Modele liniowe - raport 5

Łukasz Rębisz

04.02.2023

Celem raportu jest **porównanie** regresji liniowej **prostej** i **wielorakiej** na podstawie przykładowych danych.

Pierwszy zbiór danych zawiera informacje medyczne 46 pacjentów opisujące wiek, wskaźnik opisujący poważność przebytych chorób, poziom lęku oraz poziom satysfakcji.

Zadanie 1

Naszym zadaniem jest stowrzenie i przeanalizowanie modelu regresji liniowej wyjaśniającego wpływ wieku, przebytych chorób i poziomu lęku na poziom satysfakcji pacjentów.

Otrzymujemy następujący model liniowy:

- równanie regresji liniowej: Y = 1.053 -0.006 X_1 + 0.002 X_2 + 0.03 X_3, gdzie X_1 oznacza wiek, X_2 poważność przebytych chorób, X_3 poziom lęku. Otrzymane współczynniki wskazują, że największy wpływ na poziom satysfakcji ma poziom lęku.
- Współczynnik determinacji $R^2 = 0.542$, zmodyfikowany współczynnik wynosi z kolei $R^2_{adj} = 0.509$. Wartości równe ok. 0.5 wskazują na średnie dopasowanie modelu regresji liniowej wielorakiej o trzech zmiennych do badanych danych.

Zbadajmy następujący problem testowania:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0, \quad H_1 : \beta_1 \neq 0 \text{ lub } \beta_2 \neq 0 \text{ lub } \beta_3 \neq 0,$$

gdzie β_k oznacza współczynnik dla zmiennej X_k.

Przeprowadźmy **ogólny test F**:

$$F = \frac{MSM}{MSE} = \frac{SSM/dfM}{SSE/dfE}.$$

Przy prawdziwości H_0 $F \sim F(dfM, dfE)$, gdzie F oznacza rozkład Fishera-Snedecora.

```
F_test_full <- function(LinMod, alpha){
  dfM <- length(anova(LinMod)$'Sum Sq')-1
  dfE <- anova(LinMod)$'Df'[dfM+1]

  SSM <- sum(anova(LinMod)$'Sum Sq'[1:dfM])
  SSE <- anova(LinMod)$'Sum Sq'[dfM+1]

  F_test <- (SSM/dfM)/(SSE/dfE)
  q <- qf(alpha, dfM, dfE)
  return(c(F_test, q))
}
```

```
F_stat1 <- F_test_full(LinMod1, 0.05)
```

- Statystyka $F = 16.538 > 0.116 = F^{-1}(\alpha = 0.05, dfM = 3, dfE = 42)$, ($p$ - wartość $= 3.0431098 \times 10^{-7}$). zatem odrzucamy hipotezę zerową na poziomie istotności $\alpha = 0.05$. Zatem **odrzucamy** hipotezę mówiącą, że **żadna z badanych zmiennych nie ma wpływu** na poziom satysfakcji pacjentów.

Zadanie 2

Otrzymujemy następujące 95% przedziały ufności dla współczynników regresji:

- $\beta_1 \in$ [-0.0121, 4×10^{-4}], wartość 0 należy do przedziału ufności. Nie odrzucamy więc hipotezy mówiącej, że zmienna X_1 - wiek pacjentów nie ma wpływu na poziom satysfakcji.
- $\beta_2 \in$ [-0.0097, 0.0136], wartość 0 należy do przedziału ufności. Nie odrzucamy więc hipotezy mówiącej, że zmienna X_2 - poważność przebytych chorób nie ma wpływu na poziom satysfakcji.
- $\beta_3 \in$ [0.0115, 0.0488], wartość 0 NIE należy do przedziału ufności. Odrzucamy więc hipotezę mówiącą, że zmienna X_3 - poziom lęku pacjentów nie ma wpływu na poziom satysfakcji.

