Analiza i predykcja wieku w zależności od czynników wpływających na kondycje serca

Ewa Bojke, Jarosław Kuczyński

25 stycznia 2020

Zbiór danych

W naszym projekcie będziemy używali danych ze strony https://kaggle.com. Ta baza danych zawiera 76 atrybutów, ale wszystkie opublikowane eksperymenty odnoszą się do korzystania z podzbioru 14 z nich. W szczególności baza danych Cleveland jest jedyną, z której do tej pory korzystali naukowcy ML.

Zmiennne

zmienna	opis
`age`	wiek
`sex`	płeć
`cp`	indykator bólu w klatce piersiowej(od 0 do 1 wartości)
`trestbps`	spoczynkowe ciśnienie krwi
`chol`	surowica cholestoral w mg / dl
`fbs`	indykator czy cukier na czczo(0=nie, 1=tak)
`restecg`	indykator spoczynkowych wyników elektrokar(0,1)
`thalach`	maksymalne osiągnięte tętno
`exang`	indykator dusznicy bolesnej wywołanej wysiłkiem fizycznym(0=nie,1=tak)
`oldpeak`	depresja ST wywołana wysiłkiem w stosunku do odpoczynku

Sformułowanie problemu badawczego oraz agenda do zadań

W naszym projekcie

- przygotujemy dane do interpretacji tzw. czyszczenie danych i ich wczytanie
- przeanalizujemy zmienne i zależności między nimi tzw. eksploracyjna analiza danych
- stworzymy model predykcyjny wieku osób chorych
- sprawdzimy jaki model najbardziej pasuje do danych

Wczytywanie danych

```
heart = read.csv("C:/Users/Ewci/Desktop/wnioskowanie II/projekt/heart.csv")
colnames(heart)=c('wiek', 'plec', 'bol', 'cisnienie','cholesterol', 'cukier','spo.wyn','max tetno','dusznica','depresja ST')
```

Zamiana odpowiednich zmiennych na czynniki

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
heart <- heart %>% mutate(plec = factor(plec, levels = c(man=0, woman = 1), labels = c("man","woman"))) %>%
mutate(cukier = factor(cukier, levels = c(cukrzyk= 1, zdrowy = 0), labels = c("cukrzyk", "zdrowy")))%>% muta
te(dusznica = factor(dusznica, levels = c(dusznica= 1, `bez dusznicy` = 0),labels = c("dusznica", "bez duszn
icy")))%>% mutate(spo.wyn = factor(spo.wyn, levels = c(duze= 1, `male` = 0),labels = c("duze", "male")))%>%
mutate(bol = factor(bol, levels = c(bol= 1, `brak bolu` = 0),labels = c("bol", "brak bolu")))
```

Wnioski i obserwacje:

Zamieniliśmy zmienne numeryczne plec, cukier, dusznica, bol, spo.wyn na zmienne kategoryczne. Użyjemy ich do analizy wariancji w kolejnych etapach projektu.

Podstawowe statystyki dla zmiennych w naszym zbiorze danych

```
lapply(heart,summary)
```

```
## $wiek
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   47.50   55.00   54.37   61.00   77.00
##
## $plec
##   man woman
##    96   207
##
## $bol
##       bol brak bolu
##       180       123
##
## $cisnienie
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    94.0   120.0   130.0   131.6   140.0   200.0
##
## $cholesterol
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   126.0   211.0   240.0   245.4   274.5   409.0
##
## $cukier
## cukrzyk  zdrowy
##      45     258
##
## $spo.wyn
## duze male NA's
##  152  147    4
##
## $`max tetno`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.0   133.5   153.0   149.6   166.0   202.0
##
## $dusznica
##     dusznica bez dusznicy
##            99          204
##
## $`depresja ST`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.80    1.04    1.60    6.20
```

Wnioski i obserwacje:

Możemy zaobserwować, że wiek badanych ludzi waha się od 29 do 77 lat. Przy czym połowa ludzi ma wiek w przedziale 47.5 - 61 lat z medianą 55 lat. W badaniach mamy więcej kobiet niż mężczyzn.

Jeśli chodzi o ciśnienie skurczowe krwi, to możemy zauważyć, że wahania występują między 94mm a 200mm. Większa część badanych ma wyniki ciśnienia skurczowego między 120 a 140mm z medianą 130mm. Ogólnie O nadciśnieniu mówimy już w momencie, gdy ciśnienie skurczowe przekracza 140 mm .

Natomiast wyniki cholesterolu u zdrowego człowieka nie powinny przekraczać 190mg. W naszych danych wahania występują między 126mg a 409mg. Większość badanych ma wyniki w przedziale 211 mg - 274.5mg.

Ostatnią zmienną, którą zbadamy to tętno. Wahania występują między 71 a 202 uderzeń serca na minutę. Większa część osób ma tętno w przedziale (134,166) z medianą równą 153. Wynikiem zagrażającym zdrowiu jest przekroczenie 100 uderzeń serca na minutę.

Korelacje między zmiennymi

