Zadanie 3

Pierwszym krokiem będzie wygenerowanie macierzy $X_{100\times 2}$, której wiersze będą iid losowymi wektorami z rozkładu $N(0, \Sigma/100)$, gdzie:

$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

oraz wygenerowanie wektora $Y = \beta_1 X_1 + \epsilon$, gdzie $\beta_1 = 3$, X_1 jest pierwszą kolumną X, $\epsilon \sim N(0, I)$.

```
sigma = matrix(c(1, 0.9, 0.9, 1), 2, 2)
X = mvrnorm(100, c(0, 0), sigma/100)
B1 = 3
Y = B1 * X[,1] + rnorm(100)
```

Następnie, skonstruuję przedział ufności 95% dla β_1 dla modelu $Y = \beta_0 + \beta_1 X_1 + \epsilon$

```
base_model = lm(Y~X[,1])
confint(base_model)

                2.5 %    97.5 %
(Intercept) -0.1922508 0.2177333
X[, 1]       2.5006446 6.4853741
```

i dla modelu $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
full_model = lm(Y~X[,1] + X[,2])
confint(full_model)

                 2.5 %     97.5 %
(Intercept)  -0.1840347  0.2166482
X[, 1]        5.1988565 16.5184162
X[, 2]      -13.0361805 -1.1733241
```

Zauważając, że w przypadku obu modelu, 95% przedział ufności dla β_1 nie zawiera wartości 0, możemy stwierdzić, że dla testu:

$$H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$$

należy odrzucić hipotezę H_0. Oznacza to, że istnieje liniowa relacja pomiędzy X_1 a wektorem odpowiedzi Y.
Zarówno β_0 (w obu modelach) jak i β_2 zawierają w swoich przedziałach ufności wartość 0, więc nie możemy odrzucić hipotez o nieistotności interceptu oraz X_2.

Kolejnym krokiem będzie obliczenie odchylenia standardowego estymatora β_1 dla obu modeli.

Odchylenie dla modelu $Y = \beta_0 + \beta_1 X_1 + \epsilon$:

```
s2 = 1/(100-2)*sum(base_model$residuals^2)
v = sum((X[,1] - mean(X[,1]))^2)
sigma_base = sqrt(s2/v)
sigma_base

[1] 1.00398
```

Odchylenie dla modelu $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$:

```
s2 = 1/(100-3)*sum((Y-predict(full_model))^2)
v = s2 * solve(t(X)%*%X)
sigma_full = sqrt(v[2,2])
sigma_full

[1] 2.988212
```

Następnie, policzymy moce testów dla β_1 dla obu modeli.

```
delta = B1/sigma_base
tc = qt(1-0.05/2, 100-2)
power_base = pt(-tc, 100-2, delta) + 1 - pt(tc, 100-2, delta)
```

Otrzymujemy wartości 0.841 dla zredukowanego modelu, 0.168 dla pełnego modelu. Wysoka wartość mocy testu dla zredukowanego modelu wynika ze sposobu generowania wektora odpowiedzi Y. Moc testu dla β_1 w pełnym modelu jest znacznie mniejsza, co spowodowane jest uwzględnieniem X_2 w modelu.

Ostatnim krokiem w trzecim zadaniu będzie wygenerowanie 1000 niezależnych kopii wektora ϵ , a następnie 1000 związanych z ϵ wektorów odpowiedzi.
Dla każdego z wektorów odpowiedzi estymujemy β_1 oraz wykonujemy test istotności β_1 (dla obu modeli).

```
base_model_b1_rejections = 0
full_model_b1_rejections = 0
b1s_base = c()
b1s_full = c()

for (i in 1:1000){
  error = rnorm(100, 0, 1)
  Y = 3 * X[, 1] + error
  base_model = lm(Y~X[, 1])
  full_model = lm(Y~X[, 1] + X[, 2])
  interval_base = confint(base_model)
  interval_full = confint(full_model)
  if(interval_base[2,1] >= 0 | interval_base[2,2] <=0 ){
    base_model_b1_rejections = base_model_b1_rejections + 1
  }
  if(interval_full[2,1] >= 0 | interval_full[2,2] <=0 ){
    full_model_b1_rejections = full_model_b1_rejections + 1
  }
  b1s_base[i] = base_model$coefficients[2]
  b1s_full[i] = full_model$coefficients[2]
}
```

Porównamy estymację odchylenia standardowego $s(\beta_1)'$ z obliczoną wcześniej wartością $s(\beta_1)$:

	model zredukowany	**model pełen**
$s(\beta_1)'$	0.96784	2.816504
$s(\beta_1)$	1.00398	2.988212

Oraz estymację mocy testu (π') z otrzymanymi wcześniej wartościami (π):

	model zredukowany	**model pełen**
π'	0.862	0.174
π	0.841	0.168

Możemy zauważyć, że zarówno estymowane wartości odchylenia standardowego β_1, jak i estymowane wartości mocy testu dla β_1, są bardzo zbliżone do obliczonych wcześniej wartości teoretycznych.

Zadanie 4

Zaczniemy od wygenerowania macierzy $X_{1000\times 950}$, zawierającej elementy będące iid zmiennymi losowymi z rozkładu $N(0, \sigma = 0.1)$

```
X = matrix(rnorm(1000*950, 0, 0.1), 1000, 950)
```

Następnie, wyznaczymy wektor odpowiedzi $Y = X\beta + \epsilon$, dla $\beta = (3, 3, 3, 3, 3, 0, \dots)^T$.

```
B = rep(0, 950)
B[1:5] = 3
Y = X%*%B + rnorm(1000)
```

Korzystając z kolejno 1, 2, 5, 10, 50, 100, 500, 950 pierwszych kolumn macierzy planu, budować będziemy modele regresji liniowej. Będziemy chcieli znaleźć model, który będzie najlepiej dopasowany do danych. Obliczymy SSE, MSE, AIC, p-wartości odpowiadające pierwszym dwóm regresorom oraz liczbę zmiennych, które nie miały wpływu na wyznaczony wektor odpowiedzi Y, ale dla których hipoteza zerowa o braku istotności nie mogłaby zostać odrzucona na podstawie testu t Studenta (oznaczone jako FD).

```
k = c(1, 2, 5, 10, 50, 100, 500, 950)
SSE = rep(0, length(k))
MSE = rep(0, length(k))
AIC = rep(0, length(k))
p_vals = matrix(0, length(k), 2)
FD = rep(0, length(k))

for(i in 1:length(k)){
  model = lm(Y~X[,1:k[i]])
```

```
  SSE[i] = sum(model$residuals^2)
  MSE[i] = sum((model$fit - X%*%B)^2)
  AIC[i] = AIC(model)
  if (i==1){
    p_vals[i, 1] = summary(model)$coefficient[2,4]
  }
  else{
    p_vals[i,] = summary(model)$coefficient[2:3, 4]
  }
  if(k[i] > 5){
    FD[i] = sum(summary(model)$coefficient[7:k[i], 4] < 0.05)
  }
}
```

Liczba kolumn	1	2	5	10	50	100	500	950
SSE	1338	1263	999	992	944	910	517	60
MSE	391	288	6	13	61	95	488	945
AIC	3135	3079	2851	2854	2884	2947	3182	1942
FD	0	0	0	0	2	2	20	19

Wszystkie p-wartości dla dwóch pierwszych zmiennych wyniosły mniej niż 0.05.

Na podstawie kryterium AIC, możemy stwierdzić, że model zawierający wszystkie 950 kolumn jest najlepiej dopasowany do danych. Wraz ze wzrostem liczby kolumn, maleje wartość SSE, co oznacza, że wzrasta dopasowanie modelu do danych. Dla MSE, wartości początkowo maleją, aż do minimum równego 6, osiąganego dla modelu otrzymanego z pierwszych pięciu kolumn. Następnie, wartości MSE wzrastają aż do 945 dla modelu zbudowanego na 950 kolumnach. FD wzrasta wraz z dodawaniem kolejnych kolumn, za wyjątkiem ostatniego modelu, gdzie wartość FD spadła z 20 do 19.

W kolejnym kroku dokonamy identycznej analizy, ale dla modeli, które będą budowane na podstawie kolumn, których estymowane współczynniki regresji są największe. Musimy więc zbudować model pełen i posortować współczynniki w kolejności malejącej ze względu na ich wielkość:

```
model = lm(Y~X[,1:950])
best_vars_idxs = order(abs(summary(model)$coefficient[2:951]), decreasing=TRUE)
```

Ponownie budujemy osiem różnych modeli. Otrzymujemy tabelę:

Liczba kolumn	1	2	5	10	50	100	500	950
SSE	1338	1336	1334	1241	1102	950	322	60
MSE	391	391	394	284	243	203	683	945
AIC	3135	3136	3140	3077	3039	2991	2710	1942
FD	0	0	0	0	7	15	295	17

Ponownie, p-wartości dla dwóch pierwszych zmiennych wyniosły mniej niż 0.05 dla każdego z modeli.

Tym razem również, według kryterium AIC, model zawierający wszystkie kolumny jest modelem najlepiej dopasowanym do danych. SSE znów jest malejące. MSE nie jest monotoniczne, a minimum osiąga dla modelu zbudowanego ze 100 zmiennych. FD ponownie osiągnęło maksymalną wartość dla modelu zbudowanego z 500 zmiennych.

W kolejnych zadaniach analizować będziemy dane z pliku CH06PR15.txt, które zawierają kolumny z: wiekiem, drastycznością przebiegu choroby, poziomem lęku oraz poziomem zadowolenia.

Zadanie 5

Zaczniemy od zbudowania modelu regresji liniowej, gdzie zmiennymi objaśniającymi są wiek (X_1), drastyczność przebiegu choroby (X_2) oraz poziom lęku (X_3), a zmienną objaśnianą jest poziom zadowolenia.

```
model = lm(satisfaction ~ age + severity + anxiety, data)
summary(model)

Call:
lm(formula = satisfaction ~ age + severity + anxiety, data = data)

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 .
age         -0.005861   0.003089  -1.897  0.06468 .
severity     0.001928   0.005787   0.333  0.74065  
anxiety      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
```

Otrzymujemy równanie $Y = 1.053245 - 0.005861X_1 + 0.001928X_2 + 0.030148X_3$.
Wartość R^2 wyniosła 0.5415, zatem jedynie około 54% wariancji w Y jest wyjaśniona poprzez zmienne objaśniające.
Przetestujemy hipotezę:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs } H_1 : \exists_{i \in 1,2,3} \beta_i \neq 0$$

```
msm = sum((model$fit - mean(data$satisfaction))^2)/3
mse = sum(model$residuals^2) / 42
f_stat = msm/mse
f_stat
tc = qf(1-0.05, 3, 42)
tc
```

```
p_val = 1 - pf(f_stat, 3, 42)
```

Wartość statystyki F wyniosła 16.53756. F^* będące kwantylem rzędu 0.95 z rozkładu Fishera-Snedecora o 3, 42 stopniach swobody wyniosło 2.82. Jest ono o wiele mniejsze od statystyki testowej F, więc możemy odrzucić hipotezę zerową o braku wpływu pomiędzy każdym z regresorów, a wektorem odpowiedzi. Istotnie, p-wartość wyniosła $3.04311e-07$, co jest znacznie mniejsze od 0.05.

Zadanie 6

Korzystając z modelu zbudowanego w poprzednim zadaniu, podamy 95% przedziały ufności dla regresorów.

```
confint(model)[2:4,]

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

Następnie, sprawdzimy statystyki testowe oraz p-wartość dla testów każdego z regresorów.

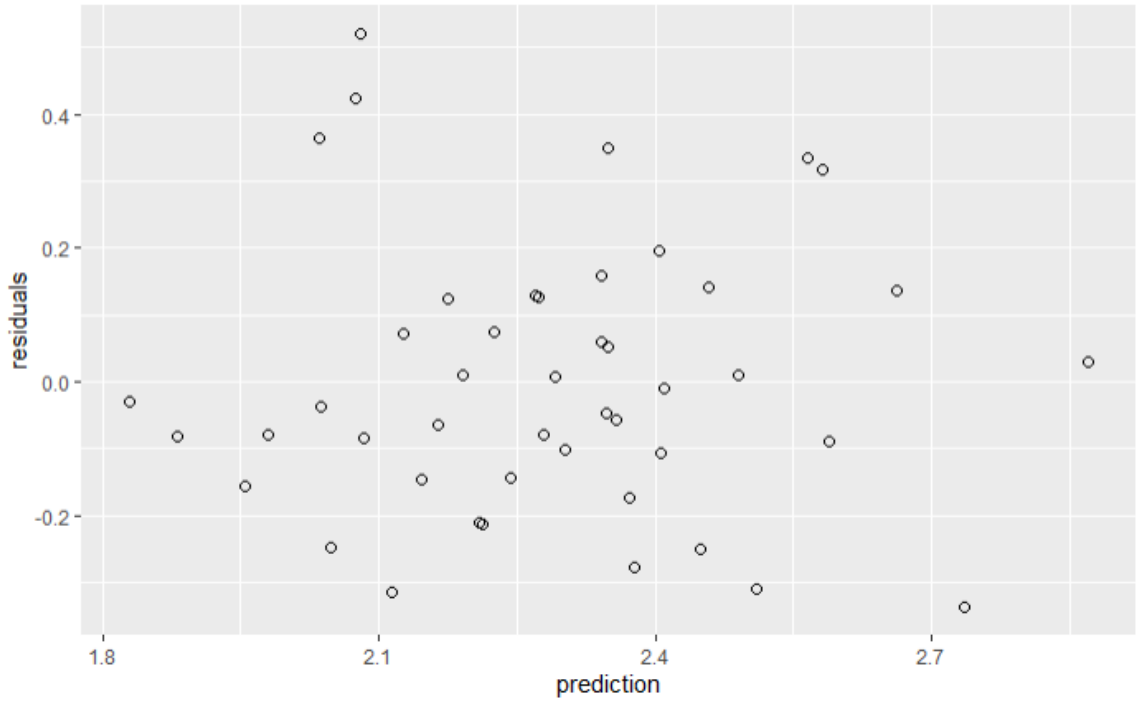
test	statystyka T	p-wartość
$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$	-1.897	0.065
$H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$	0.333	0.741
$H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$	3.257	0.002

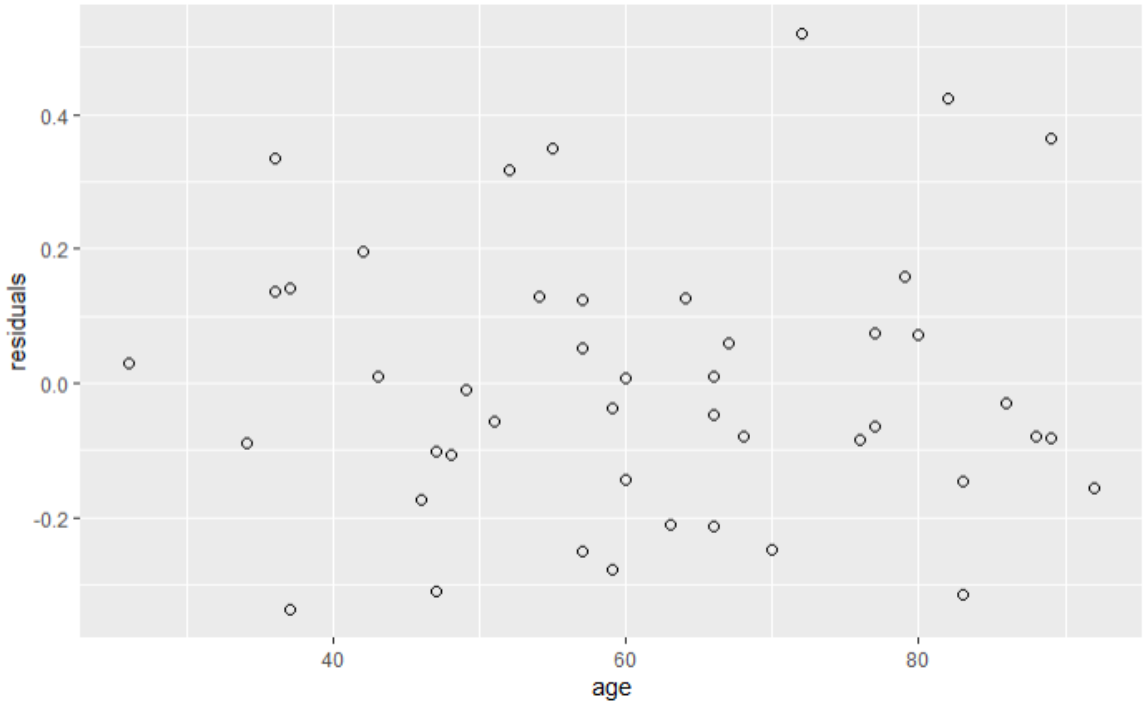
Możemy wywnioskować, że zależność liniowa pomiędzy zmienną objaśniającą a zmienną objaśnianą zachodzi wyłącznie w przypadku zmiennej *anxiety* opisującej poziom lęku. Potwierdzają to zarówno przedziały ufności - jedynie przedział ufności dla poziomu lęku nie zawierał 0 - oraz wykonane testy, w przypadku których jedynie test $H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$ dawał p-wartość mniejszą od 0.05.