Zbadajmy istotność każdej ze zmiennych indywidualnie, wykonując **t-test** dla każdej ze zmiennych (testy analizowane na poprzednim raporcie):

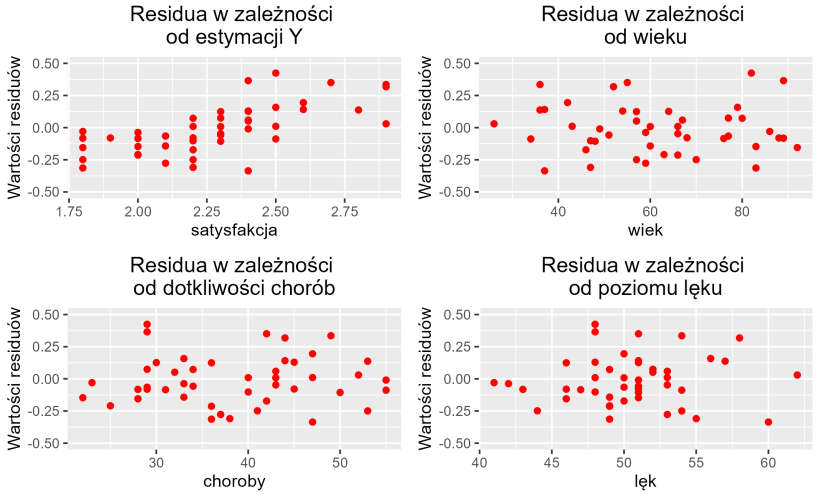
$$T = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \sim t(n-p) = t(42) \quad \text{(przy prawdziwości hipotezy zerowej)}.$$

- T_1 = -1.8972967, (p - wartość = 0.0647),
- T_2 = 0.3331876, (p - wartość = 0.7407),
- T_3 = 3.2568922, (p - wartość = 0.0022).

Kwantyl $t^{-1}(1 - \frac{0.05}{2}, 42) = 2.0181$, zatem odrzucamy brak wpływu jedynie zmiennej X_3 (na poziomie istotności $\alpha = 0.05$) na poziom satysfakcji pacjentów. Jest to wynik analogiczny do powyższej analizy przedziałów ufności dla współczynników regresji.

Zadanie 3

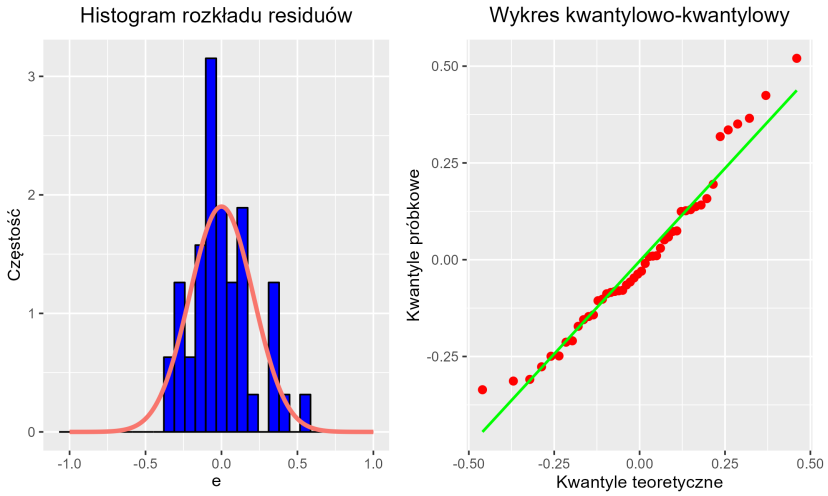
Zależność wartości residuów od zmiennej objaśnianej oraz każdej ze zmiennych objaśniających.



Analiza powyższych wykresów nie wskazuje na występowanie obserwacji odstających dla residuów. Rozkład wyników jest ponadto chaotyczny - nie obserwujemy, by residua układały się w niechaotyczne wzory. W przypadku zależności od wyników zmiennej objaśnianej występujące nad sobą wyniki są spowodowane małą liczbą obserwacji.

Zadanie 4

Normalność residuów:



Powyższy histogram wskazuje na w przybliżeniu normalny rozkład residuów. Jedynie wartość bliska średniej jest osiągana przez większą liczbę obserwacji, niż wynikałoby z rozkładu normalnego. Brak natomiast ciężkich ogonów rozkładu.

Wykres kwantylowo-kwantylowy wskazuje na dobre dopasowanie rozkładu wartości residuów do rozkładu normalnego. Jedynie skrajne wartości residuów wykazują nieznaczne odstępstwa od rozkładu normalnego.

Test Shapiro-Wilka:

P-wartość $= 0.1481172$. Wartość świadczy o tym, że nie mamy podstaw do odrzucenia hipotezy zerowej mówiącej, że rozkład danych jest normalny.

