Modele Liniowe - Lista 3

Jakub Kuciński 309881

Grudzień 2021

Spis treści

1 Zadanie 1

2 Zadanie 2*

Zadanie dodatkowe - oddane oddzielnie.

3 Zadanie 3

3.1 a)

Szukane wartości wyznaczyłem przy pomocy funkcji *summary* oraz *confint*. Wyestymowane równanie regresji: $Y = 0.10102 \cdot X - 3.55706$. Wyznaczona wartość R^2 to 0.4016146, czyli 40% zmienności zmiennej GPA stanowi stanowi zmienność wyjaśniona przez model. Testowana hipoteza zerowa H_0: $\beta_1 = 0$. Statystyka testowa ma postać: $T = \frac{\bar{\beta_1} - 0}{s(\bar{\beta_1})}$. Pochodzi z rozkładu t-studenta z $78 - 2 = 76$ stopniami swobody. Odpowiadająca p-wartość: $4.74e - 16$. Widzimy więc, że przy założeniu prawdziwości hipotezy zerowej, prawdopodobieństwo pojawiania się zdarzenia co najmniej tak rzadkiego jak nasze wynosi mniej niż $4.74e - 16$. Prawdopodobieństwo to jest bardzo bliskie zeru, więc można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną H_1: $\beta_1 \neq 0$.

3.2 b)

Oczekiwana wartość PGA dla IQ równego 100 wynosi 6.545114, a odpowiadający 90% przedział predykcyjny to $[3.79753, 9.292698]$.

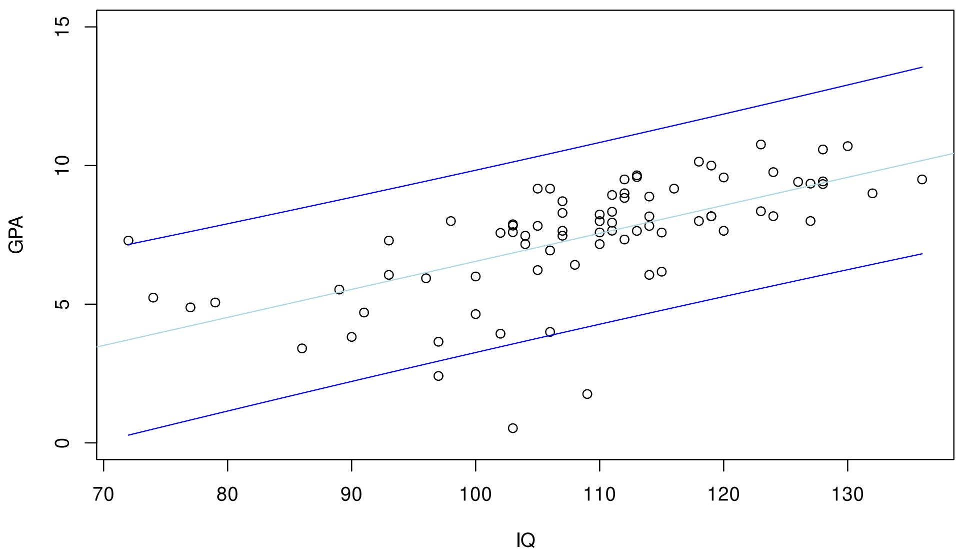
3.3 c)

Rysunek 1 przedstawia 95% pasmo predykcyjne. 4 obserwacje wypadają poza pasmem. Jest to zgodne z naszymi oczekiwaniami, bo zgodnie z teorią około 5% wszystkich obserwacji ($0.05 \cdot 78 = 3.9$) powinno wypadać poza przedziałami predykcyjnymi.

4 Zadanie 4

4.1 a, b)

Szukane wartości wyznaczyłem przy pomocy funkcji *summary* oraz *confint*. Wyestymowane równanie regresji: $Y = 0.09165 \cdot X + 2.22588$. Wyznaczona wartość



Rysunek 1: 95% pasmo predykcyjne

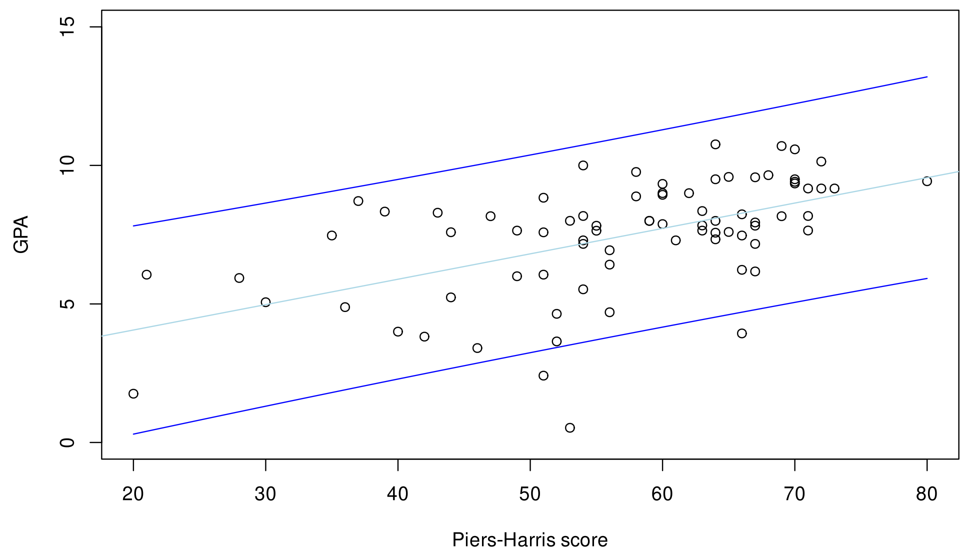
R^2 to 0.2935829, czyli 29% zmienności zmiennej GPA stanowi stanowi zmienność wyjaśniona przez model. Testowana hipoteza zerowa H_0: $\beta_1 = 0$. Statystyka testowa ma postać: $T = \dfrac{\bar{\beta_1} - 0}{s(\bar{\beta_1})}$. Pochodzi z rozkładu t-studenta z $78 - 2 = 76$ stopniami swobody. Odpowiadająca p-wartość: $3.01e - 07$. Widzimy więc, że przy założeniu prawdziwości hipotezy zerowej, prawdopodobieństwo pojawiania się zdarzenia co najmniej tak rzadkiego jak nasze wynosi mniej niż $3.01e - 07$. Prawdopodobieństwo to jest bardzo bliskie zeru, więc można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną H_1: $\beta_1 \neq 0$.

