Sprawozdanie 5

Katarzyna Botulińska

2025-01-31

Spis treści

zad 1a

Podgląd na plik 'tabela1_6.txt'.

Tablica 1: Plik z danymi

Age	Sickness	Anxiety	Satisfaction
48	50	51	2.3
57	36	46	2.3
66	40	48	2.2
70	41	44	1.8
89	28	43	1.8

Tablica 2: Porównanie wyników obliczonych za pomocą poleceń wbudownaych w R i wzorów teoretycznych

	R	teoretycznie
b_0	1.0532451	1.0532451
b_1	-0.0058605	-0.0058605
b_2	0.0019280	0.0019280
b_3	0.0301477	0.0301477
R^2	0.5415482	0.5411407

Równanie regresji: $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \epsilon$, czyli $Y = 1.0532 + \text{-}0.0059 X_1 + 0.0019 X_2 + 0.0301 X_3$.

Wnioski: Zauważyć można, ze wartości współczynników regresji oraz współczynnika R^2 wyliczone teoretycznie i za pomocą wzorów wbudowanych w R dają takie same wyniki.

zad 1b

Tablica 3: Porównanie wyników obliczonych za pomocą poleceń wbudownaych w R i wzorów teoretycznych

	R	teoretycznie
F-statystyka	16.5375621	16.5104442
P-wartość	0.0000003	0.0000003

Testowana hipoteza:

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

$H_1 : \beta_i \neq 0$ przynajmniej dla jednego i

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

Liczba stopni swobodwy $dfE(R) - dfE(F) = 3$

Wnioski: P-wartość jest mniejsza niż $\alpha = 0.05$, czyli test F jest istotny statystycznie. Wysoka $F - statystyka$ sugeruje, że model ma duży wpływ na zmienność danych. Oznacza to, że odrzucamy H_0. Wyciągmay tym samym wniosek, że przyanjmniej jeden z predyktorów (wiek, poziom niepokoj, ciężkość choroby) istotnie wpływa na zmienną zależną (poziom satysfakcji).

zad 1c

• Testowanie hipotezy, że poziom satysfakcji pacjentów nie zależy od wieku

Tablica 4: Porównanie wyników obliczonych za pomocą poleceń wbudowanych w R i wzorów teoretycznych

	R	teoretycznie
staytyka t	-1.8972967	-1.8972967
p-wartość	0.0646781	0.0646781

Testowana hipoteza:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Statystyka testowa $t = \frac{\hat{\beta}}{SSE(\beta)}$

Liczba stopni swobodwy $df = 42$

Wnioski: P-wartość jest większa niż $\alpha = 0.05$, w związku z tym nie możemy odrzucić hipotezy zerowej. Na podstawie przeprowadzonego testu nie możemy stwierdzić, że poziom satysfkacji nie zależy od wieku.

• Testowanie hipotezy, że poziom satysfakcji pacjentów nie zależy od ciężkości choroby

Tablica 5: Porównanie wyników obliczonych za pomocą poleceń wbudowanych w R i wzorów teoretycznych

	R	teoretycznie
staytyka t	0.3331876	0.3331876
p-wartość	0.7406503	0.7406503

Testowana hipoteza:

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

Statystyka testowa $t = \frac{\hat{\beta}}{SSE(\beta)}$

Liczba stopni swobodwy $df = 42$

Wnioski: P-wartość jest większa niż $\alpha = 0.05$, w związku z tym nie możemy odrzucić hipotezy zerowej. Na podstawie przeprowadzonego testu nie możemy stwierdzić, że poziom satysfkacji nie zależy od ciężkości choroby.

• Testowanie hipotezy, że poziom satysfakcji pacjentów nie zależy od poiozmu niepokoju

Tablica 6: Porównanie wyników obliczonych za pomocą poleceń wbudowanych w R i wzorów teoretycznych

	R	teoretycznie
staytyka t	3.2568922	3.2568922
p-wartość	0.0022323	0.0022323

Testowana hipoteza:

$H_0 : \beta_3 = 0$

$H_1 : \beta_3 \neq 0$

Statystyka testowa $t = \frac{\hat{\beta}}{SSE(\beta)}$

Liczba stopni swobodwy $df = 42$

Wnioski: P-wartość jest mniejsza niż $\alpha = 0.05$, w związku z tym odrzucamy hipotezę zerową. Na podstawie przeprowadzonego testu z dużym prawdopodobieństwem możemy stwierdzić, że poziom satysfkacji pacjentów zależy od poziomu niepokoju.

zad 1d

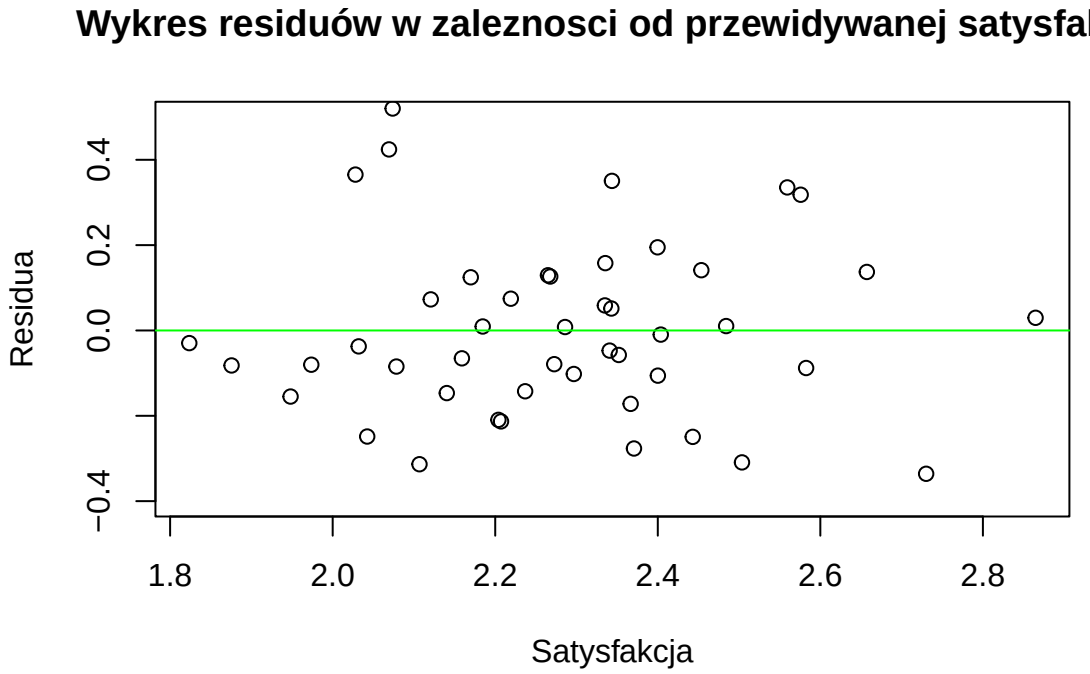
95% przedziały ufności dla współczynnika regresji przy wieku [-0.0121, 0]