Wnioski:

Analiza danych dotyczących wpływu wieku, dotkliwości przebytych chorób i poziomu lęk na poziom satysfakcji pacjentów oparta na modelu regresji liniowej wielorakiej wskazała, że dane spełniają założenia modelu (analiza residuów) oraz jedyną zmienną mającą istotny wpływ na poziom satysfkacji pacjentów jest poziom lęku.

Przetestujmy następnie **dokładność** modelu regresji liniowej wielorakiej w zależności od **wyboru zmiennych objaśniających**. Analizie poddane różne modele badające następujące dane:

Dane dotyczą wyników studentów informatyki na jednej z amerykańskich uczelni, zmiennymi są:

GPA - odpowiednik średniej z 3 semestrów nauki,

HSM - odpowiednik średniej z ocen z matematyki z liceum (high school math),

HSS - odpowiednik średniej z ocen z przedmiotów ścisłych z liceum (high school science),

HSE - odpowiednik średniej z ocen z j. angielskiego (high school english),

SATM - ustandaryzowany test dla uczniów szkół średnich w USA z matematyki,

SATV - ustandaryzowany test dla uczniów szkół średnich w USA ze zdolności językowych (verbal),

GEN - płeć.

Zadanie 5

Porównajmy następujące modele:

i) model zredukowany: GPA~HSM+HSS+HSE,

ii) model pełny (poza zmienną binarną - płcią): GPA~SATM+SATV+HSM+HSS+HSE.

Porównanie błędów *SSE*

$SSE(R) = 107.75$, $SSE(F) = 106.82$

wskazuje na mniejszy błąd SSE w przypadku modelu zredukowanego. Jest to zgodne z przewidywaniami, ponieważ łączny błąd $SST = SSM + SSE$ nie zmienia się w przypadku redukcji modelu. W modelu pełnym większą rolę odgrywa SSM, stąd wartość SSE jest mniejsza.

Ogólny test *F*

Zbadajmy następujący problem testowy (w modelu o pięciu zmiennych):

$$H_0 : \beta_{SATM} = \beta_{SATV} = 0 \quad \text{vs} \quad H_1 : \beta_{SATM} \neq 0 \text{ lub } \beta_{SATV} \neq 0.$$

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

```
F_statistic <- function(lm_R, lm_F, alpha){
  SSE_R <- sum((summary(lm_R)$residuals)^2)
  SSE_F <- sum((summary(lm_F)$residuals)^2)

  dfE_R <- as.numeric(summary(lm_R)$fstatistic[3])
  dfE_F <- as.numeric(summary(lm_F)$fstatistic[3])

  MSE_F <- SSE_F/dfE_F

  F_st <- ((SSE_R-SSE_F)/(dfE_R-dfE_F))/MSE_F
  return(c(F_st,  qf(1-alpha, dfE_R-dfE_F, dfE_F)))
}

F_stat5 <- F_statistic(LinMod_5i, LinMod_5ii, 0.05)
```

Statystyka $F = 0.95 < 3.037 = F^{-1}(\alpha = 0.05, dfE(R) - dfE(F), dfE(F))$. Zatem Nie odrzucamy hipotezy zerowej na poziomie istotności $\alpha = 0.05$. Zatem **NIE odrzucamy** hipotezy mówiącej, że zmienne $SATM$ i $SATV$ nie mają wpływu na średnią ocen GPA.

Ogólny test *F* - funkcja *anova*

```
anova(LinMod_5i, LinMod_5ii)
```

```
## Analysis of Variance Table
##
## Model 1: GPA ~ HSM + HSS + HSE
## Model 2: GPA ~ SATM + SATV + HSM + HSS + HSE
##   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
```

Wyniki w powyższej tabeli wskazują na analogiczną wartość statystyki $F = 0.9503$, p-wartość równa 0.3882 powoduje brak odrzucenia hipotezy zerowej mówiącej, że pierwszy model (model zredukowany, o trzech zmiennych) lepiej oddaje charakter danych niż model o pięciu zmiennych. Jest to analogiczny wniosek do powyższego mówiącego, że dodatkowe zmienne (*SATM, SATV*) nie wnoszą istotnych nowych informacji do modelu opartego na trzech zmiennych (*HSM, HSS, HSE*).