```
cor(heart$cisnienie,heart$cholesterol)
```

```
## [1] 0.1393339
```

```
cor(heart$cisnienie,heart$`max tetno`)
```

```
## [1] -0.04669773
```

```
cor(heart$cholesterol,heart$`max tetno`)
```

```
## [1] -0.01750271
```

```
cor(heart$cisnienie,heart$`depresja ST`)
```

```
## [1] 0.1932165
```

```
cor(heart$cholesterol,heart$`depresja ST`)
```

```
## [1] 0.05336997
```

```
cor(heart$`max tetno`,heart$`depresja ST`)
```

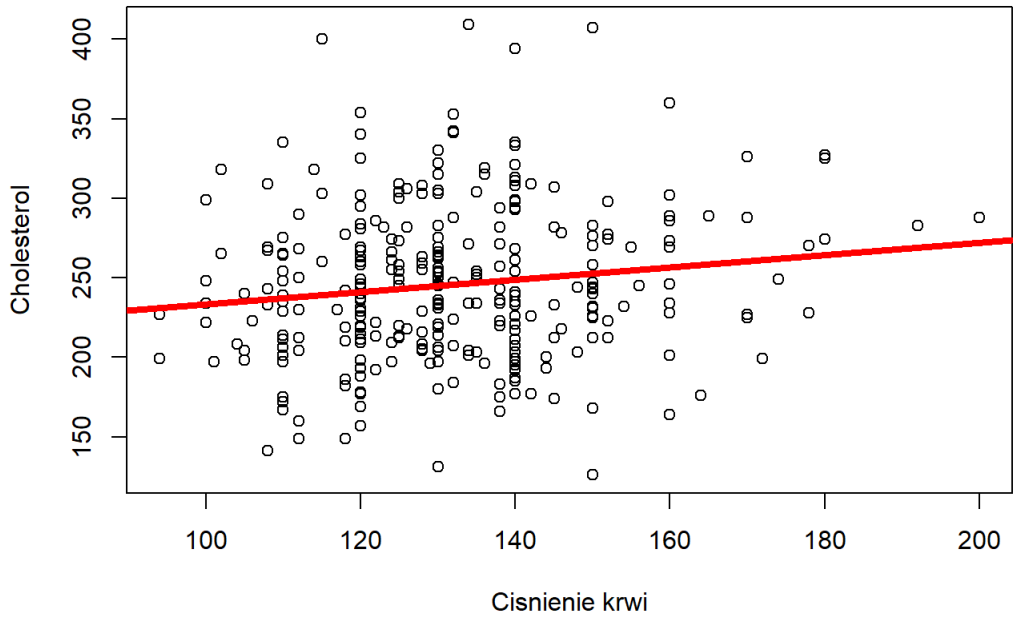
```
## [1] -0.3441869
```

Wnioski i obserwacje:

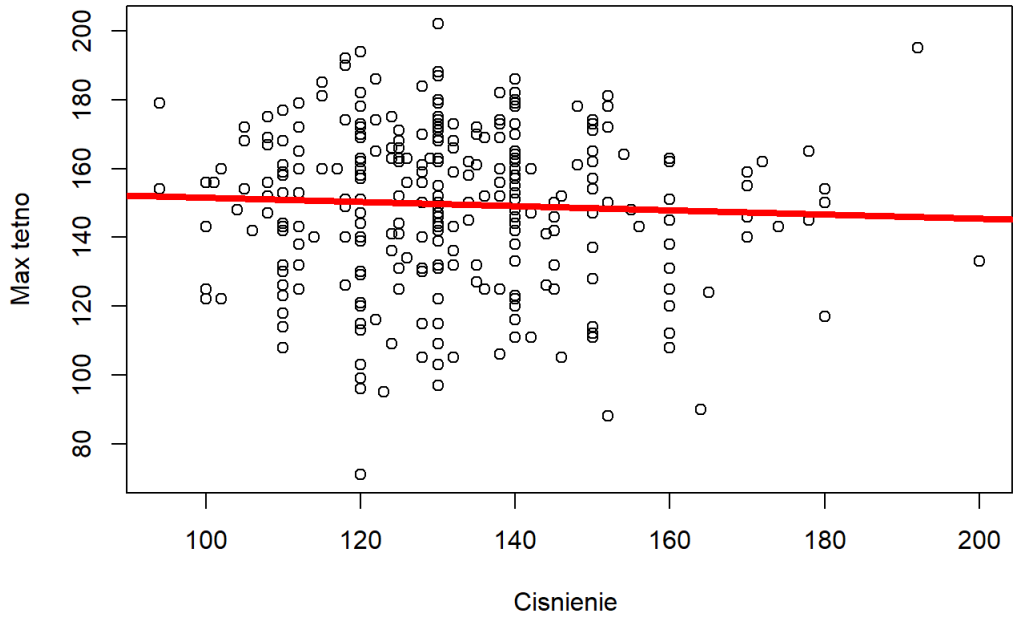
Większość zmiennych nie jest skorelowana. Jedynie mamy większą korelację przy zmiennych ciśnienie i depresja ST, przy zmiennych ciśnienie i cholesterol oraz przy zmiennych max tetno i depresja ST. Korelacja wynosi odpowiednio około 0.2, 0.14 i -0.34.

Przykładowe zależności między zmiennymi i modele liniowe dla zmiennych

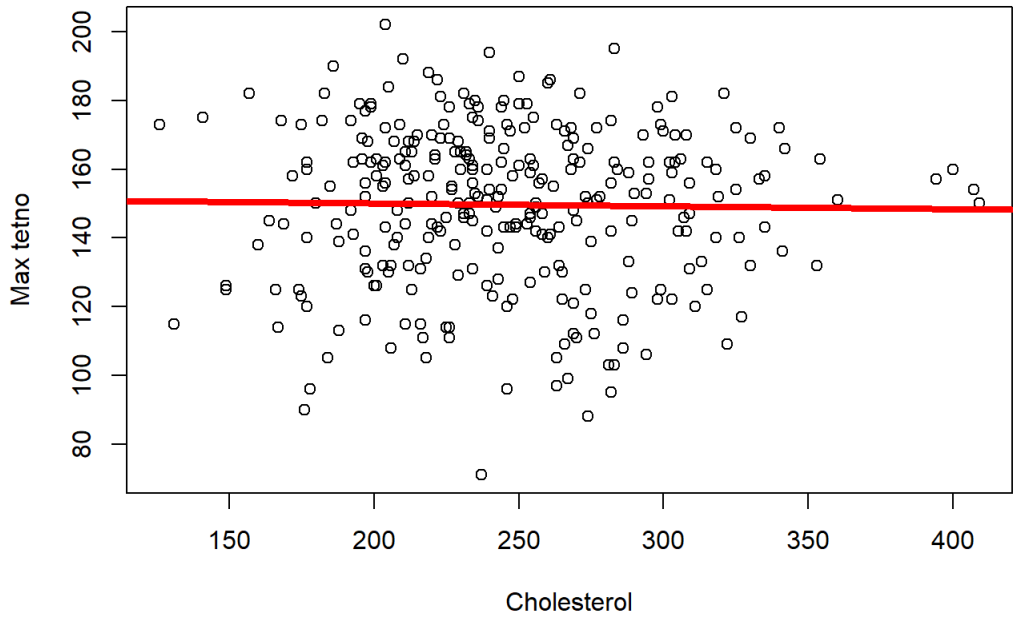
```
mod1=lm(heart$cholesterol~heart$cisnienie)
plot(heart$cisnienie,heart$cholesterol, xlab="Cisnienie krwi", ylab="Cholesterol")
abline(mod1,lwd=4,col='red')
```



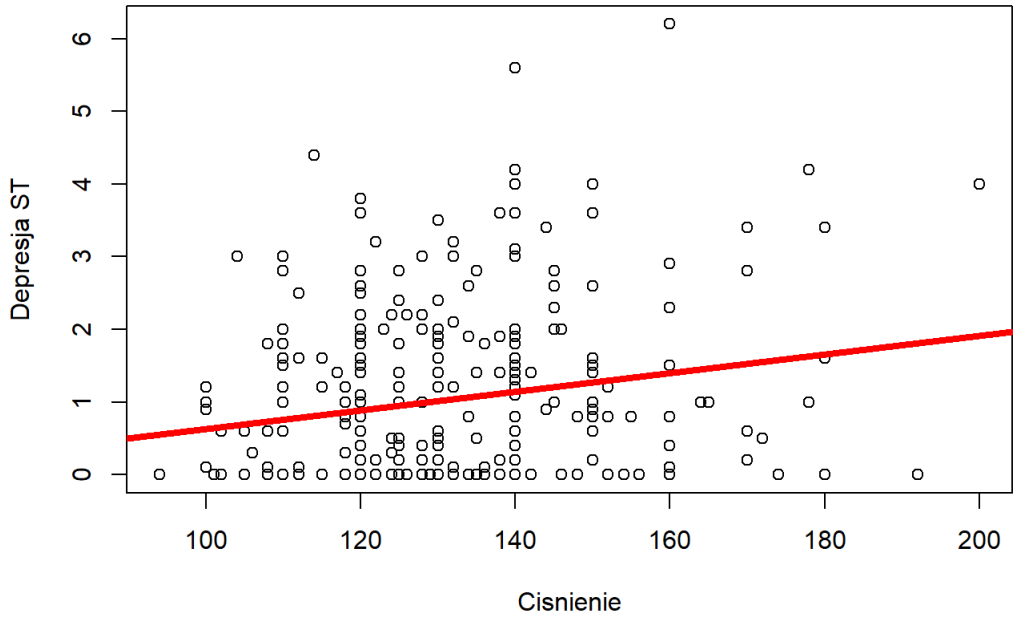