4.2 c)

Oczekiwana wartość PGA dla wyniku Piers-Harris równego 60 wynosi 7.72502, a odpowiadający 90% przedział predykcyjny to $[4.747302, 10.70274]$.

4.3 d)

Rysunek 2 przedstawia 95% pasmo predykcyjne. 3 obserwacje wypadają poza pasmem. Jest to zgodne z naszymi oczekiwaniami, bo zgodnie z teorią około 5% wszystkich obserwacji $(0.05 \cdot 78 = 3.9)$ powinno wypadać poza przedziałami predykcyjnymi.



Rysunek 2: 95% pasmo predykcyjne

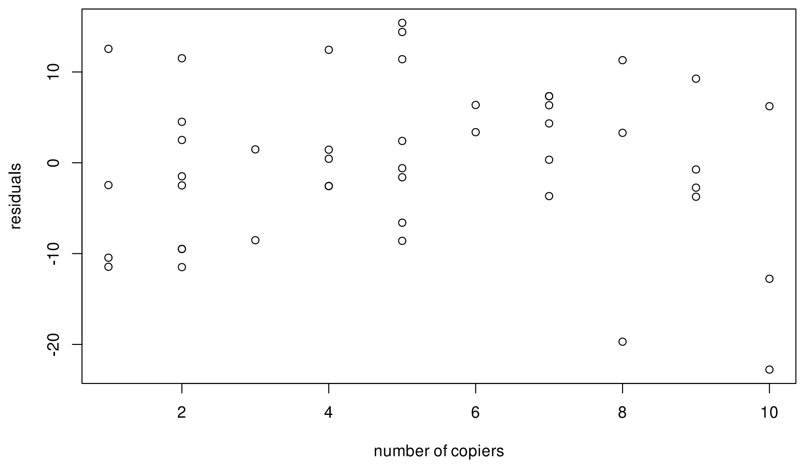
4.4 e)

P-wartości wskazują, że prawdopodobieństwo, że GPA jest niezależne od IQ jest mniejsze niż, że jest niezależne od wyniku Piers-Harris aczkolwiek obie p-wartości są niemal zerowe, więc GPA jest zależna od obu z nich. Na podstawie współczynnika R^2 widzimy, że model ze zmienną niezależną IQ wyjaśnia więcej zmienności zmiennej GPA (40%) niż model ze zmienną niezależną Piers-Harris (29%). Pasmo predykcyjne ze zmienną IQ również wydaje się węższe od pasma z Piers-Harris. Na podstawie tych obserwacji możemy stwierdzić, że IQ jest lepszym predyktorem zmiennej GPA.

5 Zadanie 5

5.1 a)

Suma residuów wynosi -1.176836e-14. Jest to liczba niemal równa 0. Drobny błąd może być wynikiem błędów numerycznych. Otrzymana wartość jest zgodna z faktem, że w modelu liniowym średnia błędu jest równa 0.



Rysunek 3: Wykres residuów względem zmiennej objaśniającej.

5.2 b)

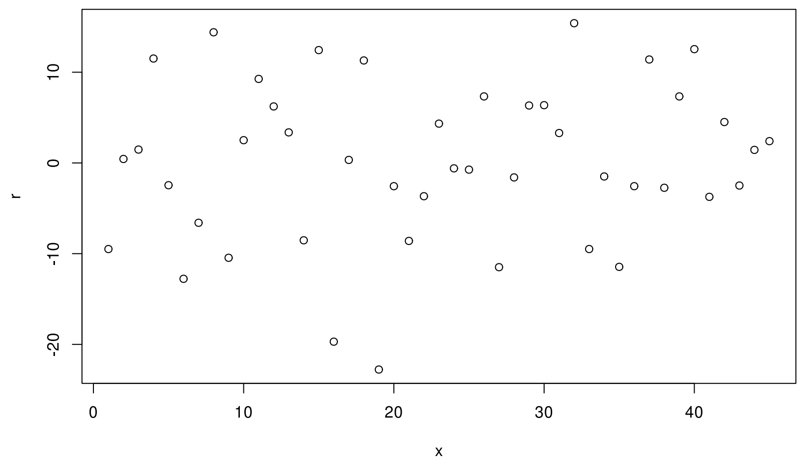
Rysunek 3 przedstawia wartości residuów względem zmiennej objaśniającej. Widzimy, że ich rozmieszczenie wokół wartości 0 wygląda losowo, a wariancja wygląda na stałą względem wartości zmiennej objaśniającej. Dwa residua nieco odstają od pozostałych przyjmując wartości w okolicach -20.