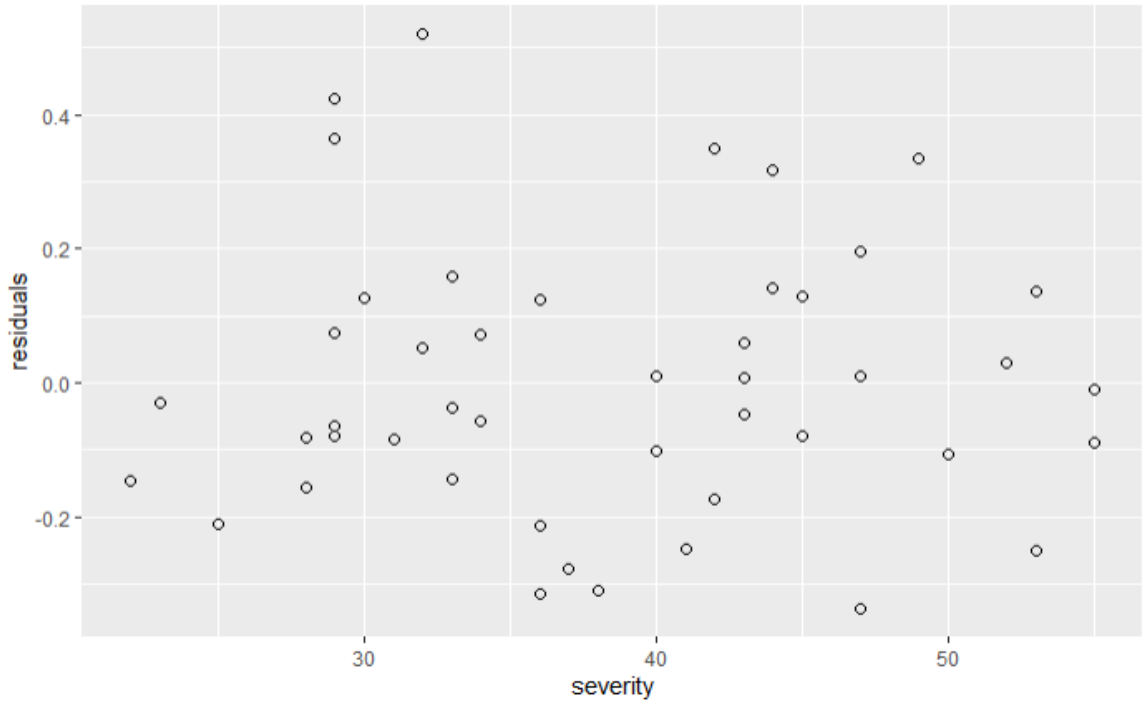
Zadanie 7

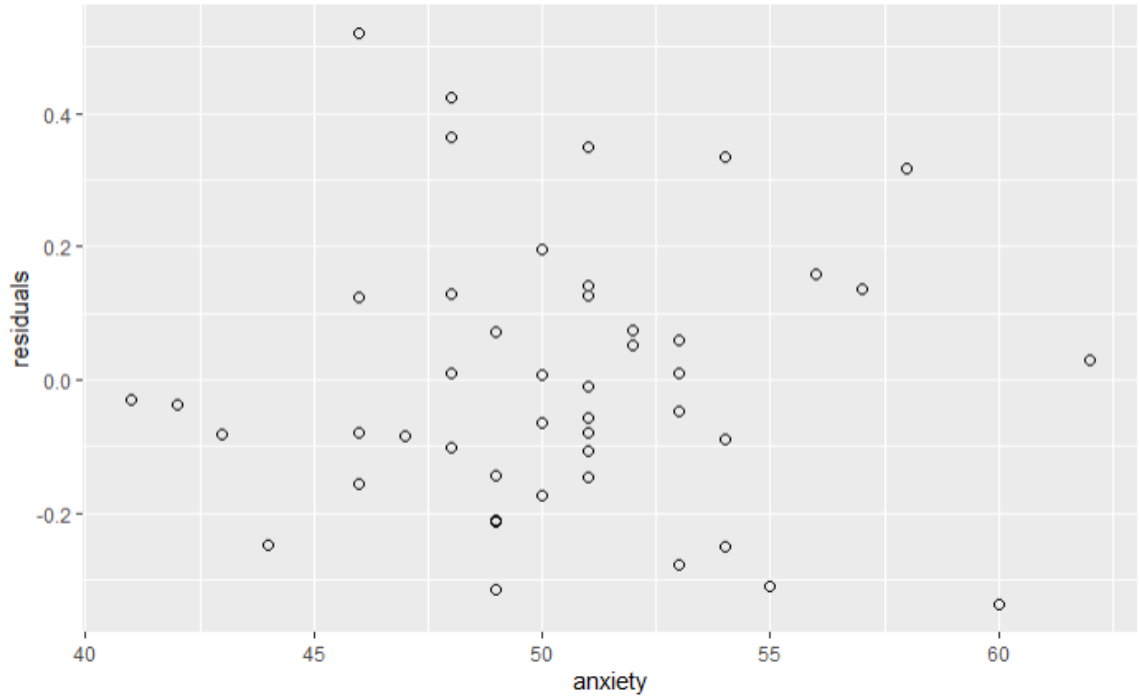
W tym zadaniu, wygenerujemy wykresy resiuduów vs satysfakcji oraz każdej ze zmiennych objaśniających.

```
data$residuals = model$residuals
data$prediction = model$fit
satisfaction_plot = ggplot(data) + geom_point(aes(prediction, residuals), shape
= 1, size = 2, )
age_plot = ggplot(data) + geom_point(aes(age, residuals), shape = 1, size = 2)
severity_plot = ggplot(data) + geom_point(aes(severity, residuals), shape = 1,
size = 2)
anxiety_plot = ggplot(data) + geom_point(aes(anxiety, residuals), shape = 1,
size = 2)
```









Możemy zauważyć brak zależności pomiędzy zmiennymi a wektorem residuów. Mają one strukturę losową, skupioną wokół 0, z niewielką liczbą obserwacji odstających.

Zadanie 8

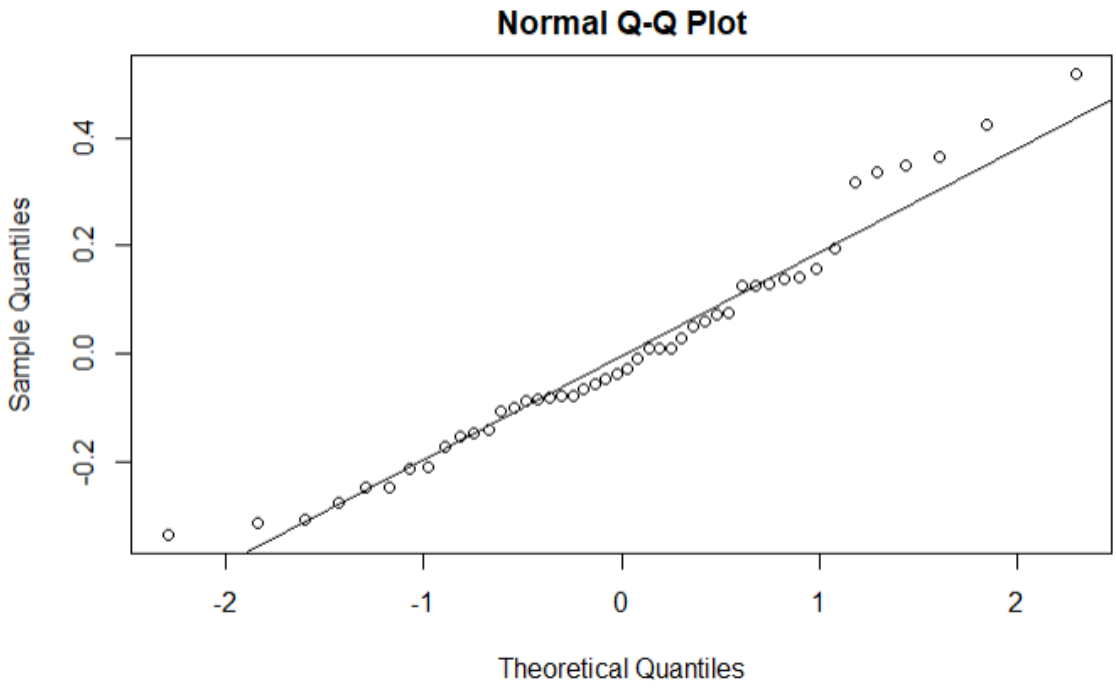
Ostatnią częścią analizy danych pacjentów, będzie wykonanie testu Shapiro-Wilka w celu ustalenia, czy residua pochodzą z rozkładu normalnego.

```
shapiro.test(model$residuals)


    Shapiro-Wilk normality test

data:  model$residuals
W = 0.96286, p-value = 0.1481
```

P-wartość wyniosła wartość 0.1481, więc nie możemy odrzucić hipotezy o normalności residuów.