```
mod2=lm(heart$`max tetno`~heart$cisnienie)
plot(heart$cisnienie, heart$`max tetno`, xlab="Cisnienie ", ylab="Max tetno")
abline(mod2,lwd=4,col="red")
```



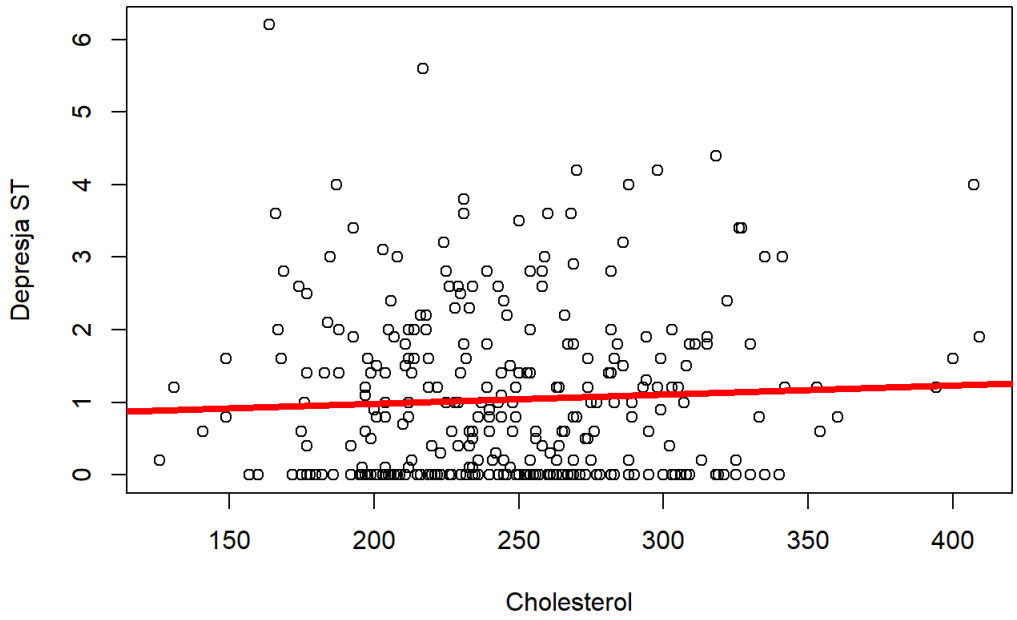

```
mod3=lm(heart$`max tetno`~heart$cholesterol)
plot(heart$cholesterol,heart$`max tetno`, xlab="Cholesterol", ylab="Max tetno")
abline(mod3,lwd=4,col="red")
```



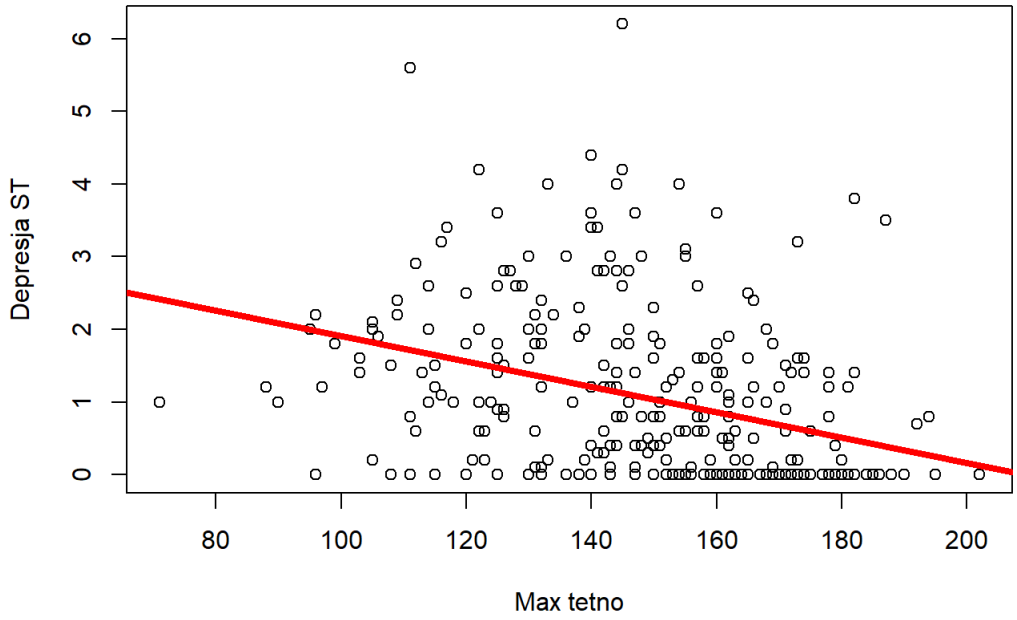

```
mod4=lm(heart$`depresja ST`~heart$cisnienie)
plot(heart$cisnienie,heart$`depresja ST`, xlab="Cisnienie",ylab="Depresja ST")
abline(mod4,lwd=4,col="red")
```




```
mod5=lm(heart$`depresja ST`~heart$cholesterol)
plot(heart$cholesterol,heart$`depresja ST`, xlab="Cholesterol", ylab="Depresja ST")
abline(mod5,lwd=4,col='red')
```




```
mod6=lm(heart$`depresja ST`~heart$`max tetno`)
plot(heart$`max tetno`,heart$`depresja ST`, xlab="Max tetno", ylab="Depresja ST")
abline(mod6,lwd=4,col='red')
```



Wnioski i obserwacje:

Widzimy,że jest zależność między zmiennymi, ale bardzo mała. Największą zależność możemy zauważyć między zmiennymi depresja ST i max tetno.

Badanie zależności

```
summary(mod1)
```

```
##
## Call:
## lm(formula = heart$cholesterol ~ heart$cisnienie)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.535  -32.476   -4.905   28.823  162.638
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     194.6659    20.9846   9.277   <2e-16 ***
## heart$cisnienie   0.3858     0.1580   2.441   0.0152 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48.17 on 301 degrees of freedom
## Multiple R-squared:  0.01941,    Adjusted R-squared:  0.01616
## F-statistic: 5.959 on 1 and 301 DF,  p-value: 0.01522
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = heart$`max tetno` ~ heart$cisnienie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -79.356 -15.051   2.742  16.425  52.254
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     157.67436    9.98472  15.792   <2e-16 ***
## heart$cisnienie  -0.06099    0.07520  -0.811    0.418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.92 on 301 degrees of freedom
## Multiple R-squared:  0.002181,   Adjusted R-squared:  -0.001134
## F-statistic: 0.6578 on 1 and 301 DF,  p-value: 0.418
```