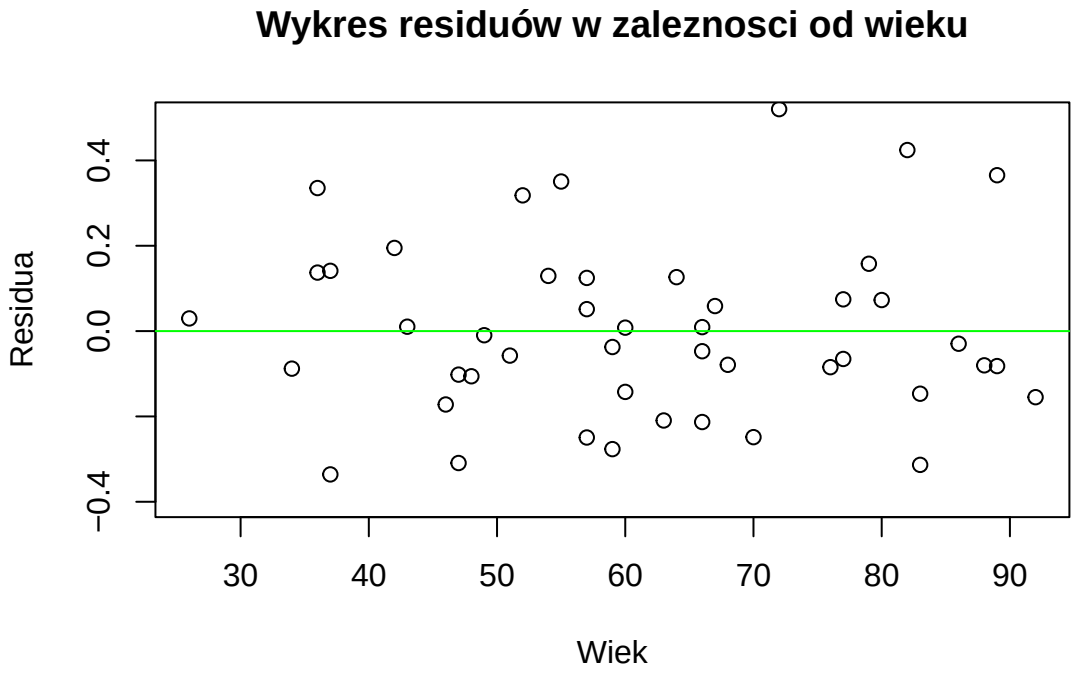
95% przedziały ufności ufnośco dla współczynnika regresji przy ciężkości choroby [-0.0098, 0.01]

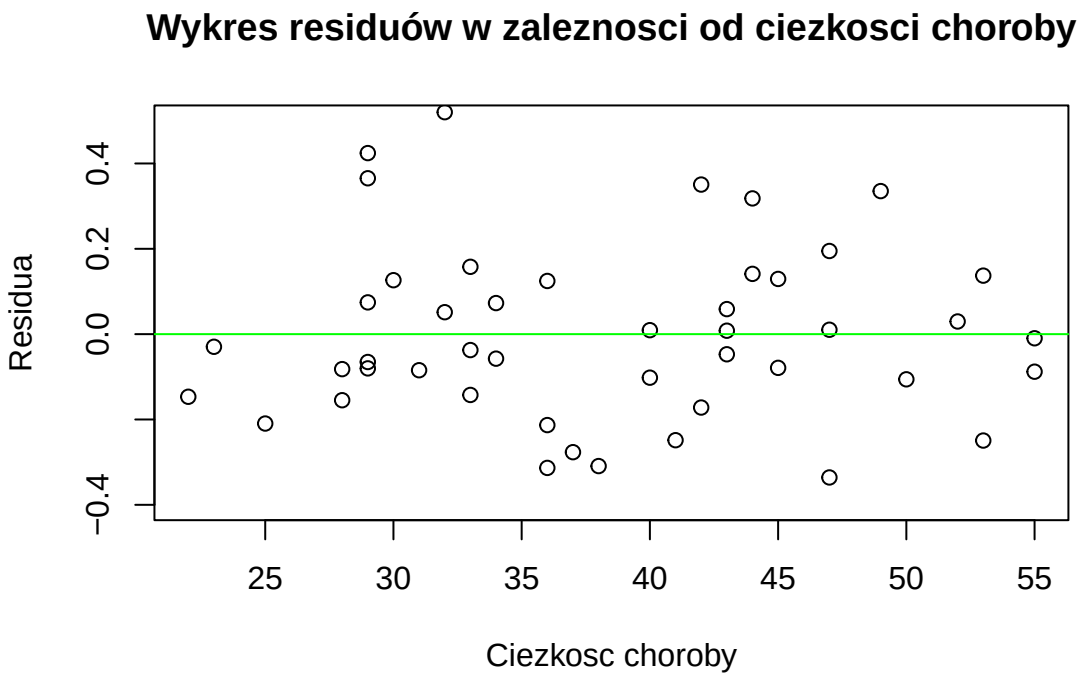
95% przedziały ufności dla współczynnika regresji przy poziomie niepokoju [0.0114, 0.05]