5.3 c)

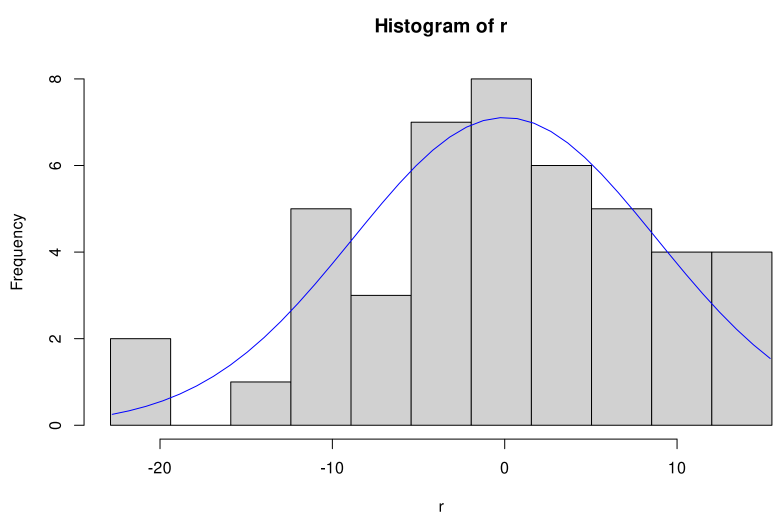
Rysunek 3 przedstawia wartości residuów względem zmiennej objaśniającej. Widzimy, że ich rozmieszczenie wokół wartości 0 wygląda losowo, wariancja nie zmienia się. Dwa residua nieco odstają od pozostałych przyjmując wartości w okolicach -20.

5.4 d)

Wykres (nr 5) prawdopodobieństwa normalnego w większości zgadza się z otrzymanym histogramem, możemy więc podejrzewać, że błędy pochodzą z rozkładu normalnego. Widzimy jednak, że powinniśmy zgodnie z rozkładem normalnym powinniśmy otrzymać znacznie mniej residuów odstających na poziomie -20 niż w rzeczywistości zaobserwowaliśmy (lewa strona wykresu).



Rysunek 4: Wykres residuów względem pojawienia się w zbiorze danych.



Rysunek 5: Histogram i wykres prawdopodobieństwa normalnego.

6 Zadanie 6

6.1 a)

	Oryginalne dane	Zmienione dane
Dopasowane równanie	$Y = 15.0352 \cdot X - 0.5802$	$Y = -3.059 \cdot X + 135.900$
T-test	31.123	-0.193
P-wartość	<2e-16	0.848
R^2	0.9575	0.000863
$Estymowane \sigma$	8.914	292.8

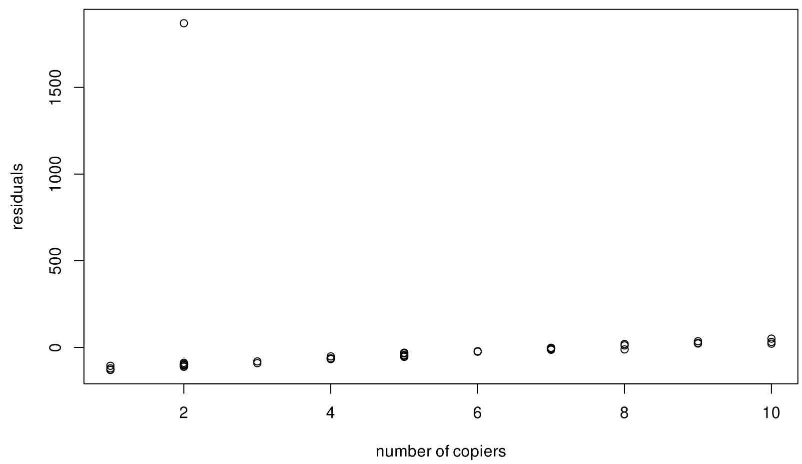
Dodanie odstającej obserwacji znacząco zmieniło równanie dopasowanej prostej. Widzimy również, że testowana hipoteza zerowa ($H_0 : \beta_1 = 0$) zostaje stanowczo odrzucona w oryginalnych danych (p-wartość bardzo bliska 0), natomiast nie ma podstaw do odrzucenia hipotezy zerowej w zmienionych danych (p-wartość bardzo wysoka 0.848), czyli w drugim przypadku zmienna wynikowa wydaje się niezwiązana ze zmienną niezależną. Estymowana wariancja błędu w modelu ze zmienionymi danymi znacząco zwiększą się względem oryginalnego (do 292.8 z 8.914). Widzimy też, że o ile dla oryginalnych danych model wyjaśniał aż 0.9575 zmienności zmiennej zależnej, to dla zmienionych danych model wyjaśnia już tylko 0.000863 zmienności, czyli niemal w ogóle nie wyjaśnia. Widzimy więc, że pojedyncza, mocno odstająca obserwacja może całkowicie uniemożliwić nam zaobserwowanie liniowych zależności w danych.

