Modele Liniowe - Lista 4

Jakub Kuciński 309881

Styczeń 2022

Spis treści

1 Zadanie 1

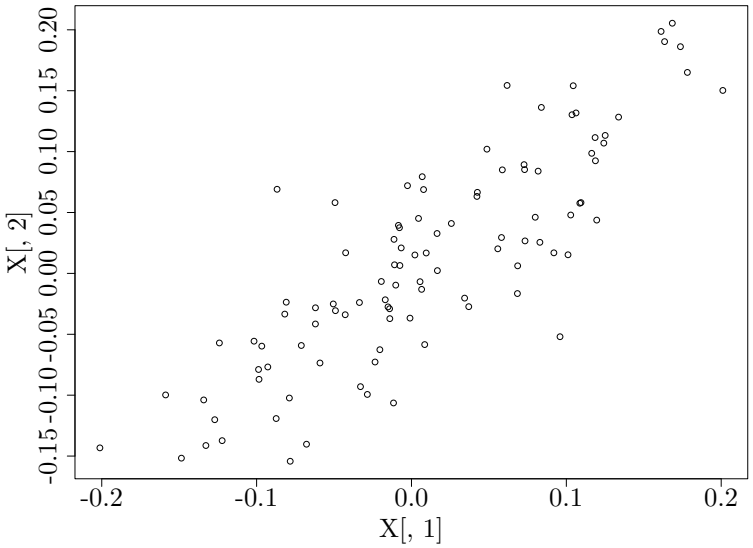
Zadanie dodatkowe - oddane oddzielnie.

2 Zadanie 2

Zadanie dodatkowe - oddane oddzielnie.

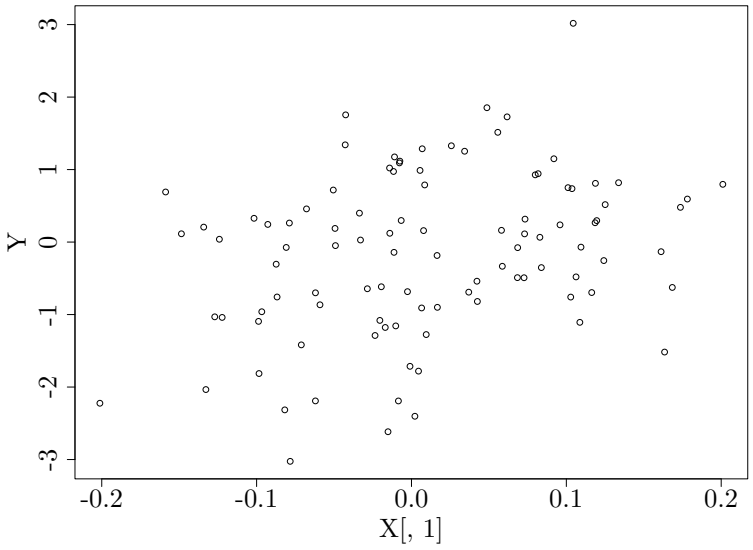
3 Zadanie 3

3.1 a)



Rysunek 1: Dane z macierzy X.

Wykres 1 przedstawia 100 wygenerowanych punktów z podanego w zadaniu rozkładu (macierz X). Rysunek 2 przedstawia Y względem pierwszej kolumny X.



Rysunek 2: Wykres Y względem pierwszej kolumny X.

3.2 b)

95% przedział ufności dla modelu regresji $Y = \beta_0 + \beta_1 X_1 + \epsilon$ wyniósł $[1.4439, 6.3509]$, a dla modelu $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ - $[-2.4770, 6.6331]$. Widzimy, że pierwszy model posiada węższy przedział ufności od drugiego modelu. Można więc powiedzieć, że mamy większą pewność co do wartości parametru β_1 dla pierwszego modelu. Widzimy też, że 0 nie należy do 95% przedziału ufności pierwszego modelu, odrzucamy zatem hipotezę zerową $\beta_1 = 0$ i przyjmujemy hipotezę alternatywną $\beta_1 \neq 0$. Przeciwnie jest w przypadku drugiego modelu, gdyż 0 należy do 95% przedziału ufności, czyli nie mamy podstaw do odrzucenia hipotezy zerowej.

3.3 c)

Wiemy że $\hat{\beta}$ pochodzi z rozkładu $N(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$. Wystarczy zatem wyliczyć macierz $\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$ i wyjąć z niej wartość odpowiadającą wariancji $\hat{\beta}_1$ (w R będzie to pole $[2, 2]$). Naszą odpowiedzią będzie oczywiście pierwiastek z otrzymanej wariancji. Dla pierwszego modelu wyniosła 1.18905, a dla drugiego 2.21734.

Moc testu binarnego to prawdopodobieństwo odrzucenia H_0, gdy prawdziwa jest hipoteza H_1. Moc testu dla pierwszego modelu oraz $\alpha = 0.05$, $\beta_1 = 3$ wyniosła 0.70480. Oznacza to, że z prawdopodobieństwem 0.70480 odrzucona zostanie hipoteza $H_0 : \beta_1 = 0$. Dla drugiego modelu moc testu wyniosła 0.26798.

3.4 d)

Testujemy, czy β_1 jest różna od 0: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. Odrzucamy hipotezę zerową, gdy $0 \notin confint(\beta_1)$. Parametr α został ustalony na poziomie 0.05

W pierwszym przypadku hipoteza zerowa została odrzucona 715 razy, a w drugim 277 razy z 1000 testów. Obserwowane częstotliwości są zgodne z obliczonymi teoretycznymi mocami testów (0.70480 oraz 0.26798).

Wyestymowane odchylenie standardowe wyniosło dla pierwszego modelu 1.17538, a dla drugiego 2.20229, co również zgadza się z otrzymanymi teoretycznymi wartościami (1.18905 oraz 2.21734).

4 Zadanie 4

4.1 a)

nvars	sse	mse	aic	pval1	pval2	false_discoveries
1	1375.3462	0.002413066	3162.583	6.697481e-12	NA	0
2	1279.7538	0.002275317	3092.545	9.059257e-13	2.398713e-17	0
5	962.9687	0.004818304	2814.143	2.268619e-17	6.592735e-20	0
10	949.3208	0.018465650	2809.869	7.565170e-17	2.495270e-20	1
50	923.4464	0.044361838	2862.235	1.669502e-16	1.084515e-19	1
100	876.2572	0.091616915	2909.781	1.869848e-15	1.079141e-18	3
500	432.0719	0.535734688	3002.714	1.848352e-13	1.663913e-15	37
950	32.7059	0.935126596	1321.677	2.201033e-02	5.520591e-03	85

Rysunek 3: Wyznaczone eksperymentalnie wartości dla modeli o podanej liczbie pierwszych zmiennych.