W celu dalszego zbadania rozkładu residuów, możemy wykonać wykres kwantylowo-kwantylowy.

```
qqnorm(model$residuals)
qqline(model$residuals)
```



Zauważamy, że punkty układają się wokół prostej w sposób dość dobrze dopasowany, więc rozkład residuów jest bliski rozkładowi normalnego.

W kolejnych zadaniach przeanalizujemy dane pochodzące z pliki *csdata.dat* dotyczące studentów. Zawierają one kolumny: GPA, HSM, HSS, HSE, SATM, SATV, SEX.

Zadanie 9

Zbudujemy dwa modele regresji liniowej:

1) Model zredukowany - przewidywanie GPA na podstawie HSM, HSS i HSE.
2) Model pełen - przewidywanie GPA na podstawie HSM, HSS, HSE, SATM, SATV, SEX.

```
reduced_model = lm(GPA ~ HSM + HSS + HSE, data)
full_model = lm(GPA ~ HSM + HSS + HSE + SATM + SATV, data)
```

Przetestujemy hipotezę:

$$H_0 : \beta_{SATM} = \beta_{SATV} = 0 \text{ vs } H_1 : \beta_{SATM} \neq 0 \vee \beta_{SATV} \neq 0$$

```
sse_r = sum(reduced_model$residuals^2)
sse_f = sum(full_model$residuals^2)
dfe_r = length(data$id) - 4
dfe_f = length(data$id) - 6
mse_f = sse_f / dfe_f

F_stat = (sse_r - sse_f) / (dfe_r - dfe_f) / mse_f
F_stat
fc = qf(1-0.05, 2, dfe_f)
fc
```

Wyznaczyliśmy statystykę testową F, która wyniosła 0.95. Wartość F^* będącą kwantylem rzędu 0.95 ze stopniami swobody $2, 218$, wyniosła 3.04, co jest znacznie większe od F, więc nie możemy odrzucić hipotezy zerowej o braku istotności obu regresorów: SATM i SATV.

Porównamy otrzymane wyniki z wynikami funkcji *anova*.

```
anova(reduced_model, full_model)


Analysis of Variance Table

Model 1: GPA ~ HSM + HSS + HSE
Model 2: GPA ~ 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
```

Statystyka testowa F obliczona przez funkcje *anova* jest taka sama, jak wartość wyznaczona wcześniej. P-wartość wynosi 0.3882, więc ponownie - nie możemy odrzucić hipotezy zerowej o braku istotności regresorów SATM i SATV. Stopnie swobody wynoszą $2, 218$.

Zadanie 10

Zbudujemy model przewidujący GPA na podstawie kolejno: SATM, SATV, HSM, HSE, HSS. Następnie, korzystając z funkcji *anova* i *Anova*, obliczymy sumy typu I i II.

```
model = lm(GPA ~ SATM + SATV + HSM + HSE + HSS, data)
anova(model)
Anova(model)
```

Zmienna	Suma typu I	Suma typu II
SATM	8.583	0.928
SATV	0.001	0.233
HSM	17.726	6.772
HSE	1.891	0.957
HSS	0.442	0.422

Powyższą wartość sumy typu I dla zmiennej HSM zweryfikujemy obliczając różnicę SSM dwóch modeli:
1) GPA przewidywane przez SATM, SATV, HSM
2) GPA przewidywane przez SATM, SATV

```
sat_hsm_model = lm(GPA ~ SATM + SATV + HSM, data)
sat_model = lm(GPA ~ SATM + SATV, data)

hsm_ssm = sum((mean(data$GPA) - sat_hsm_model$fit)^2) - sum((mean(data$GPA)-
sat_model$fit)^2)
hsm_ssm
```

Otrzymujemy wartość 17.72647, czyli taką samą, jaką otrzymaliśmy przy pomocy funkcji *anova*.

Analizując tabelę sum typu I i II możemy zauważyć, że dla zmiennej HSS oba typy sum są takie same. Wynika to wprost z definicji, oba typy sum dla HSS zdefiniowane są jako $\text{SSM}(HSS|SATM, SATV, HSM, HSE)$.

Zadanie 11

Wyznaczmy nową zmienną SAT, będącą sumą zmiennych SATM i SATV, a następnie zbudujmy model przewidujący GPA na podstawie SAT, SATM i SATV.

```
data$SAT = data$SATM + data$SATV
model = lm(GPA ~ SATM + SATV + SAT, data)
summary(model)

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

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
```

Możemy zauważyć, że zmienna SAT nie wprowadza żadnych nowych informacji do modelu. Wynika to z faktu, że jest ona kombinacją liniową pozostałych zmiennych. Zmieniając kolejność regresorów w wywołaniu funkcji *lm* na SAT, SATM, SATV, możemy zauważyć podobne zjawisko:

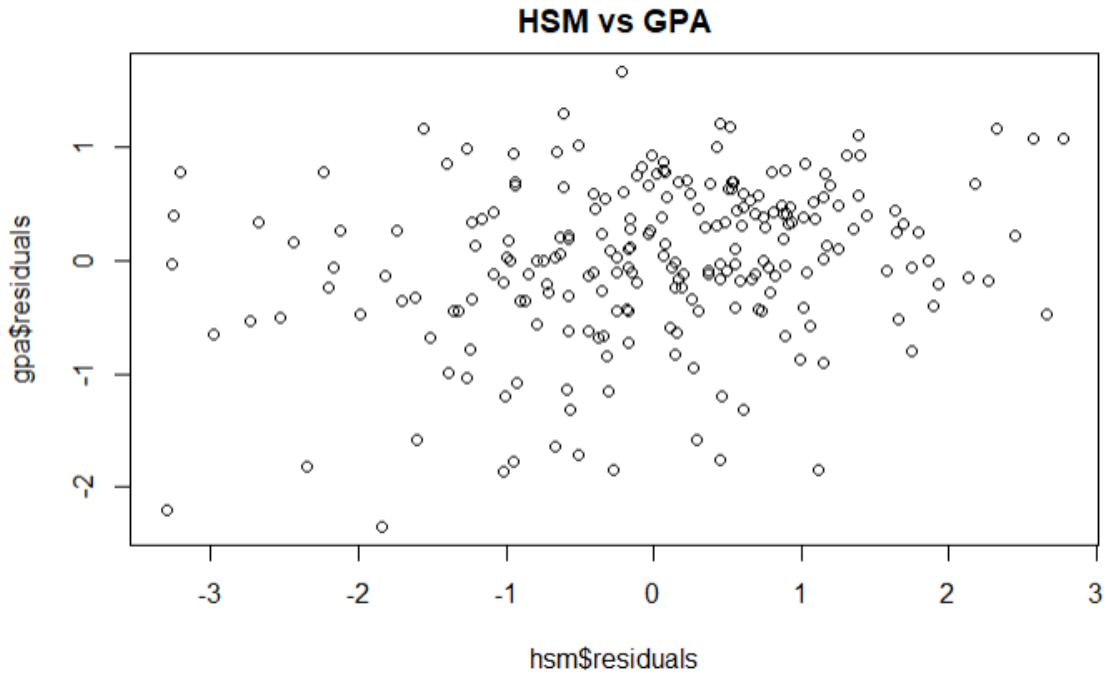
```
model = lm(GPA ~ SAT + SATM + SATV, data)
summary(model)

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

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
```

Tym razem, zmienna SATV nie wnosi żadnych nowych informacji do modelu zawierającego zmienne SAT i SATM. Ponownie, wynika to z faktu, że jest ona kombinacją liniową zmiennych SAT i SATM ($SATV = SAT - SATM$).

