera, wobec tego wartości te są takie same. Uzasadnijmy tę obserwację teoretycznie.
Przypomnijmy, że gęstość rozkładu t-studenta to:

$$f_X(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \cdot \Gamma(\frac{n}{2})(\frac{x^2}{n} + 1)^{-\frac{n+1}{2}}}$$

Gdzie:

- $x \in \mathbb{R}$
- $n \in \mathbb{N}$ to liczba stopni swobody

Weźmy dowolną zmienną losową $X \sim t(n)$. Niech Y będzie zadane przekształceniem $Y = g(X) = X^2$.
Wówczas przekształcenie odwrotne to $g^{-1}(Y) = \sqrt{Y}$.
Jakobian takiego przekształcenia to $J = \frac{dX}{dY} = \frac{1}{2\sqrt{Y}}$.
Wówczas gęstość Y to:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot |J| = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi} \cdot \Gamma(\frac{n}{2})(\frac{x^2}{n} + 1)^{-\frac{n+1}{2}}} \cdot |\frac{1}{2\sqrt{y}}| = \frac{\Gamma(\frac{n+1}{2}) \cdot (\frac{1}{n})^{\frac{1}{2}} \cdot y^{\frac{1}{2}-1} \cdot (1 + \frac{y}{n})^{-\frac{n+1}{2}}}{\Gamma(\frac{1}{2})\Gamma(\frac{n}{2})} = \frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \left(\frac{1}{n}\right)^{\frac{1}{2}} \cdot y^{\frac{1}{2}-1} \cdot (1$$

Zaś $h(y) = \frac{1}{B(\frac{1}{2}, \frac{n}{2})} \cdot \left(\frac{1}{n}\right)^{\frac{1}{2}} \cdot y^{\frac{1}{2}-1} \cdot (1 + \frac{y}{n})^{-\frac{n+1}{2}}$ jest gęstością rozkładu $F(1, n)$, co chcieliśmy pokazać.

Zadanie 2

a)

W pliku znajdują się 22 obserwacje. ### b) Zauważmy, że $s^2 = \frac{SSE}{dfE}$, zatem wartość s to:

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

c)

Statystyka testowa: $F = \frac{MSM}{MSE} = \frac{\frac{SSM}{dfM}}{\frac{SSE}{dfE}}$ $F(1, n-2) = F(1, 20)$
Wartość tej statystyki:

```
## [1] 5
```

```
## [1] 4.351244
```

Wartość krytyczna:

```
## [1] 4.351244
```

$F > F_c$, zatem odrzucamy hipotezę zerową. ### d) Jaką część zmienności zmiennej odpowiedzi wyjaśnia model?

Część całkowitej zmienności w wektorze Y (SST), którą stanowi zmienność wyjaśniona przez model (SSM). Mówi o tym współczynnik determinancji modelu R^2 (tu próbkowy współczynnik korelacji między zmienną odpowiedzi a zmienną objaśniającą).

e)

Policzmy próbkowy współczynnik korelacji między zmienną odpowiedzi a zmienną objaśniającą.
Skorzystamy ze wzoru $R^2 = \frac{SSM}{SST}$. Zatem wartość R wynosi:

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