Zgodnie z oczekiwaniami wraz z dodawaniem kolejnych zmiennych SSE maleje (rysunek 3), czyli coraz lepiej dopasowujemy się do danych. Widzimy również, że p-wartości początkowo maleją, a następnie (od $nvars = 10$) zaczynają rosnąć. W szczególności dla $nvars = 950$ p wartości są bardzo wysokie w porównaniu do mniejszych modeli. Oznacza to, że stajemy się coraz mniej pewni czy prawdziwie niezerowe współczynniki przy pierwszych dwóch zmiennych są niezerowe. Liczba fałszywych odkryć również rośnie wraz ze wzrostem liczby zmiennych, co oznacza, że coraz częściej odrzucamy hipotezę, że dany współczynnik jest zerowy, pomimo tego, że w rzeczywistości był faktycznie zerowy. Podobnie rośnie wartość MSE, czyli wraz ze wzrostem zmiennych coraz bardziej oddalamy się od predykcji rzeczywistego modelu. Na podstawie AIC powinniśmy wybrać model

z największą liczbą zmiennych pomimo faktu, że zawiera on olbrzymią liczbę nadmiarowych zmiennych.

4.2 b)

nvars	sse	mse	aic	pval1	pval2	false_discoveries
1	1375.3462	0.002413066	3162.583	6.697481e-12	NA	0
2	1279.7538	0.002275317	3092.545	9.059257e-13	2.398713e-17	1
5	962.9687	0.004818304	2814.143	2.268619e-17	6.592735e-20	3
10	949.3208	0.018465650	2809.869	7.565170e-17	2.495270e-20	4
50	923.4464	0.044361838	2862.235	1.669502e-16	1.084515e-19	4
100	876.2572	0.091616915	2909.781	1.869848e-15	1.079141e-18	6
500	432.0719	0.535734688	3002.714	1.848352e-13	1.663913e-15	39
950	32.7059	0.935126596	1321.677	2.201033e-02	5.520591e-03	87

Rysunek 4: Wyznaczone eksperymentalnie wartości dla modeli o podanej liczbie "największych" zmiennych.

Dla modelów powstałych ze zmiennych o największych współczynnikach wnioski odnośnie SSE, MSE, AIC oraz p-wartości są analogiczne jak w przypadku modelów powstałych z pierwszych $nvars$ zmiennych. Widzimy za to, że zmieniła się liczba fałszywych odkryć, które zaczęły pojawiać się już w najmniejszych modelach i ich stosunek względem ogólnej liczby zmiennych (dla mniejszych modeli) jest znaczący. Widzimy więc, że wybór zmiennych na podstawie wielkości współczynnika niekoniecznie zwróci nam istotne w rzeczywistości zmienne.

5 Zadanie 5

Szukane wartości wyznaczyłem przy pomocy funkcji *summary*. Wyestymowane równanie regresji:

$$satisfaction = 1.053245 - 0.005861 \cdot age + 0.001928 \cdot severity + 0.030148 \cdot anxiety$$

Wyznaczona wartość R^2 to 0.5415, czyli około 54 % zmienności zmiennej satisfaction stanowi zmienność wyjaśniona przez model. Testowana hipoteza zerowa

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0.$$

Hipoteza alternatywna

$$H_1 : \beta_i \neq 0 \ \textit{dla przynajmniej jednego} \ i \in \{1, 2, 3\}.$$

Statystyka testowa ma postać:

$$F = \frac{MSM}{MSE} = 16.54$$

przy H_0 ma rozkład Fishera–Snedecora z $dfM = p - 1 = 3$ i $dfE = n - p = 42$ stopniami swobody. Odpowiadająca p-wartość wyniosła $3.043e - 07$. Widzimy więc, że przy założeniu prawdziwości hipotezy zerowej, prawdopodobieństwo pojawiania się zdarzenia co najmniej tak rzadkiego jak nasze wynosi mniej niż $3.043e - 07$. Prawdopodobieństwo to jest bardzo bliskie zeru, więc można odrzucić hipotezę zerową i przyjąć hipotezę alternatywną H_1, czyli przynajmniej jedna zmienna objaśniająca ma istotny wpływ na zmienną wynikową.

6 Zadanie 6

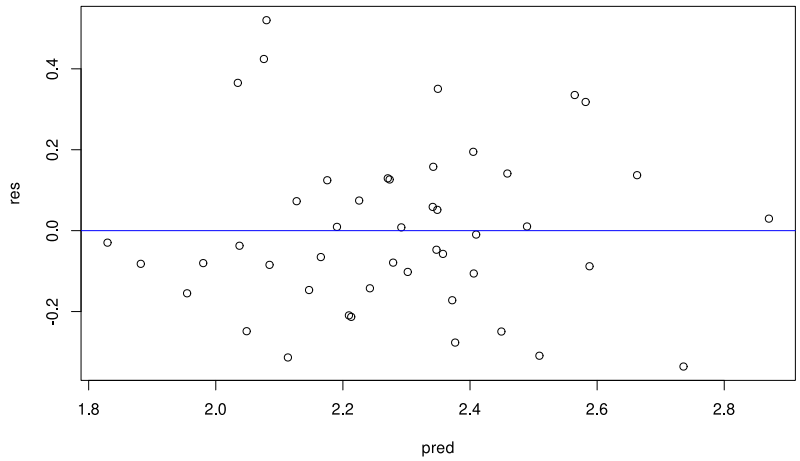
95% przedziały ufności dla wyestymowanych współczynników poszczególnych zmiennych niezależnych:

	2.5%	97.5%
age	-0.01209411	0.0003730895
severity	-0.00974994	0.0136060385
anxiety	0.01146717	0.0488283055

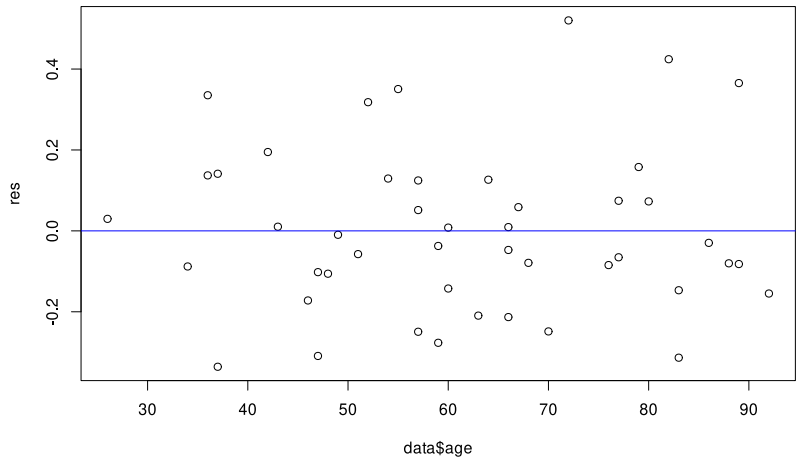
We wszystkich przypadkach testowana statystyka zerowa pochodzi z rozkładu t-studenta z $n - p = 46 - 4 = 42$ stopniami swobody. Przyjmiemy poziom istotności $\alpha = 0.05$. Przyjmujemy hipotezę zerową (nie mamy podstaw do jej odrzucenia), gdy *p-wartość* $> \alpha$. Jeśli *p-wartość* $< \alpha$ odrzucamy hipotezę H_0 i przyjmujemy hipotezę alternatywną H_1.

zm. niezależna	H_0	H_1	statystyka	p-wartość	odrzucamy H_0
age	$\beta_1 = 0$	$\beta_1 \neq 0$	$T = \frac{\beta_1}{s(\beta_1)}$	0.06468	Nie
severity	$\beta_2 = 0$	$\beta_2 \neq 0$	$T = \frac{\beta_2}{s(\beta_2)}$	0.74065	Nie
anxiety	$\beta_3 = 0$	$\beta_3 \neq 0$	$T = \frac{\beta_3}{s(\beta_3)}$	0.00223	Tak

