# Labolatorium 4

Modele Regresji liniowej i ich zastosowania

Adrian Siwak, album 242084

2023-03-23

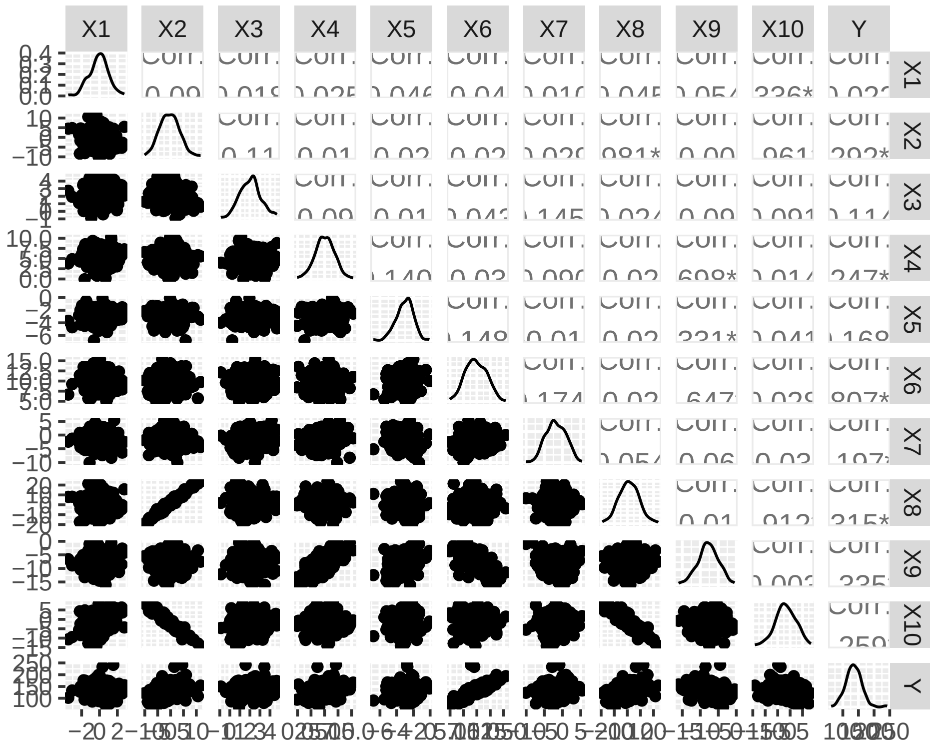
## Spis treści

```
library(xtable)
library(openxlsx)
library(corrplot)
library(regclass)
library(GGally)
dane<-read.xlsx("regresja_wielokrotna.xlsx")
```

# 1

Dla każdej z par utworzonych ze zmiennych $X1 \cdots X10$ wykonaj wykres rozrzutu i po przeanalizowaniu tych rysunków odpowiedz na następujące pytania

```
ggpairs(dane)
```



### (a)

Które ze zmiennych objaśniających mogą mieć najmocniejszy liniowy wpływ na zmienną objaśnianą $Y$?

Wykres rozrzutu zmiennej objaśniającej $X_6$ i zmiennej objaśnianej $Y$ wskazuje na zależność liniową.

### (b)

Czy pojawia się problem współliniowości, to znaczy, czy istnieje choć jedna para silnie ze sobą skorelowanych zmiennych objaśniających?

Tak, takie pary to $X_2$ i $X_8$ oraz $X_2$ i $X_{10}$ oraz $X_8$ i $X_{10}$.

### (c)

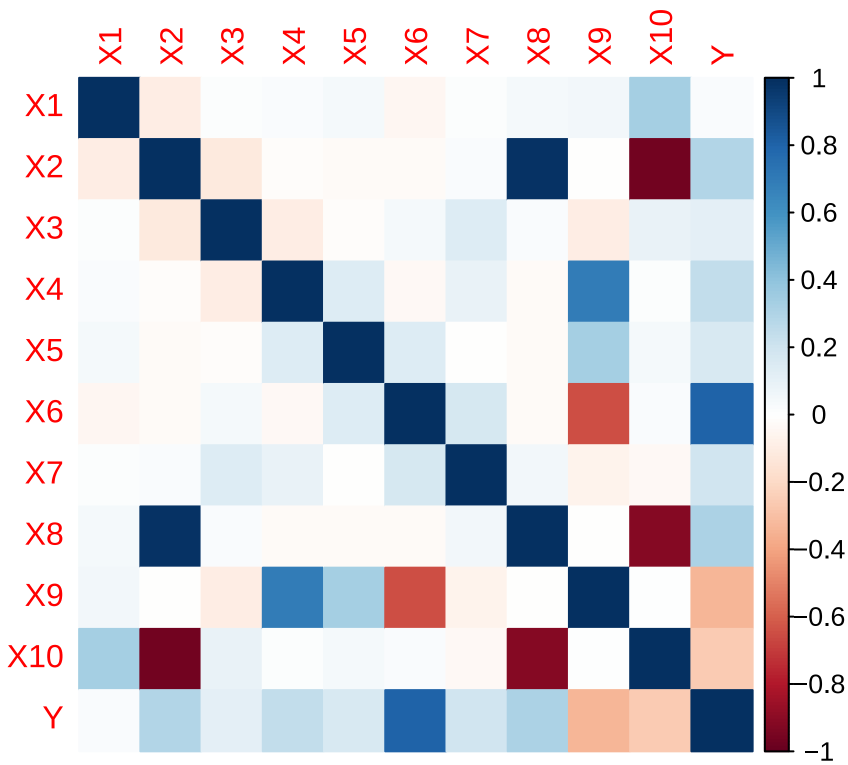
Czy pojawiają się obserwacje odstające?