Zadanie 6

Stwórzmy model regresji liniowej $GPA \sim SATM + SATV + HSM + HSE + HSS$ (w takiej kolejności zmiennych). Obliczmy sumy typu I i II:

Sumy I typu dodają po jednej zmiennej w każdym kroku, tzn.:

- Wpływ pierwszej zmiennej = $SSM(X_1)$.
- Wpływ drugiej zmiennej (po uwzględnieniu pierwszej) = $SSM(X_2|X_1)$.
- Wpływ trzeciej zmiennej (po uwzględnieniu pierwszej i drugiej) = $SSM(X_3|X_1, X_2)$.
- $SSM(X_4|X_1, X_2, X_3)$ itd.

Z kolei **sumy II typu** opisują, ile zmienności modelu objaśnia dana zmienna X_i po uwzględnieniu wpływu pozostałych zmiennych:

$$SSM(X_i|X_1, ..., X_{i-1}, X_{i+1}, ..., X_{p-1}).$$

Sumy typu I - funkcja *anova*

```
##            Sum Sq
## SATM        8.583
## SATV        0.001
## HSM        17.726
## HSE         1.891
## HSS         0.442
## Residuals 106.819
```

Sumy typu II - funkcja *Anova* z biblioteki *car*

```
##    var Sum_sq
## 1 SATM  0.928
## 2 SATV  0.233
## 3  HSM  6.772
## 4  HSE  0.957
## 5  HSS  0.442
```

a)

Sprawdźmy, że:

$$SSM(HSM|SATM, SATV) = SSM(SATM, SATV, HSM) - SSM(SATM, SATV).$$

W tym celu stwórzmy następujące modele liniowe:

i) GPA~SATM+SATV+HSM

ii) GPA~SATM+SATV.

Obliczmy wartości SSM dla tych modeli.

```
LinMod6i <- lm(GPA~SATM+SATV+HSM, data_2)
LinMod6ii <- lm(GPA~SATM+SATV, data_2)

y_hat_i=predict(LinMod6i)
SSM_i=sum((y_hat_i-mean(data_2$GPA))^2)

y_hat_ii=predict(LinMod6ii)
SSM_ii=sum((y_hat_ii-mean(data_2$GPA))^2)

results_6a <- SSM_i-SSM_ii
```

Różnica $SSM(SATM, SATV, HSM) - SSM(SATM, SATV) = 17.726 = SSM(HSM|SATM, SATV)$ (wartość w tabeli błędów typu I powyżej).

b)

Zauważmy, że zgodnie z definicjami sum I i II typu dla ostatniej zmiennej (HSS) obie wartości są równe (wpływ zmienności ostatniej zmiennej po uwzględnieniu zmienności wszystkich pozostałych zmiennych).