6.2 b)

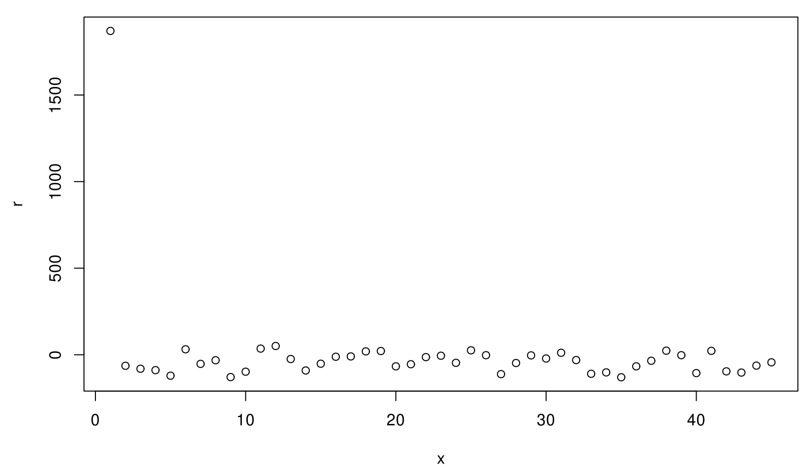
Na wykresach 6 i 6 widzimy, że wprowadzona zmienne odstająca znacznie odstaje od wszystkich pozostałych, które z kolei są skoncentrowane w okolicach wartości 0. Na wykresie 6 widzimy, że pojawienie się wartości odstającej zakłóca dopasowanie się modelu do pozostałych obserwacji - dla mniejszej liczby kopiarek dostajemy residua ujemna, a dla większych dodatnie, a powinniśmy dostawać symetryczne i losowe odchylenia wokół wartości 0. Przez nią również histogram 8 znacząco odbiega od otrzymanego wykresu prawdopodobieństwa normalnego - wokół zera mamy zdecydowanie więcej residuów niż powinniśmy mieć, ponad wartością 1500 nie powinniśmy zaobserwować żadnej obserwacji (a mamy jedną), natomiast pomiędzy nimi nie obserwujemy żadnej, chociaż powinniśmy.

7 Zadanie 7

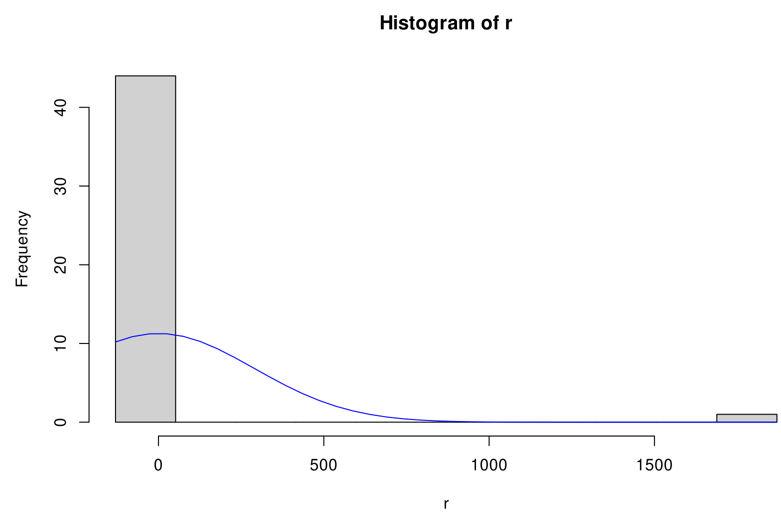
Wyestymowane równanie regresji: $Y = -0.3240 \cdot X + 2.5753$. Testowana hipoteza zerowa H_0: $\beta_1 = 0$. Statystyka testowa ma postać: $T = \frac{\bar{\beta_1} - 0}{s(\bar{\beta_1})}$ i pochodzi z rozkładu t-studenta z 13 stopniami swobody. Odpowiadająca p-wartość: $4.61e-06$.



Rysunek 6: Wykres residuów względem zmiennej objaśniającej.



Rysunek 7: Wykres residuów względem pojawienia się w zbiorze danych.



Rysunek 8: Histogram i wykres prawdopodobieństwa normalnego.

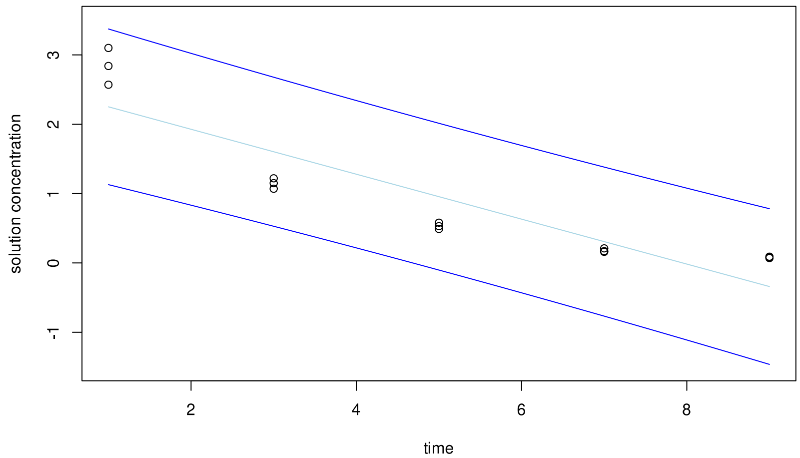
Widzimy więc, że przy założeniu prawdziwości hipotezy zerowej, prawdopodobieństwo pojawiania się zdarzenia co najmniej tak rzadkiego jak nasze wynosi mniej niż $4.61e-06$. Prawdopodobieństwo to jest bardzo małe (w szczególności mniejsze niż zazwyczaj przyjmowane 0.05), więc można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną H_1: $\beta_1 \neq 0$, czyli że zmienna wynikowa jest liniowo zależna od zmiennej objaśniającej. Współczynnik R^2 wynosi 0.8116, więc około 81% zmienności zmiennej zależnej jest tłumaczona przez model.