Tak, po wykresach rozrzutów można wnioskować że pojawiają się obserwacje odstające.

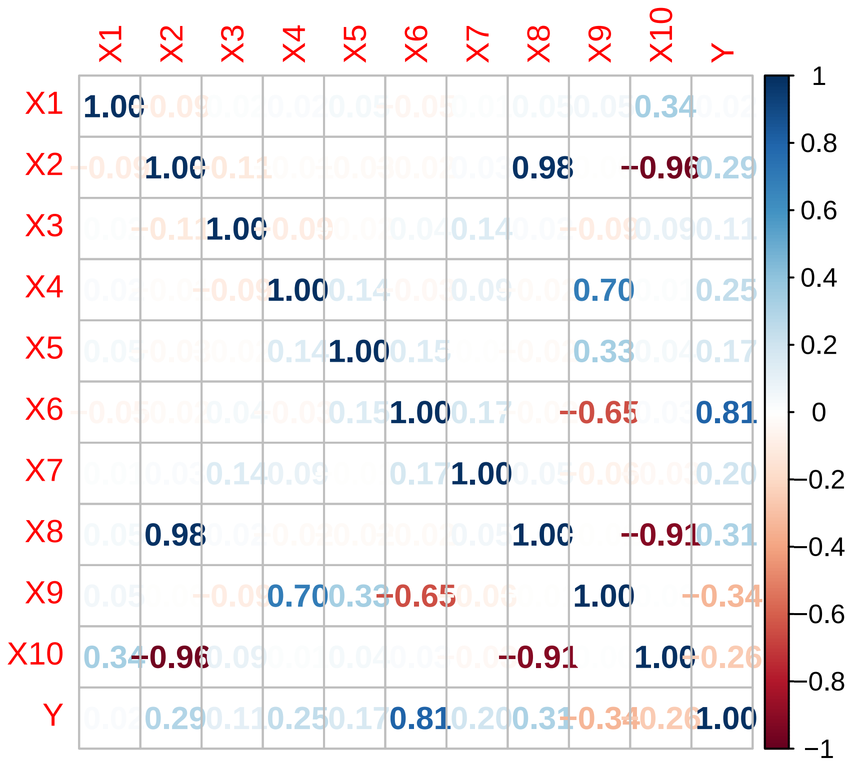
## 2

Wyznacz macierz korelacji próbkowych dla zmiennych $Y, X_1 \ldots X_{10}$ i po przeanalizowaniu tak otrzymanych współczynników ponownie odpowiedz na pytania (a) i (b) z poprzedniego zadania.

```
cor=cor(dane)
corrplot(cor,method = "color")
```




```
corrplot(cor,method = "number")
```




Ad (a) odpowiedź nie zmienia się. Ad (b) odpowiedź nie zmienia się.

# 3

Skonstruuj model regresji liniowej opisujący zależność między zmienną $Y$ a zmiennymi objaśniającymi

```
Y<-dane$Y
model<-lm(Y~.,dane)
```

## (a)

Wyznacz estymator najmniejszych kwadratów $\hat{\beta}$

```
beta=model$coefficients
print(beta)
```

```
## (Intercept)          X1          X2          X3          X4          X5
##  0.52260442  2.84867658  1.82514674  3.64879728  3.95371546  0.21928094
##          X6          X7          X8          X9         X10
## 11.00583662 -0.03279499 -0.14514568  0.13848213 -0.73124055
```

### (b)

Czy którakolwiek ze zmiennych objaśniających z tego (pełnego) modelu ma liniowy wpływ na zmienną objaśnianą? Odpowiedź uzasadnij podając p-wartość testu F.

```
summary<-summary(model)
p_value<- pf(summary$fstatistic[1],summary$fstatistic[2],summary$fstatistic[3],lower.tai
```

Ponieważ p-wartość testu F wynosi 3.21244213951607e-73 możemy wnioskować że któraś ze zmiennych objaśniających ma liniowy wpływ na zmienną objaśnianą $Y$.