Zadanie 7

Naszym zadaniem jest stworzenie nowej zmiennej

$$SAT = SATM + SATV,$$

a następnie przeanalizowanie modelu regresji liniowej

$$GPA \sim SATM + SATV + SAT.$$

```
data_2$SAT <- data_2$SATM+data_2$SATV
LinMod7 <- lm(GPA~SAT+SATM+SATV, data_2)
summary(LinMod7)
```

```
##
## Call:
## lm(formula = GPA ~ SAT + SATM + SATV, data = data_2)
##
## 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 ***
## SAT         -2.456e-05  6.185e-04  -0.040 0.968357
## SATM         2.307e-03  1.097e-03   2.104 0.036486 *
## SATV               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 w przypadku, gdy jedna ze zmiennych jest **kombinacją liniową** pozostałych, nie są spełnione założenia modelu regresji liniowej. Macierz $X'X$ jest wówczas nieodwracalna (gdzie X oznacza macierz planu). W powyższych wynikach objawia się to wartościami **NA** dla ostatniej zmiennej objaśniającej ($SATV$).

Zadanie 8

Stwórzmy pełny model regresji liniowej

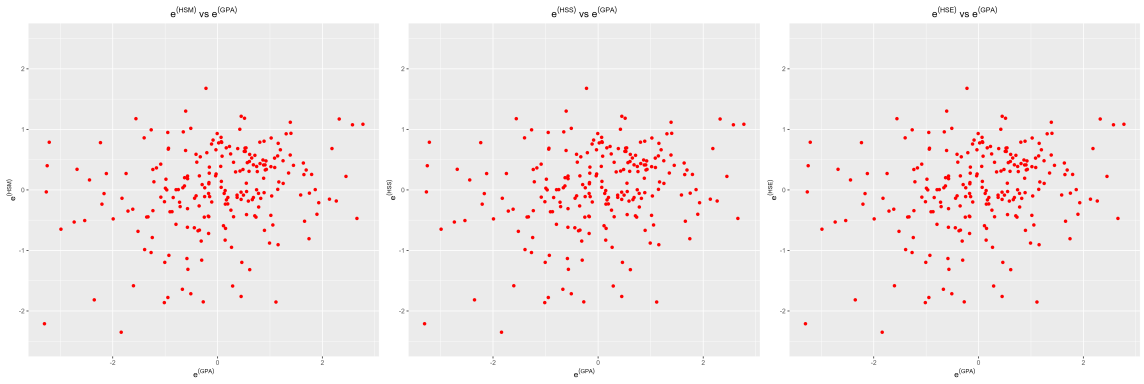
$$GPA \sim HSM + HSS + HSE + SATM + SATV + SEX.$$

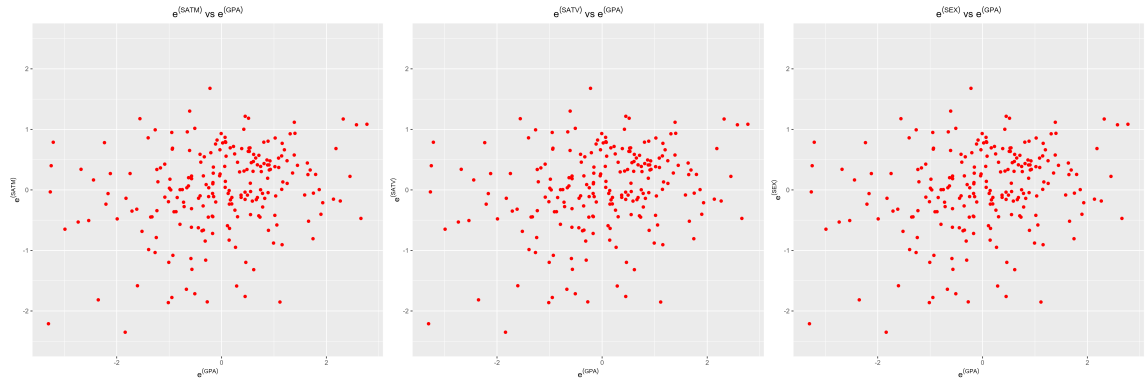
Przeanalizujmy *partial regression plots* - wykresy ukazujące wpływ, jaki wywiera dodanie nowej zmiennej objaśniającej $\tilde{X}_i$ do modelu, który już zawiera inne zmienne niezależne. W tym celu:

- Obliczmy wektor residuów $e^{(Y)}$ dla modelu liniowego, w którym zmienną objaśnianą jest Y, a zmiennymi objaśniającymi są wszystkie X-y oprócz X_i.
- Obliczmy wektor residuów $e^{(X_i)}$ dla modelu liniowego, w którym zmienną objaśnianą jest X_i, a zmiennymi objaśniającymi są wszystkie X-y oprócz X_i.
- Stwórzmy wykres rozrzutu $e^{(X_i)}$ vs $e^{(Y)}$.

Ponieważ z definicji wektor rozrzutu opisuje to, czego nie wyjaśniły zmienne objaśniające, zatem wykres rozrzutu $e^{(X_i)}$ vs $e^{(Y)}$ opisuje relację między X_i a Y po uwzględnieniu wpływu pozostałych X-ów.