8 Zadanie 8

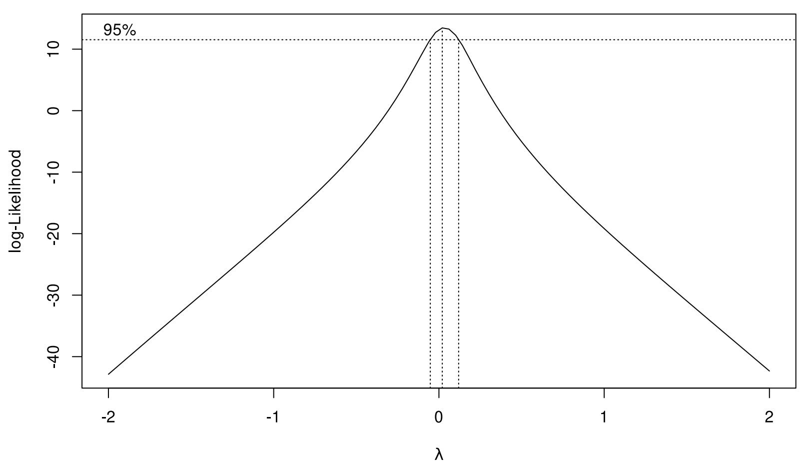
Na 95% paśmie predykcyjnym 9 widzimy, że prosta regresji nie dopasowuje się dobrze do danych, a pasmo predykcyjne jest bardzo szerokie. Po sposobie ułożenia punktów możemy podejrzewać, że relacja jest nieliniowa. Wyliczony współczynnik korelacji wynosi wysoki i wynosi 0.9008759 (pierwiastek z współczynnika R^2).

9 Zadanie 9

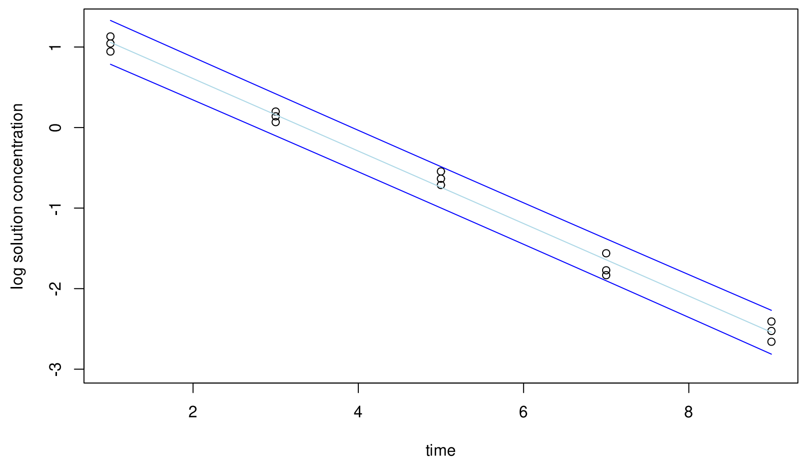
W 95% przedziale ufności (w okolicach środka) na rysunku 10 dla λ mamy wartość 0, która odpowiada przekształceniu $Y' = log(Y)$.



