Modele 5

Katarzyna Stasińska

2024-01

Zadanie 1

a)

Polecenia wbudowane:

```
## Y= 1.053245 + -0.005860509 X1 + 0.001928049 X2 + 0.03014774 X3
```

```
## Współczynnik R^2 0.5415482
```

Wzory teoretyczne:

```
X = as.matrix(dane[,1:3])
nowa_kolumna = rep(1, nrow(X))
X = cbind(nowa_kolumna, X)
Y = as.matrix(dane[,4])
Bety = solve(t(X) %*% X) %*% (t(X)) %*% Y
cat("Y=", Bety[1],"+",Bety[2], "X1 +", Bety[3],"X2 +", Bety[4],"X3")
```

```
## Y= 1.053245 + -0.005860509 X1 + 0.001928049 X2 + 0.03014774 X3
```

```
SSM = sum((predict(model) - mean(Y))^2)
SST = sum((Y - mean(Y))^2)
R2 = SSM/SST
cat("Współczynnik R^2", R2)
```

```
## Współczynnik R^2 0.5415482
```

b)

Rozważmy hipotezę $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ przeciwko $H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0$.

F - Statystyka testowa z rozkładu Fishera-Snedecora z 3 i 46 - 4 = 42 stopniami swobody.

Przyjmijmy, że $\alpha = 0.05$.

Wzory teoretyczne:

```
dfM = 3
dfE = 42
SSE = SST - SSM
MSE = SSE/dfE
MSM = SSM/dfM
F = MSM/MSE
pval = 1 - pf(F,3,42)
cat("statystyka testowa:", F, "pvalue:",pval)
```

```
## statystyka testowa: 16.53756 pvalue: 3.04311e-07
```

Polecenia wbudowane

```
##
## Call:
## lm(formula = dane[, 4] ~ ., data = dane[, 1:3])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33589 -0.13333 -0.03347  0.12599  0.52022
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.053245   0.613791   1.716  0.09354 .
## wiek        -0.005861   0.003089  -1.897  0.06468 .
## ciężkość     0.001928   0.005787   0.333  0.74065
## niepokój     0.030148   0.009257   3.257  0.00223 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2098 on 42 degrees of freedom
## Multiple R-squared:  0.5415, Adjusted R-squared:  0.5088
## F-statistic: 16.54 on 3 and 42 DF,  p-value: 3.043e-07
```

P-wartość jest mniejsza od poziomu istotności, zatem możemy odrzucić hipotezę zerową.

c) Wiek

Rozważmy hipotezę $H_0 : \beta_1 = 0$ przeciwko $H_1 : \beta_1 \neq 0$.

F - Statystyka testowa z rozkładu Fishera-Snedecora z 1 i 42 stopniami swobody.

Przyjmijmy, że $\alpha = 0.05$.

Wzory teoretyczne:

```r
modelR = lm(dane[,4] ~., dane[,2:3])

SSM_R = sum((predict(modelR) - mean(Y))^2)
SSE_R = SST - SSM_R

F = (SSE_R - SSE)/MSE
pval = 1 - pf(F,1,42)
cat("statystyka testowa:", F, "pvalue:",pval)
```

```
## statystyka testowa: 3.599735 pvalue: 0.06467813
```

Polecenia wbudowane

```
## Analysis of Variance Table
##
## Model 1: dane[, 4] ~ ciężkość + niepokój
## Model 2: dane[, 4] ~ wiek + ciężkość + niepokój
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 2.0070
## 2     42 1.8486  1   0.15844 3.5997 0.06468 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P-wartość jest większa od poziomu istotności, zatem nie możemy odrzucić hipotezy zerowej.

c) ciężkość

Rozważmy hipotezę $H_0 : \beta_2 = 0$ przeciwko $H_1 : \beta_2 \neq 0$.

F - Statystyka testowa z rozkładu Fishera-Snedecora z 1 i 42 stopniami swobody.

Przyjmijmy, że $\alpha = 0.05$.

Wzory teoretyczne:

```
modelR = lm(dane[,4] ~ dane[,1] + dane[,3])

SSM_R = sum((predict(modelR) - mean(Y))^2)
SSE_R = SST - SSM_R

F = (SSE_R - SSE)/MSE
pval = 1 - pf(F,1,42)
cat("statystyka testowa:", F, "pvalue:",pval)
```

```
## statystyka testowa: 0.111014 pvalue: 0.7406503
```

Polecenia wbudowane

```
## Analysis of Variance Table
##
## Model 1: dane[, 4] ~ dane[, 1] + dane[, 3]
## Model 2: dane[, 4] ~ wiek + ciężkość + niepokój
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     43 1.8534
## 2     42 1.8486  1 0.0048861 0.111 0.7407
```

P-wartość jest większa od poziomu istotności, zatem nie możemy odrzucić hipotezy zerowej.

c) Niepokój

Rozważmy hipotezę $H_0 : \beta_3 = 0$ przeciwko $H_1 : \beta_3 \neq 0$.

F - Statystyka testowa z rozkładu Fishera-Snedecora z 1 i 42 stopniami swobody.

Przyjmijmy, że $\alpha = 0.05$.

Wzory teoretyczne:

```
modelR = lm(dane[,4] ~., dane[,1:2])

SSM_R = sum((predict(modelR) - mean(Y))^2)
SSE_R = SST - SSM_R

F = (SSE_R - SSE)/MSE
pval = 1 - pf(F,1,42)
cat("statystyka testowa:", F, "pvalue:",pval)
```

```
## statystyka testowa: 10.60735 pvalue: 0.002232272
```

Polecenia wbudowane

```
## Analysis of Variance Table
##
## Model 1: dane[, 4] ~ wiek + ciężkość
## Model 2: dane[, 4] ~ wiek + ciężkość + niepokój
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
```

```
## 1     43 2.3154
## 2     42 1.8486  1   0.46686 10.607 0.002232 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