### (c)

Wyznacz współczynniki determinacji $R^2$ i $AdjR^2$

```
R_2<-summary$r.squared
Adj_R_2<-summary$adj.r.squared
```

Współczynnik $R^2$ wynosi 0.853354393031298 Współczynnik $AdjR^2$ wynosi 0.845595366207557

## 4

Rozwiąż problem współliniowości

### (a)

Spośród zmiennych objaśniających, dla których $VIF$ przekracza 10 (lub równoważnie $TOL := 1/VIF < 0,1$), wybierz tę z największą wartością $VIF$ i usuń ją z modelu.

| | VIF.model. |
|---|---|
| X1 | 35.15 |
| X2 | 1497.68 |
| X3 | 27.26 |
| X4 | 36.09 |
| X5 | 11.76 |
| X6 | 42.03 |
| X7 | 1.09 |
| X8 | 1469.49 |
| X9 | 89.58 |
| X10 | 73.67 |

Należy usunąć zmienną $X2$.

## (b)

Oblicz *wskaźniki podbicia wariancji* w modelu regresji zawierającym *pozostałe* zmienne objaśniające. Jeśli któryś z tych wskaźników jest większy od 10 wróć do poprzedniego punktu.

```
tabela2<-data.frame(VIF(model2))

while(max(tabela2$VIF.model2.)>10){
index<-which.max(tabela2$VIF.model2.)
dane2<-dane2[-index]
model2<-lm(Y~.,dane2)
tabela2<-data.frame(VIF(model2))
}
xtab<-xtable(tabela2)

print(xtab,title="VIF", type = "latex", table.placement = "H", comment=FALSE)
```

| | VIF.model2. |
|---|---|
| X1 | 1.01 |
| X3 | 1.03 |
| X4 | 1.05 |
| X5 | 1.05 |
| X6 | 1.07 |
| X7 | 1.07 |
| X8 | 1.01 |

# 5

Zidentyfikuj i ewentualnie usuń z próby obserwacje, które mogą być wpływowe. W tym celu przeanalizuj

## (a)

*wpływy (leverages)* kolejnych obserwacji, czyli liczby $h_{11}, ..., h_{nn}$ tworzące główną przekątna macierzy $H$,

```
levreges<-lm.influence(model2)$hat
p<-7
n<-200
levreges.indexes<-which(levreges>(3*p)/n)
```

Indeksy obserwacji odstających to 175

### (b)

*odległości Cooke'a* $D_1, \dots D_n$

```
cook<-cooks.distance(model2)
cook.indexes<-which(cook>4/(n-p))
```

Indeksy obserwacji odstających to 100, 200

### (c)

*standaryzowane rezydua* $r_1 \dots r_n$,

```
stand.res<-abs(rstandard(model2))
stand.res.indexes<-which(stand.res>2)
```

Indeksy obserwacji odstających to 100, 200

### (d)

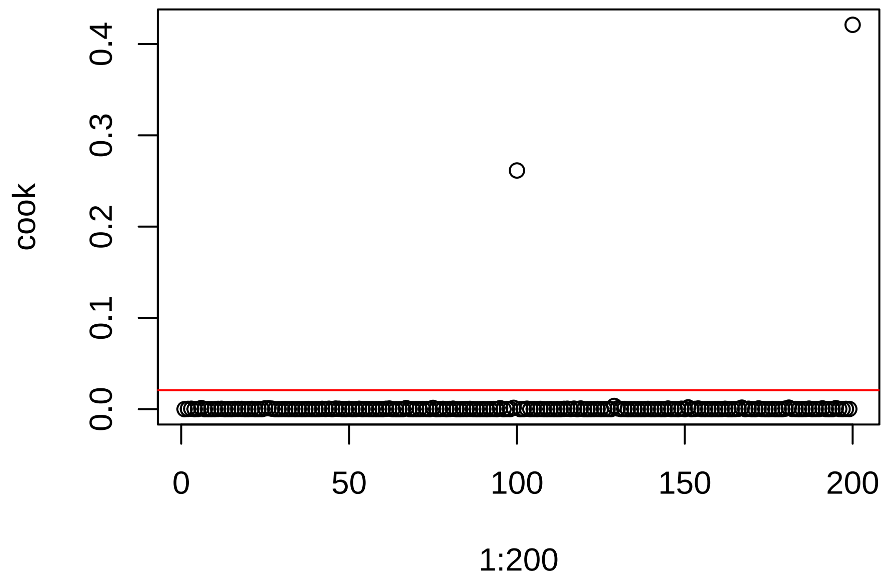
$DFFITS_1, \dots, DFFITS_n$.

```
dffits<-dffits(model2)
dffits.indexes<-which(dffits>2*sqrt((p+1)/(n-p-1)))
```

Indeksy obserwacji odstających to 100, 200

### wykres

```
plot(1:200,cook)
abline(h=4/(n-p),col='red')
```




### usunięcie obserwacji

```
dane2<-dane2[-100,]
dane2<-dane2[-199,]#200 obserwacja po usunięciu obserwacji 100 ma indeks 199
```

## 6

Wykorzystując zmienne i obserwacje, które nie zostały usunięte, ponownie zbuduj model regresji liniowej opisujący zależność między zmienną $Y$ a zmiennymi objaśniającymi.