Zadanie 12

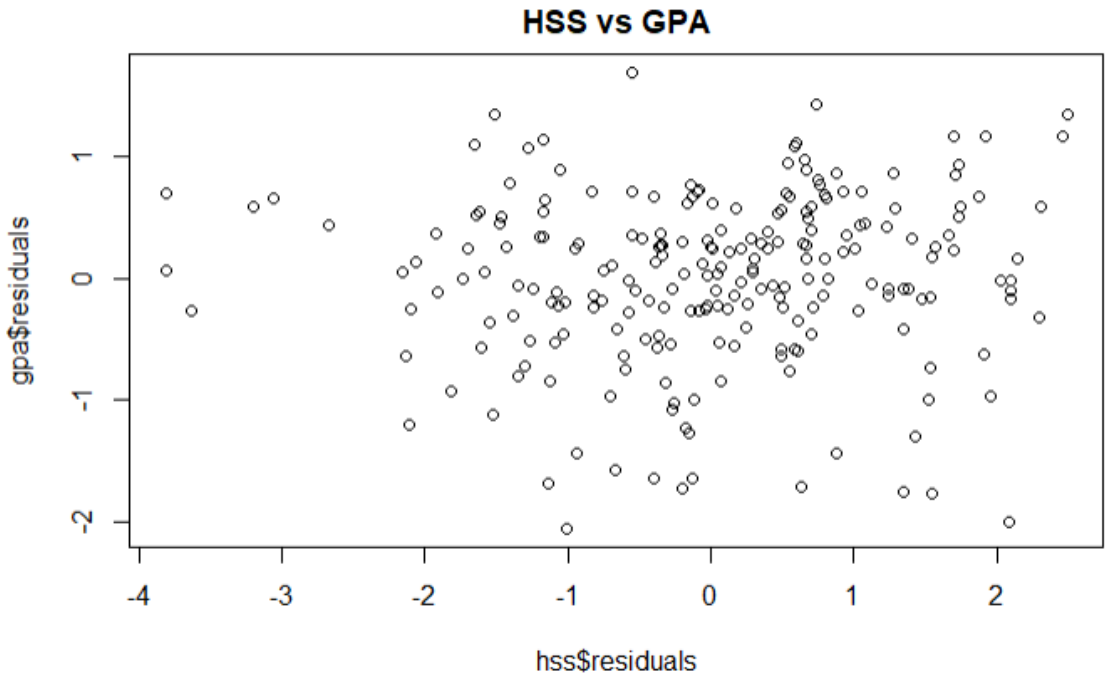
Zbudujemy model dla zmiennej objaśnianej GPA oraz zmiennych objaśniających HSM, HSS, HSE, SATM, SATV, SEX oraz przeanalizujemy **partial regression plots**. Są one używane do przeanalizowania wpływu dodania nowej zmiennej X_j do modelu, który zawiera już inne zmienne niezależne. Metoda ta polega na obliczeniu wektora residuów $e^{(Y)}$ dla modelu, w którym zmienną objaśnianą jest Y, a zmiennymi objaśnianymi są wszystkie X-y poza X_j oraz wektora residuów $e^{(X_j)}$ dla modelu, w którym zmienną objaśnianą jest X_j, a zmiennymi objaśniającymi są wszystkie X-y poza X_j. Następnie wykonuje się wykres $e^{(X_j)}$ vs $e^{(Y)}$. Brak wyraźnej struktury na wykresie wskazuje na brak nowych, istotnych informacji wnoszonych przez X_j do modelu. Natomiast relacja liniowa pomiędzy residuami wskazuje na fakt, że X_j wnosi dodatkową informację do modelu.

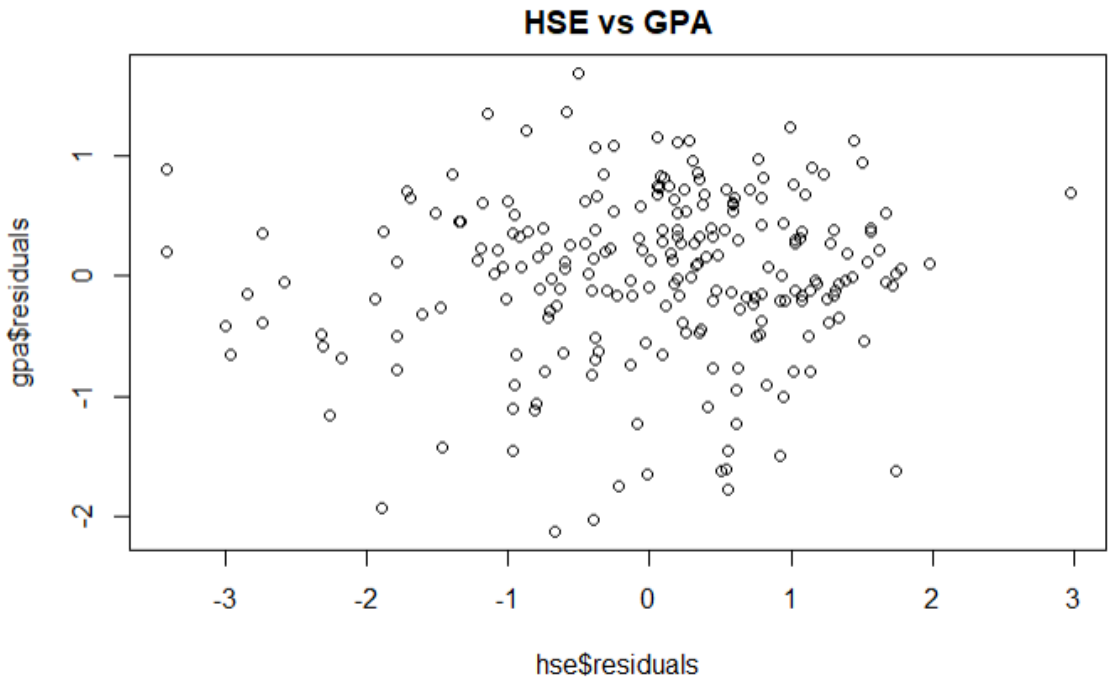
```r
# HSM
gpa = lm(GPA ~ HSS + HSE + SATM + SATV + SEX, data)
hsm = lm(HSM ~ HSS + HSE + SATM + SATV + SEX, data)
plot(hsm$residuals, gpa$residuals, main="HSM vs GPA")
```