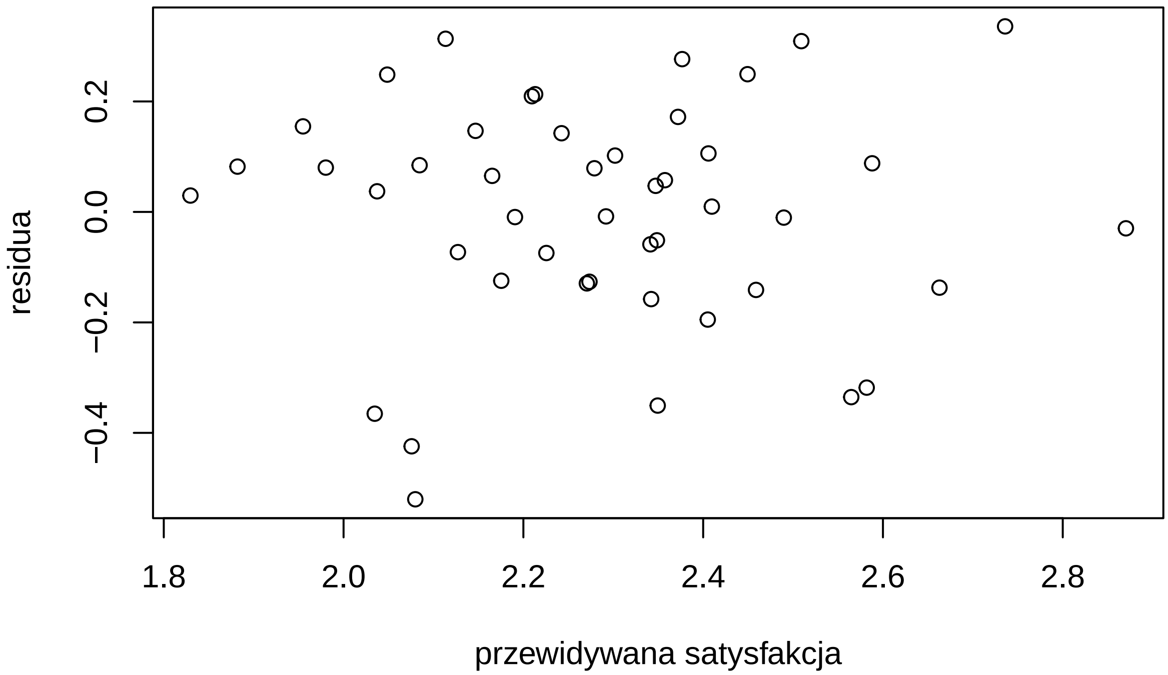
P-wartość jest mniejsza od poziomu istotności, zatem możemy odrzucić hipotezę zerową.

d)

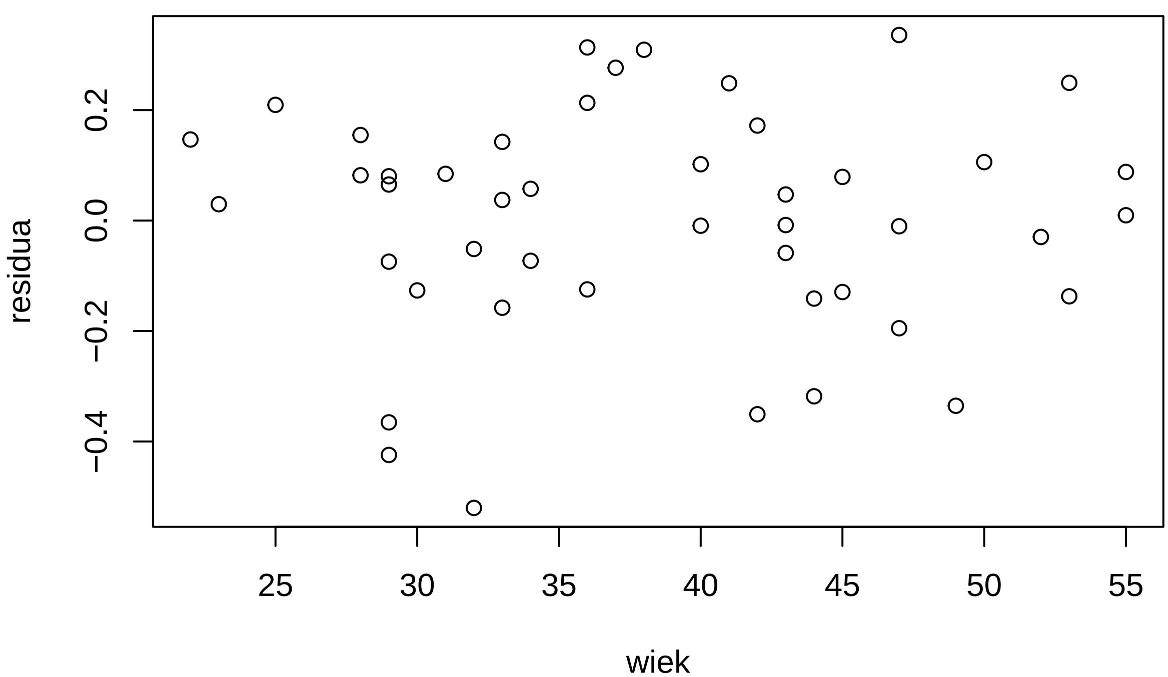
```
## Przedział ufności dla [ -0.01209411 , 0.0003730895 ]
## Przedział ufności dla [ -0.00974994 , 0.01360604 ]
## Przedział ufności dla [ 0.01146717 , 0.04882831 ]
```

Możemy zwrócić uwagę, że jedynie trzeci przedział ufności nie zawiera 0. I jedynie w przypadku trzecim odrzuciliśmy hipotezę zerową.