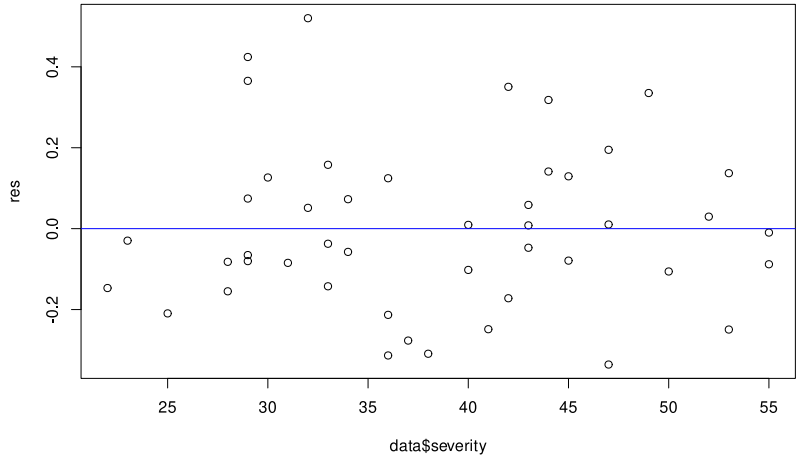
Nie mamy podstaw do odrzucenia hipotez zerowych w przypadku zmiennych *age* oraz *severity*. Odrzucamy natomiast hipotezę zerową dla *anxiety*. Oznacza to, że zmienna *anxiety* w istotny sposób wpływa na zmienną zależną. Widzimy, że otrzymane wyniki są zgodne z 95% przedziałami ufności. W przypadku, gdy wartość 0 znajdowała się w środku przedziału ufności nie mieliśmy podstaw do odrzucenia hipotezy zerowej, a gdy leżała poza, odrzuciliśmy hipotezę zerową.



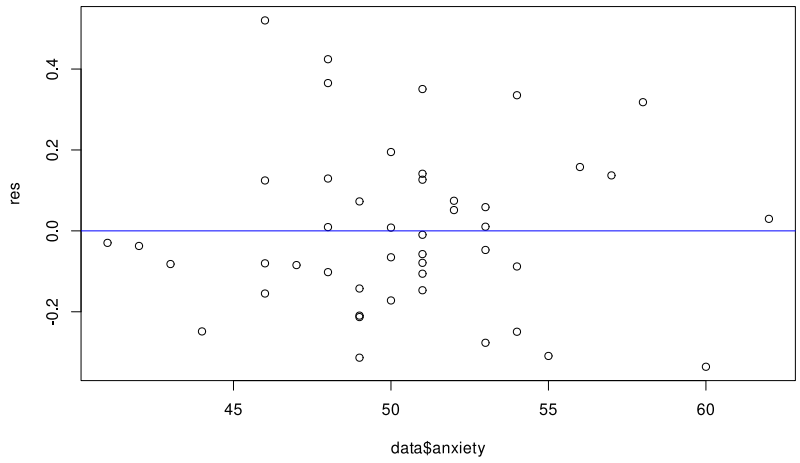
Rysunek 5: Residua vs przewidywane satisfaction



Rysunek 6: Residua vs age



Rysunek 7: Residua vs severity

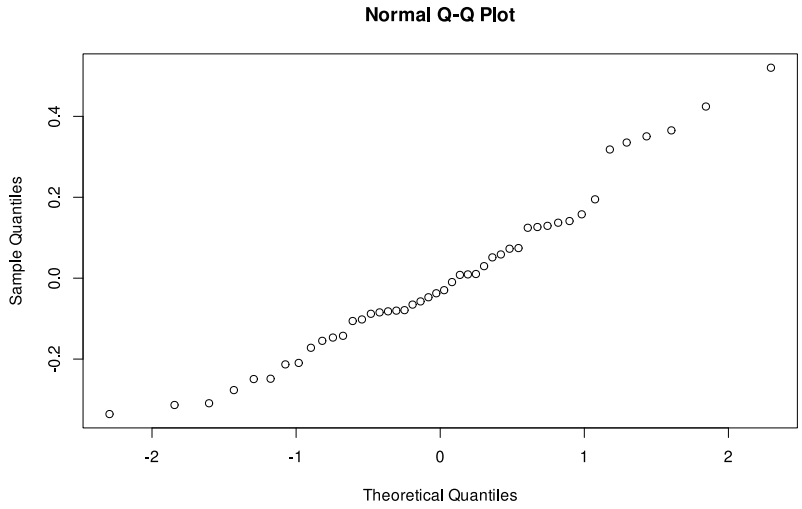


Rysunek 8: Residua vs anxiety

7 Zadanie 7

Na rysunkach 5, 6, 7, 8 nie widzimy żadnych nietypowych wzorów. Wariancja residuów wydaje się niezależna od wartości zmiennych objaśniających i przewidywanego satisfaction. Widzimy jednak, że istnieje kilka punktów nieco odstających od pozostałych (residua o wartościach około 0.4).

8 Zadanie 8



Rysunek 9: Q-Q plot

Na rysunku 9 widzimy, że kwantyle układają się w przybliżeniu liniowo, przy czym linia przechodzi przez punkt $(0, 0)$. Możemy więc podejrzewać, że residua pochodzą z rozkładu normalnego. Widzimy też, że istnieje kilka obserwacji, które odstają nieco od pozostałych i leżą nieco powyżej oczekiwanej prostej (jest to zgodne z naszymi obserwacjami z poprzedniego zadania). Shapiro-Wilk test zwrócił wynik $W = 0.96286$ oraz odpowiadającą *p-wartość* $= 0.1481$. Nie mamy zatem podstaw do odrzucenia hipotezy zerowej, że residua pochodzą z rozkładu normalnego.

9 Zadanie 9

Modelem pełnym jest w naszym przypadku model ze zmiennymi zależnymi SATM, SATV, HSM, HSS and HSE, a modelem zredukowanym - model ze zmiennymi zależnymi HSM, HSS and HSE. W obu przypadkach zmienną objaśnianą jest GPA. Testujemy hipotezę $H_0 : \beta_{SATM} = \beta_{SATV} = 0$. Hipoteza

alternatywna $H_1 : \beta_{SATM} \neq 0 \vee \beta_{SATV} \neq 0$. Statystyka testowa ma postać

$$F = \frac{(SSE(R) - SSE(F))/(dfE(R) - dfE(F))}{MSE(F)},$$

gdzie R oznacza model zredukowany, a F model pełny. Wartości SSE można obliczyć sumując kwadraty residuów dopasowanych modeli. Stopnie swobody $dfE(R)$ oraz $dfE(F)$ wynoszą odpowiednio $224 - 4 = 220$ i $224 - 6 = 218$, czyli liczba stopni swobody licznika wynosi $220 - 218 = 2$, a mianownika 218. Stąd też statystyka F pochodzi z rozkładu Fishera–Snedecora z 2 i 218 stopniami swobody. Ostatecznie $F = 0.95032$, a odpowiadająca p-wartość to 0.38821. Nie mamy zatem podstaw do odrzucenia hipotezy zerowej. Widzimy zatem, że zmienne SATM i SATV nie mają znaczącego wpływu na zmienną objaśnianą.

