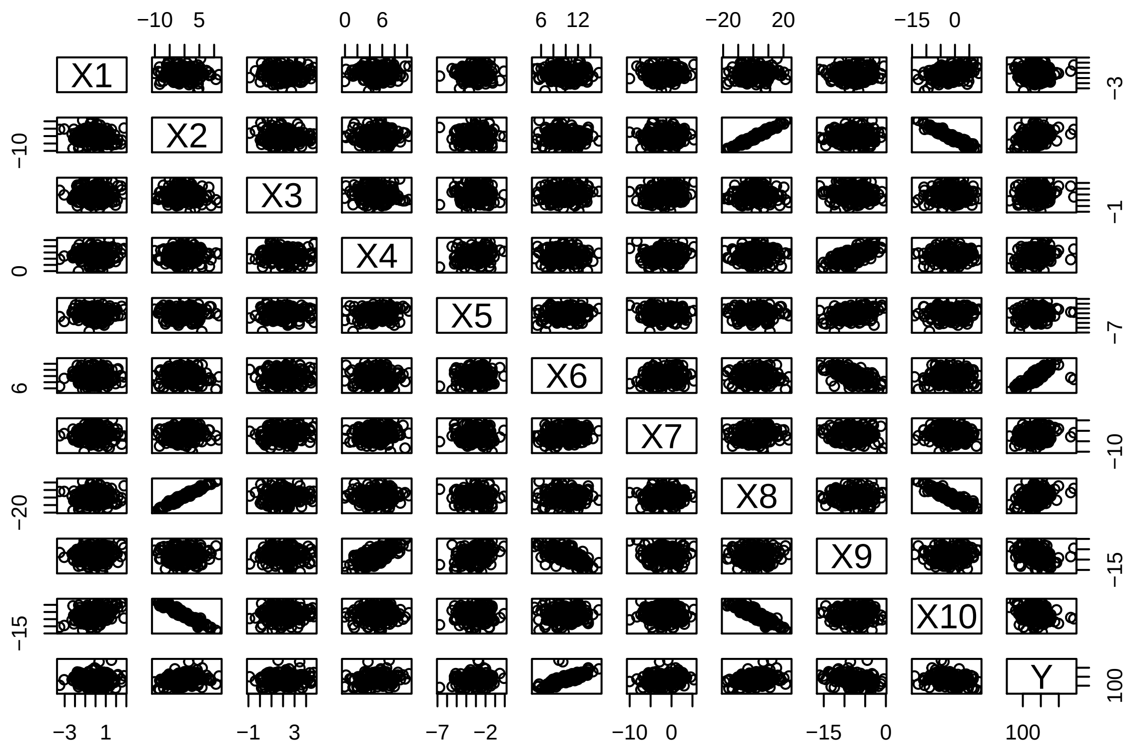
Modele regresji raport 2

Paweł Matławski

15 04 2021

Zadanie nr 1

```
library(readxl)
library(xtable)
regresja_wielokrotna <- read_excel("C:/Users/48795/Desktop/Modele Regresji/regresja wielokrotna.xlsx")
attach(regresja_wielokrotna)
pairs(regresja_wielokrotna)
```



Analizując wykresy rozrzutów par dla zmiennych $Y, X_1, X_2, ..., X_{10}$ zmienna objaśniająca X_6 wydaje się mieć największy liniowy wpływ na zmienną Y. Możemy się spodziewać problemu współliniowości (silnej korelacji) pomiędzy parami: $(X_2, X_8), (X_2, X_{10})$ oraz (X_8, X_{10}). Na wykresie rozrzutu (X_6, X_{10}) widzimy pojedyncze wartości odstające dla zmiennej X_{10}.

Zadanie nr 2

```
xtable(cor(regresja_wielokrotna))
```

% latex table generated in R 4.0.5 by xtable 1.8-4 package % Wed Apr 21 21:08:03 2021

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Y
X1	1.00	-0.09	0.02	0.02	0.05	-0.05	0.01	0.05	0.05	0.34	0.02
X2	-0.09	1.00	-0.11	-0.01	-0.03	-0.02	0.03	0.98	-0.00	-0.96	0.29
X3	0.02	-0.11	1.00	-0.09	-0.02	0.04	0.14	0.02	-0.09	0.09	0.11
X4	0.02	-0.01	-0.09	1.00	0.14	-0.03	0.09	-0.02	0.70	0.01	0.25
X5	0.05	-0.03	-0.02	0.14	1.00	0.15	-0.01	-0.02	0.33	0.04	0.17
X6	-0.05	-0.02	0.04	-0.03	0.15	1.00	0.17	-0.02	-0.65	0.03	0.81
X7	0.01	0.03	0.14	0.09	-0.01	0.17	1.00	0.05	-0.06	-0.03	0.20
X8	0.05	0.98	0.02	-0.02	-0.02	-0.02	0.05	1.00	-0.01	-0.91	0.31
X9	0.05	-0.00	-0.09	0.70	0.33	-0.65	-0.06	-0.01	1.00	0.00	-0.34
X10	0.34	-0.96	0.09	0.01	0.04	0.03	-0.03	-0.91	0.00	1.00	-0.26
Y	0.02	0.29	0.11	0.25	0.17	0.81	0.20	0.31	-0.34	-0.26	1.00

Macierz korelacji potwierdza nasze spostrzeżenia z wykresów rozrzutu z zadania pierwszego. Występuje silna korelacja pomiędzy parami: $(X_2, X_8), (X_2, X_{10}) oraz (X_8, X_{10})$.

Zadanie nr 3

```
model1 <- lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10, data = regresja_wielokrotna)
summary(model1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
##     X10, data = regresja_wielokrotna)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.810 -1.972 -1.048  0.070 97.059
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.52260    6.65182   0.079   0.9375
## X1           2.84868    4.02874   0.707   0.4804
## X2           1.82515    7.20785   0.253   0.8004
## X3           3.64880    3.61818   1.008   0.3145
## X4           3.95372    2.37698   1.663   0.0979 .
## X5           0.21928    2.46797   0.089   0.9293
## X6          11.00584    2.38215   4.620 7.08e-06 ***
## X7          -0.03279    0.27664  -0.119   0.9058
## X8          -0.14515    3.59911  -0.040   0.9679
## X9           0.13848    2.34504   0.059   0.9530
## X10         -0.73124    1.50367  -0.486   0.6273
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.14 on 189 degrees of freedom
```

```
## Multiple R-squared:  0.8534, Adjusted R-squared:  0.8456
## F-statistic:   110 on 10 and 189 DF,  p-value: < 2.2e-16
```

Na podstawie p-value dla dopasowanego modelu odrzucamy hipotezę zerową, która mówi o braku zależności liniowej z którąkolwiek ze zmiennych.