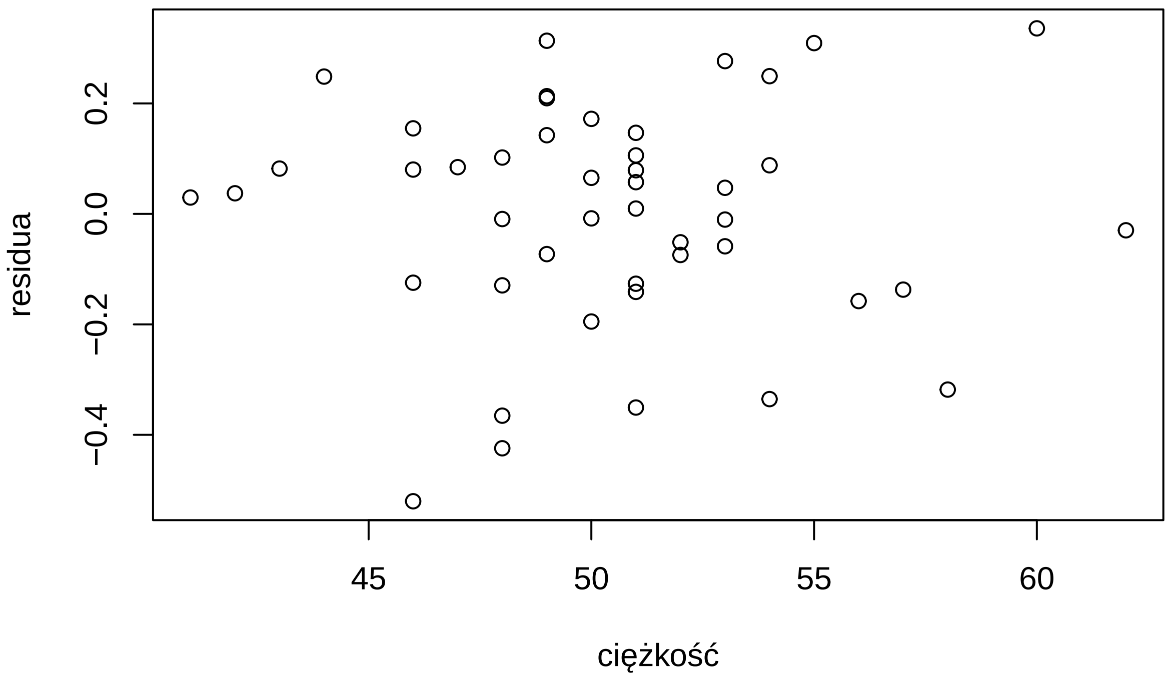
Zadanie 2



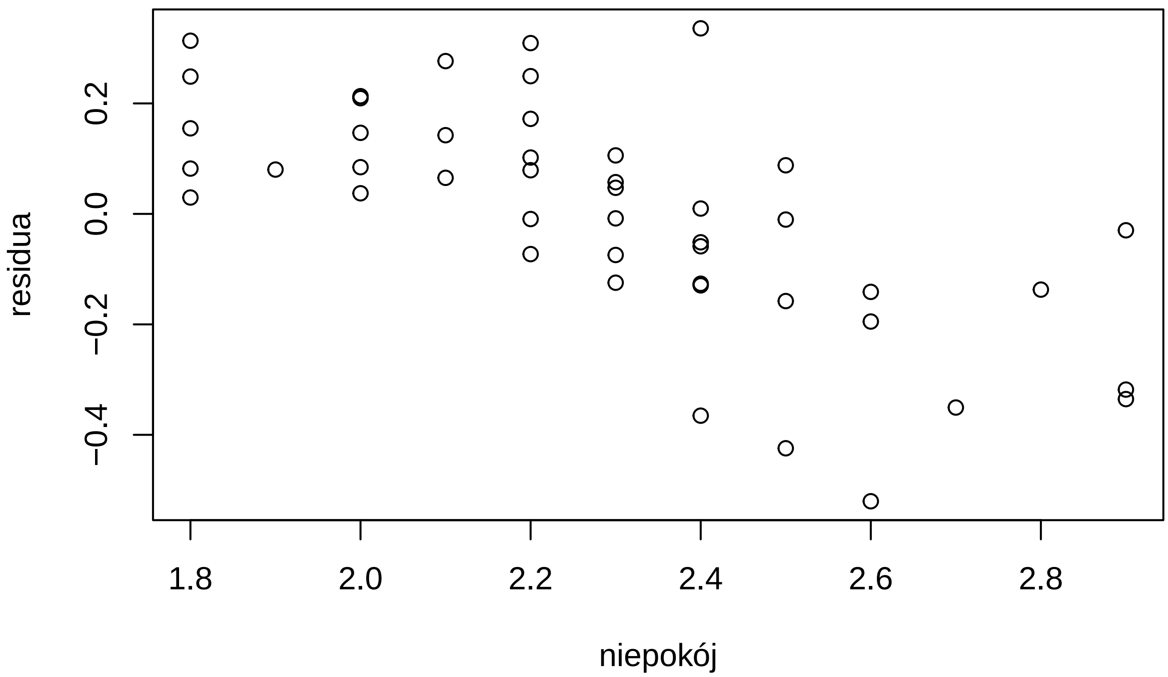
residua
0.2
0.0
−0.2
−0.4
1.8
2.0
2.2
2.4
2.6
2.8
przewidywana satysfakcja