Wyniki otrzymane za pomocą wywołania $anova(reg1, reg2)$ (rysunek 10) są takie same jak w przypadku wartości wyliczonych ręcznie.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	220	107.7505	NA	NA	NA	NA
2	218	106.8191	2	0.9313136	0.9503276	0.38821

Rysunek 10: Funkcja *anova*.

10 Zadanie 10

Rysunek 11 przedstawia otrzymane sumy pierwszego i drugiego typu.

	Df	Sum Sq
	<int>	<dbl>
SATM	1	8.582934e+00
SATV	1	9.054942e-04
HSM	1	1.772647e+01
HSE	1	1.891193e+00
HSS	1	4.421433e-01

	Sum Sq
	<dbl>
SATM	0.9279988
SATV	0.2326519
HSM	6.7724312
HSE	0.9568040
HSS	0.4421433

Rysunek 11: Sumy pierwszego i drugiego typu.

10.1 a)

Suma kwadratów dla modelu ze zmienną HSM wyniosła 109.152, a dla modelu bez tej zmiennej 126.878. Ich różnica wynosi 17.726, czyli dokładnie tyle, ile suma typu pierwszego dla zmiennej HSM, co widzimy na rysunku 11.

10.2 b)

Ostatnie sumy typów I i II są zawsze równe, opisują je dokładnie te same wzory (w obu przypadkach porównujemy model bez ostatniej zmiennej z modelem ze wszystkimi zmiennymi).

11 Zadanie 11

Po wywołaniu summary na naszym modelu (rysunek 12) widzimy, że nie ma podstaw do odrzucenia H_0 dla SATV, ale możemy odrzucić H_0 dla SATM. Zmienna SAT nie została w ogóle użyta przez model, bo jest kombinacją liniową SATM i SATV i nie wnosi żadnej nowej informacji. Widzimy, że ogólnie model jest bardzo słaby, bo wyjaśnia jedynie 6.337% zmienności zmiennej objaśnianej.

```
Call:
lm(formula = GPA ~ SATM + SATV + SAT, data = dt)

Residuals:
     Min       1Q   Median       3Q      Max
-2.59483 -0.37920  0.08263  0.55730  1.39931

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.289e+00  3.760e-01   3.427 0.000728 ***
SATM         2.283e-03  6.629e-04   3.444 0.000687 ***
SATV        -2.456e-05  6.185e-04  -0.040 0.968357
SAT                 NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7577 on 221 degrees of freedom
Multiple R-squared:  0.06337,   Adjusted R-squared:  0.05489
F-statistic: 7.476 on 2 and 221 DF,  p-value: 0.0007218
```

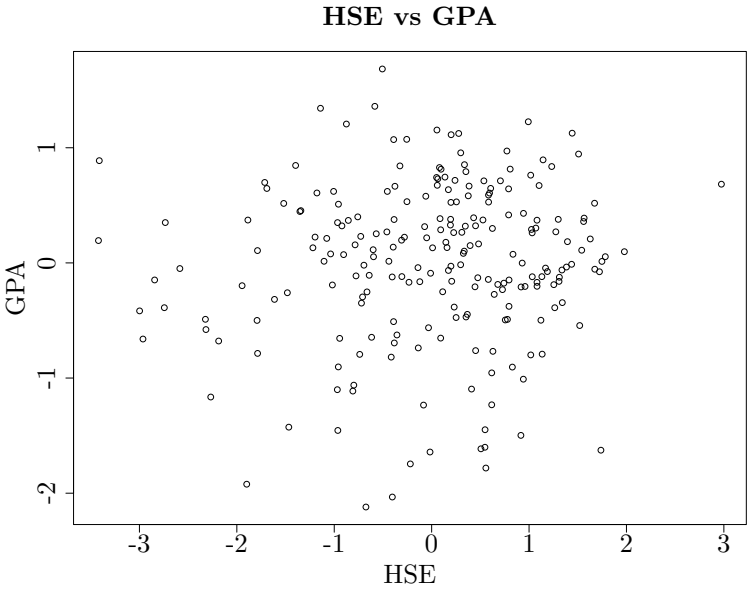
Rysunek 12: $summary(reg1)$

12 Zadanie 12

Partial regression plot ukazuje wpływ, jaki wywiera dodanie nowej zmiennej objaśniającej do modelu, który już zawiera kilka zmiennych niezależnych. Powstaje poprzez zestawienie residuów dwóch modeli, których zmiennymi objaśniającymi są wszystkie zmienne X_i oprócz badanej zmiennej, a zmienną objaśnianą jest w jednym przypadku Y a w drugim badany X_i. Wykres pokazuje relację między badanym X_i a Y po uwzględnieniu pozostałych X_i.

Na wykresie HSS vs GPA widzimy, że istnieje kilka obserwacji odbiegających od pozostałych na osi HSS (wartości poniżej -2.5). Na wykresie SEX vs

GPA widzimy, że istnieją dwa klastry punktów, jeden ze średnią około -0.3 dla SEX a drugi z około 0.5. Oba klastry oscylują wokół wartości 0 dla GPA. Pozostałe wykresy nie zawierają żadnych widocznych nieregularności. Widzimy więc, że żadna ze zmiennych nie wnosi do modelu istotnej informacji ponad to co objaśniły pozostałe zmienne.



13 Zadanie 13

Residua zachowują się normalnie i oscylują wokół 0 (rysunki 13 i 14). Wariancja nie wydaje się być zależna od numeru obserwacji, co zgadza się z modelem teoretycznym. Nie widzimy również żadnych obserwacji odstających.

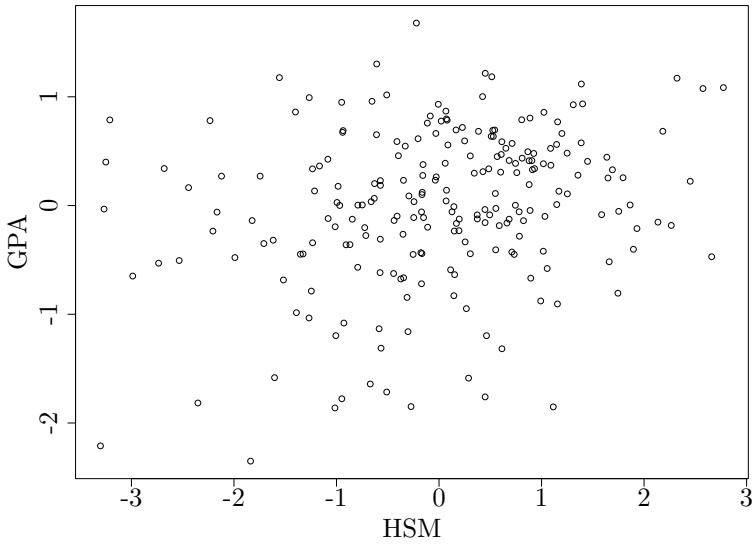
14 Zadanie 14

DFFITS pozwala zbadać wpływ obserwacji Y_i na predykcję $\hat{Y}_i$. Ma ona postać