Zadanie nr 4

```
library(car)
```

```
## Loading required package: carData
```

```
vif(model1)
```

```
##          X1          X2          X3          X4          X5          X6
##   35.145297 1497.679681   27.256636   36.089952   11.758465   42.033020
##          X7          X8          X9         X10
##    1.093435 1469.489960   89.575184   73.671270
```

```
#pozbywamy się X2
model2 <- lm(Y~X1+X3+X4+X5+X6+X7+X8+X9+X10, data = regresja_wielokrotna)
vif(model2)
```

```
##        X1        X3        X4        X5        X6        X7        X8        X9
## 11.331815  1.852655 35.844845 11.605348 41.780951  1.074035 64.180276 88.932719
##       X10
## 72.840327
```

```
#pozbywamy się X9
model3 <- lm(Y~X1+X3+X4+X5+X6+X7+X8+X10, data = regresja_wielokrotna)
vif(model3)
```

```
##        X1        X3        X4        X5        X6        X7        X8       X10
## 11.276908  1.842674  1.048843  1.051652  1.098026  1.072346 63.522539 72.147455
```

```
#pozbywamy się X10
model4 <- lm(Y~X1+X3+X4+X5+X6+X7+X8, data = regresja_wielokrotna)
vif(model4)
```

```
##       X1       X3       X4       X5       X6       X7       X8
## 1.008049 1.034192 1.047199 1.050985 1.065881 1.071434 1.007002
```

Na podstawie współczynników VIF (podbicia wariancji), pozbyliśmy się następujących zmiennych objaśniających: X_2, X_9, X_{10}, tym samym rozwiązując problem współliniowości.