residua
0.2
0.0
−0.2
−0.4
25
30
35
40
45
50
55
wiek



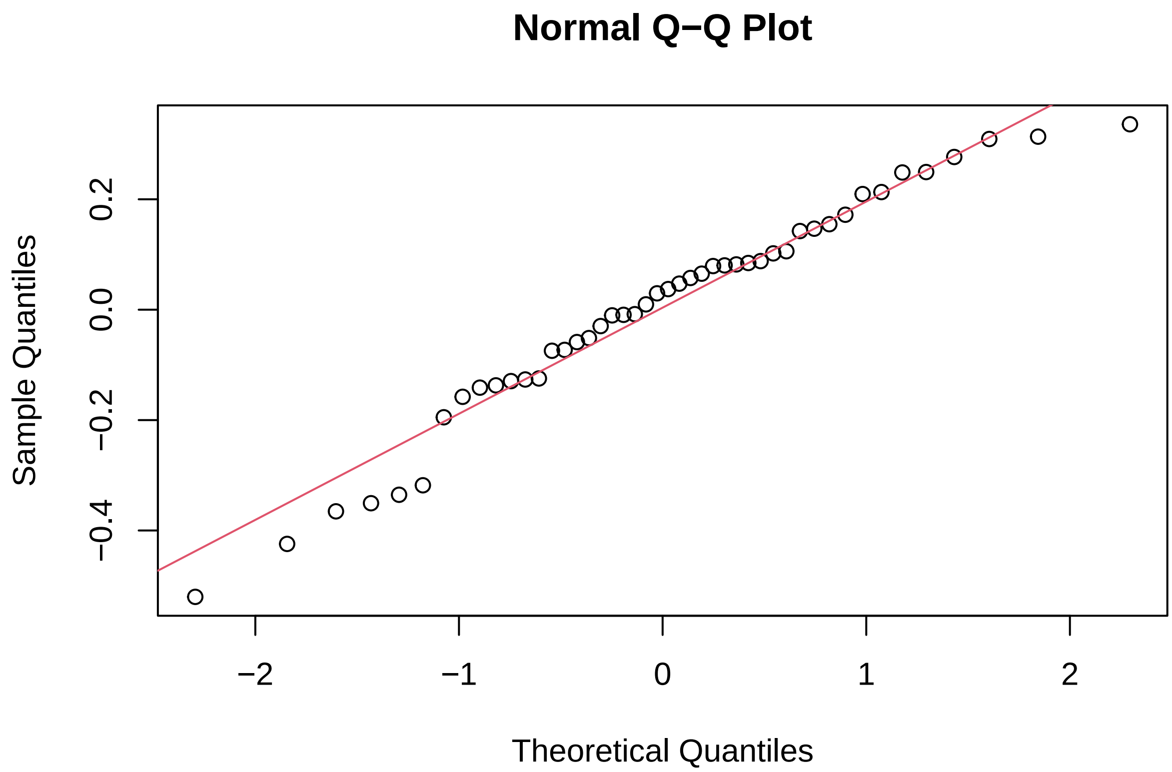
residua
0.2
0.0
−0.2
−0.4
45
50
55
60
ciężkość



Wraz ze wzrostem niepokoju, wartości residuów maleją.

Zadanie 3

```
##
##  Shapiro-Wilk normality test
##
## data:  residua
## W = 0.96286, p-value = 0.1481
```



Rozkład residuów może być w przybliżeniu normalny, test Shapiro-Wilka nie pozwala nam odrzucić hipotezy zerowej mówiącej o normalności tego rozkładu. Patrząc na wykres qqnorm możemy zauważyć, że ogony z obu stron odstają.

Zadanie 4

a)

```
## Różnica w SSE = 0.9313136
```

Rozważmy hipotezę $H_0 : \beta_4 = \beta_5 = 0$ przeciwko $H_1 : \beta_4 \neq 0 \vee \beta_5 \neq 0$.

F - Statystyka testowa z rozkładu Fishera-Snedecora z 2 i 224-6 $= 218$ stopniami swobody.

Przyjmijmy, że $\alpha = 0.05$.

```
## Statystyka testowa F wynosi =  0.9503276
```

b)

```
## Analysis of Variance Table
##
## Model 1: dane[, 2] ~ HSM + HSS + HSE
## Model 2: dane[, 2] ~ HSM + HSS + HSE + SATM + SATV
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    220 107.75
## 2    218 106.82  2   0.93131 0.9503 0.3882
```

Korzystając z funkcji anova mamy $F = 0,9503$ z 2 i 218 stopniami swobody, $p_{wartość} = 0.3882$. $p_{wartość} > \alpha$ zatem nie możemy odrzucić hipotezy zerowej.

Zadanie 5

a)

```
## Ładowanie wymaganego pakietu: carData
```

```
## Sumy kwadratów typu I
```

```
## 8.582934 0.0009054942 17.72647 1.891193 0.4421433
```

```
## Sumy kwadratów typu II
```

```
## 0.9279988 0.2326519 6.772431 0.956804 0.4421433
```

Jeśli znamy wartości sum typu I, to znamy też sumę kwaratów pełnego modelu, jest to ich suma. Sumy kwadratów typu II są używane do testowania hipotez, które badają istotność parametru β.

b)

```
## Suma kwaratów typu I dla HSM =  17.72647
```

```
## SSM modelu1 - SSM modelu2 = 17.72647
```

c)