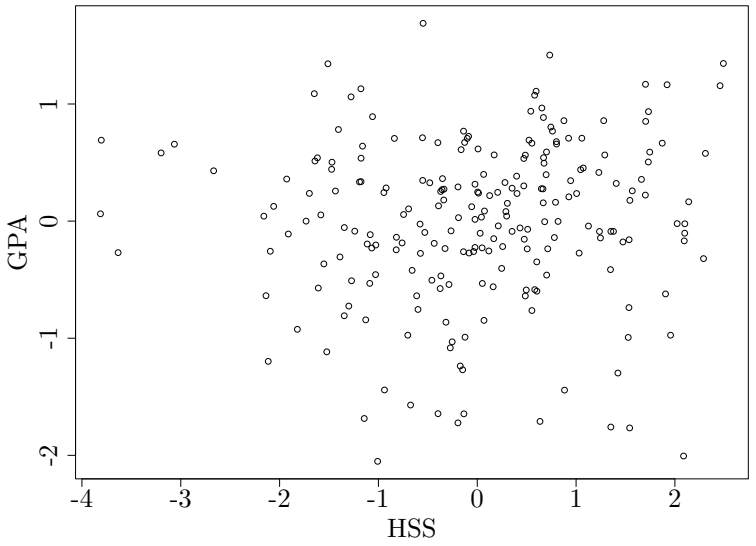
$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)i}}{\sqrt{s^2_{(i)} H_{ii}}}$$

Wykres 15 przedstawia wartości DFFITS dla kolejnych obserwacji z naszego zbioru danych dla modelu pełnego. Obserwacje leżące poza przedziałem $2\sqrt{p/n}$ mają znaczący wpływ na predykcję. Widzimy, że zdecydowana większość obserwacji leży wewnątrz przedziału, ale jest kilkanaście obserwacji wykraczających poza ten zakres i mogą być one obserwacjami odstającymi lub wpływowymi.