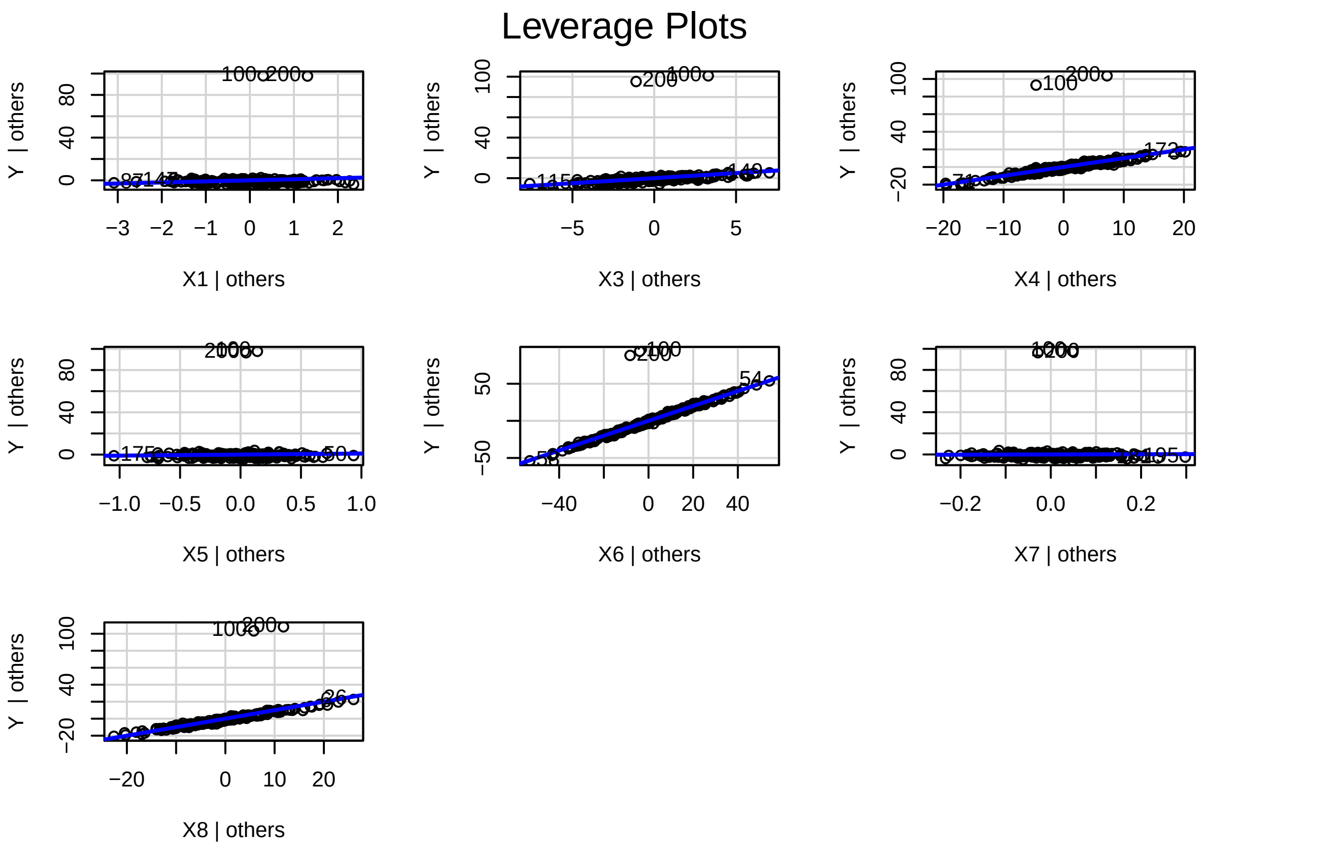
Zadanie nr 5

```
library(olsrr)
```

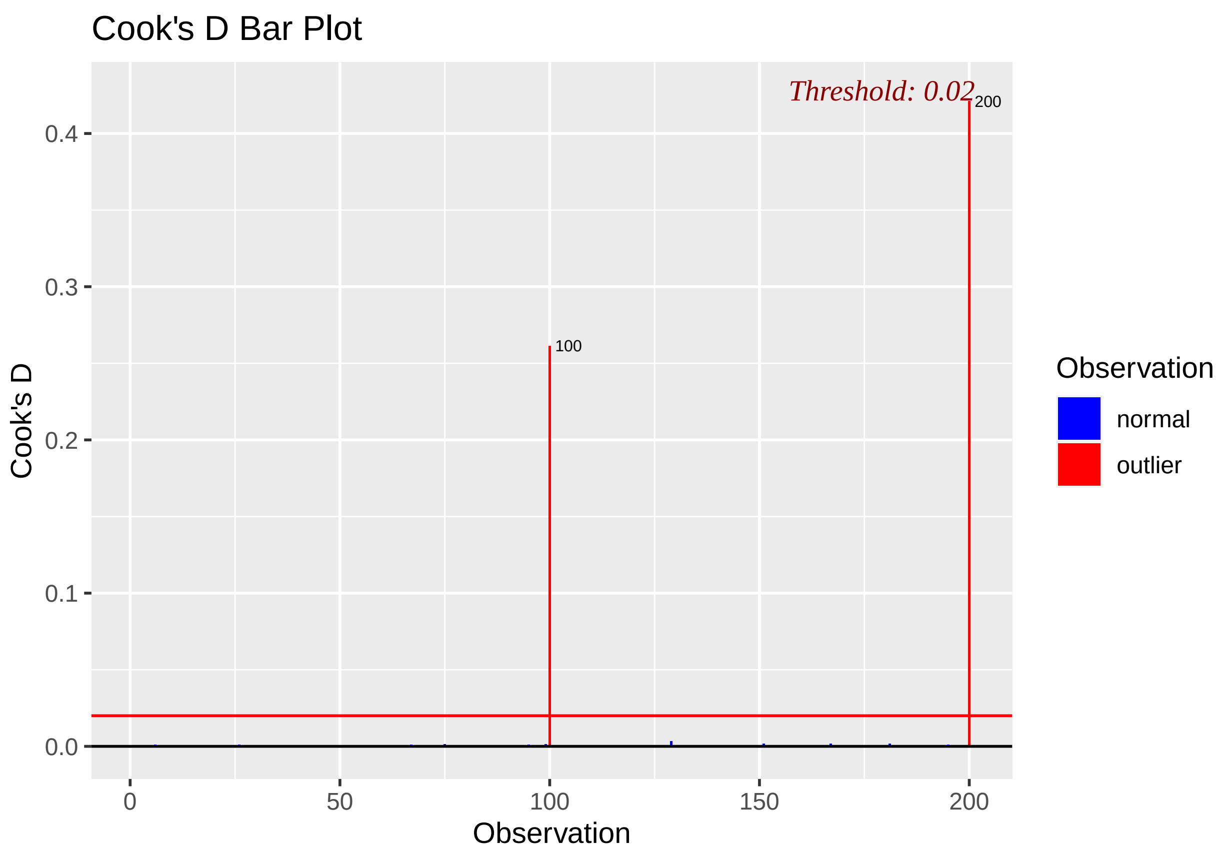
```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

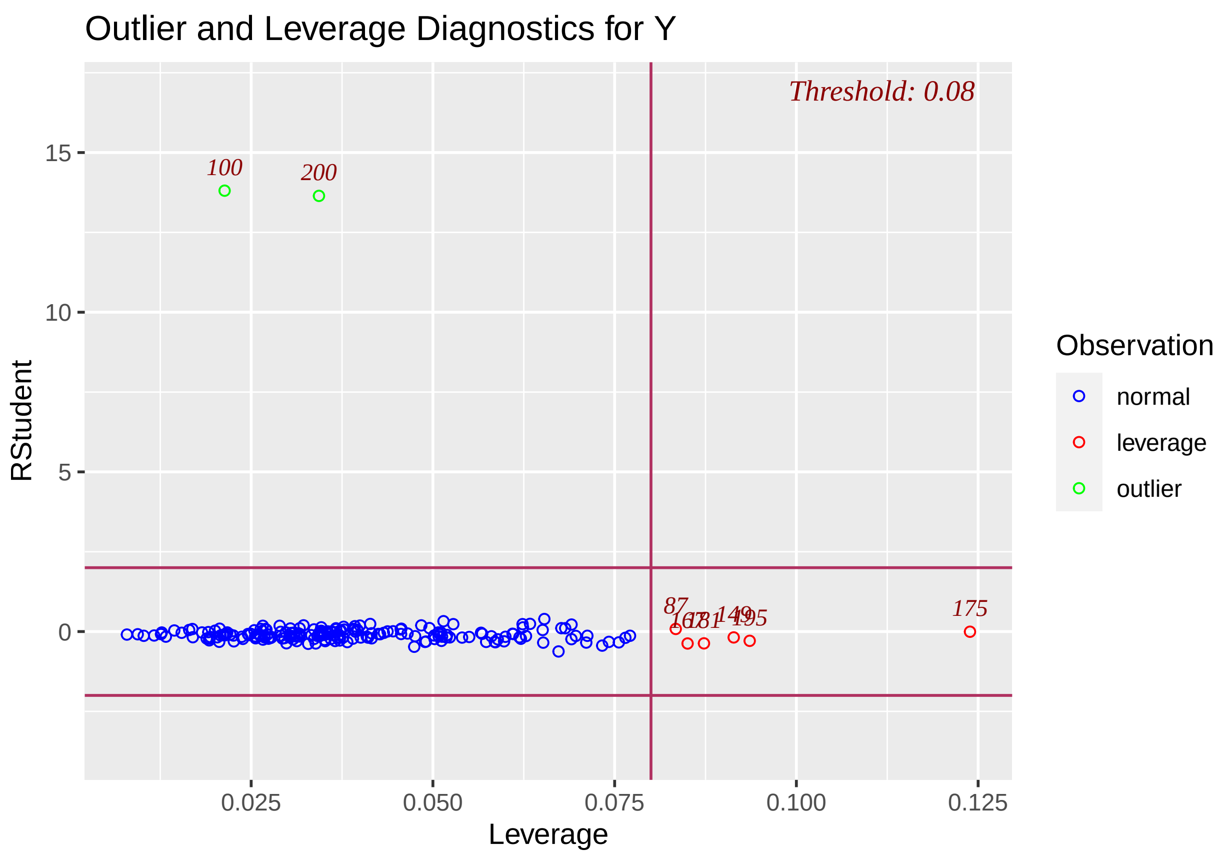
```
#wpływy
leveragePlots(model4)
```



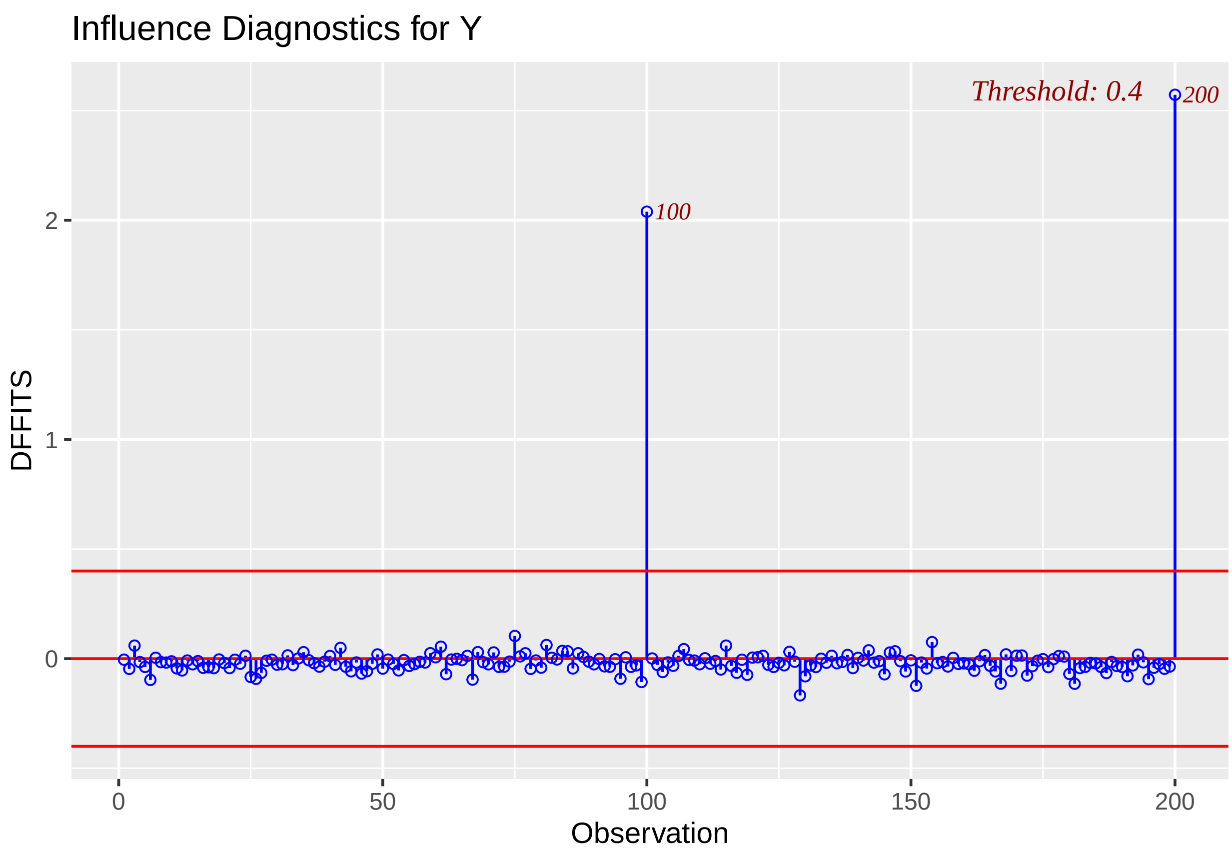