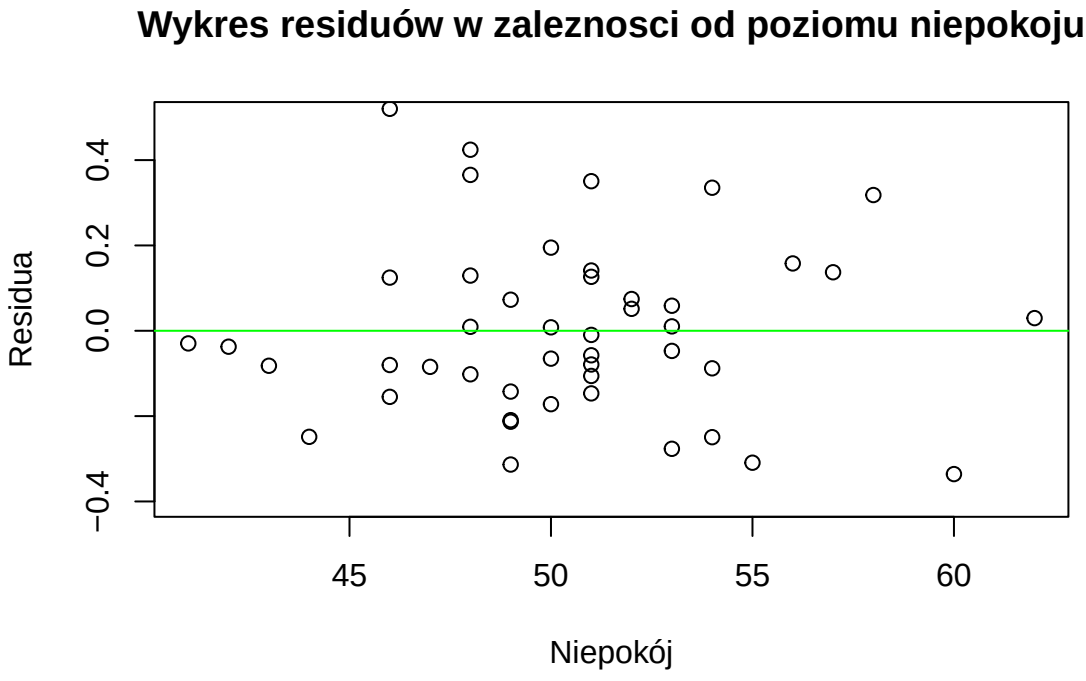
Wnioski: Związek między tymi wynikami, a wynikami z punktu *c*) jest bardzo dobrze widoczny. Ponieważ 95% przedziały ufności $\hat{\beta}_1$ oraz $\hat{\beta}_2$ zawierają 0, a przdział dla $\hat{\beta}_3$ nie zawiera 0. Dlatego, tylko dla $\hat{\beta}_3$ możemy spokojnie odrzucić H_0, a dla pozostałych β nie.

zad 2


Wykres residuów w zaleznosci od przewidywanej satysfakcji
Residua
0.4
0.2
0.0
−0.4
1.8
2.0
2.2
2.4
2.6
2.8
Satysfakcja

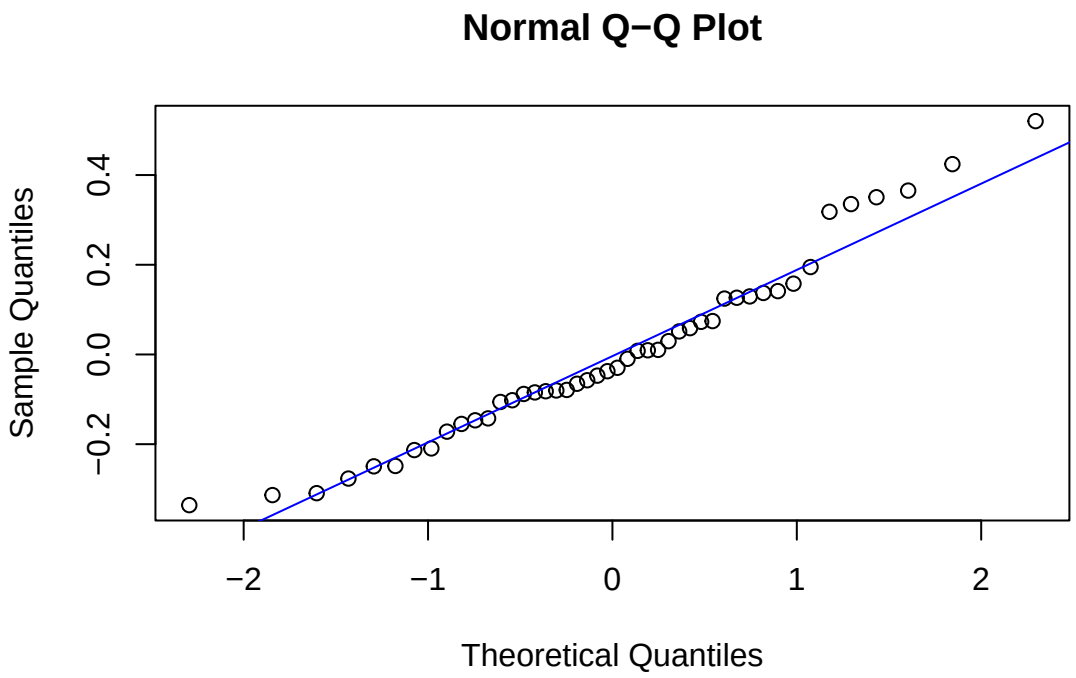

Wykres residuów w zaleznosci od wieku
Residua
0.4
0.2
0.0
−0.4
30
40
50
60
70
80
90
Wiek





Wnioski: Na wszystkich wykresach punkty są losowo rozrzucone wokół 0, nie występują żadne nietypowe wzory, ani ekstremalnie odstające wartości.

zad 3



H_0 : dane pochodzą z rozkładu normalnego

H_1 : dane nie pochodzą z rozkładu normlanego

Wnioski: Wykres residuów jest w przybliżeniu normalny, ponieważ staytsyka $W = 0.96$ jest bliska 1, a p-wartość testu Shapiro-Wilka wynosi $0.15 > 0.05$. W związku z tym nie ma dowodów na istotne odstępstwa od normalności. Wykres kwantylowo-kwantylowy także potwierdza normalność rozkładu, chociaż możemy zauważyć, że "na ogonach" wyniki są bardziej rozproszone.

zad 4

Różnica między statystykami SSE dla dwóch modeli. Dla modelu z HSM, HSS, HSE statytystyka $SSE(R) = 107.75$, a dla drugiego modelu $SSE(F) = 106.82$.

Różnica $SSE(R) - SSE(F) = 107.75 - 106.82 = 0.93$.

Konstrukcja statystyki $F = \frac{SSE(R)-SSE(F)/df(R)-df(F)}{MSE(F)} = 0.9503$.

$H_0 : \beta_4 = \beta_5 = 0$

H_1 : przyanjmniej dla jednej z β_4 , $\beta_5 \neq 0$

Statystyka F obliczona za pomocą funkcji *anova*: $F = 0.9503$.

Liczba stopni swobody: $df_1 = 2$, $df_2 = dfE(F) = 218$.

P-wartość: 0.39.