```
summary(mod3)
```

```
##
## Call:
## lm(formula = heart$`max tetno` ~ heart$cholesterol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.717 -15.981   3.267  16.870  52.011
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       151.673203   6.800917  22.302   <2e-16 ***
## heart$cholesterol  -0.008256   0.027183  -0.304    0.762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.94 on 301 degrees of freedom
## Multiple R-squared:  0.0003063,  Adjusted R-squared:  -0.003015
## F-statistic: 0.09224 on 1 and 301 DF,  p-value: 0.7616
```

```
summary(mod4)
```

```
##
## Call:
## lm(formula = heart$`depresja ST` ~ heart$cisnienie)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8119 -0.9037 -0.3747  0.7091  4.7974
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.644059   0.497136  -1.296 0.196127
## heart$cisnienie  0.012791   0.003744   3.417 0.000721 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.141 on 301 degrees of freedom
## Multiple R-squared:  0.03733,    Adjusted R-squared:  0.03413
## F-statistic: 11.67 on 1 and 301 DF,  p-value: 0.0007214
```

```
summary(mod5)
```

```
##
## Call:
## lm(formula = heart$`depresja ST` ~ heart$cholesterol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1603 -0.9874 -0.2952  0.6363  5.2643
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.726398   0.344304   2.110   0.0357 *
## heart$cholesterol 0.001276   0.001376   0.927   0.3545
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.161 on 301 degrees of freedom
## Multiple R-squared:  0.002848,   Adjusted R-squared:  -0.0004644
## F-statistic: 0.8598 on 1 and 301 DF,  p-value: 0.3545
```

```
summary(mod6)
```

```
##
## Call:
## lm(formula = heart$`depresja ST` ~ heart$`max tetno`)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9756 -0.7683 -0.3356  0.5625  5.0793
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.650496   0.415282    8.79  < 2e-16 ***
## heart$`max tetno` -0.017447   0.002743   -6.36 7.48e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.092 on 301 degrees of freedom
## Multiple R-squared:  0.1185, Adjusted R-squared:  0.1155
## F-statistic: 40.45 on 1 and 301 DF,  p-value: 7.482e-10
```

Wnioski i obserwacje:

Jak widać większość zmiennych jest słabo zależne i są słabo skorelowane. Jedynie w ostanim modelu widzimy pewną zależność. Błąd standardowy reszt wynosi w tym przypadku tylko 1.141. Mod6 wyjaśnia około 11.5% wariancji depresji ST biorąc pod uwagę maksymalne tętno.

Modele w regresji wieloliniowej

Regresja wieloliniowa jest to wyjaśnianie wartości zmiennej zależnej za pomocą więcej niż jednego predyktora. Stworzymy model nie biorący pod uwagę zależności pomiędzy zmiennymi

```
model = lm(heart$wiek~+cholesterol+`max tetno`+cisnienie, heart)
summary(model)
```

```
##
## Call:
## lm(formula = heart$wiek ~ +cholesterol + `max tetno` + cisnienie,
##     data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.455  -5.689   0.275   6.044  23.518
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.324418   5.007422  10.649  < 2e-16 ***
## cholesterol  0.031023   0.009435   3.288  0.00113 **
## `max tetno` -0.152454   0.019829  -7.688 2.16e-13 ***
## cisnienie    0.123395   0.026149   4.719 3.65e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.884 on 299 degrees of freedom
## Multiple R-squared:  0.2539, Adjusted R-squared:  0.2464
## F-statistic: 33.92 on 3 and 299 DF,  p-value: < 2.2e-16
```

Wnioski i obserwacje:

Każdy wzrost o 1 mg cholesterolu sprawia, że średnia wartość wieku będzie wzrastała o 0.03 lat, czyli niecałe 11 dni.(zakładając, że pozostałe zmienne się nie zmieniają)

Każdy wzrost o 1mm ciśnienia krwi sprawia, że średnia wartość wieku będzie wzrastała o 0.12 lat, czyli niecałe 44 dni.(zakładając, że pozostałe zmienne się nie zmieniają)

Każdy wzrost o 1 uderzenie serca na minutę sprawia, że średnia wartość wieku będzie malała o 0.15lat, czyli niecałe 54 dni.(zakładając, że pozostałe zmienne się nie zmieniają)

Współczynnik determinacji wynosi 0.2464. Natomiast błąd standardowy reszt 7.884. Model wyjaśnia 24% wariancji wieku. W większości wyników mamy taką sytuację, że osoby mające coraz słabszą kondycję serca są starsze. Jeżeli osoby badane mają podwyższone tętno to wystąpi to przeważnie w młodszym wieku.

Dodajemy zmienną depresja ST i tworzymy model.

```
model2ST = lm(heart$wiek~+cholesterol+`max tetno`+cisnienie+`depresja ST`, heart)
summary(model2ST)
```

```
##
## Call:
## lm(formula = heart$wiek ~ +cholesterol + `max tetno` + cisnienie +
##     `depresja ST`, data = heart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.5365  -5.7926   0.3072   6.0870  23.6825
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   52.91945    5.08028  10.417  < 2e-16 ***
## cholesterol    0.03091    0.00945   3.271   0.0012 **
## `max tetno`   -0.14888    0.02113  -7.047 1.27e-11 ***
## cisnienie      0.12097    0.02664   4.542 8.11e-06 ***
## `depresja ST`  0.20977    0.42446   0.494   0.6215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.894 on 298 degrees of freedom
## Multiple R-squared:  0.2545, Adjusted R-squared:  0.2445
## F-statistic: 25.44 on 4 and 298 DF,  p-value: < 2.2e-16
```

Wnioski i obserwacje:

Tak jak wcześniej widzimy przeważnie średni wzrost zmiennej wiek. Zauważmy że nowa zmienna zmniejszyła współczynnik determinacji, który teraz wynosi 0.2445 i zwiększyła błąd standardowy reszt, który teraz wynosi 7.894. Możemy powiedzieć, że ten model nie jest lepszy od poprzedniego. Depresja ST jest skorelowana z ciśnieniem i z max tetno i dlatego model zbytnio się nie różni chociaż błąd standardowy reszt jest większy. Informacje, które są przechowywane w zmiennej ciśnienie i max tetno zawiera również zmienna z nimi skorelowana depresja ST.

Model nie biorący pod uwagę zależności pomiędzy wszystkimi zmiennymi

```
model.model= lm(heart$wiek~`max tetno`+cholesterol, heart)
summary(model.model)
```

```
##
## Call:
## lm(formula = heart$wiek ~ `max tetno` + cholesterol, data = heart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.6789  -6.3058   0.1261   6.3754  22.3905
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.676482   3.939287  17.434  < 2e-16 ***
## `max tetno` -0.156637   0.020500  -7.641 2.92e-13 ***
## cholesterol  0.037198   0.009669   3.847 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.159 on 300 degrees of freedom
## Multiple R-squared:  0.1984, Adjusted R-squared:  0.193
## F-statistic: 37.12 on 2 and 300 DF,  p-value: 3.949e-15
```

Wnioski i obserwacje:

Błąd standardowy reszt jest mniejszy w pierwszym modelu o nazwie “model” i wynosi 7.884 lat, natomiast w modelu “model.model” błąd standardowy reszt wynosi 8.159 lat i współczynnik determinacji jest sporo mniejszy w drugim modelu. Możemy stwierdzić, że model o nazwie “model.model” nie jest lepszy. Szukamy dalej bardziej dopasowanego modelu do naszych danych.

Analiza wariancji

Możemy starać się przewidywać wartość zmiennej wiek za pomocą zmiennej cukier. Może doprowadzić nas to do pewnych wniosków, np takich ,że im człowiek jest starszy tym jego podatność na cukrzyce wzrasta.

Zwróćmy jednak uwagę na istotną różnicę w stosunku do regresji prostej. Mianowicie, zmienna cukier nie jest liczbowa. W takiej sytuacji, gdy zmienną numeryczną przewidujemy za pomocą zmiennej kategorycznej, mówimy, że mamy do czynienia z analizą wariancji(ANOVA).

Hipoteza Ho:

Model i model.model nie różnią się statystycznie.