```
#wykres wartości D
ols_plot_cooksd_bar(model4)
```




```r
#studentyzowane rezydua
ols_plot_resid_lev(model4)
```




```
#dffits
ols_plot_dffits(model4)
```



Analizując wykresy wpływów, odległości Cooke'a, studentyzowanych rezyuów i dffits, możemy usunąć z naszego modelu obserwacje o numerach: 100, 200.

Zadanie nr 6

```
regresja_wielokrotna2 <- regresja_wielokrotna[-c(100,200),]
attach(regresja_wielokrotna2)
```

```
## The following objects are masked from regresja_wielokrotna:
##
##     X1, X10, X2, X3, X4, X5, X6, X7, X8, X9, Y
```

```
model5 <- lm(Y~X1+X3+X4+X5+X6+X7+X8, data = regresja_wielokrotna2)
summary(model5)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4 + X5 + X6 + X7 + X8, data = regresja_wielokrotna2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66738 -0.22918  0.00776  0.21741  0.94441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.221466   0.185415   1.194   0.2338
## X1          0.039981   0.022143   1.806   0.0726 .
```

```
## X3           2.007676   0.022861  87.822   <2e-16 ***
## X4           3.998568   0.013115 304.893   <2e-16 ***
## X5           0.027732   0.023827   1.164   0.2459    
## X6          10.986335   0.012250 896.863   <2e-16 ***
## X7           0.001580   0.008846   0.179   0.8584    
## X8           0.996283   0.003059 325.668   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3274 on 190 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998 
## F-statistic: 1.483e+05 on 7 and 190 DF,  p-value: < 2.2e-16
```

Po usunięciu z modelu wartości odstających wpływowych i po redukcji współliniowości, widzimy że dalej nasz model jest zależny liniowo przynajmniej od jednej zmiennej, a dodatkowo analizując współczynniki R-squared i Adjusted R-squared zauważamy że model jest znacznie lepiej dopasowany do danych.

Zadanie nr 7

```r
library(MASS)
```

```
## 
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':
## 
##     cement
```

```r
modelFORWARD <- stepAIC(model5, direction = 'forward', trace = FALSE)
modelFORWARD
```

```
## 
## Call:
## lm(formula = Y ~ X1 + X3 + X4 + X5 + X6 + X7 + X8, data = regresja_wielokrotna2)
## 
## Coefficients:
## (Intercept)           X1           X3           X4           X5           X6  
##     0.22147      0.03998      2.00768      3.99857      0.02773     10.98633  
##          X7           X8  
##     0.00158      0.99628
```

```r
modelBACKWARD <- stepAIC(model5, direction = 'backward', trace = FALSE)
modelBACKWARD
```

```
## 
## Call:
## lm(formula = Y ~ X1 + X3 + X4 + X6 + X8, data = regresja_wielokrotna2)
## 
## Coefficients:
## (Intercept)           X1           X3           X4           X6           X8  
##     0.09591      0.04129      2.00798      4.00102     10.98893      0.99624
```

```r
modelBOTH <- stepAIC(model5, direction = 'both', trace = FALSE)
modelBOTH
```

```
## 
## Call:
```

```
## lm(formula = Y ~ X1 + X3 + X4 + X6 + X8, data = regresja_wielokrotna2)
##
## Coefficients:
## (Intercept)           X1           X3           X4           X6           X8
##     0.09591      0.04129      2.00798      4.00102     10.98893      0.99624
```

```
ols_mallows_cp(modelFORWARD, model5)
```

```
## [1] 8
```

```
ols_mallows_cp(modelBACKWARD, model5)
```

```
## [1] 5.369195
```

```
ols_mallows_cp(modelBOTH, model5)
```

```
## [1] 5.369195
```

Po zastosowaniu krokowych metod wyboru modelu (metoda wprowadzania postępującego, metoda eliminacji wstecz oraz metoda łącząca obie poprzednie) i analizy współczynników Mallowa C_p, wybieramy model otrzymany zarówno z 'eliminacji', jak i połączenia dwóch metod, czyli uzależniony od zmiennych objaśniających: X_1, X_3, X_4, X_6, X_8. Współczynnik Mallowa jest mniejszy i bardziej zbliżony do ilości predyktorów.