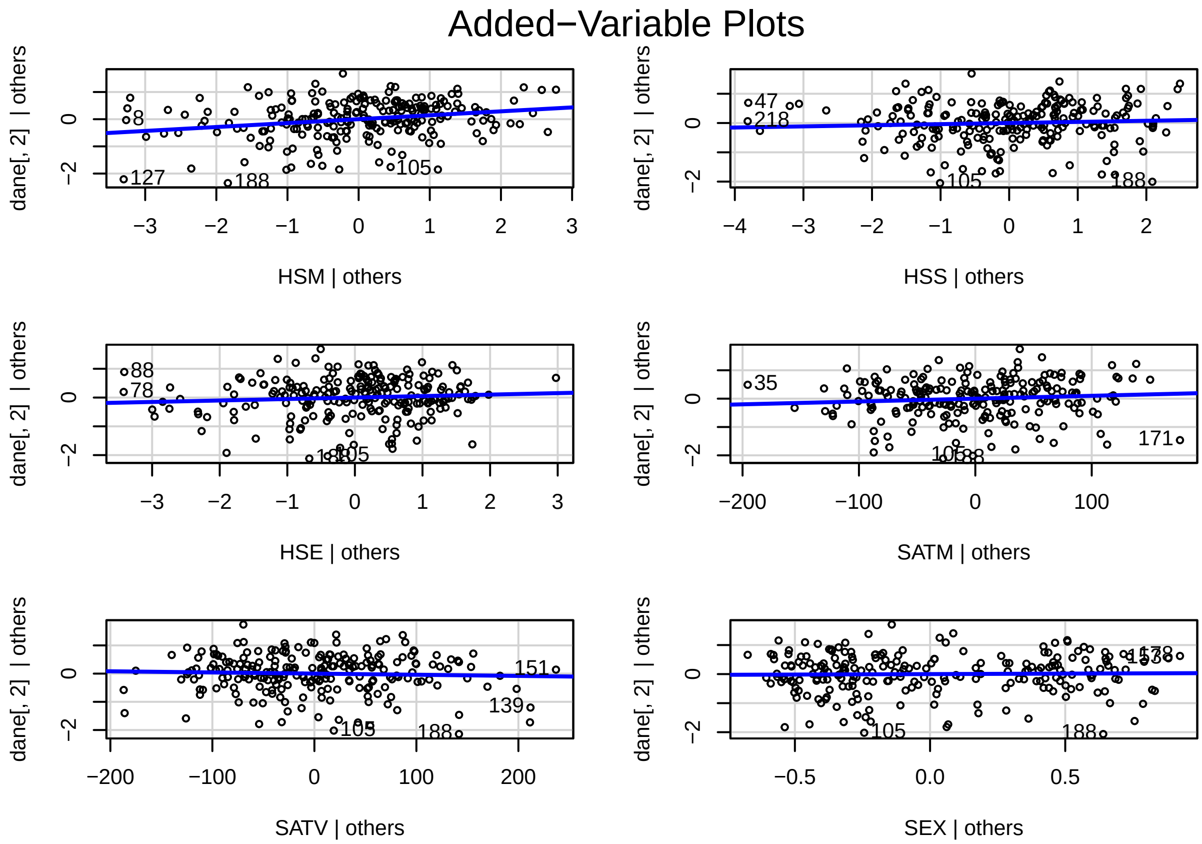
Tak, sumy kwadratów typu I i II są takie same dla ostatniego predykatora. Sumy kwadratów typu I definiujemy jako wpływ i-tej zmiennej po uwzględnieniu i-1 poprzednich zmiennych. Zatem dla ostatniego predykatora suma kwadratów typu I opisuje wpływ ostatniego predykatora po uwzględnieniu wszystkich poprzednich. Z kolei sumy kwadratów typu II zefiniowane są jako wpływ i-tej zmiennej po uwzględniniu wszystkich pozostałych w tym modelu.

Zadanie 6

```
##
## Call:
## lm(formula = dane2[, 2] ~ dane2[, 6] + dane2[, 7] + dane2[, 9])
##
## 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 ***
## dane2[, 6]   2.283e-03  6.629e-04   3.444 0.000687 ***
## dane2[, 7]  -2.456e-05  6.185e-04  -0.040 0.968357
## dane2[, 9]          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
```

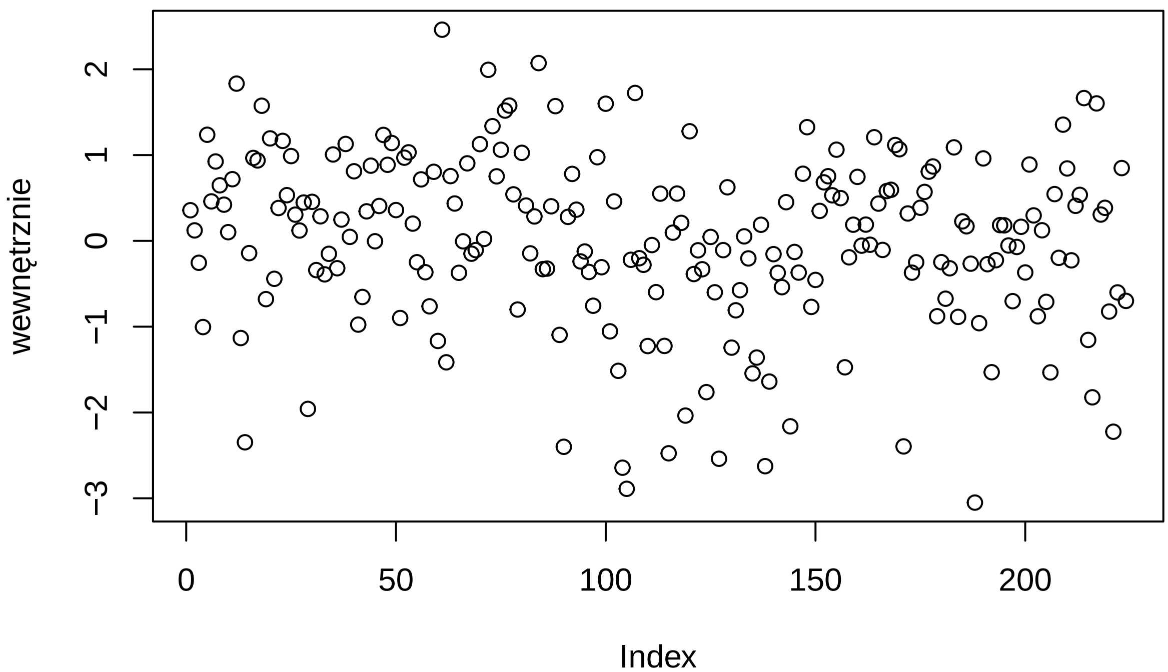
Zauważmy, że tak skonstruowany model nie definiuje nam współczynnika przy zmiennej SAT, bo jest ona kombinacją liniową pozostałych zmiennych.