### (a)

Wyznacz estymator najmniejszych kwadratów $\hat{\beta}$.

```
Y<-dane2$Y
model2<-lm(Y~.,dane2)
beta=model2$coefficients
```

$\hat{\beta}$ = 0.221466019895436, 0.0399808705552782, 2.00767562041835, 3.99856811088558, 0.0277321006720695, 10.9863346696194, 0.00157988372238295, 0.996283303910784

## (b)

Czy którakolwiek ze zmiennych objaśniających z tego (pełnego) modelu ma liniowy wpływ na zmienną objaśnianą? Odpowiedź uzasadnij podając p-wartość testu $F$.

```
summary<-summary(model2)
p_value<- pf(summary$fstatistic[1],summary$fstatistic[2],summary$fstatistic[3],lower.tai
```

p-wartość jest równa 0.

Ponieważ p-wartość testu $F$ obliczana jest w liczbach zmiennoprzecinkowych, została zaokrąglona do zera, możemy zatem stwierdzić że któraś ze zmiennych objaśniających ma liniowy wpływ na zmienną objaśnianą.

## (c)

Wyznacz współczynniki determinacji $R^2$ i $AdjR^2$. Czy po usunięciu niektórych zmiennych lub obserwacji polepszyło się dopasowanie modelu do danych?

```
R_2.2<-summary$r.squared
Adj_R_2.2<-summary$adj.r.squared
```

$R^2$ jest równa 0.999817004730811 > 0.853354393031298 , $AdjR^2$ jest równa 0.999810262799841 > 0.845595366207557, dopasowanie modelu poprawiło się.

# 7

Wykorzystaj *regresję krokową*, opcje *forward* i *backward* (w pakiecie do wyboru podzbioru zmiennych objaśniających „najlepiej" opisującego liniowy wpływ zmiennych objaśniających na zmienną $Y$ . Wykorzystaj także inne opcje regresji krokowej, dostępne w używanym przez Ciebie pakiecie. Oczywiście, przy tej analizie użyj zmodyfikowanych danych, powstałych po usunięciu niektórych zmiennych i niektórych obserwacji.

*Uwaga*: W ten sposób można otrzymać różne modele, więc do dalszej analizy *wybierz jeden z nich* (za pomocą współczynnika $C_p$ Mallowsa albo skorygowanego $R^2$) i nazwij go *modelem M*.

```
backward <- step(model2, direction='backward', scope=formula(model2), trace=0)
forward <- step(model2, direction='forward', scope=formula(model2), trace=0)
summary_back<-summary(backward)
summary_for<-summary(forward)
Adj_R_2._back<-summary_back$adj.r.squared
Adj_R_2._for<-summary_for$adj.r.squared
```

$Adj R^2$ w modelu *forward* wynosi 0.999810262799841. $Adj R^2$ w modelu *backward* wynosi 0.999810886170365. Modelem M zostaje model *backward*.

## (a)

Wyznacz estymator najmniejszych kwadratów $\hat{\beta}$ w modelu $M$.

```
model_M<-backward
beta_M=model_M$coefficients
```

$\hat{\beta}$ jest równy 0.0959075277385715, 0.0412929949552652, 2.00798285930852, 4.00101775452038, 10.9889282244157, 0.996242951373832.

## (b)

Czy którakolwiek ze zmiennych objaśniających z *modelu M* ma liniowy wpływ na zmienną objaśnianą? Odpowiedź uzasadnij podając p-value (p-wartość) odpowiedniego testu.

```
summary_M<-summary(model_M)
p_value_M<- pf(summary_M$fstatistic[1],summary_M$fstatistic[2],summary_M$fstatistic[3],lo
```

p-wartość testu $F$ wynosi 0 , (została zaokrąglona przy obliczeniach na liczbach zmiennoprzecinkowych), więc można wnioskować że któraś zmienna objaśniająca ma liniowy wpływ na zmienną objaśnianą.