```
M <- modelBOTH
summary(M)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4 + X6 + X8, data = regresja_wielokrotna2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7262 -0.2291  0.0024  0.2152  0.9577
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.095908   0.145407    0.66    0.510
## X1           0.041293   0.022078    1.87    0.063 .
## X3           2.007983   0.022559   89.01   <2e-16 ***
## X4           4.001018   0.012869  310.91   <2e-16 ***
## X6          10.988928   0.011876  925.27   <2e-16 ***
## X8           0.996243   0.003049  326.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3269 on 192 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 2.083e+05 on 5 and 192 DF,  p-value: < 2.2e-16
```

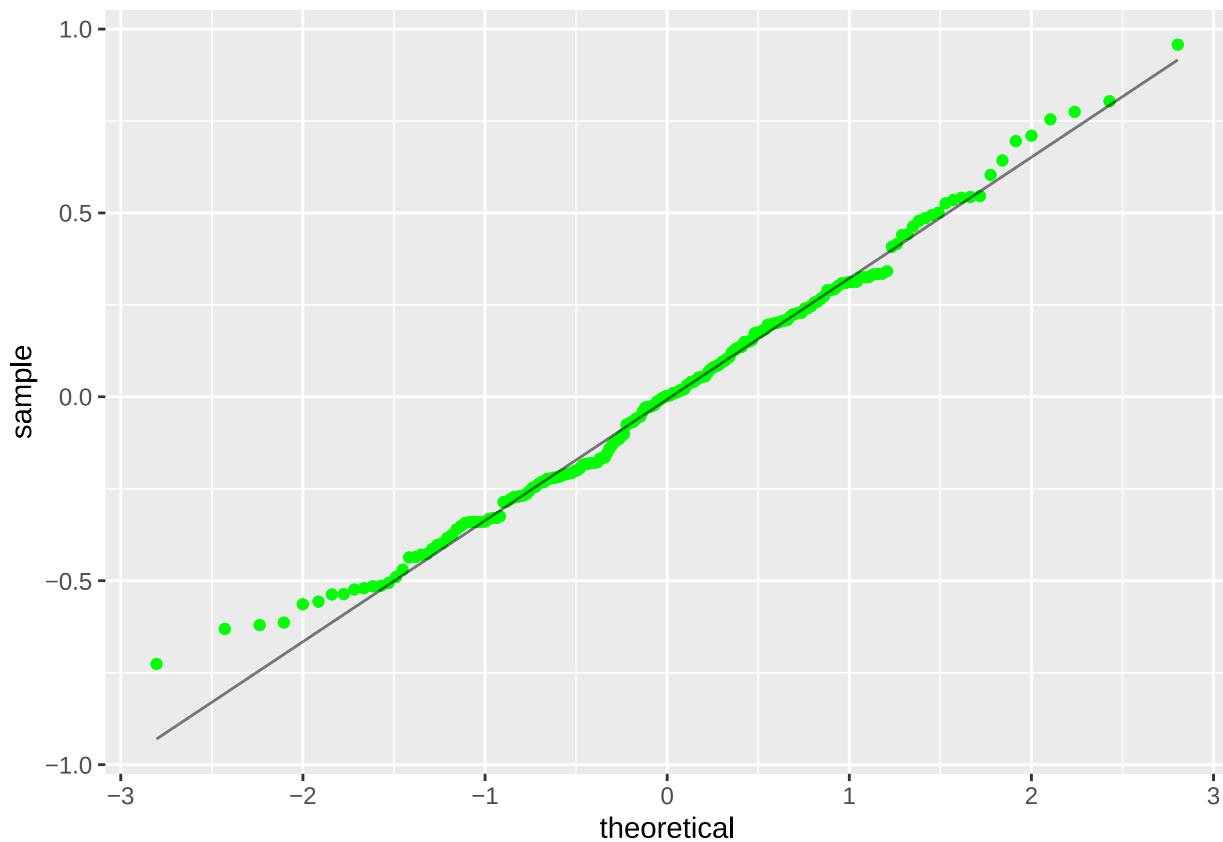
```
confint(M)
```