```
LinMod_Y1 <- lm(GPA~HSM+HSS+HSE+SATM+SATV, data_2)
LinMod_X1 <- lm(SEX~HSM+HSS+HSE+SATM+SATV, data_2)
residuals_Y1 <- summary(LinMod_Y1)$residuals
residuals_X1 <- summary(LinMod_X1)$residuals
```





Na żadnym z powyższych wykresów nie obserwujemy żadnej wyraźnej struktury. Wskazuje to na fakt, że żadna pojedyncza zmienna nie wnosi do modelu istotnej informacji ponad to, co objaśniły pozostałe zmienne.

Nie obserwujemy też żadnych silnych odstępstw od założeń modelu, które objawiałyby się poprzez obserwacje odstające czy też brak stałości wariancji.

Zadanie 9

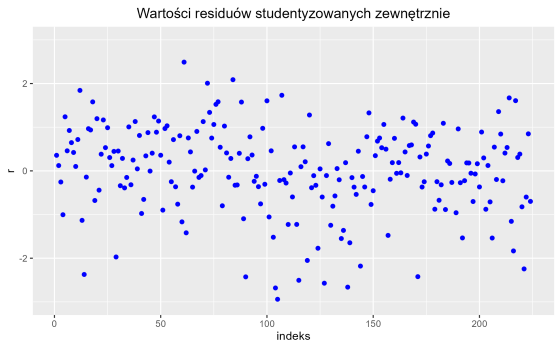
Naszym zadaniem jest zbadanie residuów studentyzowanych zewnętrznie:

$$\tilde{e}_i = \frac{Y_i - \hat{Y}_{(i)i}}{\sqrt{s^2_{(i)}(1 - H_{(i)i})}},$$

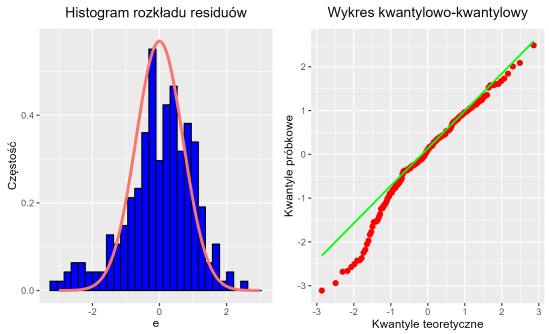
gdzie Y_i jest niezależną obserwacją dla modelu zbudowanego z pominięciem i-tej obserwacji, zatem licznik i mianownik są niezależne, więc

$$\tilde{e}_i \sim t(df = n - 1 - p).$$

Zbadajmy residua studentyzowane zewnętrznie:



Powyższy wykres wskazuje na występowanie istotnych obserwacji odstających dla wartości residuów studentyzowanych zewnętrznie osiągających wartości poniżej -2. Zauważmy, że takie wartości są przyjmowane stosunkowo często w porównaniu do przeciwnych, dużych wartości. W celu potwierdzenia tej hipotezy zbadajmy normalność rozkładu residuów.



Analiza powyższego histogramu i wykresu kwantylowo-kwantylowego wskazuje na to, że rozkład residuów studentyzowanych zewnętrznie jest w przybliżeniu normalny poza residuami, które osiągają skrajnie małe wartości (poniżej -2). Dla tych residuów występuje ciężki ogon rozkładu - liczba residuów o skrajnie małych wartościach NIE jest zgodna z rozkładem normalnym. Wartości odstające dla residuów wsazują na odstępstwa od założeń dotyczących błędów ϵ i występowaniu obderwacji odstających.

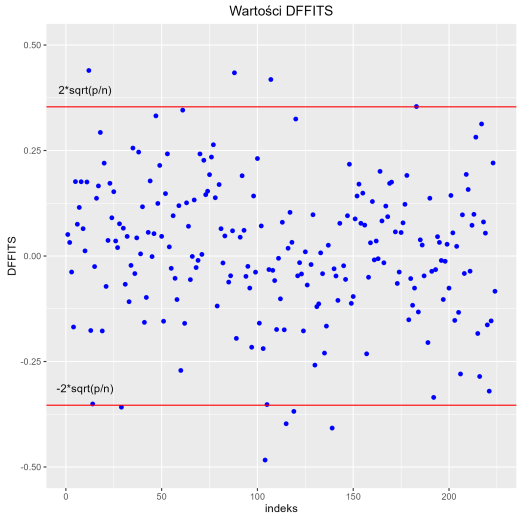
Zadanie 10

Miara *DFFITS* opisuje wpływ danej obserwacji na predykcję wektora odpowiedzi:

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

Powyższa definicja wskazuje, że miara ta dla i-tej obserwacji jest standaryzowaną różnicą pomiędzy predykcjami wartości Y_i uzyskanymi na podstawie dwóch modeli skonstruowanych na danych, odpowiednio z/bez obserwacji Y_i.