Zadanie 7

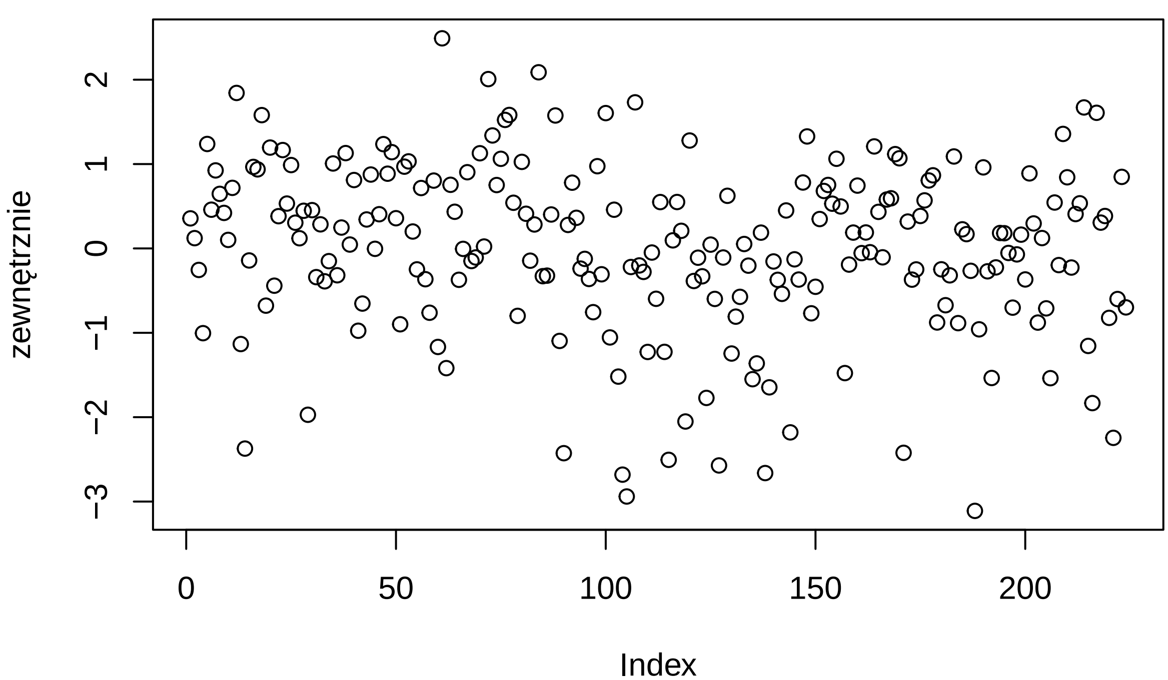


Partial regression plot przedstawia efekt dodania kolejnej zmiennej do modelu, który zawiera już jedną lub więcej zmiennych niezależnych. Nachylenie niebieskiej linii jest równe wartości estymatora danej zmiennej objaśniającej w modelu regresji wielorakiej. Im mniejsza wartość bezwzgledna nachylenia niebieskiej prostej, tym mniejsza informacja wniesiona do modelu przez daną zmienną. Wszystkie niebieskie wykresy są liniowe, więc nie jest wymagana transformacja danych.

b)



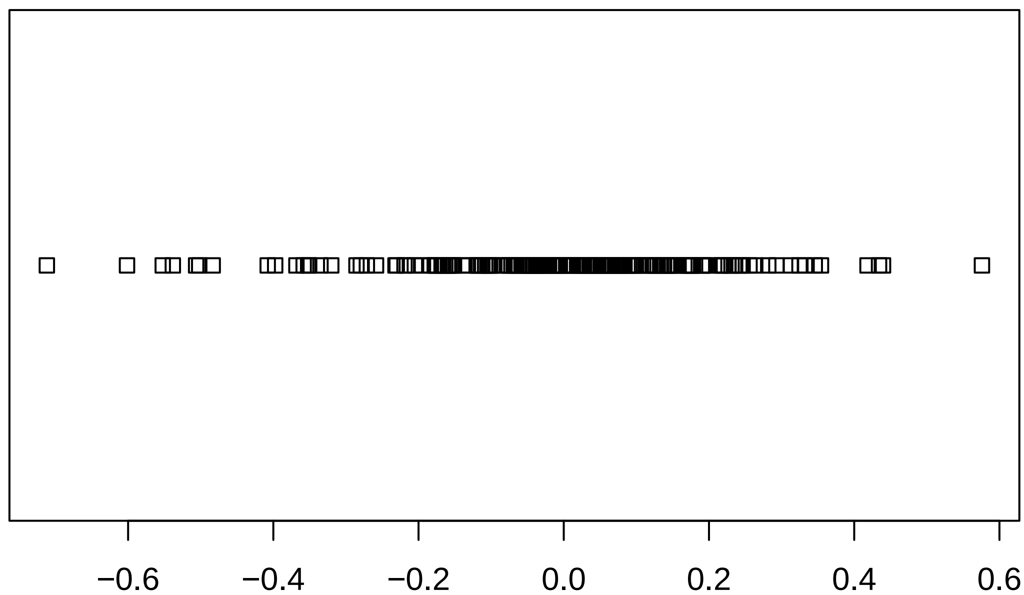
wewnętrznie
2
1
0
−1
−2
−3
0
50
100
150
200
Index