```
##                    2.5 %      97.5 %
## (Intercept) -0.190893382  0.38270844
## X1          -0.002253668  0.08483966
## X3           1.963487470  2.05247825
## X4           3.975635474  4.02640003
## X6          10.965503193 11.01235326
## X8           0.990229563  1.00225634
```

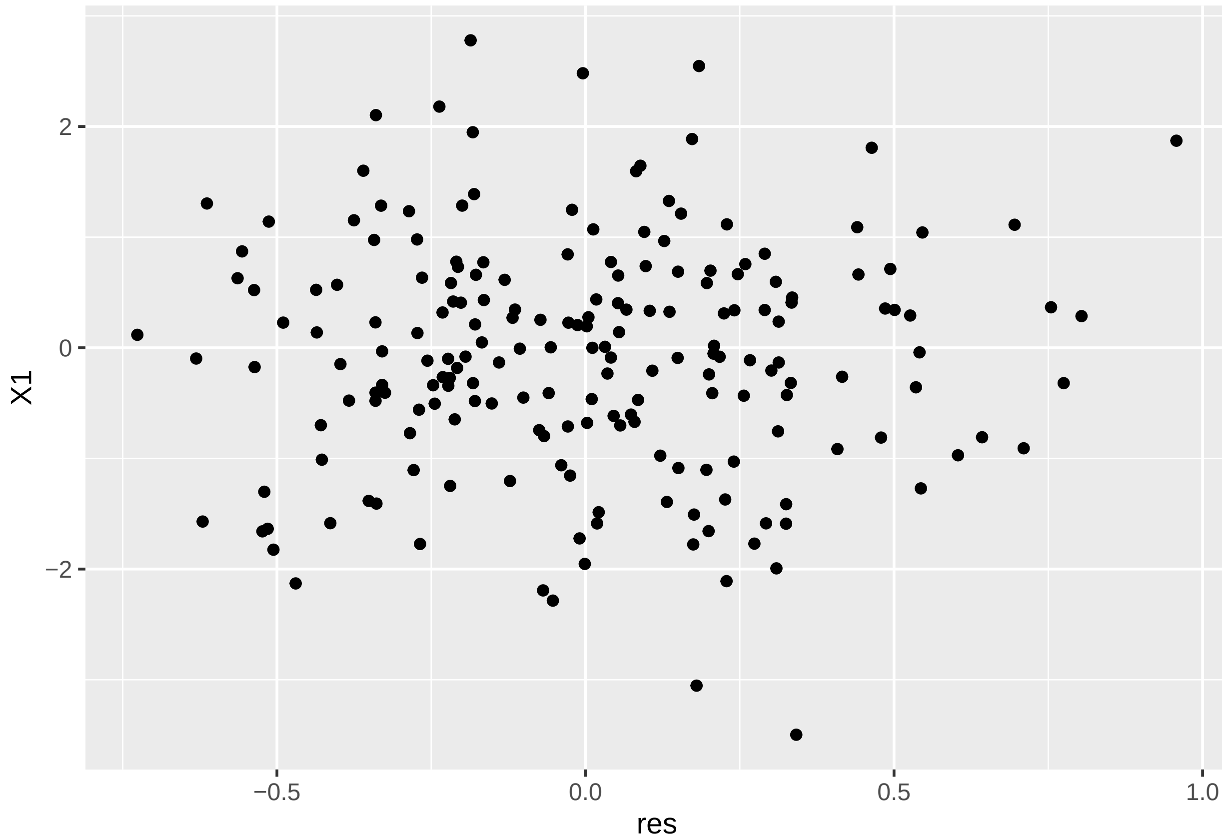
Na podstawie p-value ze statystyki F, odrzucamy hipotezę zerową, przyjmując że model jest zależny liniowo przynajmniej od jednej zmiennej objaśniającej. Dodatkowo na na podstawie wartości p możemy uznać, że na poziomie istotności 0.05 wszystkie zmienne oprócz stałej (X_0) i X_1 mają liniowy wpływ na zmienną objaśnianą. Wyznaczyliśmy także przedziały ufności dla zmiennych z modelu. Na podstawie wskaźników R-squared i Adjusted R-squared widzimy, że nasz model jest bardzo dobrze dopasowany do danych.

Zadanie nr 8

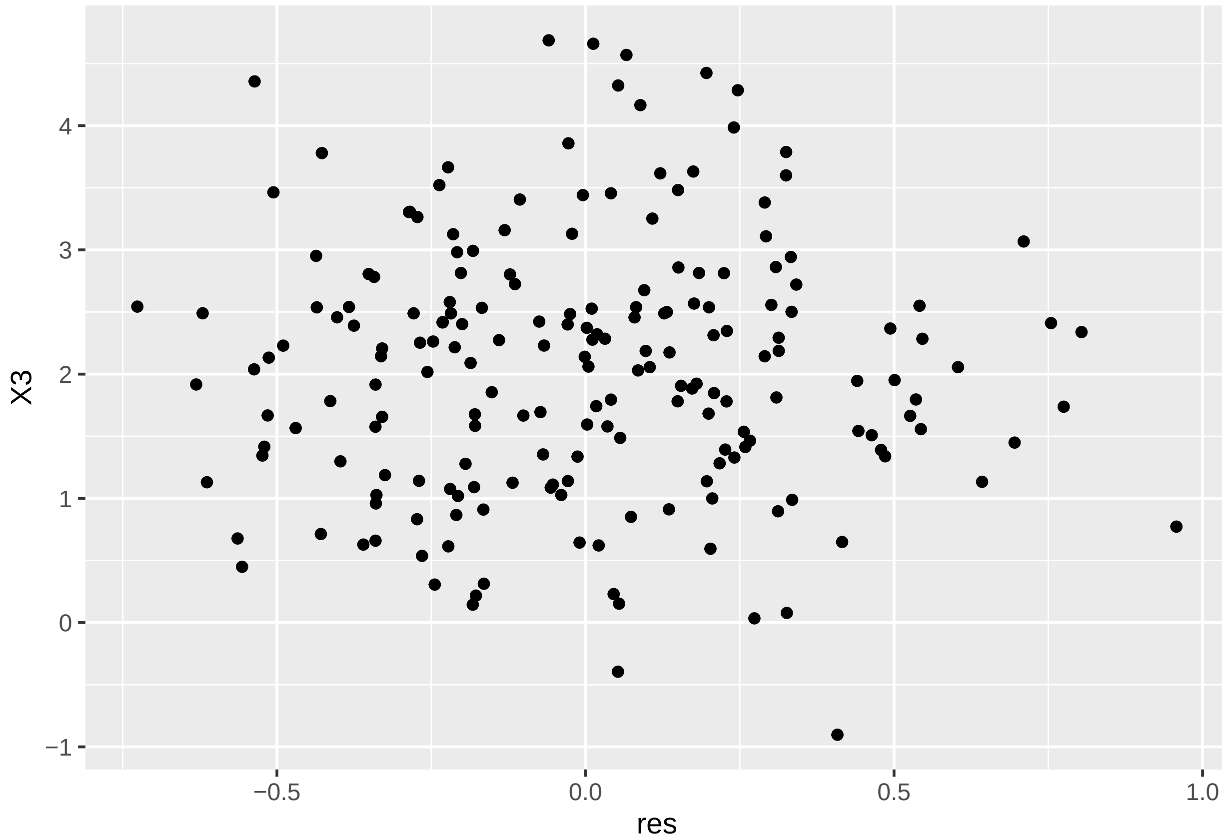
```
library(ggplot2)
df8 <- data.frame(res = M$residuals, X1 = M$model$X1, X3 = M$model$X3, X4 = M$model$X4, X6 = M$model$X6
ggplot(as.data.frame(M$residuals), aes(sample=M$residuals))+geom_qq(col='green')+geom_qq_line(alpha=0.5
```



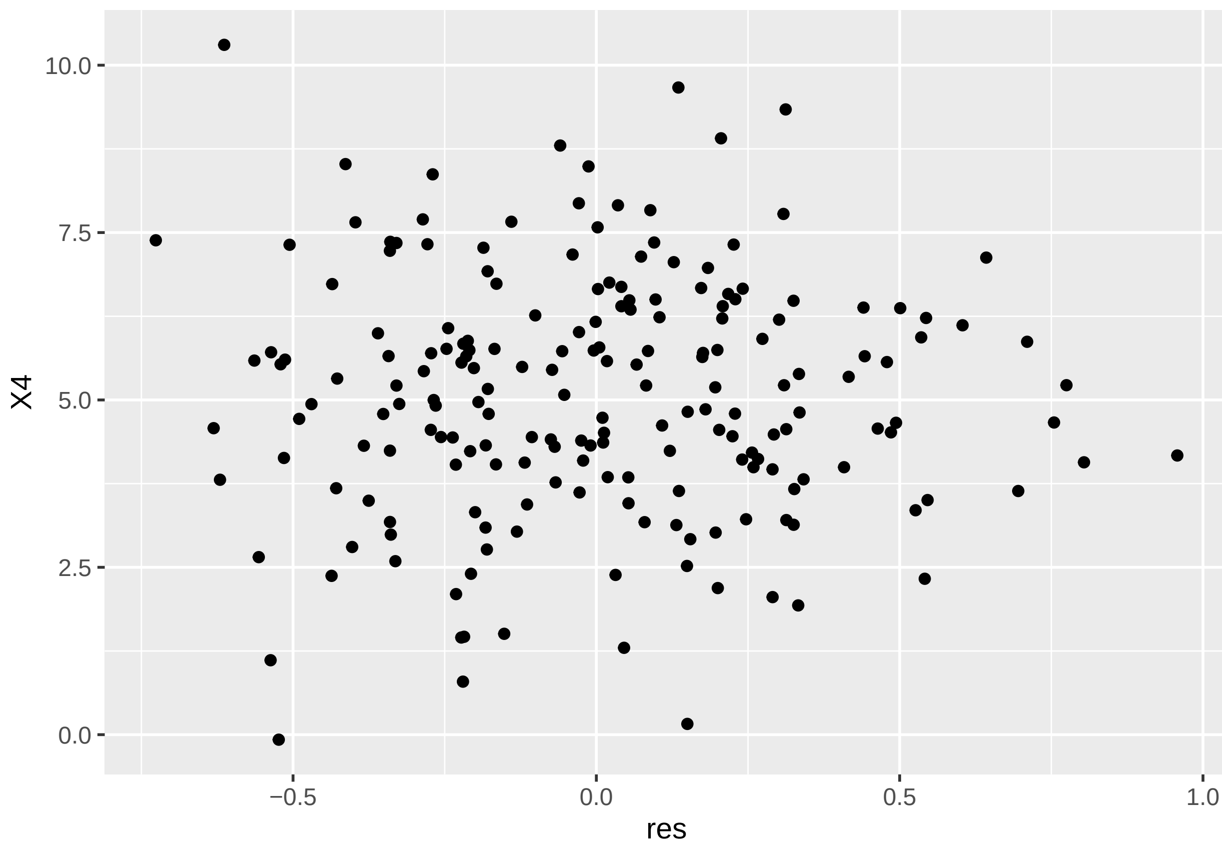