## (c)

Dla każdej ze zmiennych objaśniających, które znalazły się w *modelu M*, sprawdź, czy ma ona liniowy wpływ na zmienną objaśnianą, gdy w modelu uwzględnione zostały pozostałe zmienne. Podaj p-wartość odpowiedniego testu i sformułuj wniosek.

```
print(anova(model_M))
```

```
## Analysis of Variance Table
##
## Response: Y
##            Df Sum Sq Mean Sq   F value    Pr(>F)
## X1          1     41      41    386.62 < 2.2e-16 ***
## X3          1   1296    1296  12126.36 < 2.2e-16 ***
## X4          1   8206    8206  76804.49 < 2.2e-16 ***
## X6          1  90327   90327 845413.51 < 2.2e-16 ***
```

```
## X8          1  11409   11409 106777.67 < 2.2e-16 ***
## Residuals 192     21       0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Można wnioskować że każda ze zmiennych objaśniających w *modelu M* ma liniowy wpływ na zmienną objaśnianą, ponieważ wartości $F$ testów -kolumna Pr(>F)- są mniejsze od 0.05.

## (d)

Wyznacz przedział ufności na poziomie ufności 0.95 dla współczynników regresji, odpowiadających zmiennym z *modelu M.*

```
con<-confint(model_M)
xtable(con)
```

% latex table generated in R 4.2.2 by xtable 1.8-4 package % Sun Apr 16 01:35:41 2023

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | -0.19 | 0.38 |
| X1 | -0.00 | 0.08 |
| X3 | 1.96 | 2.05 |
| X4 | 3.98 | 4.03 |
| X6 | 10.97 | 11.01 |
| X8 | 0.99 | 1.00 |

## (e)

Wyznacz współczynniki determinacji $R^2$ i $AdjR^2$ w *modelu M.*