zewnętrznie
2
1
0
−1
−2
−3
0
50
100
150
200
Index

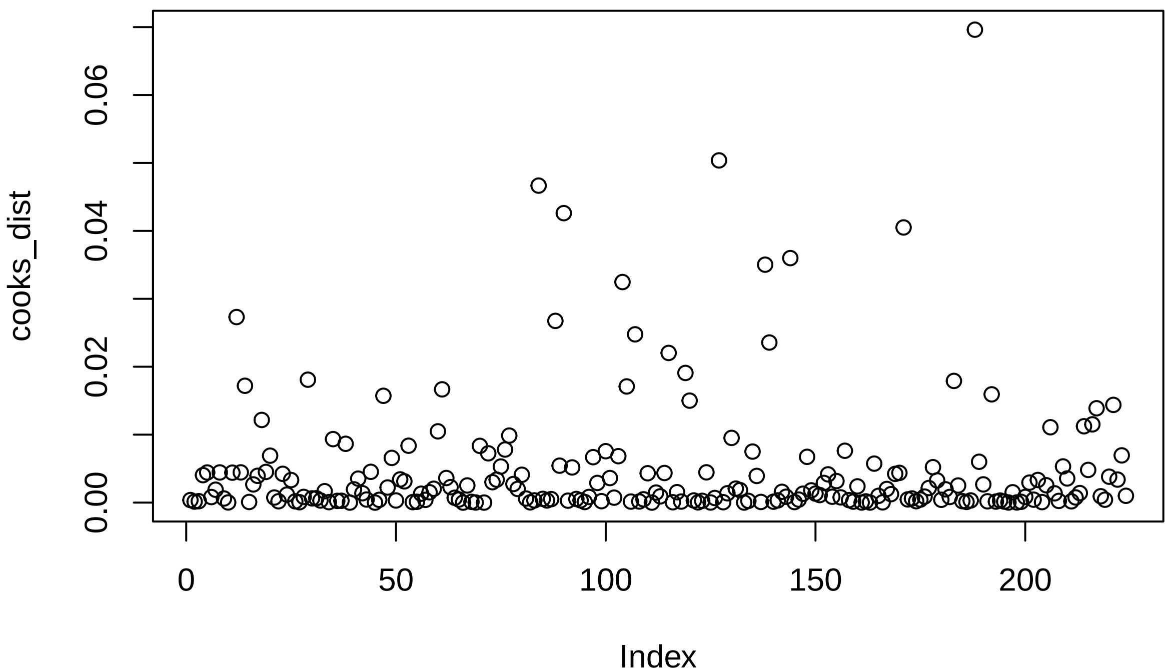
W residuach studentyzowanych wewnętrznie korzystamy z klasycznego modelu (wykorzystującego wszystkie obserwacje). O zewnętrznej studentyzacji residuów mówimy, kiedy korzystamy z takiego samego modelu, ale z pominięciem i-tej obserwacji, do wyznaczenia wartości i-tego residuum. Zaletą wewnętrznie studentyzowanych residuów jest to, że określają one, jak duże są reszty w jednostkach odchylenia standardowego, a zatem można je łatwo wykorzystać do identyfikacji wartości odstających. Dlatego powinniśmy lepiej się przyjrzeć tym residuuom, których bezwględne wartości są najwyższe.

c)



DFFITS dla i–tej obserwacji jest standaryzowaną różnicą pomiędzy predykcjami wartości Y_i uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, pełnych i bez obserwacji Y_i. Spodziewamy się, że obie predykcje będą przyjmowały podobne wartości, wtedy DFFITS przyjmuje małe wartości. Powinniśmy się lepiej przyjrzeć tym obserwacjom, dla których $|DFFITS_i| > 2\sqrt{p/n} = 2 * \sqrt{5/224} = 0.2988072$

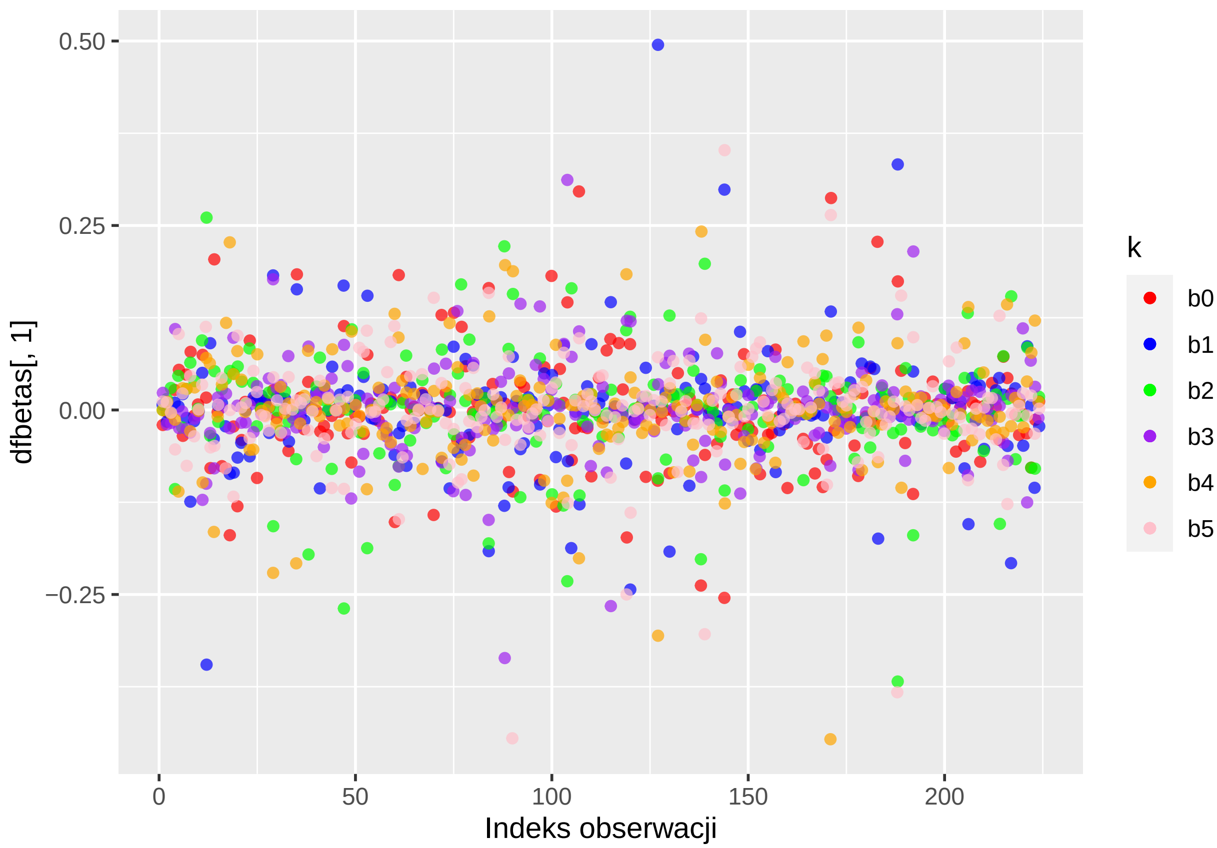
d)



Odległość Cook'a (D_i) dla i–tej obserwacji również jest standaryzowaną różnicą pomiędzy predykcjami wektora Y uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, pełnych i bez obserwacji Y_i. Analogicznie jak w przypadku wyżej, im mniejsza odległość cook'a tym lepiej. Powinniśmy lepiej przyjrzeć się obserwacjom, dla których $|D_i| > 1$.