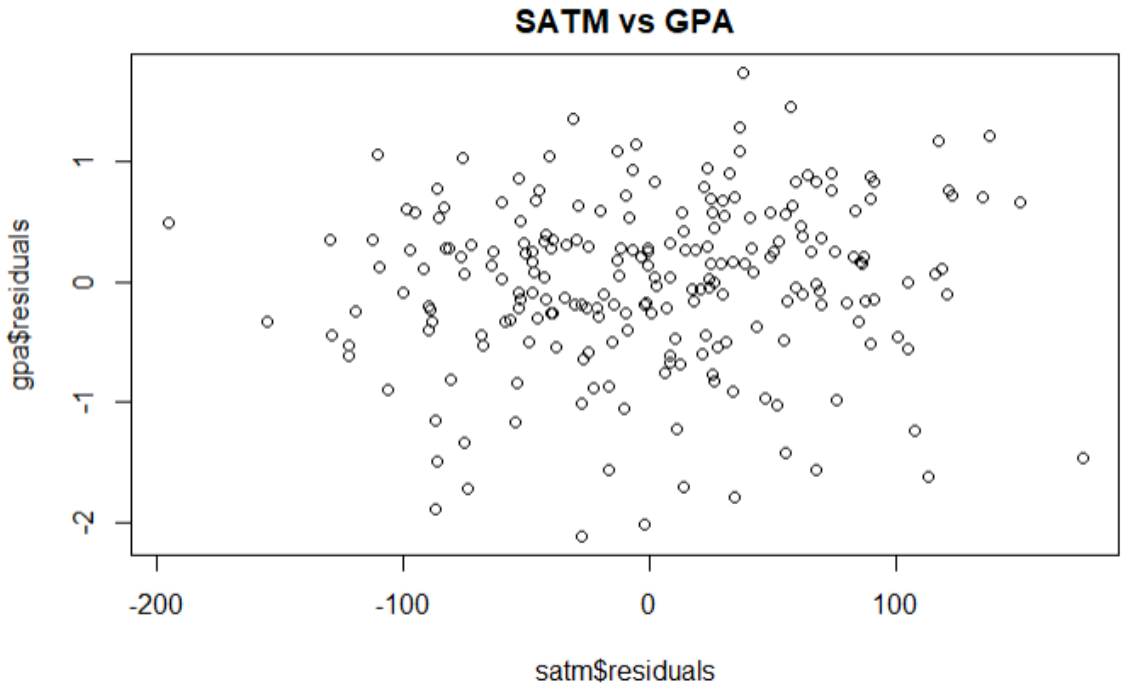


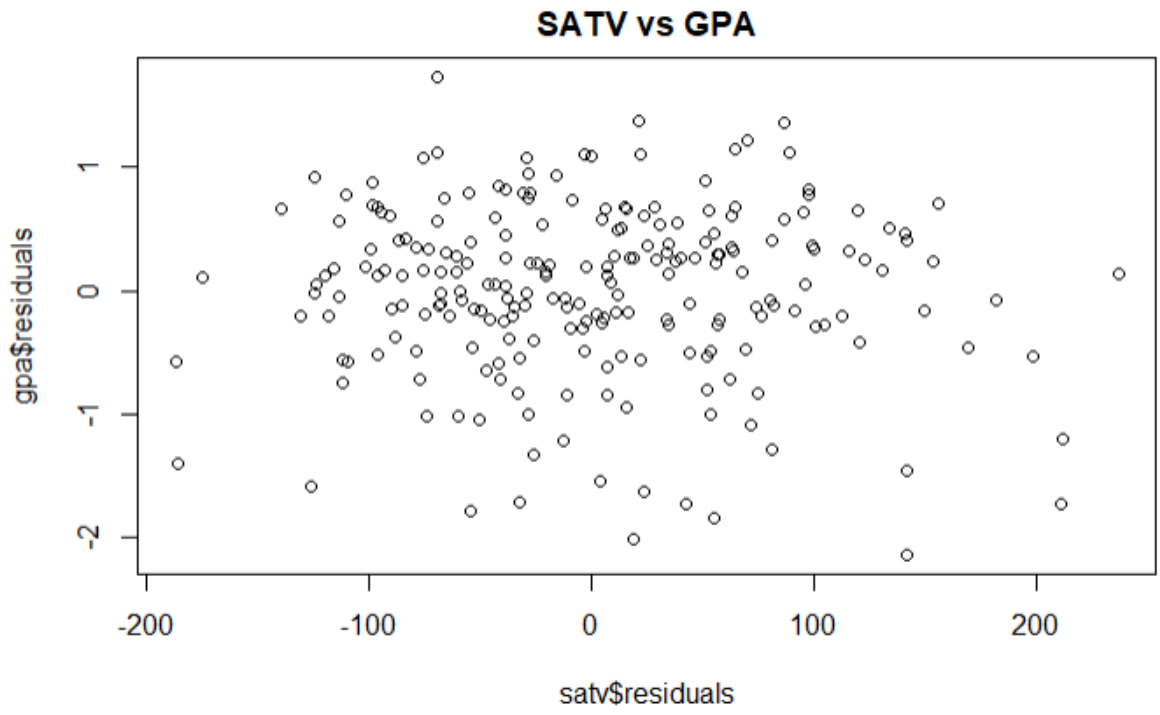
Obserwacje są skupione wokół punktu $(0, 0)$. Dane nie mają struktury losowej, relacja liniowa występuje, ale jest bardzo mało wyraźna. Może to sugerować, że zmienna HSM dostarcza istotnych informacji do modelu zawierającego pozostałe zmienne.

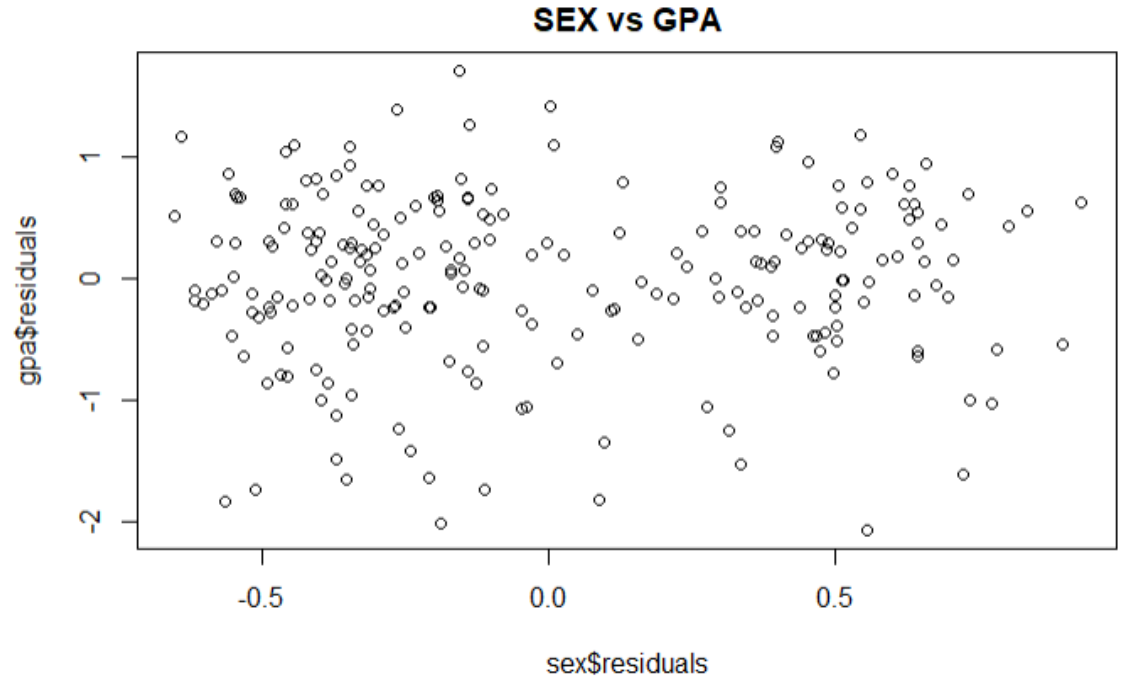
W analogiczny sposób wyznaczymy **partial regression plots** dla pozostałych zmiennych.