Wnioski: Jeśli statystyka $F > F^*$ to odrzucamy H_0. Natomiast $F = 0.9503 < F^* = 3.04$ i p-wartość 0.39 > 0.05, to znaczy, że nie możemy odrzucić H_0. Zmienne $SATM$ i $SATV$ mogą być nieistotne przy opisie GPA studentów. Jednkaże wynik jest niekonkluzywny, więc nie mamy takiej pewności.

zad 5a

	Sumy kwadratów typu I	Sumy kwadratów typu II
SATM	8.5829	0.9280
SATV	0.0009	0.2327
HSM	17.7265	6.7724
HSE	1.8912	0.9568
HSS	0.4421	0.4421

Wnioski:

Zmienna HSM ma dużą wartość zarówno dla sum typu I, jak i II, co świadczy o tym, że w dużym stopniu wyjaśnia zmienność w modelu zarówno przed jak i po uwzględnieniu innych zmiennych modelu.

Zmienna $SATM$ ma dużą wartość sumy typu I, ale małą wartość sumy typu II. Możemy na tej podstawie wywnioskować, że $SATM$ na początku jest ważna dla modelu, ale po uwzględnieniu innych zmiennych jej wpływ maleje.

Zmienna HSE ma średni wkład w wyjaśnienie zmienności modelu, a po uwzględnieniu wszystkich zmiennych jej wpływ nieco maleje.

Zmienna HSS ma niewielki wpływ na model w przypadku obu sum, a zmienna $SATV$ w znikomym stopniu wyjaśnia zmienność modelu.

Podumowując największy wpływ na zmienność modelu ma zmienna HSM, nastepnie zmienne $SATM$ i HSE. Zmienne HSS i $SATV$ mają najmniejszy wpływ.

zad 5b

Zweryfikujemy teraz, że suma kwadratów typu I dla zmiennej HSM jest równa różnicy SSM dla modelu opisującego GPA za pomocą $SATM$, $SATV$ oraz HSM (model 1) oraz SSM dla modelu opisującego GPA za pomocą $SATM$ oraz $SATV$ (model 2).

Wniosek: Suma I typu dla zmiennej HSM wynosi 17.7265, a różnica statystyk, z którą porównujemy tą sumę jest równa 17.7265, co dowodzi temu, że obie wartości są sobie równe. Dzieje się tak, dlatego, że zachodzi wzór:

$$SSM(X_3) = SSM(X_3|X_1, X_2) = SSM(X_1, X_2, X_3) - SSM(X_1, X_2) = model_1 - model_2$$

zad 5c

$$SSM_2(X_p) = SSM_2(X_p|X_1, X_2, \ldots, X_{p-1}, X_{p+1}, \ldots X_k)$$

$$\text{gdy p to ostatnia zmienna to } = SSM_2(X_p|X_1, X_2, \ldots, X_{p-1}) = SSM_1(X_p|X_1, X_2, \ldots, X_{p-1})$$

SSM_x, gdzie $x \in \{1, 2\}$ to suma odpowiednio I lub II typu.

Wnioski: Sumy typu I i sumy typu II są sobie równe zawsze dla ostatniego predyktora, ponieważ sumy typu I, biorą predyktory po kolei o 1 więcej, a sumy typu II wszystkie pozostałe, to dla ostatniego prdyktora otrzymujmey zawsze taką samą sumę.

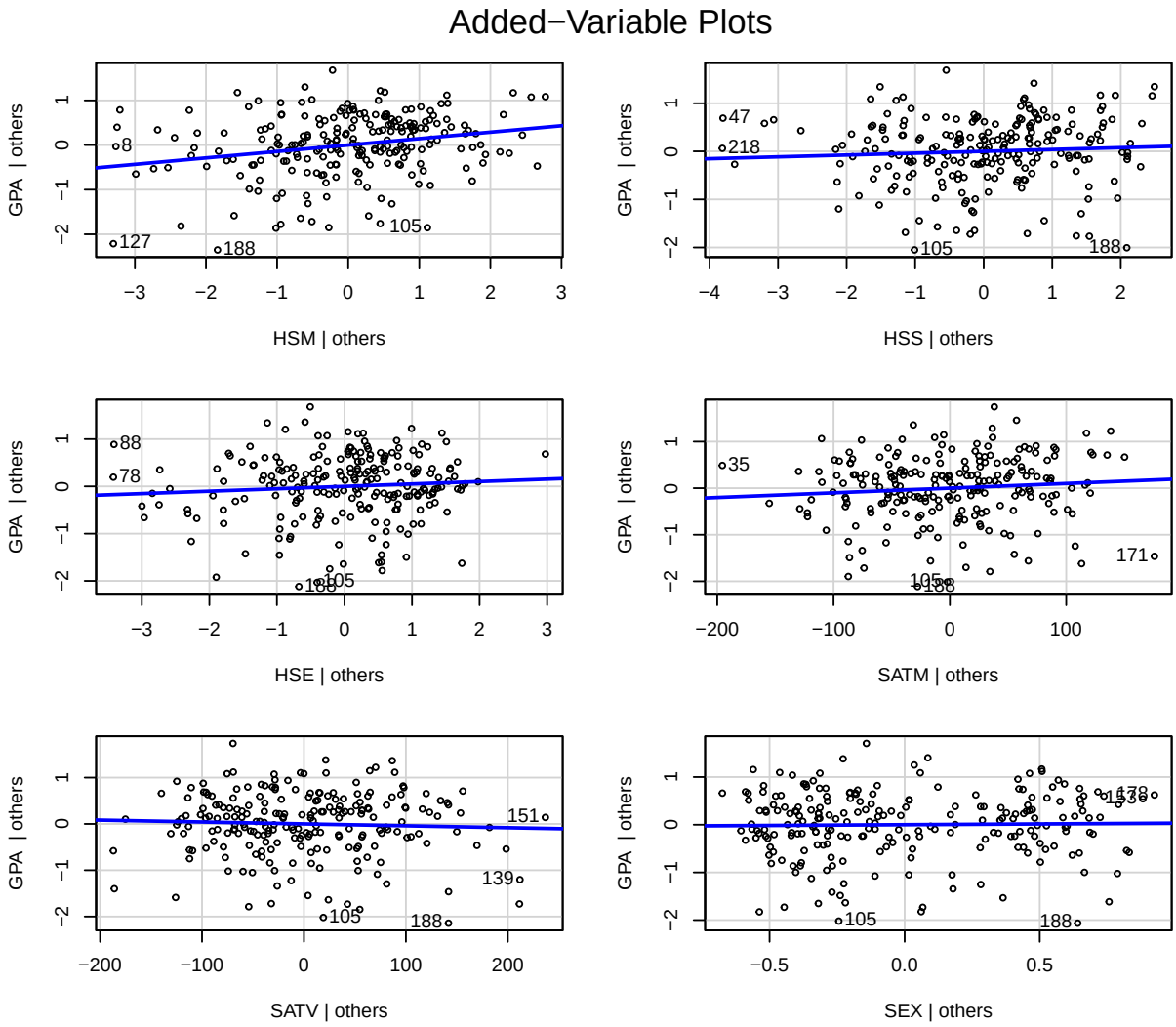
zad 6

Wygenerowano nową zmienną SAT jako sumę dwóch testów $SATM$ i $SATV$.