```
model1 = lm(heart$wiek~cholesterol+`max tetno`+cisnienie, heart)
model.model1 = lm(heart$wiek~cholesterol*`max tetno`*cisnienie, heart)
anova(model1,model.model1)
```

```
## Analysis of Variance Table
##
## Model 1: heart$wiek ~ cholesterol + `max tetno` + cisnienie
## Model 2: heart$wiek ~ cholesterol * `max tetno` * cisnienie
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    299 18585
## 2    295 17784  4    800.93 3.3215 0.01109 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wnioski i obserwacje:

P-value jest małe więc nie akceptujemy hipotezy Ho.

Analiza kowariancji

Zmienną numeryczną możemy wyjaśniać za pomocą innej zmiennej numerycznej i zmiennej kategorycznej jednocześnie.

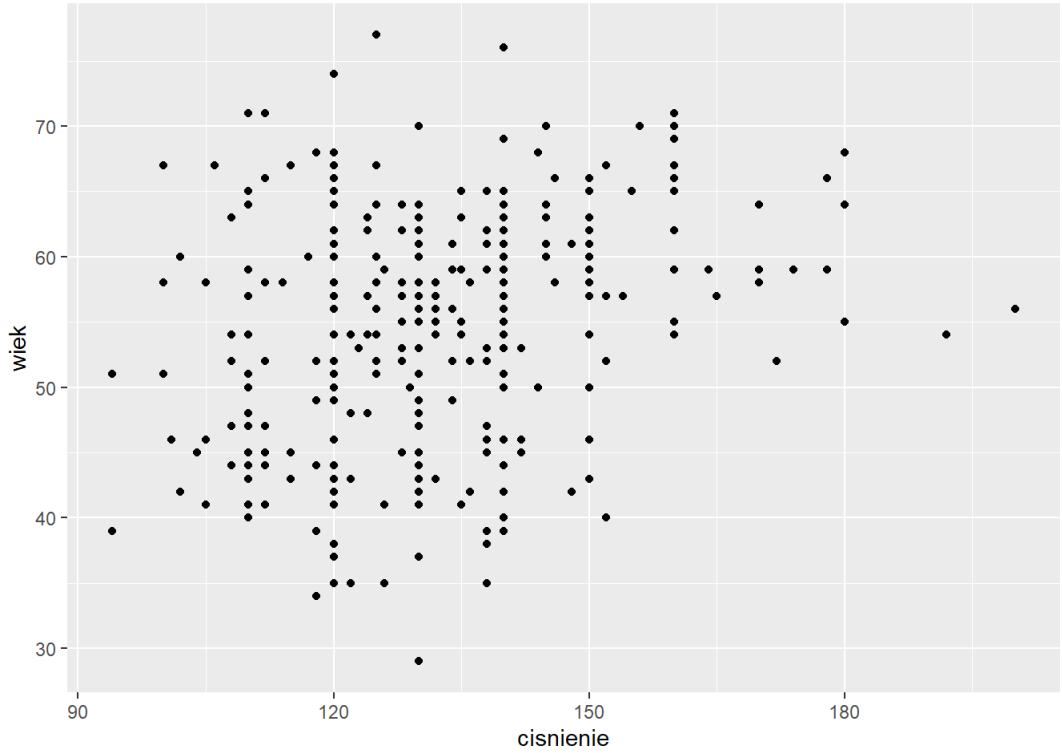
Jakie mamy predyktory?

Jednym z dobrych predyktorów wieku osoby jest cisnienie. Rzecz jasna, że osoby starsze mają większe ciśnienie niż osoby młodsze.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
ggplot(heart, aes(x =cisnienie,y = wiek)) + geom_point()
```




```
heart %>% filter(!is.na(cisnienie) & !is.na(wiek)) %>%
summarise(correlation = cor(cisnienie, wiek))
```

```
##   correlation
## 1   0.2793509
```

Wnioski i obserwacje:

Widzimy, że współczynnik korelacji jest istotny i wynosi około 0.3, jednakże nie jest bardzo duży, co sugeruje, że na wiek wpływają również inne czynniki.

Korelacje wieku z innymi zmiennymi numerycznymi

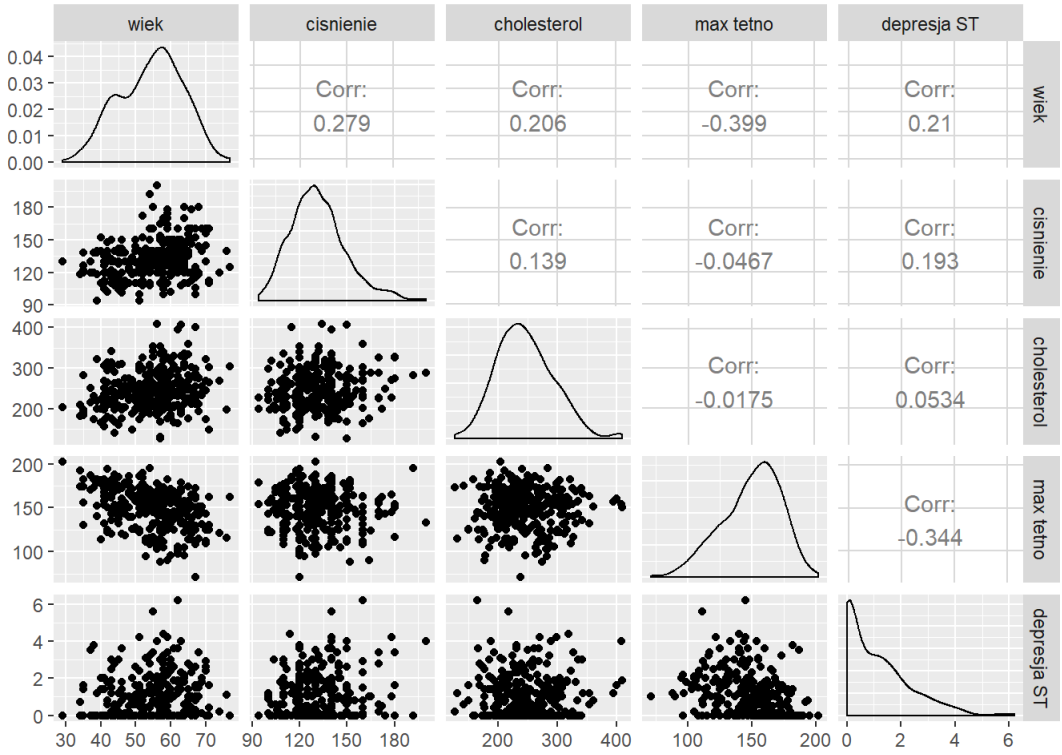
```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.5.3
```