```
R_2_m<-summary_M$r.squared
Adj_R_2_m<-summary_M$adj.r.squared
```

$R^2$ w modelu *modelu M* wynosi 0.999815686013756. $AdjR^2$ w modelu *modelu M* wynosi 0.999810886170365.

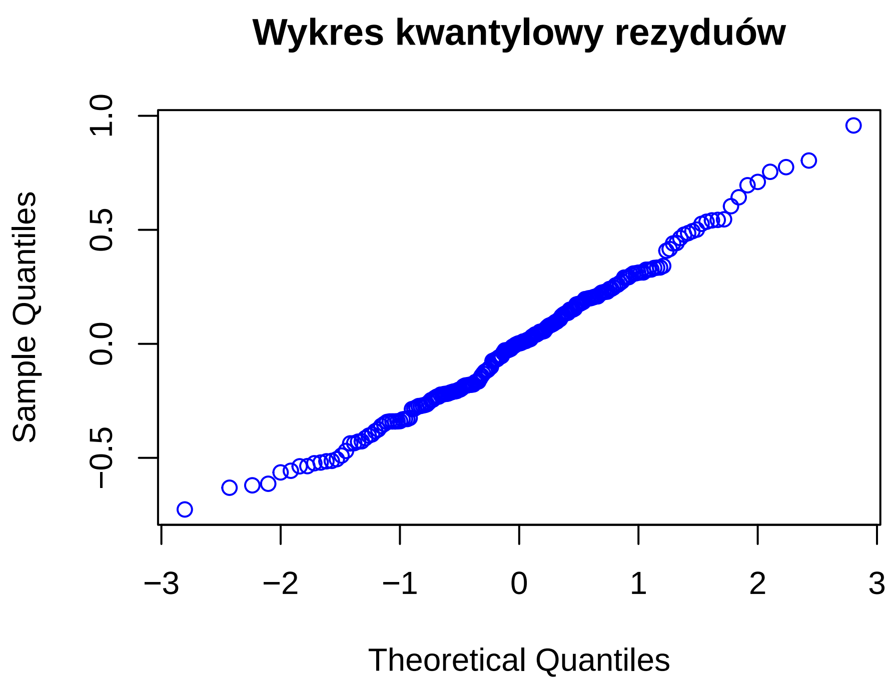
# 8

Przeanalizuj zachowanie reszt w *modelu M*, by sprawdzić czy spełnione są założenia występujące w modelu regresji liniowej. W tym celu należy wykonaj

### (a)

wykresy kwantylowe dla reszt,

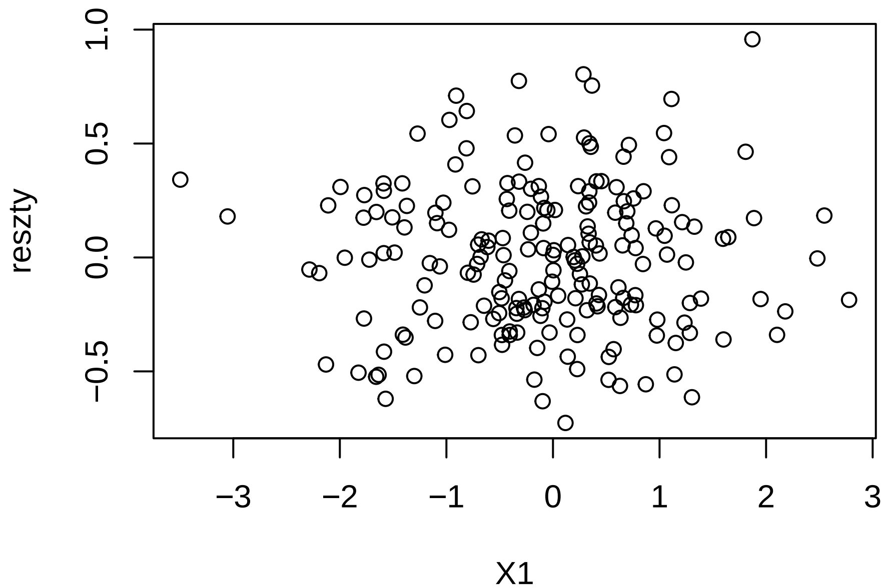
```
qqnorm(model_M$residuals,main="Wykres kwantylowy rezyduów",col = 'blue')
```



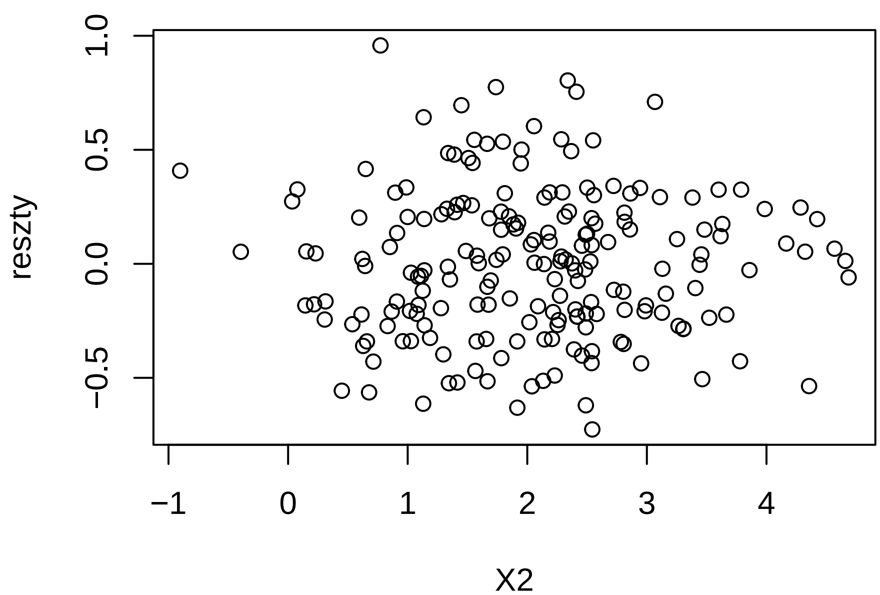

### (b)

wykresy reszt względem każdej ze zmiennych objaśniających,