Rysunek 9: 95% pasmo predykcyjne



Rysunek 10: Transformacja Boxa-Coxa



Rysunek 11: 95% pasmo predykcyjne

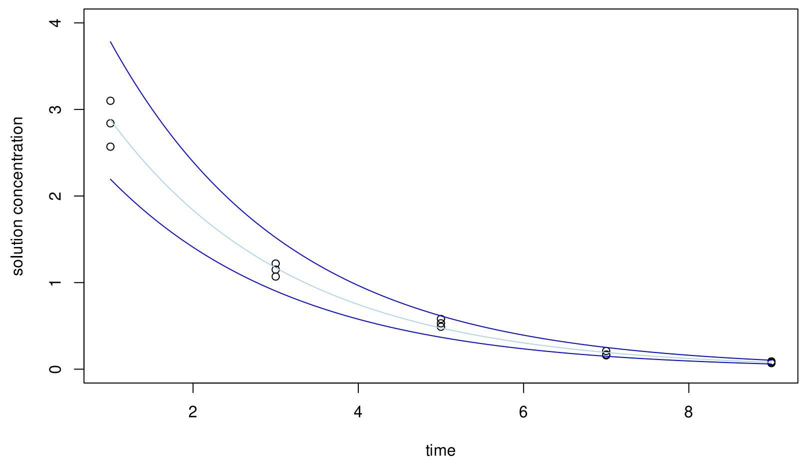
10 Zadanie 10

Wyestymowane równanie regresji: $Y = -0.44993 \cdot X + 1.50792$. Testowana hipoteza zerowa H_0: $\beta_1 = 0$. Statystyka testowa ma postać: $T = \frac{\bar{\beta_1} - 0}{s(\bar{\beta_1})}$ i pochodzi z rozkładu t-studenta z 13 stopniami swobody. Odpowiadająca p-wartość: $2.19e - 15$. Widzimy więc, że przy założeniu prawdziwości hipotezy zerowej, prawdopodobieństwo pojawiania się zdarzenia co najmniej tak rzadkiego jak nasze wynosi mniej niż $2.19e - 15$, czyli niemal zero, więc można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną H_1: $\beta_1 \neq 0$, czyli że zmienna wynikowa jest liniowo zależna od zmiennej objaśniającej. Współczynnik R^2 wynosi 0.993, więc aż 99%, czyli prawie cała zmienność zmiennej zależnej jest tłumaczona przez model.

Na 95% paśmie predykcyjnym 11 widzimy, że nowa prosta regresji zdecydowanie lepiej dopasowuje się do danych niż ta przed nałożeniem logarytmu na zmienną wynikową. Wyliczony współczynnik korelacji jest wysoki i wynosi 0.9964826 (pierwiastek z R^2), czyli prawie pełne 1 i zdecydowanie więcej niż dla poprzedniego modelu.

11 Zadanie 11

Rysunek 12 przedstawia 95% pasmo predykcyjne w wejściowych jednostkach dla modelu ze zmienną wynikową *logy*. Widzimy, że model ten zdecydowanie lepiej dopasowuje się do danych i pasmo jest zdecydowanie węższe niż dla mo-



Rysunek 12: 95% pasmo predykcyjne

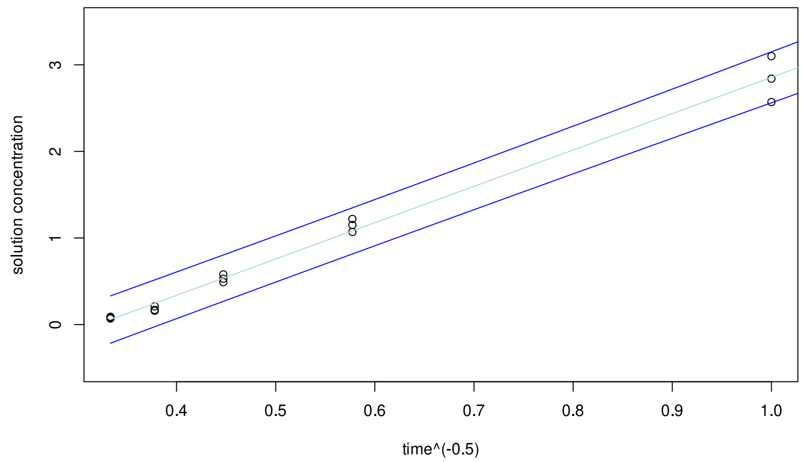
delu zaaplikowanego na wejściowych zmiennych. Współczynnik korelacji wyniósł 0.9945587, czyli sporo więcej niż dla pierwszego modelu.

12 Zadanie 12

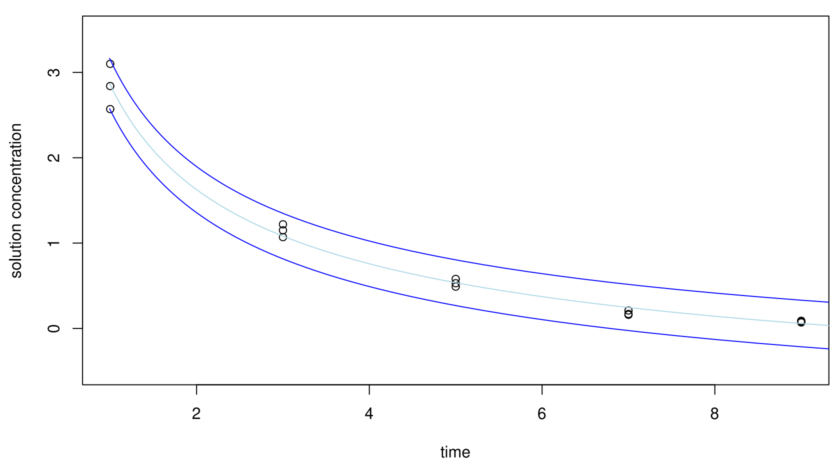
Wyestymowane równanie regresji: $Y = 4.19632 \cdot X - 1.34078$. Testowana hipoteza zerowa H_0: $\beta_1 = 0$. Statystyka testowa ma postać: $T = \dfrac{\bar{\beta}_1 - 0}{s(\bar{\beta}_1)}$ i pochodzi z rozkładu t-studenta z 13 stopniami swobody. Odpowiadająca p-wartość: $6.90e-14$. Widzimy więc, że przy założeniu prawdziwości hipotezy zerowej, prawdopodobieństwo pojawiania się zdarzenia co najmniej tak rzadkiego jak nasze wynosi mniej niż $6.90e - 14$, czyli niemal zero, więc można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną H_1: $\beta_1 \neq 0$, czyli że zmienna wynikowa jest liniowo zależna od zmiennej objaśniającej. Współczynnik R^2 wynosi 0.9881, więc około 99%, czyli prawie cała zmienność zmiennej zależnej jest tłumaczona przez model.