Wnioski: W wyniku nie uzyskaliśmy nic sensownego, gdyż model nie był w stanie wyznaczyć współczynnika $\hat{\beta}_3$. Jest to spowodowane tym, że macierz planu $\mathbb{X}$ jest singularna, czyli nie istnieje jej odwrotność.

Ponieważ $\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'Y$, stąd nie jesteśmy w stanie jej wyznaczyć. Singularność macierzy $\mathbb{X}$ wynika z tego, że wśród jej kolumn występuje kombinacja liniowa innych kolumn. Jest nią oczywiście ostatnia kolumna ze zmienną SAT, która jest równa sumie dwóch wcześniejszych kolumn macierzy $\mathbb{X}$.

zad 7a



Partial regression plots to wykresy, które pokazują jaki wpływ ma dodanie nowej zmiennej objaśniającej $\tilde{X}_i$ do modelu, który ma już inne zmienne niezależne. Opisuje relacje X_i vs Y uwzgledniając jaki wpływ na model mają pozostałe $X - y$.

Informacje jakie przekazują wykresy:

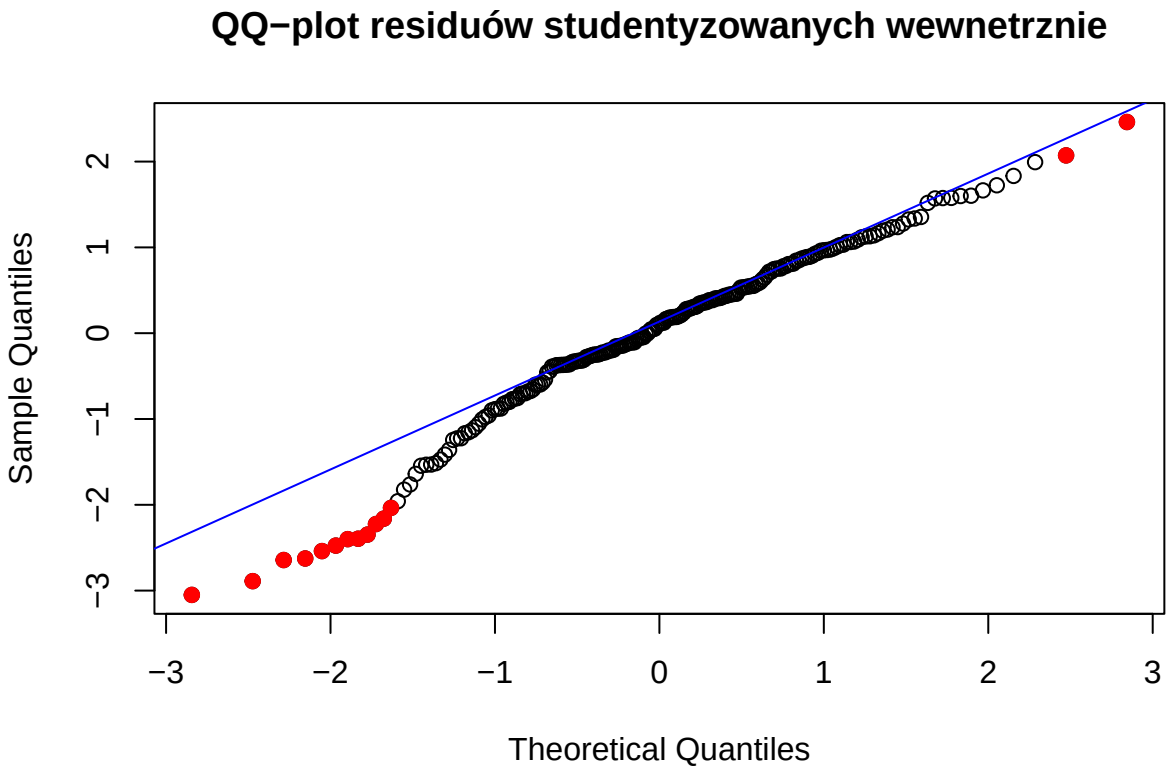
• jeśli wykres e^{X_i} vs e^Y nie ma jakiejś konkretnej struktury to daj enam informacje, że zmienna X_i nie wnosi do modelu istotnejt informacji, ponad to co objasniały pozostałę $X - y$

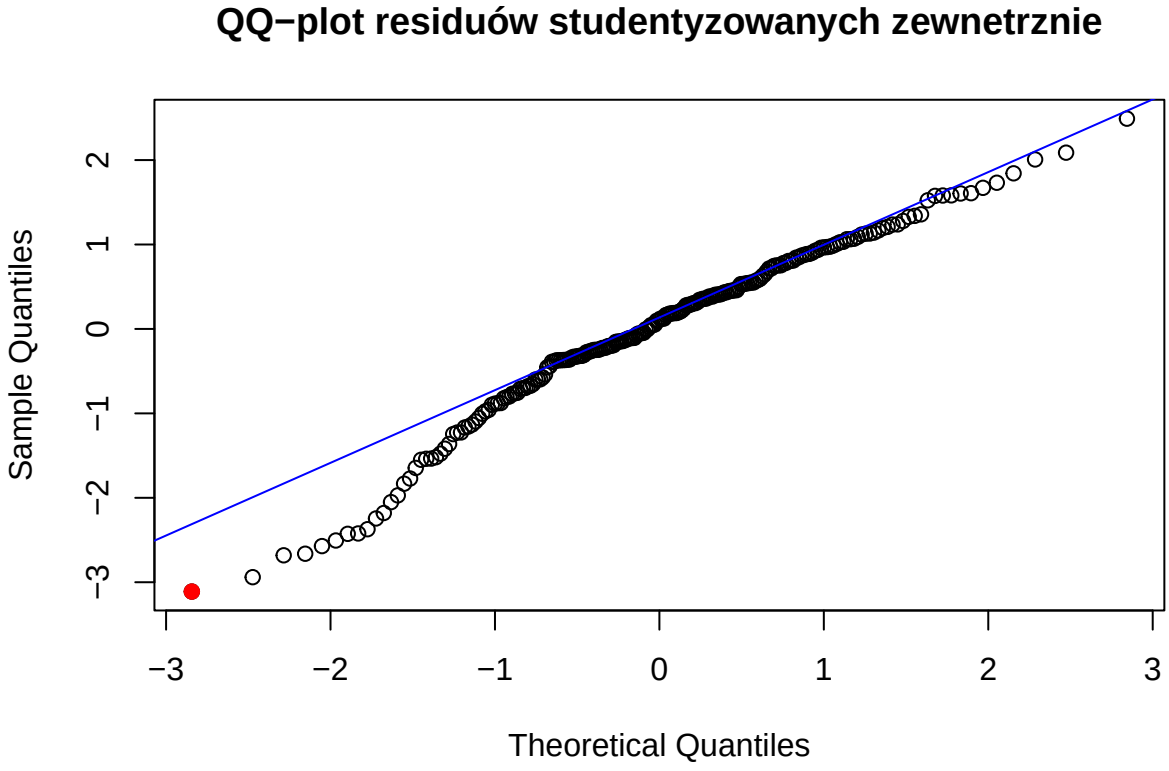
• jesli wykres ma struktrue liniową to (współczynnik kierunkowy $\neq$ 0) to zmienna wnosi dodatkową informację do modelu