e)

```
## Ładowanie wymaganego pakietu: lattice
```



Miara DFBETAS służy do badania wpływu Y_i na estymację parametru β_k. Dla k–tego parametru jest różnicą pomiędzy dwoma estymatorami parametru β_k uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, pełnych i bez obserwacji Y_i podzielonymi przez estymator odchylenia standardowego estymatora uzyskanego na podstawie niepełnego modelu. Analogicznie jak wyżej, im mniejsza miara $|DFBETA_k|$ tym lepiej. Powinniśmy lepiej przyjrzeć się obserwacjom, dla których $|DFBETA_k| > 2/\sqrt{n} = 0.1336306$.

f)

Tolerance jest odwrotnością Variance Inflation Factor (VIF). VIF bada, dla k-tej zmiennej, w jakim stopniu zmienna X_k jest objaśniana przez wszystkie pozostałe zmienne objaśniające. Gdy Tolarance < 0.1 to ma miejsce problem z multikolinearnością.

Możemy zauważyć, że w modelu z zadania 7 nie występuje problem z multikolinearnością.

```
## Tolerance 0.5188628 0.5088203 0.5429546 0.5745498 0.7310535 0.7742519
```

Natomiast w modelu z zadania 6 występuje ten problem, dlatego spodziewam się tolerancji < 0.1. Wbudowane funkcje zwracają błąd przy próbie liczenia dokładnej wartości tolerancji, bo model składa sie z liniowozależnych zmiennych.

g)

Kryteria AIC oraz BIC są modyfikacjami metody największej wiarogodności i są skonstruowane w taki sposób, by znaleźć balans pomiędzy dopasowaniem modelu do danych i nadmierną złożonością modelu. Statystyka Cp Mallowsa opisuje łączne zachowanie obciążeń.

```
## (Intercept)         HSM         HSS         HSE        SATM        SATV
##        TRUE        TRUE       FALSE       FALSE       FALSE       FALSE
```

```
##          SEX
##        FALSE
```

Najlepszy model według Cp:

```
## (Intercept)          HSM          HSS          HSE         SATM         SATV
##        TRUE         TRUE        FALSE         TRUE        FALSE        FALSE
##          SEX
##        FALSE
```

Najlepszy model według adj R squared:

```
## (Intercept)          HSM          HSS          HSE         SATM         SATV
##        TRUE         TRUE        FALSE        FALSE        FALSE        FALSE
##          SEX
##        FALSE
```

Zadania teoretyczne

Zadanie 1

a)

$Y = 1 + 4 * 2 + 3 * 6 = 27$

b)

$s^2(pred) = s^2(\hat{\mu}_h) + s^2 = 4 + 9 = 13$

c)

Przedział ufności wyznacza $b_1 \pm t_c s(b_1)$, gdzie t_c to kwantyl rzędu $1 - \alpha/2 = 0.975$ z $n - 2 = 18$ stopniami swobody z rozkładu studenta. $t_c = 2.100922$. Zatem przedział ufności to $[1.899078, 6.100922]$

Zadanie 2

Niech $\alpha = 0.05$

a)

Suma kwadratów typu I dla X_3 ma postać $SSM(X_3|X_1, X_2)$, dokładnie taką samą jak suma kwadratów typu II. Zatem suma kwadratów typu II dla $X_3 = 20$.

b)

Rozważmy hipotezę $H_0 : \beta_1 = 0$ przeciwko $H_1 : \beta_1 \neq 0$

Wiemy, że $SSM = 360$, $SST = 760$ i $SSE = 400$ stąd $MSE = SSE/dfE = 400/20 = 20$.

$F = \frac{SSM(X_1|X_2,X_3)}{MSE(F)} = \frac{30}{20} = 1.5$

$F^*(1 - \alpha = 0.95, 1, 20) = 4.351244$. Zatem nie możemy odrzucić hipotezy zerowej, bo $F < 4.351244$.

c)

Rozważmy hipotezę $H_0 : \beta_2 = \beta_3 = 0$ przeciwko $H_1 : \beta_2 \neq 0 \vee \beta_3 \neq 0$.

Wiemy, że $SSM = SSM(X_2, X_3|X_1) + SSM(X_1)$. Zatem $SSM(X_2, X_3|X_1) = 360 - 300 = 60$. Wiemy też, że $SSM(X_2, X_3|X_1) = SSE(X_2, X_3|X_1)$. Zatem

$F = \frac{SSE(X_2,X_3|X_1)/2}{20} = 1.5$

$F^*(1 - \alpha = 0.95, 2, 20) = 3.492828$. Zatem nie możemy odrzucić hipotezy zerowej, bo $F < 3.492828$.

d)

Rozważmy hipotezę $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ przeciwko $H_1 : \beta_1 \neq \vee \beta_2 \neq 0 \vee \beta_3 \neq 0$.

$MSM = SSM/dfM = 360/3 = 120$

$F = \frac{MSM}{MSE} = \frac{120}{20} = 6$

$F^*(1 - \alpha = 0.95, 3, 20) = 3.098391$. Zatem odrzucamy hipotezę zerową, bo $F > 3.098391$.

e)

Rozważmy hipotezę $H_0 : \beta_1 = 0$ przeciwko $H_1 : \beta_1 \neq 0$. W nowym modelu $SST = 760$, $SSM = 300$, $SSE = 760 - 300 = 460$, $MSE = 460/22$, $MSM = 300/1$.

$F = \frac{MSM}{MSE} = \frac{300}{460/22} = 30 * 22/46 = 14.34783$

$F^*(1 - \alpha = 0.95, 1, 22) = 4.30095$. Zatem odrzucamy hipotezę zerową, bo $F > 4.30095$.

f)

Próbkowy współczynnik korelacji między Y a $X_1 = \sqrt{R^2} = \sqrt{\frac{SSM}{SST}} = \sqrt{\frac{300}{760}} = 0.6282809$