Na 95% paśmie predykcyjnym 13 widzimy, że nowa prosta regresji zdecydowanie lepiej dopasowuje się do danych niż model zaaplikowany na surowych danych. Wyliczony współczynnik korelacji jest wysoki i wynosi 0.9940136 (pierwiastek z R^2), czyli prawie pełne 1.

Rysunek 12 przedstawia to samo 95% pasmo predykcyjne, ale w wejściowych jednostkach dla modelu ze zmienną objaśniającą $time^{-1/2}$. Widzimy, że model ten zdecydowanie lepiej dopasowuje się do danych i pasmo jest zdecydowanie węższe niż dla modelu zaaplikowanego na wejściowych zmiennych. Współczyn-



Rysunek 13: 95% pasmo predykcyjne



Rysunek 14: 95% pasmo predykcyjne

nik korelacji wyniósł 0.9940136, czyli sporo więcej niż dla pierwszego modelu.

Pierwszy model zachowywał się zdecydowanie gorzej od dwóch pozostałych. Wyjaśniał mniej zmienności zmiennej zależnej i oglądając wykres było widać, że zdecydowanie gorzej dopasowuje się do danych. Drugi i trzeci model osiągały bardzo podobną w kontekście współczynnika R^2, korelacji i p-value. Zdecydowanie lepiej dopasowują się też do danych. Widzimy jednak, że model ze zmienną objaśnianą *logy* posiada lepsze przedziały predykcyjne od modelu ze zmienną objaśniającą $time^{-1/2}$. Dla mniejszych wartości zmiennej niezależnej obserwujemy większy rozrzut wartości niż dla większych wartości tej zmiennej, dlatego też powinniśmy preferować model, który posiada szersze przedziały predykcyjne dla mniejszych wartości i węższe dla większych wartości zmiennej niezależnej. Stąd też preferujemy drugi model nad trzecim.