HSE vs GPA
gpa$residuals
hse$residuals
-3
-2
-1
0
1
2
3
-2
-1
0
1


SATM vs GPA
gpa$residuals
satm$residuals
-200
-100
0
100
-2
-1
0
1



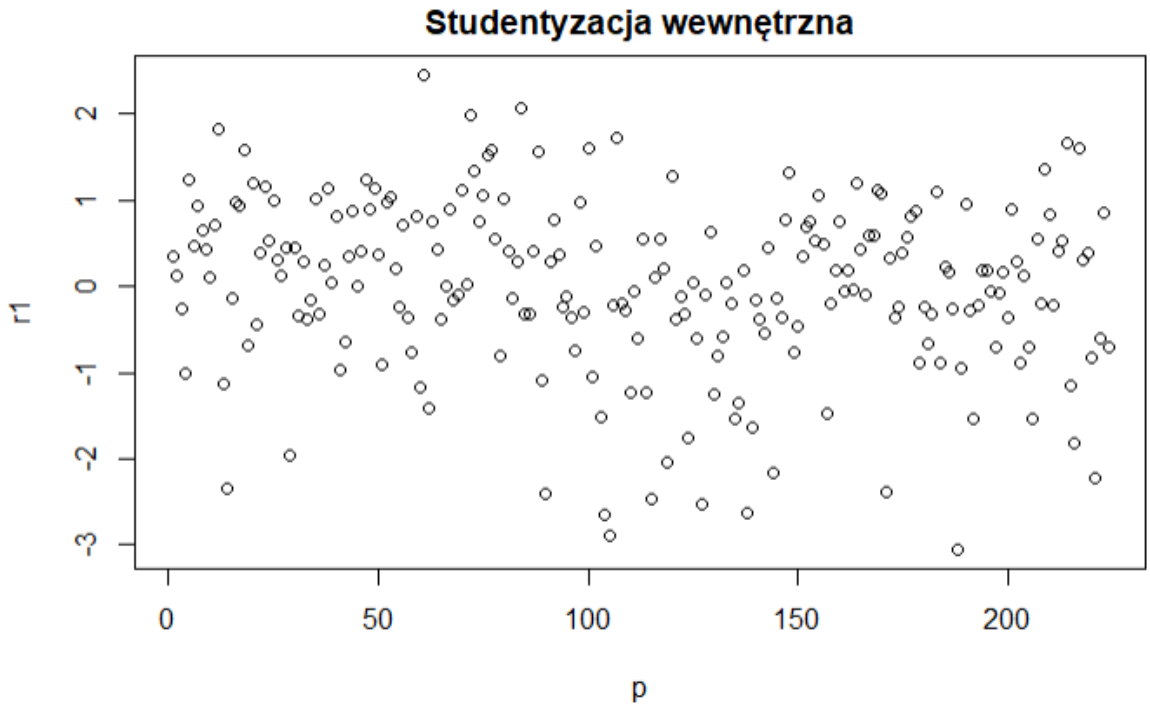


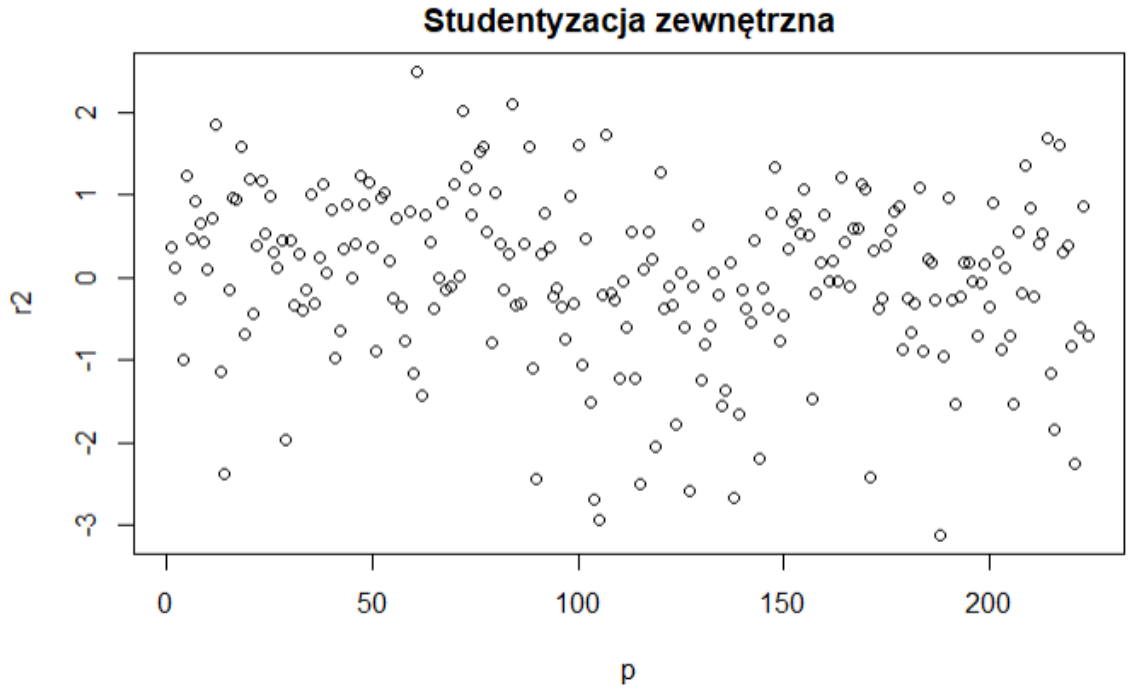
Pozostałe wykresy cechują się większą losowością w strukturze. Nie występuje relacja liniowa pomiędzy wektorami residuów. Może to sugerować, że pojedyncze zmienne, oprócz HSM, nie wnoszą żadnych istotnych informacji do modelu zawierającego pozostałe zmienne.

Zadanie 13

Przeanalizujemy studentyzowane residua - wewnętrznie i zewnętrznie.

```r
model = lm(GPA~HSM + HSS + HSE + SATM + SATV + SEX, data)
p = 1:224
r  = residuals(model)
r1 = rstandard(model) # studentyzacja wewnętrzna
r2 = rstudent(model)  # studentyzacja zewnętrzna
cbind(r, r1, r2)
plot(p, r1, main="Studentyzacja wewnętrzna")
plot(p, r2, main="Studentyzacja zewnętrzna")
```





Residua oscylują wokół zera ze stałą wariancją. Nie zauważamy widocznych obserwacji odstających.

Zadanie 14

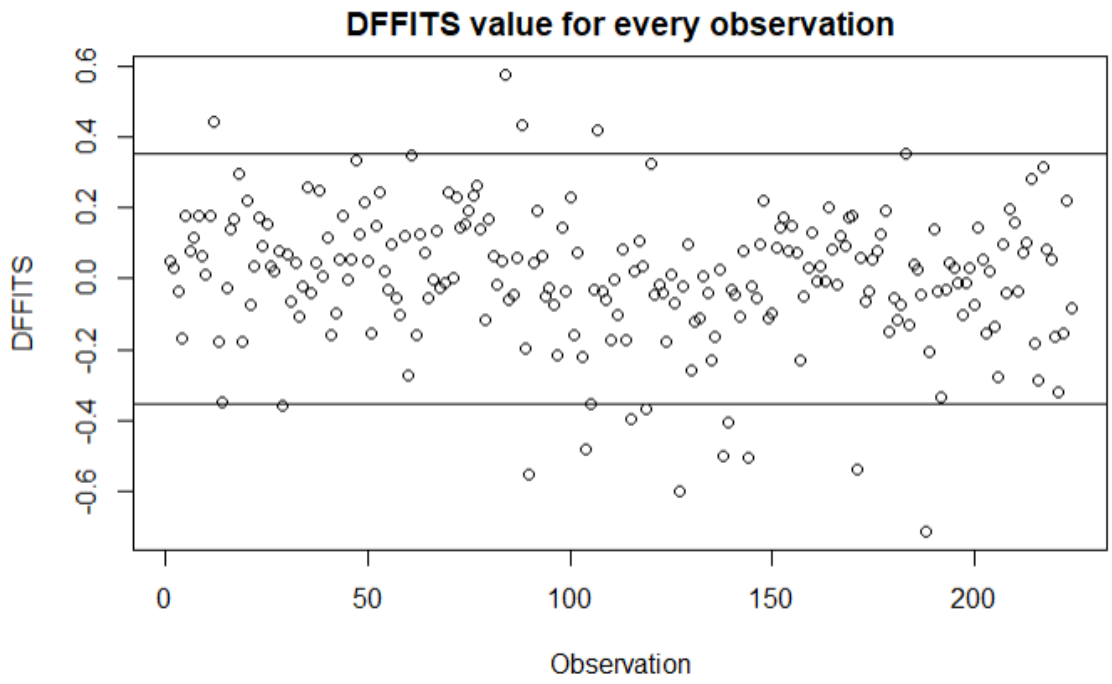
W tym zadaniu zajmiemy się miarą DFFITS. Jest ona dana wzorem:

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

co jest standaryzowaną różnicą pomiędzy predykcjami wartości Y_i uzyskanymi przez dwa modele: zbudowanym na danych zawierających i-tą obserwację, oraz na modelu niezawierającym i-tej obserwacji. DFFITS mierzy więc wpływ i-tej obserwacji na i-tą predykcję. Obserwacje mające wartość DFFITS przewyższającą $|s\sqrt{p/n}|$ należy poddać dokładniejszej analizie. Idealny zbiór danych służący do konstrukcji modelu liniowego powinien zawierać jak najmniej takich obserwacji.

Wykonamy analizę DFFITS dla modelu ze zmienną objaśnianą GPA i zmiennymi objaśniającymi HSM, HSS, HSE, SATM, SATV, SEX.

```
Observation = 1:224
DFFITS = dffits(model)
threshold = 2*sqrt(7/224)
plot(Observation, DFFITS, main="DFFITS value for every observation")
abline(h=threshold)
abline(h=-threshold)
```



Możemy zauważyć, że znaczna większość obserwacji (16 na 224, czyli 7.1%) nie przekracza progu $|s\sqrt{p/n}|$, co sugeruje, że zbiór danych jest odpowiednio zbalansowany.

Zadanie 15

Przeanalizujemy miarę tolerancji w rozpatrywanym modelu. Służy ona do badania wielkości zjawiska multikolinearności. Tolerancja mierzy, w jaki sposób zmienna X_k jest objaśniana przez pozostałe zmienne objaśniające. Na problemy z multikolinearnością wskazują wartości tolerancji mniejsze od 0.1.

```
vif_val = vif(model)
tolerance = 1/vif_val
tolerance
```

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

Tolerancja dla każdej ze zmiennych jest znacznie większa od 0.1, co sugeruje, że w naszych danych nie występuje problem multikolinearności.