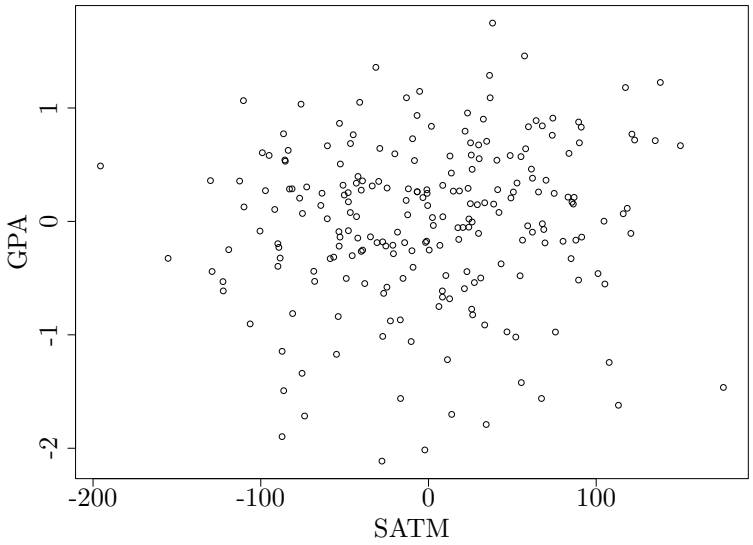
HSM vs GPA



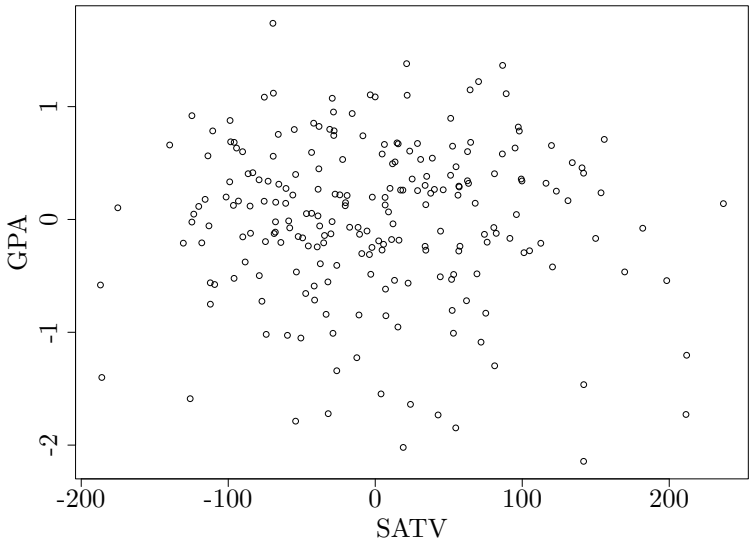
HSS vs GPA

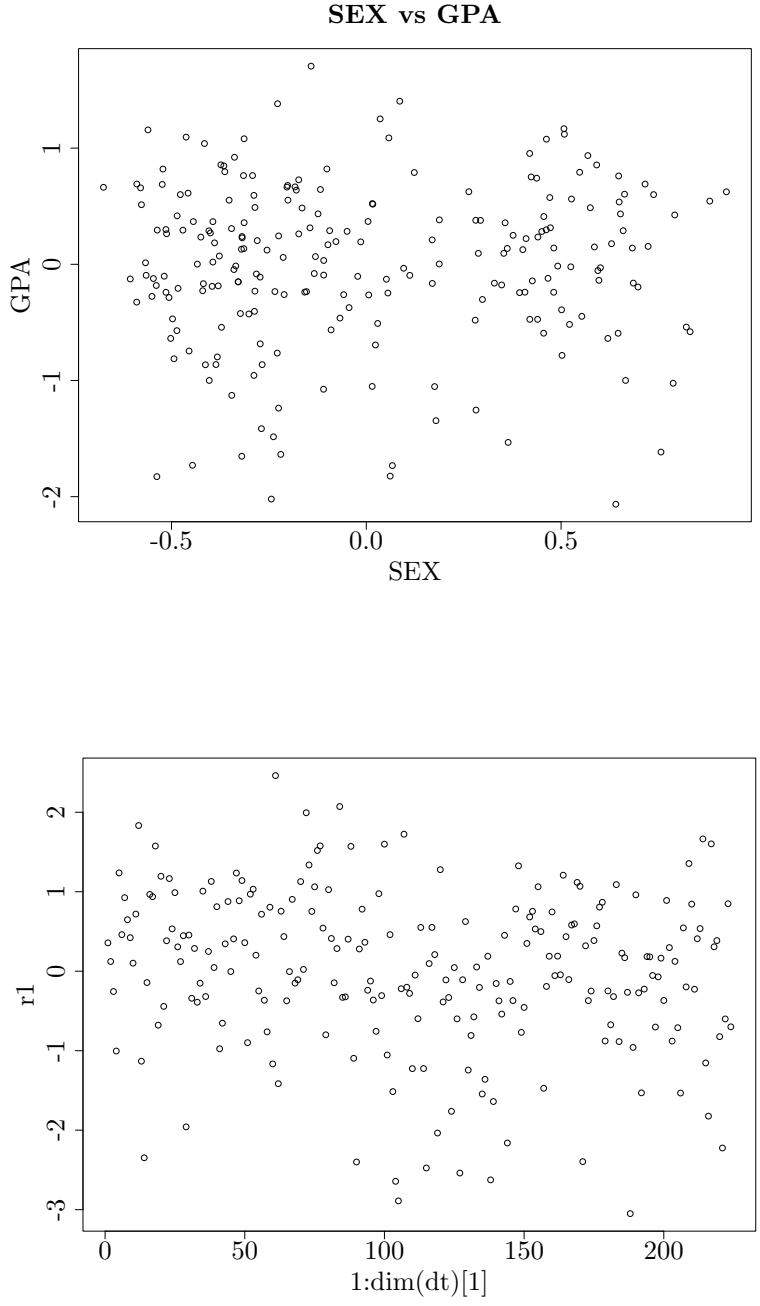


SATM vs GPA

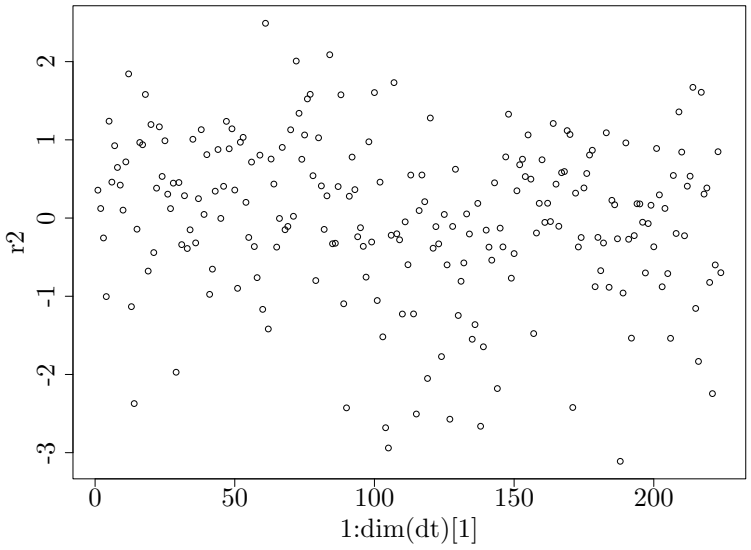


SATV vs GPA

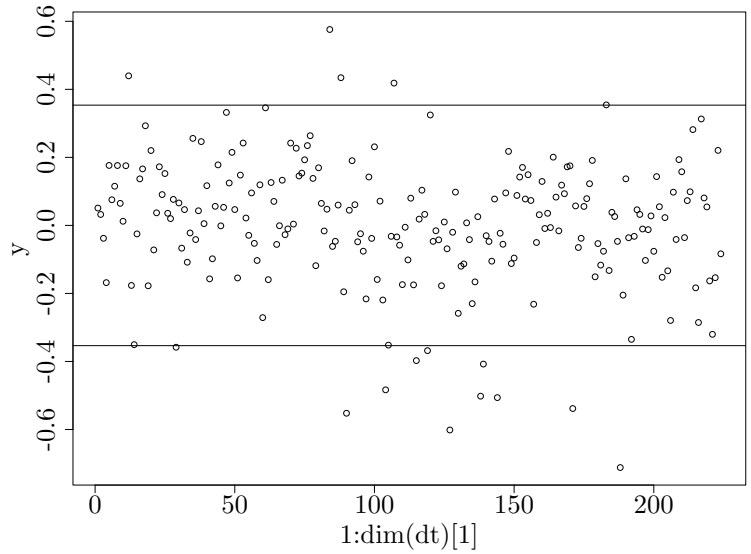




Rysunek 13: Studentyzacja wewnętrzna



Rysunek 14: Studentyzacja zewnętrzna



Rysunek 15: DFFITS

15 Zadanie 15

Tolerancja to odwrotność VIF i pomaga nam zidentyfikować zjawisko multikolinearności. Wartości poniżej 0.1 wskazują na problem z multikolinearnością.

Zmienna niezależna	HSM	HSS	HSE	SATM	SATV	SEX
Tolerancja	0.51886	0.50882	0.54295	0.57454	0.73105	0.77425

Widzimy, że wszystkie wartości wynoszą ponad 0.1, więc w naszych danych nie występuje problem multikolinearności.

16 Zadanie 16

	bic_vals	aic_vals	(Intercept)	HSM	HSS	HSE	SATM	SATV	SEX
1	-36.52518	481.6754	1	1	0	0	0	0	0
2	-34.18564	480.6033	1	1	0	1	0	0	0
3	-30.28481	481.0925	1	1	0	1	1	0	0
4	-25.66783	482.2978	1	1	1	1	1	0	0
5	-20.74352	483.8105	1	1	1	1	1	1	0
6	-15.41891	485.7234	1	1	1	1	1	1	1

Rysunek 16: Kryteria BIC i AIC dla najlepszych modeli o danej liczbie zmiennych objaśniających.

Kryteria BIC i AIC dla modelów o takiej samej liczbie zmiennych wybierają ten sam model jako najlepszy, bo w obu kryteriach likelihood danego modelu jest taki sam, a kary za złożoności modelów są stałe (ze względu na stałe n i p). Różnica pojawia się dopiero podczas porównywania modeli o różnej liczbie zmiennych. Zarówno dla BIC jak i dla AIC wybieramy model o najmniejszej wartości kryterium. W naszym przypadku dla BIC jest to model z samą zmienną HSM, a dla AIC model ze zmiennymi HSM i HS.