```
##
## Attaching package: 'GGally'
```

```
## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ggpairs(heart %>% select_if(is.numeric))
```



Wnioski i obserwacje:

Możemy zauważyć, że dobrym predyktorem jest również cholesterol (współczynnik korelacji ponad 0.2), depresja ST (współczynik korelacji 0.21) i max tetno (wspolczynnik około -0.4).

Dwa modele liniowe i i ich porównanie

```
model.cisnienie<-lm(wiek~cisnienie,heart)
model.cukier<-lm(wiek~cukier,heart)
model.cs<-lm(wiek~cukier+cisnienie,heart)
summary(model.cisnienie)
```

```
##
## Call:
## lm(formula = wiek ~ cisnienie, data = heart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1314  -6.1441   0.5792   6.3559  23.5919
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.32545    3.80557   9.283  < 2e-16 ***
## cisnienie    0.14466    0.02866   5.048 7.76e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.735 on 301 degrees of freedom
## Multiple R-squared:  0.07804,    Adjusted R-squared:  0.07497
## F-statistic: 25.48 on 1 and 301 DF,  p-value: 7.762e-07
```

```
summary(model.cukier)
```

```
##
## Call:
## lm(formula = wiek ~ cukier, data = heart)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.907  -6.907   1.000   6.093  23.093
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    57.000      1.346   42.34   <2e-16 ***
## cukierzdrowy   -3.093      1.459   -2.12   0.0348 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.03 on 301 degrees of freedom
## Multiple R-squared:  0.01472,    Adjusted R-squared:  0.01144
## F-statistic: 4.496 on 1 and 301 DF,  p-value: 0.0348
```

```
summary(model.cs)
```

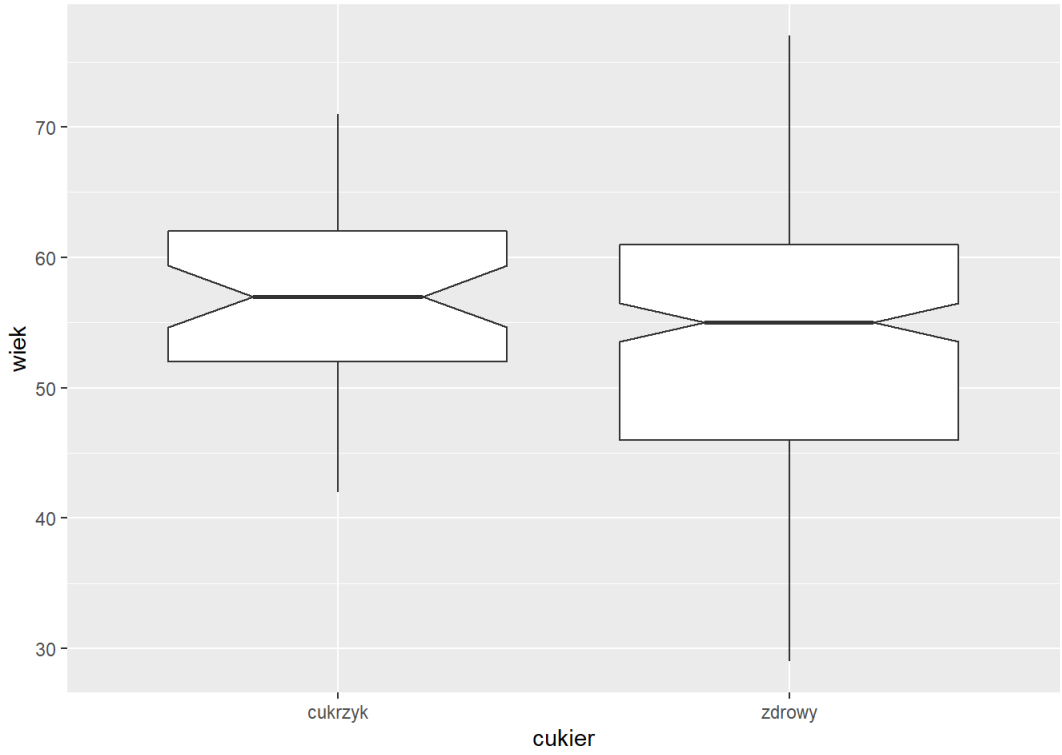
```
##
## Call:
## lm(formula = wiek ~ cukier + cisnienie, data = heart)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.8621  -6.2429   0.4928   6.0899  23.8272
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.82912    4.24896   8.903  < 2e-16 ***
## cukierzdrowy -1.88803    1.43216  -1.318    0.188
## cisnienie     0.13785    0.02909   4.739 3.32e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.724 on 300 degrees of freedom
## Multiple R-squared:  0.08335,    Adjusted R-squared:  0.07724
## F-statistic: 13.64 on 2 and 300 DF,  p-value: 2.142e-06
```

Wnioski i obserwacje:

Jak można było oczekiwać model zawierający dwie zmienne niezależne jest lepiej dopasowany od modeli z pojedynczymi zmiennymi. Współczynnik determinacji w modelu z dwoma zmiennymi niezależnymi wynosi około 8%. W modelach z pojedyńczą zmienną dużo mniej.

Jaki procent wariancji wyjaśnia model?

```
ggplot(heart %>% filter(!is.na(cukier)), aes(x = cukier, y = wiek)) +
geom_boxplot(notch = TRUE)
```



Wnioski i obserwacje:

Model z dwoma zmiennymi wyjaśnia około 8% wariancji wieku osoby badanej. Widzimy, że to czy ktoś jest cukrzykiem czy nie zależy od wieku.

Model o największej zdolności predykcji

Założenia: Staramy się, by model był jak najprostszy, gdyż nie zależy nam by dopasował się za dobrze do naszych danych.

Jednym z typowych kryteriów używanych do wyboru modelu jest zmodyfikowany współczynnik determinacji. Zawiera on w sobie informację o liczbie parametrów. Nie zawsze rośnie przy bardziej skomplikowanych modelach. Lepszą metodą jest tzw. kryterium informacyjne. AIC sprawdza się dobrze w modelach predykcyjnych.

Rozważymy pełny model, z wszystkimi możliwymi interakcjami.