13 Kod w R

```
# Zad 1
alpha = 0.05
deg = 10
# a)
tc = qt(p=1-alpha/2, df=deg)
# b)
Fc = qf(p=1-alpha, df1 = 1, df2 = deg)
# c)
Fc_sqrt = sqrt(Fc)
print(Fc_sqrt == tc)

# Zad 3
data = read.table("/home/kuba/Documents/UWr/Modele-Liniowe/List3/Tabela1_6.txt", col.names=c
# a)
reg1 = lm(GPA~IQ, data)
summary(reg1)
predictedGPA = predict.lm(reg1, data.frame(IQ=c(data$IQ)), interval = "confidence", level=0.
SST = sum((data$GPA - mean(data$GPA))^2)
SSE = sum((data$GPA - predictedGPA)^2)
R2 = 1 - SSE/ SST
R2


# b)
predict.lm(reg1, data.frame(IQ=c(100)), interval = "prediction", level=0.90)

# c)
x = data$IQ
y = data$GPA
newx = seq(min(x),max(x),by = 0.05)
```

```
conf_interval = predict(reg1, newdata=data.frame(IQ=newx), interval="prediction",level = 0.9
plot(x, y, ylim=c(0, 15),  ylab="GPA", xlab="IQ")
abline(reg1, col="lightblue")
matlines(newx, conf_interval[,2:3], col = "blue", lty=25)


# Zad 4
data = read.table("/home/kuba/Documents/UWr/Modele-Liniowe/List3/Tabela1_6.txt", col.names=c
# a,b)
reg2 = lm(GPA~pstest, data)
summary(reg2)
predictedGPA = predict.lm(reg2, data.frame(pstest=c(data$pstest)), interval = "confidence",
SST = sum((data$GPA - mean(data$GPA))^2)
SSE = sum((data$GPA - predictedGPA)^2)
R2 = 1 - SSE/ SST
R2

# c)
predict.lm(reg2, data.frame(pstest=c(60)), interval = "prediction", level=0.90)

# d)
x = data$pstest
y = data$GPA
newx = seq(min(x),max(x),by = 0.05)
conf_interval = predict(reg2, newdata=data.frame(pstest=newx), interval="prediction",level =
plot(x, y, ylim=c(0, 15),  ylab="GPA", xlab="Piers-Harris score")
abline(reg2, col="lightblue")
matlines(newx, conf_interval[,2:3], col = "blue", lty=25)

# e)


# Zad 5
data = read.table("/home/kuba/Documents/UWr/Modele-Liniowe/List3/CH01PR20.txt", col.names=c(
reg1 = lm(hours~copiers, data)

# a)
r<-residuals(reg1)
sum(r)

# b)
plot(r~data$copiers, ylab="residuals", xlab="number of copiers")
# abline(h = 0)

# c)
x<-seq(1:dim(data)[1])
```

```
plot(r~x)

# d)
h<-hist(r, breaks = seq(min(r)-0.1,max(r)+0.1,length.out = 12));
m<-mean(r);
s<-sd(r);
xfit<-seq(min(r),max(r),length=40);
d<-dnorm(xfit,m,s);
d <- d*diff(h$mids[1:2])*length(r)
lines(d~xfit, col='blue')


# Zad 6
data2 = read.table("/home/kuba/Documents/UWr/Modele-Liniowe/List3/CH01PR20.txt", col.names=c
data2[1, 1] = 2000
reg2 = lm(hours~copiers, data2)

# a)
summary(reg1)
summary(reg2)

# b)
r<-residuals(reg2)
plot(r~data$copiers, ylab="residuals", xlab="number of copiers")

x<-seq(1:dim(data)[1])
plot(r~x)

h<-hist(r, breaks = seq(min(r)-0.1,max(r)+0.1,length.out = 12));
m<-mean(r);
s<-sd(r);
xfit<-seq(min(r),max(r),length=40);
d<-dnorm(xfit,m,s);
d <- d*diff(h$mids[1:2])*length(r)
lines(d~xfit, col='blue')


# Zad 7
data = read.table("/home/kuba/Documents/UWr/Modele-Liniowe/List3/CH03PR15.txt", col.names=c(
reg1 = lm(Y~X, data)
summary(reg1)


# Zad 8
x = data$X
y = data$Y
```

```
newx = seq(min(x),max(x),by = 0.05)
conf_interval = predict(reg1, newdata=data.frame(X=newx), interval="prediction",level = 0.95
plot(x, y, xlab="time", ylim=c(-1.5, 3.5), ylab="solution concentration")
matlines(newx, conf_interval[,1], col = "lightblue", lty=25)
matlines(newx, conf_interval[,2:3], col = "blue", lty=25)

real = data$Y
pred = predict.lm(reg1, data.frame(X=data$X), interval = "prediction", level=0.95)[, 1]
cor(real, pred, method ="pearson")


# Zad 9
require(MASS)
boxcox(data$Y~data$X)


# Zad 10
data = read.table("/home/kuba/Documents/UWr/Modele-Liniowe/List3/CH03PR15.txt", col.names=c(
data$Y = log(data$Y)
reg1 = lm(Y~X, data)
summary(reg1)

x = data$X
y = data$Y
newx = seq(min(x),max(x),by = 0.05)
conf_interval = predict(reg1, newdata=data.frame(X=newx), interval="prediction",level = 0.95
plot(x, y, ylim=c(-3, 1.3), xlab="time", ylab="log solution concentration")
matlines(newx, conf_interval[,1], col = "lightblue", lty=25)
matlines(newx, conf_interval[,2:3], col = "blue", lty=25)

real = data$Y
pred = predict.lm(reg1, data.frame(X=data$X), interval = "prediction", level=0.95)[, 1]
cor(real, pred, method ="pearson")


# Zad 11
data = read.table("/home/kuba/Documents/UWr/Modele-Liniowe/List3/CH03PR15.txt", col.names=c(
data$Y = log(data$Y)
reg1 = lm(Y~X, data)
x = data$X
y = data$Y
newx = seq(min(x),max(x),by = 0.05)
conf_interval = predict(reg1, newdata=data.frame(X=newx), interval="prediction",level = 0.95
pred = predict.lm(reg1, data.frame(X=data$X), interval = "prediction", level=0.95)[, 1]

data = read.table("/home/kuba/Documents/UWr/Modele-Liniowe/List3/CH03PR15.txt", col.names=c(
```

```
reg1 = lm(Y~X, data)
x = data$X
y = data$Y
newx = seq(min(x),max(x),by = 0.05)
plot(x, y, ylim=c(0, 4), xlab="time", ylab="solution concentration")
matlines(newx, exp(conf_interval[,1]), col = "lightblue", lty=25)
matlines(newx, exp(conf_interval[,2:3]), col = "blue", lty=25)

cor(data$Y, exp(pred), method ="pearson")


# Zad 12
data = read.table("/home/s/2018/s309881/Dokumenty/Modele-Liniowe/Lista3/CH03PR15.txt", col.n
data$X = (data$X)**(-0.5)
data$Y = data$Y
reg1 = lm(Y~X, data)
summary(reg1)

x = data$X
y = data$Y
newx = seq(min(x),max(x)+0.05,by = 0.05)
conf_interval = predict(reg1, newdata=data.frame(X=newx), interval="prediction",level = 0.95
plot(x, y, ylim=c(-0.5, 3.5), xlab="time^(-0.5)", ylab="log solution concentration")
matlines(newx, conf_interval[,1], col = "lightblue", lty=25)
matlines(newx, conf_interval[,2:3], col = "blue", lty=25)

real = data$Y
pred = predict.lm(reg1, data.frame(X=data$X), interval = "prediction", level=0.95)[, 1]

cor(real, pred, method ="pearson")


# ------------------------------------------------------------------------------------------
data = read.table("/home/s/2018/s309881/Dokumenty/Modele-Liniowe/Lista3/CH03PR15.txt", col.n
data$X = (data$X)^(-0.5)
data$Y = data$Y
reg1 = lm(Y~X, data)
x = data$X
y = data$Y
newx = seq(min(x)-0.01,max(x)+0.01,by = 0.01)
conf_interval = predict(reg1, newdata=data.frame(X=newx), interval="prediction",level = 0.95
pred = predict.lm(reg1, data.frame(X=data$X), interval = "prediction", level=0.95)[, 1]

data = read.table("/home/s/2018/s309881/Dokumenty/Modele-Liniowe/Lista3/CH03PR15.txt", col.n
reg1 = lm(Y~X, data)
x = data$X
```

```
y = data$Y
newx = newx^(-2)
plot(x, y, ylim=c(-0.5, 3.5), xlab="time", ylab="solution concentration")
matlines(newx, conf_interval[,1], col = "lightblue", lty=25)
matlines(newx, conf_interval[,2:3], col = "blue", lty=25)

cor(data$Y, pred, method ="pearson")
```