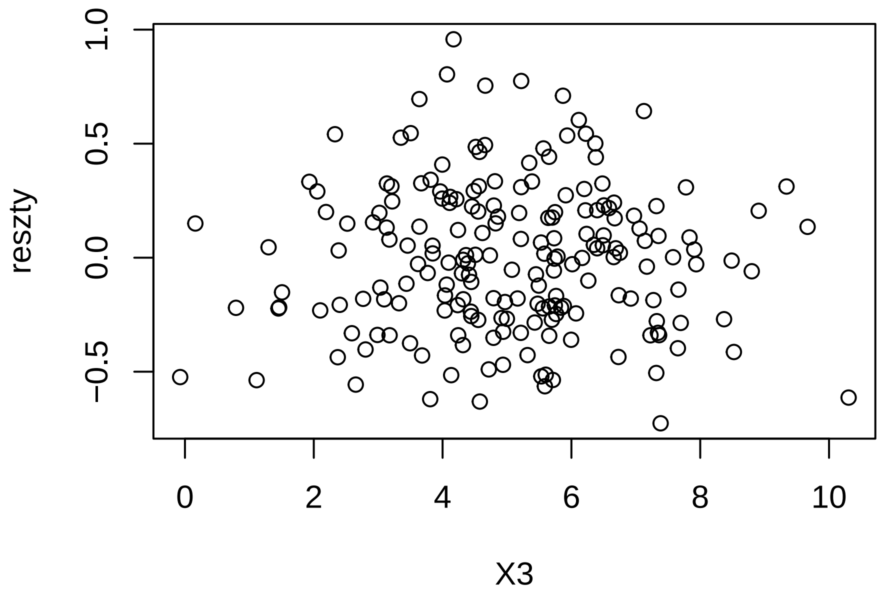
```
names<-names(dane2)
counter<-0
for (x in dane2[-8]){
  counter<-counter+1
plot(x,model_M$residuals,ylab="reszty",xlab=paste0("X",toString(counter)))}
```

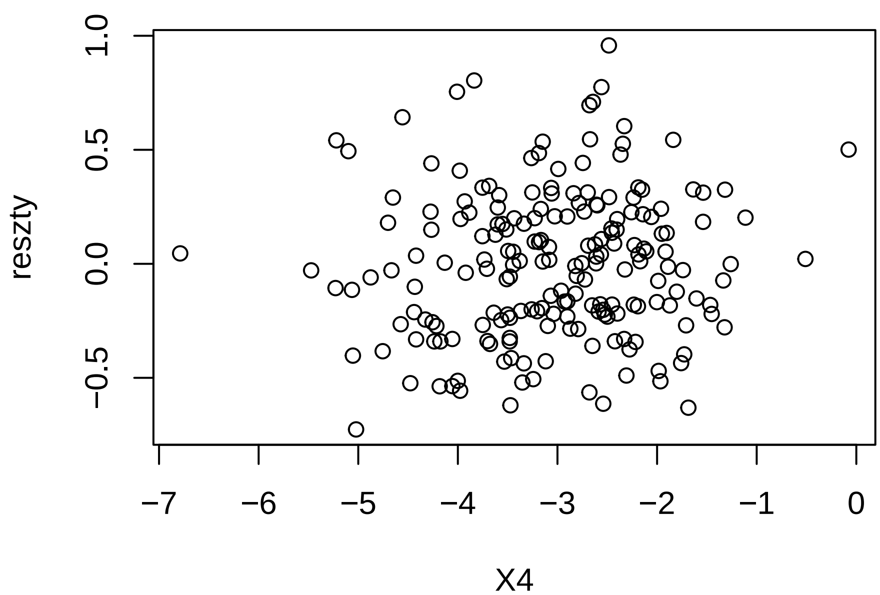


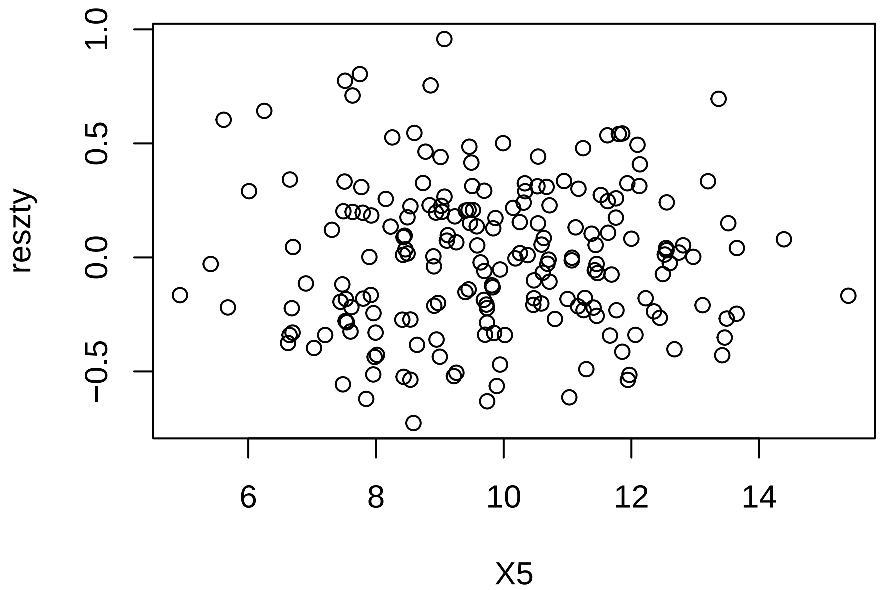
reszty
1.0
0.5
0.0
−0.5
−3
−2
−1
0
1
2
3
X1

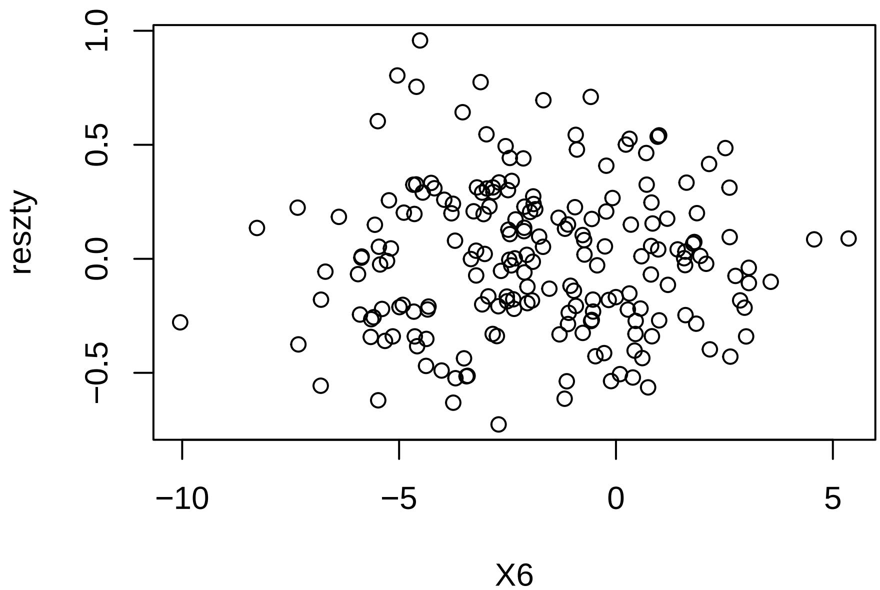



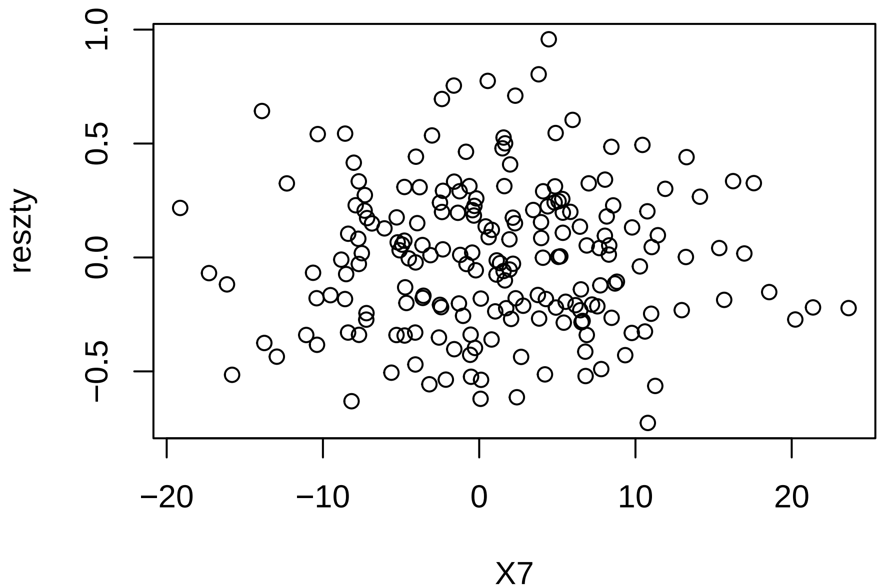
reszty
1.0
0.5
0.0
−0.5
−1
0
1
2
3
4
X2


reszty
1.0
0.5
0.0
−0.5
0
2
4
6
8
10
X3


reszty
1.0
0.5
0.0
−0.5
−7
−6
−5
−4
−3
−2
−1
0
X4


reszty
1.0
0.5
0.0
−0.5
6
8
10
12
14
X5

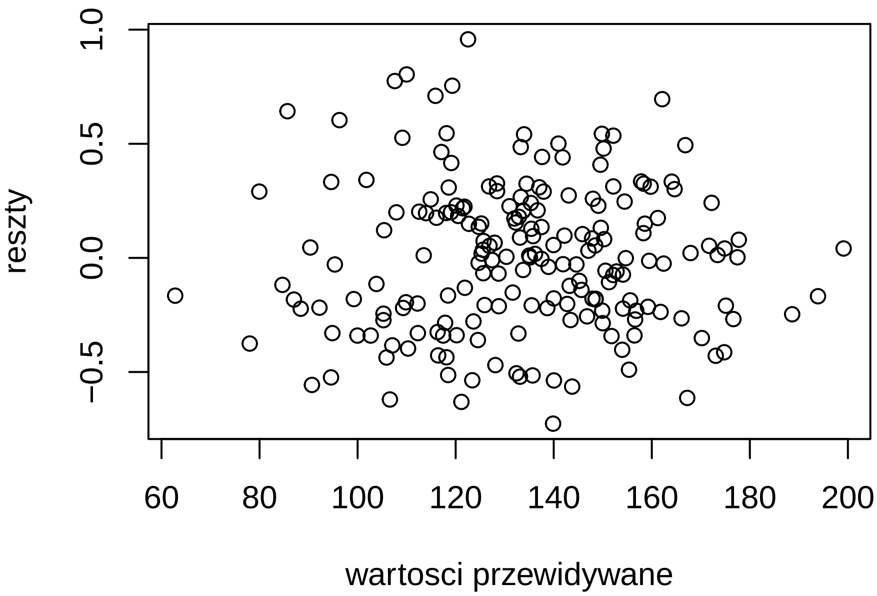

reszty
1.0
0.5
0.0
−0.5
−10
−5
0
5
X6




## (c)

wykresy reszt względem wartości przewidywanych przez model,

```
y_hat=predict(model_M,newdata=dane2)
plot(y_hat,model_M$residuals,ylab="reszty",xlab="wartości przewidywane")
```




Można wnioskować z wykresów że założenia modelu są spełnione.

# 9

Wyznacz przewidywaną przez *model M* wartość zmiennej objaśnianej $Y$, gdy zmienne objaśniające $X_1, X_2, \ldots, X_{10}$ mają wartość $1, 2, \ldots, 10$.

```
y_hat2=predict(model_M,newdata=data.frame(X1=c(1),X3=c(3),X4=c(4),X5=c(5),X6=c(6),X7=c(7)
```

Przewidywana przez *model M* wartość to 96.0687330761859.