```
model.full<-lm(wiek~bol*cisnienie*cholesterol*`max tetno`*`depresja ST`+cukier*dusznica,na.omit(heart))
summary(model.full)
```

```
##
## Call:
## lm(formula = wiek ~ bol * cisnienie * cholesterol * `max tetno` *
##     `depresja ST` + cukier * dusznica, data = na.omit(heart))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9907  -3.6565  -0.2336   3.1868  18.8583
##
## Coefficients:
##                                                          Estimate
## (Intercept)                                             5.156e+01
## bolbrak bolu                                            3.930e+02
## cisnienie                                               4.889e-02
## cholesterol                                             1.586e-01
## `max tetno`                                            -7.397e-02
## `depresja ST`                                           6.371e+01
## cukierzdrowy                                            1.996e+00
## dusznicabez dusznicy                                    3.275e+00
## bolbrak bolu:cisnienie                                 -2.901e+00
## bolbrak bolu:cholesterol                               -1.775e+00
## cisnienie:cholesterol                                  -7.395e-04
## bolbrak bolu:`max tetno`                               -2.644e+00
## cisnienie:`max tetno`                                   1.136e-04
## cholesterol:`max tetno`                                -7.945e-04
## bolbrak bolu:`depresja ST`                             -7.610e+02
## cisnienie:`depresja ST`                                 4.055e-01
```

```
## cisnienie:depresja ST                                          -4.055e-01
## cholesterol:`depresja ST`                                      -2.867e-01
## `max tetno`:`depresja ST`                                      -3.954e-01
## cukierzdrowy:dusznicabez dusznicy                              -2.677e+00
## bolbrak bolu:cisnienie:cholesterol                              1.260e-02
## bolbrak bolu:cisnienie:`max tetno`                              1.901e-02
## bolbrak bolu:cholesterol:`max tetno`                            1.169e-02
## cisnienie:cholesterol:`max tetno`                               4.416e-06
## bolbrak bolu:cisnienie:`depresja ST`                            5.879e+00
## bolbrak bolu:cholesterol:`depresja ST`                          3.014e+00
## cisnienie:cholesterol:`depresja ST`                             1.721e-03
## bolbrak bolu:`max tetno`:`depresja ST`                          5.414e+00
## cisnienie:`max tetno`:`depresja ST`                             2.614e-03
## cholesterol:`max tetno`:`depresja ST`                           1.812e-03
## bolbrak bolu:cisnienie:cholesterol:`max tetno`                 -8.317e-05
## bolbrak bolu:cisnienie:cholesterol:`depresja ST`               -2.302e-02
## bolbrak bolu:cisnienie:`max tetno`:`depresja ST`               -4.197e-02
## bolbrak bolu:cholesterol:`max tetno`:`depresja ST`             -2.156e-02
## cisnienie:cholesterol:`max tetno`:`depresja ST`                -1.115e-05
## bolbrak bolu:cisnienie:cholesterol:`max tetno`:`depresja ST`    1.652e-04
##                                                                Std. Error
## (Intercept)                                                     2.184e+02
## bolbrak bolu                                                    4.213e+02
## cisnienie                                                       1.620e+00
## cholesterol                                                     8.491e-01
## `max tetno`                                                     1.460e+00
## `depresja ST`                                                   1.889e+02
## cukierzdrowy                                                    1.745e+00
## dusznicabez dusznicy                                            1.915e+00
## bolbrak bolu:cisnienie                                          3.276e+00
## bolbrak bolu:cholesterol                                        1.755e+00
## cisnienie:cholesterol                                           6.244e-03
## bolbrak bolu:`max tetno`                                        2.694e+00
## cisnienie:`max tetno`                                           1.080e-02
## cholesterol:`max tetno`                                         5.614e-03
## bolbrak bolu:`depresja ST`                                      3.409e+02
## cisnienie:`depresja ST`                                         1.391e+00
## cholesterol:`depresja ST`                                       7.099e-01
## `max tetno`:`depresja ST`                                       1.307e+00
## cukierzdrowy:dusznicabez dusznicy                               2.110e+00
## bolbrak bolu:cisnienie:cholesterol                              1.354e-02
## bolbrak bolu:cisnienie:`max tetno`                              2.087e-02
## bolbrak bolu:cholesterol:`max tetno`                            1.125e-02
## cisnienie:cholesterol:`max tetno`                               4.120e-05
## bolbrak bolu:cisnienie:`depresja ST`                            2.590e+00
## bolbrak bolu:cholesterol:`depresja ST`                          1.417e+00
## cisnienie:cholesterol:`depresja ST`                             5.210e-03
## bolbrak bolu:`max tetno`:`depresja ST`                          2.347e+00
## cisnienie:`max tetno`:`depresja ST`                             9.596e-03
## cholesterol:`max tetno`:`depresja ST`                           4.907e-03
## bolbrak bolu:cisnienie:cholesterol:`max tetno`                  8.660e-05
## bolbrak bolu:cisnienie:cholesterol:`depresja ST`                1.068e-02
## bolbrak bolu:cisnienie:`max tetno`:`depresja ST`                1.779e-02
## bolbrak bolu:cholesterol:`max tetno`:`depresja ST`              9.766e-03
## cisnienie:cholesterol:`max tetno`:`depresja ST`                 3.591e-05
## bolbrak bolu:cisnienie:cholesterol:`max tetno`:`depresja ST`    7.357e-05
##                                                                t value
## (Intercept)                                                      0.236
## bolbrak bolu                                                     0.933
## cisnienie                                                        0.030
## cholesterol                                                      0.187
## `max tetno`                                                     -0.051
## `depresja ST`                                                    0.337
## cukierzdrowy                                                     1.144
## dusznicabez dusznicy                                             1.710
## bolbrak bolu:cisnienie                                          -0.885
## bolbrak bolu:cholesterol                                        -1.011
## cisnienie:cholesterol                                           -0.118
## bolbrak bolu:`max tetno`                                        -0.981
## cisnienie:`max tetno`                                            0.011
## cholesterol:`max tetno`                                         -0.142
## bolbrak bolu:`depresja ST`                                      -2.232
## cisnienie:`depresja ST`                                         -0.292
```