17 Kod w R

```
# # SVG graphics device
# svg("my_plot.svg")

# # Code of the plot
# plot(rnorm(20))

# # Close the graphics device
# dev.off()
```

```
require("car")
require("ggplot2")
library(svglite)
library(broom)
library(MASS)
require("leaps")

# Zad 3
#### a)
n = 100
X = matrix(rnorm(200), 2, 100)
Sigma = matrix(c(1, 0.9, 0.9, 1), 2, 2)/100
A = t(chol(Sigma))
X = A %*% X
X = t(X)
Y = 3*X[,1]+rnorm(n,0,1)
plot(X[,1], X[,2])
plot(X[,1], Y)

# svglite("myplot.svg", width = 4, height = 4)
svglite("normal1.svg")
plot(X[,1], X[,2])
dev.off()
svglite("normal2.svg")
plot(X[,1], Y)
dev.off()
# With ggsave()
# ggsave("myplot.svg", width = 8, height = 8, units = "cm")

#### b)
# reduced
reg1 = lm(Y~X[, 1])
# full
reg2 = lm(Y~X)

confint(reg1)
# odrzucamy H0

confint(reg2)
# nie odrzucamy H0

#### c)
reg1 = lm(Y~X[, 1], x = TRUE)
reg2 = lm(Y~X[, 1]+X[, 2], x = TRUE)
```

```
s2 = 1/(n-2) * sum(reg1$residuals ** 2)
s_red_beta1 = sqrt(s2*(solve(t(reg1$x) %*% reg1$x))[2, 2])
s_red_beta1
s2 = 1/(n-3) * sum(reg2$residuals ** 2)
s_full_beta1 = sqrt(s2*(solve(t(reg2$x) %*% reg2$x))[2, 2])
s_full_beta1

power = function(n, p, alpha, s_beta, beta1){
    df = n - p
    tc = qt(1-alpha/2, df)
    delta = beta1 / s_beta
    prob1 = function(delta){pt(tc, df, delta)}
    prob2 = function(delta){pt(-tc, df, delta)}
    powerOfBeta = 1 - prob1(delta) + prob2(delta)
    powerOfBeta
}
power_red = power(n, 2, 0.05, s_red_beta1, 3.0)
power_red
power_full = power(n, 3, 0.05, s_full_beta1, 3.0)
power_full

#### d)
k = 1000
x1 = 0
x2 = 0
red_betas = rep(0, k)
full_betas = rep(0, k)
for(i in 1:k){
    e = rnorm(n, 0, 1)
    Y = 3*X[, 1] + e
    reg1 = lm(Y~X[, 1])
    reg2 = lm(Y~X[, 1]+X[, 2])
    interval1 = confint(reg1)[2,]
    interval2 = confint(reg2)[2,]
    if(0>=interval1[1] && 0<=interval1[2])
        x1 = x1+1
    if(0>=interval2[1] && 0<=interval2[2])
        x2 = x2+1
    red_betas[i] = reg1$coefficients[2]
    full_betas[i] = reg2$coefficients[2]
}
sd(red_betas)
sd(full_betas)
power_red_exp = 1-x1/k
power_red_exp
power_full_exp = 1-x2/k
```

```
power_full_exp

# Zad 4
#### a)
n = 1000
X = matrix(rnorm(950000, 0, 0.1), n, 950)
e = rnorm(n)
beta = c(3, 3, 3, 3, 3, rep(0, 945))
Y = X %*% beta + e

nvars = c(1, 2, 5, 10, 50, 100, 500, 950)
sse = rep(0, 8)
mse = rep(0, 8)
mse2 = rep(0, 8)
aic = rep(0, 8)
pval1 = rep(NULL, 8)
pval2 = rep(NULL, 8)
false_discoveries = rep(0, 8)

for (i in 1:8){
    k = nvars[i]
    Xk = X[, 1:k]
    reg = lm(Y~Xk)
    sse[i] = sum(reg$residuals^2)
    aic[i] = AIC(reg)
    beta_diff = matrix(reg$coefficients[-1]-beta[1:k], k, 1)
#     mse[i] = sum((Xk%*%beta_diff)^2)/(n-k+1)
    mse[i] = mean((Xk%*%beta_diff)^2)
    pval1[i] = summary(reg)$coefficients[2, 4]
    if(i != 1)
        pval2[i] = summary(reg)$coefficients[3, 4]
#     if(i != 1)
#         false_discoveries[i] = sum(summary(reg)$coefficients[2:min(k, 6), 4]>0.05)
    if(i >= 4)
        false_discoveries[i] = sum(summary(reg)$coefficients[7:(k+1), 4]<0.05) + false_disco
}
cbind(nvars, sse, mse, aic, pval1, pval2, false_discoveries)
# cbind(nvars, sse, mse, mse2, aic, pval1, pval2, false_discoveries)

### b)

full_model = lm(Y~X)
decreasing_order = order(abs(full_model$coefficients[-1]), decreasing = TRUE)
is_true_nonzero = c(rep(TRUE, 5), rep(FALSE, 945))
is_true_nonzero = is_true_nonzero[decreasing_order]
X2 = X2[, decreasing_order]
```

```
nvars = c(1, 2, 5, 10, 50, 100, 500, 950)
sse = rep(0, 8)
mse = rep(0, 8)
mse2 = rep(0, 8)
aic = rep(0, 8)
pval1 = rep(NULL, 8)
pval2 = rep(NULL, 8)
false_discoveries = rep(0, 8)

for(i in 1:8){
    k = nvars[i]
    Xk = X[, 1:k]
    is_true_nonzero_k = is_true_nonzero[1:k]
    reg = lm(Y~Xk)
    sse[i] = sum(reg$residuals^2)
    aic[i] = AIC(reg)
    beta_diff = matrix(reg$coefficients[-1]-beta[1:k], k, 1)
#     mse[i] = sum((Xk%*%beta_diff)^2)/(n-k+1)
    mse[i] = mean((Xk%*%beta_diff)^2)
    pval1[i] = summary(reg)$coefficients[2, 4]
    if(i != 1)
        pval2[i] = summary(reg)$coefficients[3, 4]
#     true_nonzero_indices = which(is_true_nonzero_k == 1)
    true_zero_indices = which(is_true_nonzero_k == 0)
#     false_discoveries[i] = sum(summary(reg)$coefficients[2:(k+1), 4][true_nonzero_indices]
    false_discoveries[i] = sum(summary(reg)$coefficients[2:(k+1), 4][true_zero_indices]<0.05
}
cbind(nvars, sse, mse, aic, pval1, pval2, false_discoveries)
# cbind(nvars, sse, mse, mse2, aic, pval1, pval2, false_discoveries)

# Zad 5
data = read.table("./CH06PR15.txt", col.names=c("age", "severity", "anxiety", "satisfaction"
reg1 = lm(satisfaction~age+severity+anxiety, data)
summary(reg1)

# Zad 6
confint(reg1, level = 0.95)
summary(reg1)

# Zad 7
res = reg1$residuals
pred = predict(reg1)

plot(pred, res)
abline(h=0, col="blue")
```

```
plot(data$age, res)
abline(h=0, col="blue")
plot(data$severity, res)
abline(h=0, col="blue")
plot(data$anxiety, res)
abline(h=0, col="blue")

# Zad 8
shapiro.test(res)
qqnorm(res)
qqline(res)

# Zad 9
dt = read.table('./csdata.txt', col.names=c("id", "GPA", "HSM", "HSS", "HSE", "SATM", "SATV'
reg1 = lm(GPA~HSM+HSS+HSE, dt)
reg2 = lm(GPA~SATM+SATV+HSM+HSS+HSE, dt)

# H_0: b5=b6=0 (SATM and SATV)
#### a)

sse1 = sum(reg1$residuals^2)
sse2 = sum(reg2$residuals^2)
difference = sse1 - sse2
dfEdiff = (224 - 4) - (224 - 6)
dfE2 = 224 - 6