```
n <- length(LinMod_9$residuals)
p <- length(LinMod_9$coefficients)
D_dffits=dffits(LinMod_9)
nmb_out <- length(which(abs(D_dffits)>2*sqrt(p/n)))
```



Powyższy wykres obrazuje, które obserwacje są obserwacjami odstającymi/wpływowymi wg miary *DFFTIS* (oserwacje poza czerwonymi prostymi). Takich obserwacji jest 16. Indeksy obserwacji wskazują, które zmienne są wpływowe/odstające.

Zadanie 11

Tolerancja jest miarą badającą zjawisko *multikolinearności*, czyli występowanie silnej korelacji pomiędzy daną zmienną X_k a kombinacją liniową pozostałych zmiennych objaśniających.

$$Tol_k = 1 - R_k^2,$$

gdzie R_k^2 oznacza współczynnik determinacji dla modelu

$$X_k \sim X_1 + ... + X_{k-1} + X_{k+1}, ..., X_{p-1}.$$

Małe wartości współczynnika tolerancji (mniejsze niż 0.1) wskazują na silną zależność liniową pomiędzy zmienną X_k a kombinacją liniową pozostałych zmiennych.

zmienna	tolerancja
HSM	0.5188628
HSS	0.5088203
HSE	0.5429546
SATM	0.5745498
SATV	0.7310535
SEX	0.7742519

Wartości tolerancji powyżej 0.5 dla każdej ze zmiennych wskazują na to, że żadana ze zmiennych NIE jest silnie skorelowana z kombinacją liniową pozostałych zmiennych. Zatem nie występuje w tym przypadku problem zjawiska multikolinearności.

Zadanie 12

Wybierzmy najlepsze modele dla badanych danych wg kryteriów *BIC* i *AIC* będących modyfikacjami metody największej wiarogodności i są skonstruowane w taki sposób, by znaleźć balans pomiędzy dopasowaniem modelu do danych i nadmierną złożonością modelu. Karą za zbyt dużą liczbę zmiennych jest:

- $2p$ w przypadku metody *AIC*,
- $\log n \cdot p$ w przypadku metody *BIC*.

Kryterium BIC

```
data_2 <- read.table(file="csdata.txt",header = F, sep="")
colnames(data_2) <- c("id", "GPA", "HSM", "HSS", "HSE", "SATM", "SATV", "SEX")
allLM <- regsubsets(GPA~HSM+HSS+HSE+SATM+SATV+SEX, nbest=2, data_2[,2:ncol(data_2)])
all_s <- summary(allLM)
cbind(bic=all_s$bic, all_s$which)
```

```
##         bic (Intercept) HSM HSS HSE SATM SATV SEX
## 1 -36.52518           1   1   0   0    0    0   0
## 1 -14.90836           1   0   1   0    0    0   0
## 2 -34.18564           1   1   0   1    0    0   0
## 2 -33.65501           1   1   1   0    0    0   0
## 3 -30.28481           1   1   0   1    1    0   0
## 3 -29.62231           1   1   1   1    0    0   0
## 4 -25.66783           1   1   1   1    1    0   0
## 4 -25.22991           1   1   0   1    1    1   0
```

```
## 5 -20.74352           1   1   1   1    1    1   0
## 5 -20.33596           1   1   1   1    1    0   1
## 6 -15.41891           1   1   1   1    1    1   1
```

Powyższe wyniki wskazują, że najmniejsza (czyli wskazująca opytmalny model) wartość wskaźnika BIC jest osiągana dla modelu

$$GPA \sim HSM.$$

Kryterium AIC

```
AIC_step <- step(lm(GPA~1+HSM+HSS+HSE+SATM+SATV+SEX, data_2),
                 direction = 'backward')
```

```
Start:  AIC=-151.96
GPA ~ 1 + HSM + HSS + HSE + SATM + SATV + SEX

       Df Sum of Sq    RSS     AIC
- SEX   1    0.0415 106.82 -153.87
- SATV  1    0.2360 107.01 -153.47
- HSS   1    0.4801 107.26 -152.96
- HSE   1    0.7170 107.50 -152.46
<none>              106.78 -151.96
- SATM  1    0.9629 107.74 -151.95
- HSM   1    6.4635 113.24 -140.80
```

.

.

.

```
Step:  AIC=-157.08
GPA ~ HSM + HSE

       Df Sum of Sq    RSS     AIC
<none>              108.16 -157.08
- HSE   1    1.4936 109.65 -156.01
- HSM   1   15.9894 124.15 -128.20
```

Powyższe wyniki wskazują, że najniższa wartość kryterium AIC wskazująca optymalny model zostaje osiągnięta dla modelu

$$GPA \sim HSM + HSE.$$