```
ggplot(df8, aes(x=res, y=X1)) + geom_point()
```



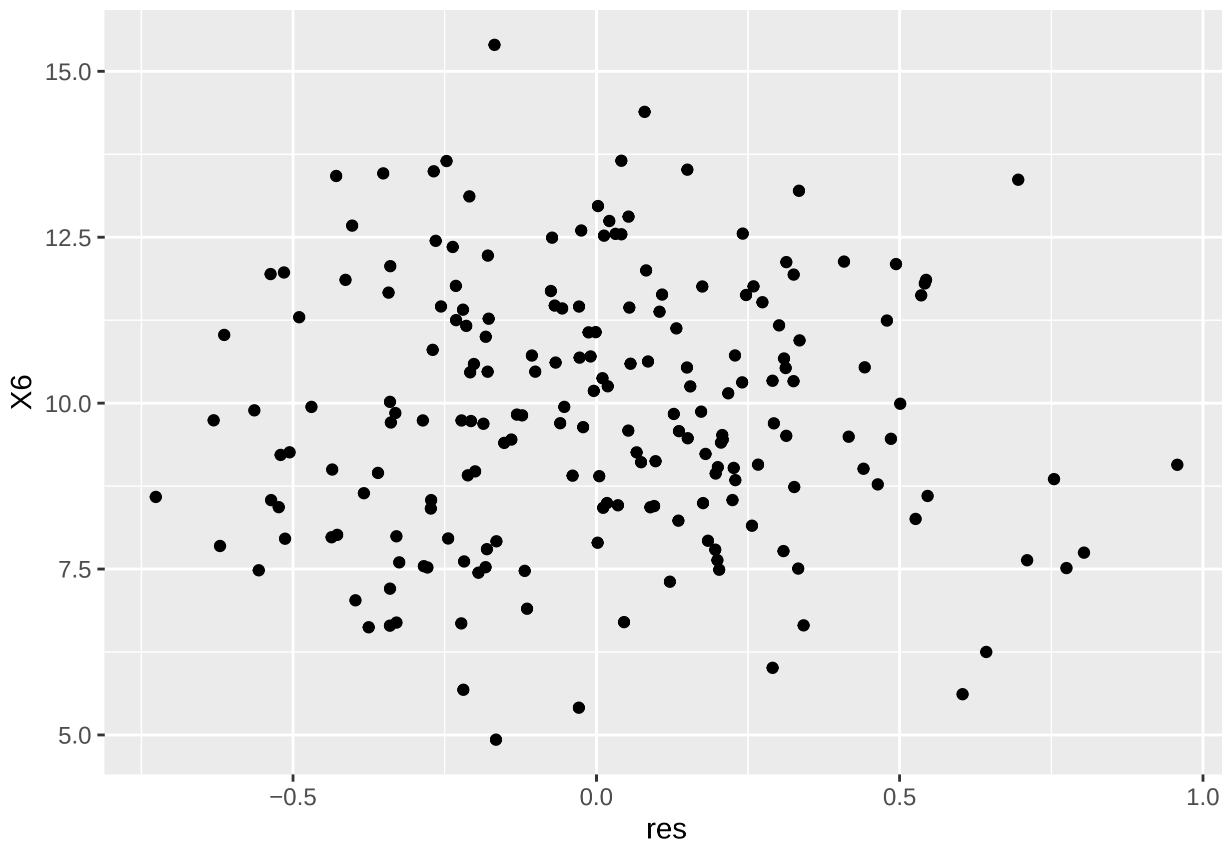

```
ggplot(df8, aes(x=res, y=X3)) + geom_point()
```



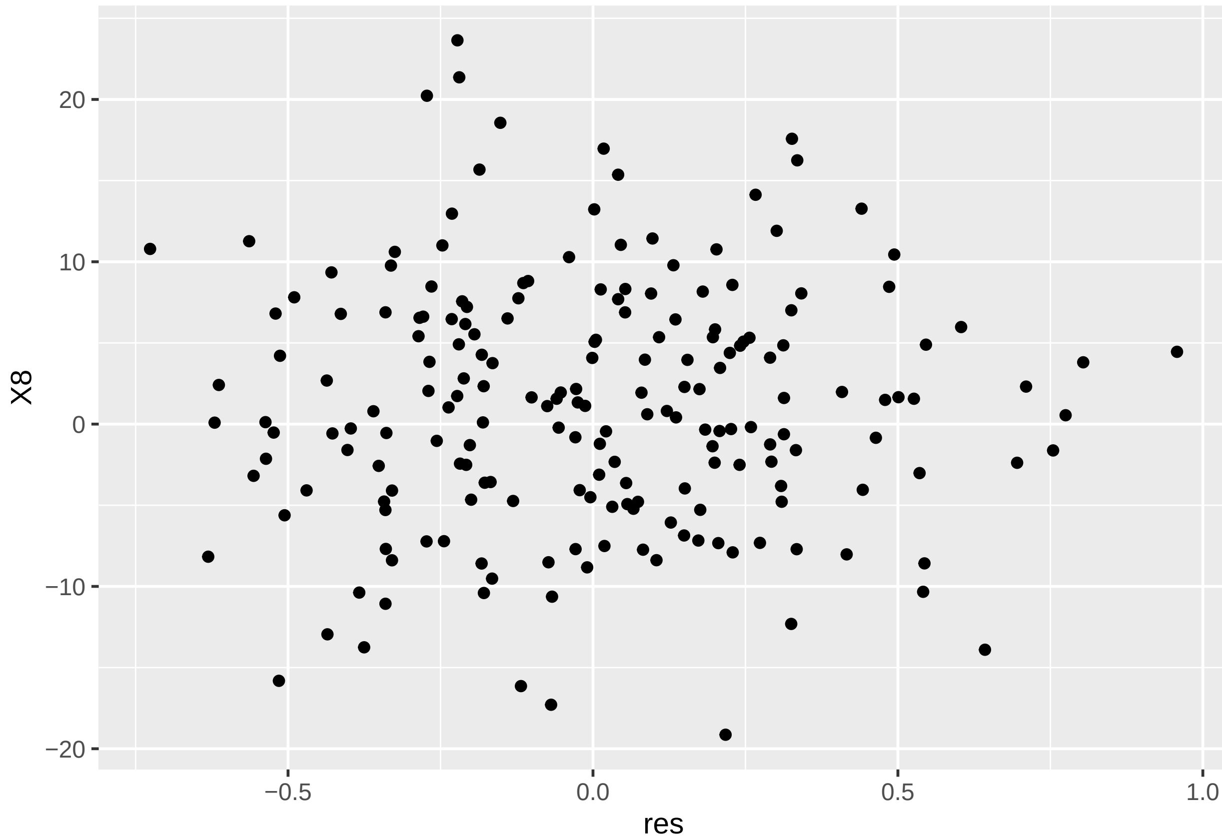