# dfEdiff = an1$`Df`[4] - an2$`Df`[6]
# dfE2 = an2$`Df`[6]
F_test = (difference / dfEdiff) / (sse2 / dfE2)
F_test
ft = qf(0.95, dfEdiff, dfE2)
ft
F_test > ft
pval = 1 - pf(F_test, dfEdiff, dfE2)
pval
pval < 0.05
# Nie odrzucamy hipotezy zerowej - uznajemy, że model zredukowany jest poprawny

#### b)
anova(reg1, reg2)
# p-wartość wynosi 0.3882 czyli tyle samo co w poprzednim podpunkcie

# Zad 10
dt = read.table('./csdata.txt', col.names=c("id", "GPA", "HSM", "HSS", "HSE", "SATM", "SATV'
reg1 = lm(GPA~SATM+SATV+HSM+HSE+HSS, dt)
# sumy typu I
anova(reg1)
```

```
# sumy typu II
Anova(reg1)

#### a)
reg1 = lm(GPA~SATM+SATV+HSM, dt)
reg2 = lm(GPA~SATM+SATV, dt)
sse1 = sum(reg1$residuals^2)
sse2 = sum(reg2$residuals^2)
sse1; sse2
type_1_SSE_HSM = sse2 - sse1
type_1_SSE_HSM
anova(reg1)$`Sum Sq`[3]

#### b)
# Ostatnie sumy typów I i II są zawsze równe, opisują je dokładnie te same wzory
# (w obu przypadkach porównujemy model bez ostatniej zmiennej z modelem ze wszystkimi zmienn

# Zad 11
dt = read.table('./csdata.txt', col.names=c("id", "GPA", "HSM", "HSS", "HSE", "SATM", "SATV"
dt$SAT = dt$SATM + dt$SATV
reg1 = lm(GPA~SATM+SATV+SAT, dt)
summary(reg1)

# Już przed wykonaniem eksperymentu możemy podejrzewać, że nic istotnego się nie wydarzy,
# bo w zadaniu 9 pokazaliśmy, że zmienne SATM i SATV nie mają istotnego wpływu na GPA.
# Po wywołaniu summary na naszym modelu widzimy, że nie ma podstaw do odrzucenia H_0 dla SAT
# ale możemy odrzucić H_0 dla SATM. Zmienna SAT nie została w ogóle użyta przez model,
# bo jest kombinacją liniową SATM i SATV i nie wnosi żandej nowej informacji.
# Widzimy, że ogólnie model jest bardzo słaby, bo wyjaśnia jedynie 0.06337 zmienności
# zmiennej objaśnianej.

# Zad 12
dt = read.table('./csdata.txt', col.names=c("id", "GPA", "HSM", "HSS", "HSE", "SATM", "SATV"

x = lm(HSM~HSS+HSE+SATM+SATV+SEX, dt)
y = lm(GPA~HSS+HSE+SATM+SATV+SEX, dt)
# svglite("HSMvsGPA.svg")
plot(x$residuals, y$residuals, xlab="HSM", ylab="GPA", main = "HSM vs GPA")
# dev.off()

x = lm(HSS~HSM+HSE+SATM+SATV+SEX, dt)
y = lm(GPA~HSM+HSE+SATM+SATV+SEX, dt)
# svglite("HSSvsGPA.svg")
plot(x$residuals, y$residuals, xlab="HSS", ylab="GPA", main = "HSS vs GPA")
# dev.off()
```

```
x = lm(HSE~HSM+HSS+SATM+SATV+SEX, dt)
y = lm(GPA~HSM+HSS+SATM+SATV+SEX, dt)
# svglite("HSEvsGPA.svg")
plot(x$residuals, y$residuals, xlab="HSE", ylab="GPA", main = "HSE vs GPA")
# dev.off()

x = lm(SATM~HSM+HSS+HSE+SATV+SEX, dt)
y = lm(GPA~HSM+HSS+HSE+SATV+SEX, dt)
# svglite("SATMvsGPA.svg")
plot(x$residuals, y$residuals, xlab="SATM", ylab="GPA", main = "SATM vs GPA")
# dev.off()

x = lm(SATV~HSM+HSS+HSE+SATM+SEX, dt)
y = lm(GPA~HSM+HSS+HSE+SATM+SEX, dt)
# svglite("SATVvsGPA.svg")
plot(x$residuals, y$residuals, xlab="SATV", ylab="GPA", main = "SATV vs GPA")
# dev.off()

x = lm(SEX~HSM+HSS+HSE+SATM+SATV, dt)
y = lm(GPA~HSM+HSS+HSE+SATM+SATV, dt)
# svglite("SEXvsGPA.svg")
plot(x$residuals, y$residuals, xlab="SEX", ylab="GPA", main = "SEX vs GPA")
# dev.off()

# Zad 13
dt = read.table('./csdata.txt', col.names=c("id", "GPA", "HSM", "HSS", "HSE", "SATM", "SATV'
reg1 = lm(GPA~HSM+HSS+HSE+SATM+SATV+SEX, dt)
r = residuals(reg1)
r1 = rstandard(reg1) # studentyzacja wewnętrzna
r2 = rstudent(reg1)  # studentyzacja zewnętrzna
# cbind(r, r1, r2)
# svglite("studentyzacjawewnetrzna.svg")
plot(1:dim(dt)[1], r1)
# dev.off()
# svglite("studentyzacjazewnetrzna.svg")
plot(1:dim(dt)[1], r2)
# dev.off()

# Zad 14
dt = read.table('./csdata.txt', col.names=c("id", "GPA", "HSM", "HSS", "HSE", "SATM", "SATV'
reg1 = lm(GPA~HSM+HSS+HSE+SATM+SATV+SEX, dt)
y = dffits(reg1)
p = dim(dt)[2] - 1
n = dim(dt)[1]
h_value = sqrt(p/n)
# svglite("dffits.svg")
```

```
plot(1:dim(dt)[1], y)
abline(h=2*h_value)
abline(h=-2*h_value)
# dev.off()

# Obserwacje leżące poza przedziałem +- 2 * h_value mają znaczący wpływ na predykcję.
# Widzimy, że zdecydowana większość obserwacji leży wewnątrz przedziału,
# ale jest kilkanaście obserwacji wykraczającymi poza ten zakres i mogą być one
# obserwacjami odstającymi lub wpływowymi.

# Zad 15
# Tolerancja to odwrotność VIF i pomaga nam zidentyfikować zjawisko multikolinearności.
# Wartości poniżej 0.1 wskazują na problem z multikolinearnością.
dt = read.table('./csdata.txt', col.names=c("id", "GPA", "HSM", "HSS", "HSE", "SATM", "SATV'
reg1 = lm(GPA~HSM+HSS+HSE+SATM+SATV+SEX, dt)
v = vif(reg1)
tolerance = 1/v
tolerance
# Widzimy, że wszystkie wartości wynosza ponad 0.1, więc w naszych danych
# nie występuje problem multikolinearności.

# Zad 16
dt = read.table('./csdata.txt', col.names=c("id", "GPA", "HSM", "HSS", "HSE", "SATM", "SATV'
n = dim(dt)[1]
reg = regsubsets(GPA~HSM+HSS+HSE+SATM+SATV+SEX, nbest=1, dt)
u = summary(reg)

aic_vals = rep(NULL, 6)
reg = lm(GPA~HSM, dt)
aic_vals[1] = AIC(reg)

reg = lm(GPA~HSM+HSE, dt)
aic_vals[2] = AIC(reg)

reg = lm(GPA~HSM+HSE+SATM, dt)
aic_vals[3] = AIC(reg)

reg = lm(GPA~HSM+HSE+SATM+HSS, dt)
aic_vals[4] = AIC(reg)

reg = lm(GPA~HSM+HSE+SATM+HSS+SATV, dt)
aic_vals[5] = AIC(reg)

reg = lm(GPA~HSM+HSS+HSE+SATM+SATV+SEX, dt)
aic_vals[6] = AIC(reg)
```

```
bic_vals = u$bic

cbind(bic_vals, aic_vals, u$which)
# Przy wyborze za pomocą BIC wybieramy model o najmniejszej wartości BIC
# W naszym przypadku jest to model z samą zmienną HSM
# Przy wyborze za pomocą AIC wybieramy model o najmniejszej wartości AIC
# W naszym przypadku jest to model z samą zmiennymi HSM i HSE
```