• dodatkowo możemy wykrywać odstępstwa od założen modelu np. brak liniwoej relacji, obserwacje odstające, brak stałości wariancji itp.

Wykresy posiadają posiadają obserwacje odstajace, które są zaznaczone na wykresach. Poza tym mozemy zauważyć, że zmienna HSM ma strukturę liniową o niezerowym współczynniku, co sugeruje, że wnosi dodatkową informację o modelu.

zad 7b


QQ−plot residuów studentyzowanych wewnetrznie
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3


QQ−plot residuów studentyzowanych zewnetrznie
Sample Quantiles
Theoretical Quantiles
−3
−2
−1
0
1
2
3

Residua studentyzowane są postaci:

$$\tilde{e_i} = \frac{Y_i - \hat{Y_i}}{\sqrt{\hat{\sigma}^2(1 - H_{ii})}}$$

Residua studentyzwane wewnętrzenie to takie, gdzie model konstruownay jest przy użyci wsystkich obserwacji. $\tilde{e_i}$ studentyzowane wewnętrznie nie ma rozkładu Studenta

$$\tilde{e_i} = \frac{e_i / \sqrt{\hat{\sigma}^2(1 - H_{ii})}}{\sqrt{s^2/\sigma^2}}$$

Residua studentyzwane zewnętrzenie to takie, gdzie model konstruowany jest z pominięicem w dnaych wartości Y_i oraz wiesza macierzy planu stowarzyszonego z Y_i. Wyłączona zostaje $i - ta$ obserwacja. $\tilde{e_i}$ studentyzowane zewnętrznie mma rozkładu Studenta

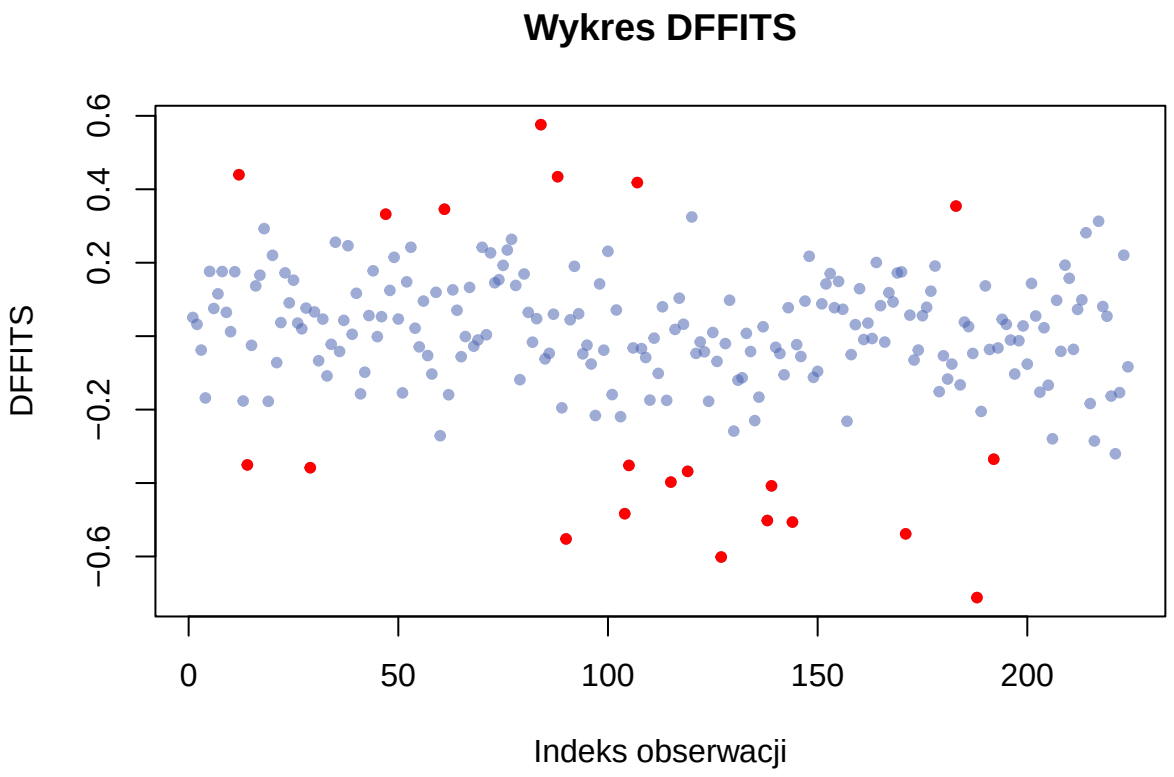
$$\tilde{e_i} = \frac{Y_i - \hat{Y_{(i)i}}}{\sqrt{\hat{s_i}^2(1 - H_{(i)ii})}}$$

Różnica między nimi polega sposobie konstrukcji oraz tym, czy zmienna $\tilde{e_i}$ ma rozkład Studneta.

Residua studentyzwane zewnętrzenie i wewnętrznie mogą informaować między innymi o obserwacjach odstających, obserwacjach znaczących i odstępstawach od założeń dostyczących błędów ε.