```
ggplot(df8, aes(x=res, y=X4)) + geom_point()
```



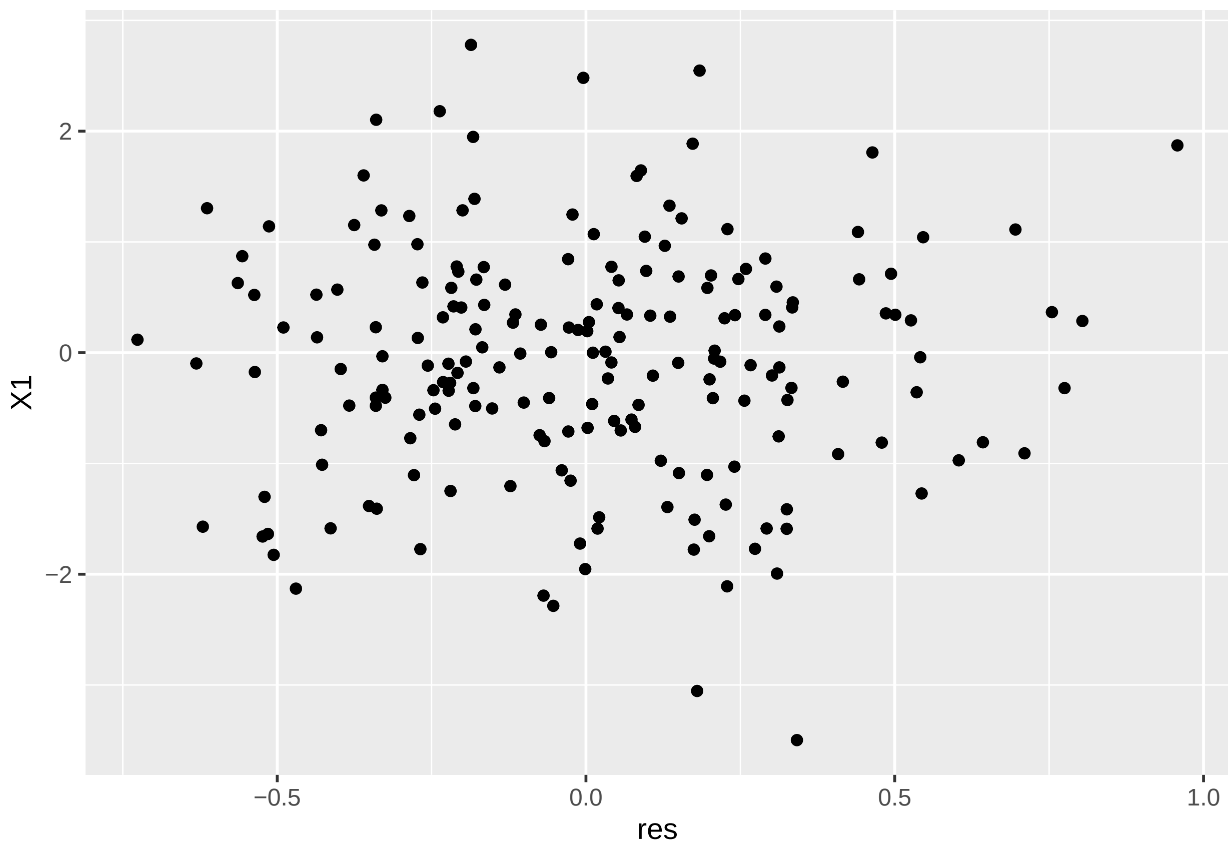

```
ggplot(df8, aes(x=res, y=X6)) + geom_point()
```



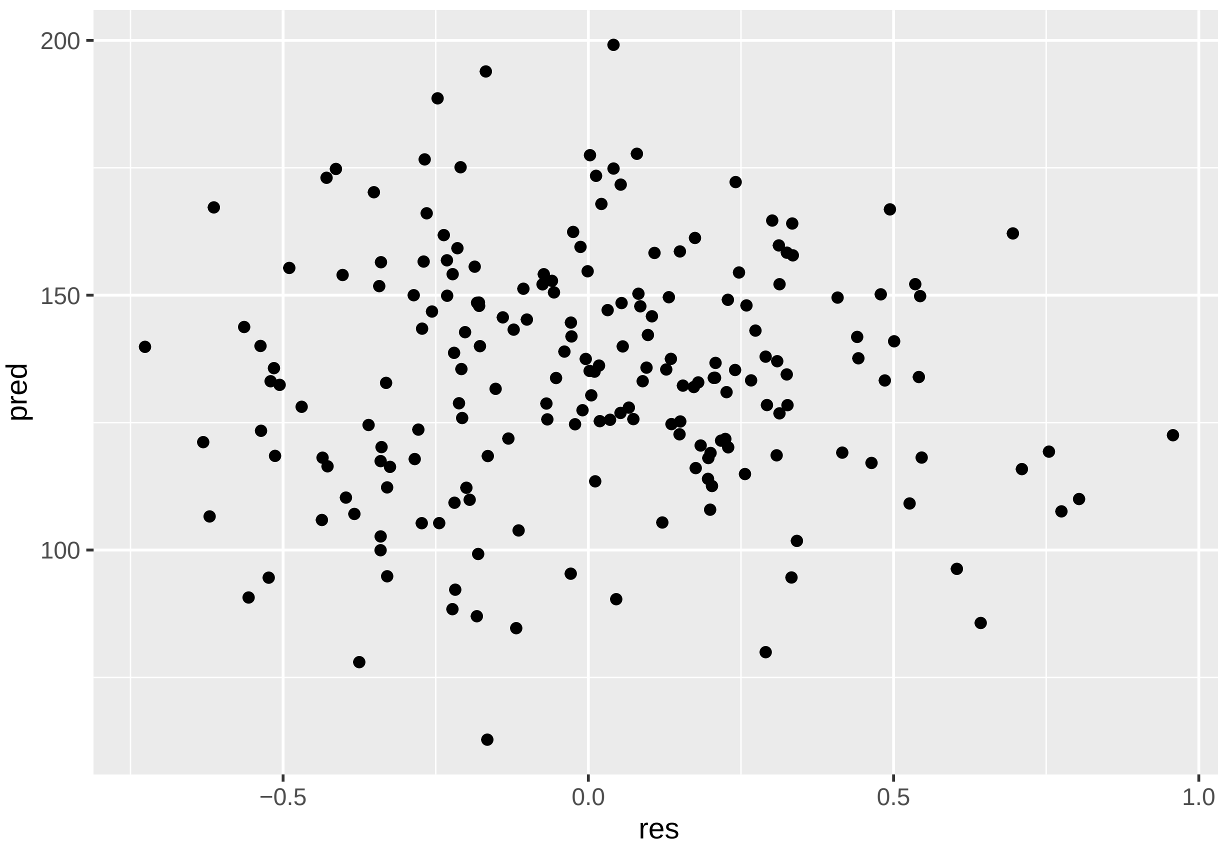

```
ggplot(df8, aes(x=res, y=X8)) + geom_point()
```




```
ggplot(df8, aes(x=res, y=X1)) + geom_point()
```




```
ggplot(df8, aes(x=res, y=pred)) + geom_point()
```



Na podstawie wykresu kwantlowego oraz scatterplotów rezyduów ze zmiennymi objaśniającymi i zmienną objaśnianą uznajemy, że są spełnione założenia występujące w modelu regresji liniowej. Wykres kwantylowy przypomina wykres dla próby z rozkładu normalnego, a w wykresach rozrzutu nie widać żadnej zależności liniowej.

Zadanie nr 9

```
newdata = data.frame(X1=1, X2=2, X3=3, X4=4, X5=5, X6=6, X7=7, X8=8, X9=9, X10=10)
predict(M, newdata = newdata)
```

```
##        1
## 96.06873
```

Przewidywana przez model M wartość zmiennej objaśnianej Y przy wartościach zmiennych objaśniających $X_1 = 1, X_2 = 2, ..., X_{10} = 10$ wynosi około 96.