Zadanie 16

W ostatnim zadaniu wybierzemy najlepszy model dla naszych danych na podstawie miar BIC i AIC (znajdziemy modele osiągające najmniejsze wartości BIC i AIC). W tym celu, skorzystam z funkcji *regsubsets* z biblioteki *leaps* oraz funkcji *dredge* z biblioteki *MuMin*.

```
require("leaps")
b<- regsubsets(GPA~HSM+HSS+HSE+SATM+SATV+SEX, nbest=1, data);
u<-summary(b);
cbind(u$bic, u$which)
```

```
            (Intercept) HSM HSS HSE SATM SATV SEX
1 -36.52518           1   1   0   0    0    0   0
2 -34.18564           1   1   0   1    0    0   0
3 -30.28481           1   1   0   1    1    0   0
4 -25.66783           1   1   1   1    1    0   0
5 -20.74352           1   1   1   1    1    1   0
6 -15.41891           1   1   1   1    1    1   1
```

Sugerując się miarą BIC, najlepszym modelem jest model zbudowany wyłącznie na podstawie zmiennej HSM. Sprawdzimy teraz, na jaki model wskazuje miara AIC.

```
require(MuMIn)

options(na.action="na.fail")
combinations <- dredge(model)

print(combinations)
```

```
Global model call: lm(formula = GPA ~ HSM + HSS + HSE + SATM + SATV + SEX, data
= data)
---
Model selection table
   (Intrc)     HSE    HSM     HSS      SATM       SATV       SEX df   logLik
 AICc delta weight
4   0.6242 0.06067 0.1827                                        4 -236.302
480.8  0.00  0.109
7   0.7404         0.1756 0.05359                                4 -236.567
481.3  0.53  0.083
```

```
12  0.3047 0.06572 0.1627         0.0007467                       5 -235.546
481.4  0.58  0.081
3   0.9077         0.2076                                         3 -237.838
481.8  1.00  0.066
8   0.5899 0.04510 0.1686 0.03432                                 5 -235.877
482.0  1.24  0.058
15  0.4843         0.1597 0.05459 0.0006383                       5 -236.009
482.3  1.51  0.051
16  0.2777 0.05060 0.1495 0.03311 0.0007323                       6 -235.149
482.7  1.90  0.042
36  0.6440 0.06405 0.1819                            -0.030240    5 -236.258
482.8  2.01  0.040
11  0.6657         0.1930         0.0006105                       4 -237.333
482.8  2.06  0.039
20  0.6212 0.06057 0.1826                    8.771e-06             5 -236.302
482.9  2.09  0.038
28  0.3484 0.07081 0.1606         0.0009279 -3.476e-04            6 -235.368
483.1  2.34  0.034
39  0.7004         0.1747 0.05402                         0.032860  5 -236.510
483.3  2.51  0.031
23  0.7329         0.1755 0.05337             2.062e-05           5 -236.566
483.4  2.62  0.029
44  0.2910 0.06476 0.1625         0.0007642               0.009563  6 -235.542
483.5  2.69  0.028
19  0.8440         0.2056                    1.590e-04            4 -237.791
483.8  2.98  0.024
35  0.8760         0.2070                                 0.026930  4 -237.800
483.8  3.00  0.024
47  0.3316         0.1538 0.05580 0.0007915               0.074940  6 -235.745
483.9  3.09  0.023
40  0.5980 0.04672 0.1686 0.03348                       -0.011100  6 -235.872
484.1  3.34  0.020
24  0.6047 0.04543 0.1688 0.03465           -4.354e-05            6 -235.874
484.1  3.35  0.020
31  0.5287         0.1580 0.05771 0.0007701 -2.654e-04            6 -235.904
484.2  3.41  0.020
32  0.3267 0.05529 0.1460 0.03591 0.0009436 -4.078e-04            7 -234.905
484.3  3.54  0.018
43  0.5340         0.1884         0.0007456               0.066390  5 -237.128
484.5  3.74  0.017
48  0.2313 0.04648 0.1479 0.03536 0.0007881               0.031020  7 -235.109
484.7  3.95  0.015
52  0.6490 0.06425 0.1820                   -1.347e-05 -0.030620  6 -236.258
484.9  4.12  0.014
27  0.6804         0.1931         0.0006454 -7.121e-05            5 -237.325
484.9  4.14  0.014
60  0.3351 0.06989 0.1604         0.0009446 -3.473e-04  0.009211  7 -235.364
485.2  4.46  0.012
55  0.6873         0.1744 0.05365             3.418e-05  0.033340  6 -236.508
485.4  4.62  0.011
63  0.3733         0.1515 0.05938 0.0009487 -2.995e-04  0.079060  7 -235.611
485.7  4.96  0.009
51  0.8041         0.2048                     1.717e-04  0.029560  5 -237.746
485.8  4.98  0.009
56  0.6164 0.04730 0.1689 0.03378           -5.121e-05 -0.012370  7 -235.867
486.3  5.47  0.007
64  0.2786 0.05103 0.1442 0.03827 0.0010030 -4.109e-04  0.032370  8 -234.862
486.4  5.61  0.007
```

```
59  0.5514         0.1884         0.0007948 -9.556e-05  0.067530  6 -237.114
486.6  5.83  0.006
14  0.2519 0.08091        0.08872 0.0016980                       5 -242.470
495.2 14.43  0.000
30  0.3366 0.08770        0.09123 0.0020200 -6.969e-04            6 -241.790
496.0 15.18  0.000
46  0.1131 0.06753        0.09365 0.0018340             0.093330  6 -242.123
496.6 15.85  0.000
45  0.2545                0.12730 0.0019010             0.162000  5 -243.404
497.1 16.30  0.000
13  0.5934                0.13090 0.0016520                       4 -244.607
497.4 16.61  0.000
62  0.1982 0.07436        0.09614 0.0021550 -6.957e-04  0.093010  7 -241.444
497.4 16.62  0.000
61  0.3332                0.13200 0.0021600 -5.511e-04  0.167200  6 -242.975
498.3 17.55  0.000
29  0.6747                0.13520 0.0018790 -4.993e-04            5 -244.257
498.8 18.00  0.000
10  0.3296 0.13690                0.0020120                       4 -245.569
499.3 18.53  0.000
26  0.4048 0.14410                0.0022980 -6.032e-04            5 -245.072
500.4 19.63  0.000
42  0.2648 0.13200                0.0020860             0.044930  5 -245.488
501.3 20.47  0.000
6   1.1130 0.07636        0.11180                                 4 -246.810
501.8 21.02  0.000
58  0.3417 0.13930                0.0023680 -6.003e-04  0.043540  6 -244.996
502.4 21.59  0.000
5   1.4130                0.15110                                 3 -248.646
503.4 22.62  0.000
22  1.0650 0.07521        0.11050            1.344e-04            5 -246.780
503.8 23.05  0.000
38  1.1140 0.07667        0.11170                      -0.002135  5 -246.809
503.9 23.11  0.000
37  1.3170                0.15080                       0.072230  4 -248.398
505.0 24.19  0.000
21  1.3140                0.14740            2.555e-04            4 -248.538
505.3 24.47  0.000
54  1.0640 0.07504        0.11060            1.351e-04  0.001172  6 -246.780
505.9 25.16  0.000
53  1.2020                0.14680            2.832e-04  0.075830  5 -248.266
506.8 26.02  0.000
2   1.4260 0.14940                                                3 -251.744
509.6 28.81  0.000
18  1.2750 0.14350                           3.942e-04            4 -251.491
511.2 30.38  0.000
34  1.4680 0.15710                                      -0.077460  4 -251.494
511.2 30.38  0.000
50  1.3310 0.15100                           3.426e-04 -0.067210  5 -251.307
512.9 32.10  0.000
41  0.8479                        0.0025540             0.197400  4 -252.528
513.2 32.45  0.000
9   1.2840                        0.0022710                       3 -254.181
514.5 33.68  0.000
57  0.8660                        0.0026050 -9.949e-05  0.198500  5 -252.515
515.3 34.52  0.000
25  1.2890                        0.0022830 -2.456e-05            4 -254.180
516.5 35.76  0.000
```

```
17  2.1490                                  9.635e-04            3 -260.034
526.2 45.39  0.000
1   2.6350                                                       2 -261.512
527.1 46.29  0.000
49  2.0120                                  9.928e-04  0.090060  4 -259.688
527.6 46.77  0.000
33  2.5300                                             0.077970  3 -261.255
528.6 47.83  0.000
Models ranked by AICc(x)
```

Miara AIC osiąga najmniejszą wartość dla modelu zbudowanego ze zmiennych HSM i HSS.