Wartości odstające, które wystepuja na wykresie są zaznaczone w kolorze czerwonym.

zad 7c



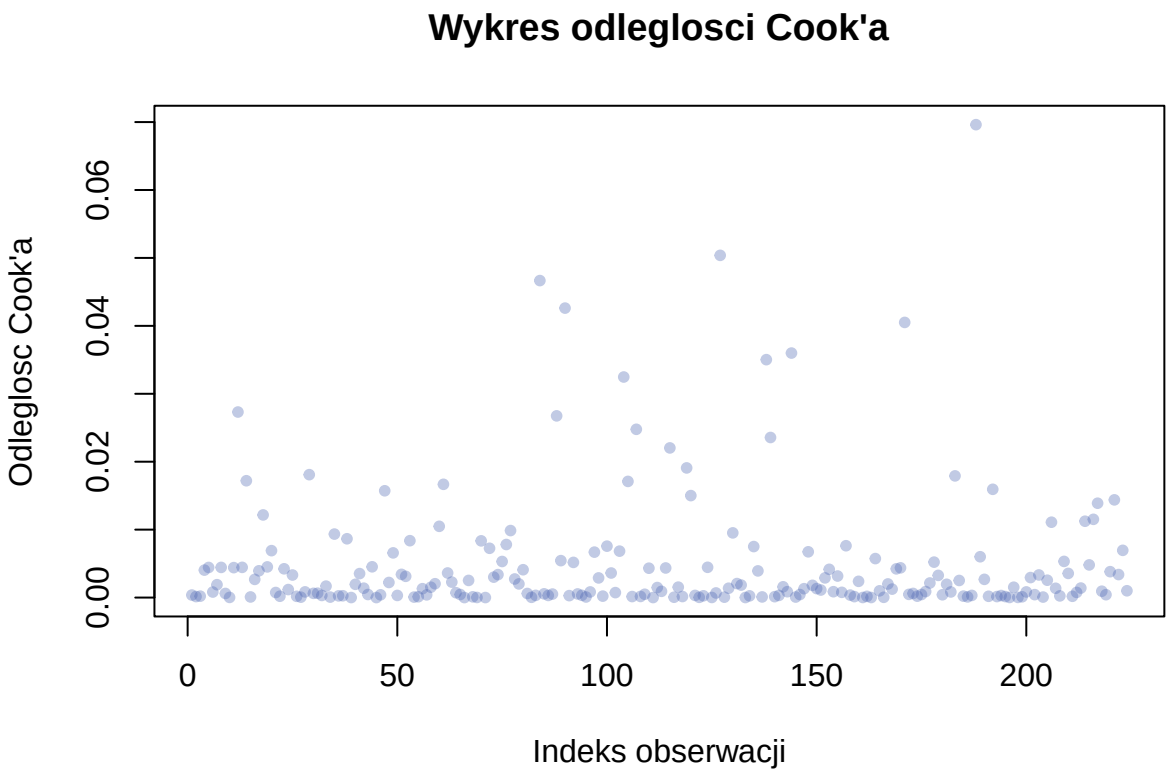
Miara $DFFITS$ dla $i-tej$ obserwacji służy do badania wpływu obserwacji Y_i na predykcję $\hat{Y}_i$.

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

Zawiera infromacje o tym, czy obserwacja Y_i ma znaczący wpływ na predykcję, jest tak gdy $\hat{Y}_i$ i $\hat{Y_{(i)i}}$ (bez $i-tej$ obserwacji) znacząco się różnią. Za taką znacząco róznicę uznaje się $|DFFITS_i| > 2\sqrt{p/n}$. Gdzie p to liczba regresorów, a n liczba obserwacji.

Możemy wywnioskować, że część obserwacjii (zaznaczonych na czerwono) ma dużą miarę $DFFITS$ i wymaga dokładniejszego zbadania.

zad 7d



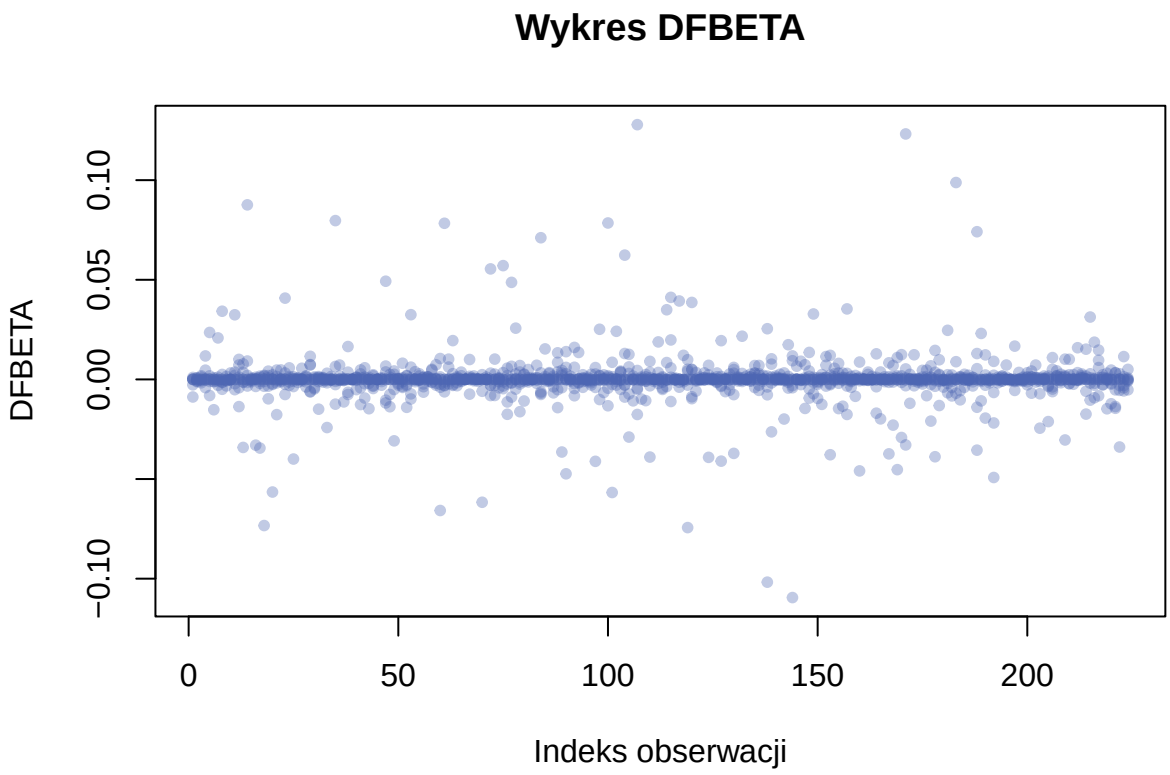
Odległość Cook'a dla $i-tej$ obserwacji służy do badania wpływu obserwacji Y_i na cały wektor predykcji $\hat{Y}$.

$$D_i = \sum_{j=1}^{n} \frac{(\hat{Y}_j - \hat{Y}_{(i)j})^2}{s^2 p}$$

Zawiera infromacje o tym, czy obserwacja Y_i ma znaczący wpływ na predykcję, jest tak gdy $\hat{Y}$ i $\hat{Y}_i$ (bez $i-tej$ obserwacji) znacząco się różnią. Za taką znacząco róznicę uznaje się $|D_i| > 1$.

Możemy wywnioskować, że predykcje Y przyjmują podobne wartości i żadna z obserwacji, nie wykazuje większego wpływu na predykcję. Jest to zgodne z naszymi oczekiwaniami.

zad 7e



Miara DFBETA dla $i-tej$ obserwacji służy do badania wpływu obserwacji Y_i na estymację parametru β_k.

$$DFBETA_k = \frac{\hat{\beta}_k - \hat{\beta}_{(i)k}}{s_{(i)}(\hat{\beta}_{(i)k})}$$

Zawiera infromacje o tym, czy $i-ta$ obserwacja ma znaczący wpływ na estymatory $\hat{\beta}_k$, $\hat{\beta}_{(i)k}$.

Możemy wywnioskować, że estymatory $\hat{\beta}_k$, $\hat{\beta}_{(i)k}$ przyjmują podobne wartości i żadna z obserwacji, nie wykazuje większego wpływu. Jest to zgodne z naszymi oczekiwaniami.

zad 7f

Tablica 8: Wyniki miary Tolerancja

HSM	HSS	HSE	SATM	SATV	SEX
0.5189	0.5088	0.543	0.5745	0.7311	0.7743

Tablica 9: Wyniki miary Tolerancja dla modelu z zadnia 6

SATM	SATV	SAT
0	0	0

Miara VIF służy do badania wielkości zjawiska multikolinearności. Dla $k-tej$ zmiennej obajśniajacej miara ta bada w jakim stopniu zmienna X_k objaśniana jest za pomocą pozostłych zmiennych objaśniających $X_1, \ldots, X_{k-1}, X_{k+1}, \ldots X_{p-1}$.

Jeśli VIF ma dużą wartośc to przekazuje infromacje o tym, że istnieje bardzo silna korelacja pomiędzy X_k i pewną kombinacją liniową pozostałych zmiennych objaśniających. Przyjmuje sie, że gdy $VIF_k > 10$ to istnieje duży problem w związku z występowaniem zjawiska multikolinearności.

Miara Tolerancja jest stosowana zamiennie z miarą VIF i jest to jej odwrtoność.

$$Tol_k = 1/VIF_k$$

W przypadku tej miary gdy $Tol_k < 0.1$ wskazuje to na problemy z multikolinearnością.

Wartośc stastytyki dla modelu w przypadku każdej zmiennej jest większa nież 0.1, więc nie spodziewamy się by model miał problem z multikolinearnością.

Wartość tej statystyki dla modelu z zadani 6 wynosi 0 dla każdej zmiennej objaśniającej, co wskazuje na to, że istnieje silna korelacja pomiędzy zmiennymi i kombinacją liniową pozostłych zmiennych objaśniających.

Jest to zgodne z oczekiwaniami, ponieważ w zadaniu 6 mieliśmy doczynienia ze zjawiskiem mulitikolinearności tworząc dodatkową zmienną SAT, tak by była kombinacją liniową $SATM$ i $SATV$.

zad 7g

Tablica 10: Dane do wyboru najlepszego modelu - kryterium BIC, Cp, R_adj

	BIC	Cp Mallows'a	R_adj	Intercept	HSM	HSS	HSE	SATM	SATV	SEX
1	-36.5252	2.8432	0.1869	1	1	0	0	0	0	0
1	-14.9084	25.4203	0.1045	1	0	1	0	0	0	0
1	-8.7133	32.3025	0.0794	1	0	0	1	0	0	0
2	-34.1856	1.8079	0.1943	1	1	0	1	0	0	0
2	-33.6550	2.3292	0.1924	1	1	1	0	0	0	0
2	-32.1225	3.8417	0.1869	1	1	0	0	1	0	0
3	-30.2848	2.3303	0.1961	1	1	0	1	1	0	0
3	-29.6223	2.9770	0.1937	1	1	1	1	0	0	0
3	-29.3585	3.2351	0.1928	1	1	1	0	1	0	0
4	-25.6678	3.5571	0.1953	1	1	1	1	1	0	0
4	-25.2299	3.9829	0.1937	1	1	0	1	1	1	0
4	-24.8811	4.3226	0.1925	1	1	0	1	1	0	1
5	-20.7435	5.0843	0.1934	1	1	1	1	1	1	0
5	-20.3360	5.4797	0.1919	1	1	1	1	1	0	1
5	-19.8256	5.9757	0.1901	1	1	0	1	1	1	1
6	-15.4189	7.0000	0.1900	1	1	1	1	1	1	1

Kryterium infromacyjne AIC i BIC są modyfikatorami metody największej wiarogodności, mierzą jak bardzo model jest dopasowany do danych uwzględniając złożoność modelu. Im mniejsza wartość staytyki AIC, BIC tym model jest lepiej dopasowany.

$$AIC(\tilde{\mathbb{X}}) = n \log\left(\frac{SSE(\tilde{\mathbb{X}})}{n}\right) + 2\tilde{p}$$

$$BIC = nlog\left(\frac{SSE(\tilde{\mathbb{X}})}{n}\right) + \log(n)\tilde{p}$$

Kyterium Cp Mallows'a jest modyfikatorem metody najmniejszych kwadratów i służy do oceny dopasowania modelu do danych, porównując dopasowanie danego modelu do modelu pełnego.

$$C_{\tilde{p}}(\tilde{\mathbb{X}}) = \frac{SSE(\tilde{\mathbb{X}})}{MSE(F)} - n + 2\tilde{p}$$

Model ma dobre własności gdy $C_{\tilde{p}} < \tilde{p}$ lub $2\tilde{p}$ lub ma najmniejszą wartość.

Wnioski: Najlepszym modelem przy użyciu kryterium BIC, Cp Mallows'a i $R^2(adj)$ jest model z interceptem, HSM i HSE. Ma najmniejszą wartość BIC, dobrą wartość dla kryterium Cp i jedną z większy dla $R^2(Adj)$.