```
## cholesterol:`depresja ST`                                      -0.404
## `max tetno`:`depresja ST`                                      -0.302
## cukierzdrowy:dusznicabez dusznicy                              -1.268
## bolbrak bolu:cisnienie:cholesterol                              0.931
## bolbrak bolu:cisnienie:`max tetno`                              0.911
## bolbrak bolu:cholesterol:`max tetno`                            1.039
## cisnienie:cholesterol:`max tetno`                               0.107
## bolbrak bolu:cisnienie:`depresja ST`                            2.270
## bolbrak bolu:cholesterol:`depresja ST`                          2.128
## cisnienie:cholesterol:`depresja ST`                             0.330
## bolbrak bolu:`max tetno`:`depresja ST`                          2.307
## cisnienie:`max tetno`:`depresja ST`                             0.272
## cholesterol:`max tetno`:`depresja ST`                           0.369
## bolbrak bolu:cisnienie:cholesterol:`max tetno`                 -0.960
## bolbrak bolu:cisnienie:cholesterol:`depresja ST`               -2.156
## bolbrak bolu:cisnienie:`max tetno`:`depresja ST`               -2.359
## bolbrak bolu:cholesterol:`max tetno`:`depresja ST`             -2.208
## cisnienie:cholesterol:`max tetno`:`depresja ST`                 -0.310
## bolbrak bolu:cisnienie:cholesterol:`max tetno`:`depresja ST`    2.245
##                                                               Pr(>|t|)
## (Intercept)                                                     0.8136
## bolbrak bolu                                                    0.3518
## cisnienie                                                       0.9759
## cholesterol                                                     0.8520
## `max tetno`                                                     0.9596
## `depresja ST`                                                   0.7362
## cukierzdrowy                                                    0.2538
## dusznicabez dusznicy                                            0.0884 .
## bolbrak bolu:cisnienie                                          0.3767
## bolbrak bolu:cholesterol                                        0.3127
## cisnienie:cholesterol                                           0.9058
## bolbrak bolu:`max tetno`                                        0.3273
## cisnienie:`max tetno`                                           0.9916
## cholesterol:`max tetno`                                         0.8876
## bolbrak bolu:`depresja ST`                                      0.0264 *
## cisnienie:`depresja ST`                                         0.7708
## cholesterol:`depresja ST`                                       0.6867
## `max tetno`:`depresja ST`                                       0.7626
## cukierzdrowy:dusznicabez dusznicy                               0.2058
## bolbrak bolu:cisnienie:cholesterol                              0.3527
## bolbrak bolu:cisnienie:`max tetno`                              0.3633
## bolbrak bolu:cholesterol:`max tetno`                            0.2996
## cisnienie:cholesterol:`max tetno`                               0.9147
## bolbrak bolu:cisnienie:`depresja ST`                            0.0240 *
## bolbrak bolu:cholesterol:`depresja ST`                          0.0343 *
## cisnienie:cholesterol:`depresja ST`                             0.7414
## bolbrak bolu:`max tetno`:`depresja ST`                          0.0219 *
## cisnienie:`max tetno`:`depresja ST`                             0.7855
## cholesterol:`max tetno`:`depresja ST`                           0.7123
## bolbrak bolu:cisnienie:cholesterol:`max tetno`                  0.3377
## bolbrak bolu:cisnienie:cholesterol:`depresja ST`                0.0320 *
## bolbrak bolu:cisnienie:`max tetno`:`depresja ST`                0.0190 *
## bolbrak bolu:cholesterol:`max tetno`:`depresja ST`              0.0281 *
## cisnienie:cholesterol:`max tetno`:`depresja ST`                 0.7565
## bolbrak bolu:cisnienie:cholesterol:`max tetno`:`depresja ST`    0.0256 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.84 on 264 degrees of freedom
## Multiple R-squared:  0.6314, Adjusted R-squared:  0.5839
## F-statistic:  13.3 on 34 and 264 DF,  p-value: < 2.2e-16
```

Wnioski i obserwacje:

Większość zmiennych jest nieistotna statystycznie. Jest tak dlatego, że informacja została podzielona pomiędzy mnóstwo predyktorów i interakcji pomiędzy nimi.

Stwórzmy sensowniejszy model. Dodajemy wszystkie dostępne zmienne. A zwłaszcza zmienną cisnienie, zmienną max tetno i zmienną cholesterol. Są one najbardziej skorelowane ze zmienną wiek.

```
model.zlozony <-lm(wiek~cholesterol+cisnienie+spo.wyn+`depresja ST`+`max tetno`+bol+dusznica+cukier:cisnieni
e,na.omit(heart))
summary(model.zlozony)
```

```
##
## Call:
## lm(formula = wiek ~ cholesterol + cisnienie + spo.wyn + `depresja ST` +
##     `max tetno` + bol + dusznica + cukier:cisnienie, data = na.omit(heart))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6509  -4.2784  -0.2256   3.4514  19.0628
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             58.607050   3.830992  15.298  < 2e-16 ***
## cholesterol              0.020423   0.007206   2.834  0.00492 **
## cisnienie                0.050915   0.020753   2.453  0.01474 *
## spo.wynmale              0.601689   0.695491   0.865  0.38768
## `depresja ST`            0.226798   0.326355   0.695  0.48765
## `max tetno`             -0.087372   0.017162  -5.091 6.41e-07 ***
## bolbrak bolu           -11.490022   0.754672 -15.225  < 2e-16 ***
## dusznicabez dusznicy     1.473818   0.795465   1.853  0.06493 .
## cisnienie:cukierzdrowy   0.002479   0.007009   0.354  0.72387
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.83 on 290 degrees of freedom
## Multiple R-squared:  0.5964, Adjusted R-squared:  0.5853
## F-statistic: 53.58 on 8 and 290 DF,  p-value: < 2.2e-16
```

Uprośćmy teraz model przy użyciu AIC.

```
model.zlozony.aic <- step(model.zlozony)
```

```
## Start:  AIC=1063.16
## wiek ~ cholesterol + cisnienie + spo.wyn + `depresja ST` + `max tetno` +
##     bol + dusznica + cukier:cisnienie
##
##                    Df Sum of Sq     RSS    AIC
## - cisnienie:cukier  1       4.3  9861.4 1061.3
## - `depresja ST`     1      16.4  9873.6 1061.7
## - spo.wyn           1      25.4  9882.6 1061.9
## <none>                           9857.2 1063.2
## - dusznica          1     116.7  9973.9 1064.7
## - cholesterol       1     273.0 10130.2 1069.3
## - `max tetno`       1     881.0 10738.2 1086.8
## - bol               1    7879.2 17736.4 1236.8
##
## Step:  AIC=1061.29
## wiek ~ cholesterol + cisnienie + spo.wyn + `depresja ST` + `max tetno` +
##     bol + dusznica
##
##                 Df Sum of Sq     RSS    AIC
## - `depresja ST`  1      16.9  9878.4 1059.8
## - spo.wyn        1      24.5  9886.0 1060.0
## <none>                        9861.4 1061.3
## - dusznica       1     118.2  9979.6 1062.8
## - cisnienie      1     215.4 10076.9 1065.8
## - cholesterol    1     277.3 10138.7 1067.6
## - `max tetno`    1     886.0 10747.5 1085.0
## - bol            1    7994.3 17855.8 1236.8
##
## Step:  AIC=1059.8
## wiek ~ cholesterol + cisnienie + spo.wyn + `max tetno` + bol +
##     dusznica
##
##               Df Sum of Sq     RSS    AIC
## - spo.wyn      1      27.4  9905.8 1058.6
## <none>                       9878.4 1059.8
## - dusznica     1     106.1  9984.5 1061.0
## - cisnienie    1     242.8 10121.2 1065.1
## - cholesterol  1     274.1 10152.5 1066.0
## - `max tetno`  1    1002.8 10881.2 1086.7
## - bol          1    8021.7 17900.0 1235.5
##
## Step:  AIC=1058.63
## wiek ~ cholesterol + cisnienie + `max tetno` + bol + dusznica
##
##               Df Sum of Sq     RSS    AIC
## <none>                       9905.8 1058.6
## - dusznica     1     101.5 10007.3 1059.7
## - cisnienie    1     263.4 10169.2 1064.5
## - cholesterol  1     305.3 10211.1 1065.7
## - `max tetno`  1    1020.5 10926.3 1086.0
## - bol          1    8079.9 17985.8 1235.0
```

Wnioski i obserwacje:

Jak widzimy aby model był dokładniejszy wymagane było usunięcie zmiennych depresja ST, cukier i spo.wyn. Ponadto wybrany model okazał się mieć niższe AIC niż model pełny.

```
extractAIC(model.full)
```

```
## [1]   35.000 1088.109
```

```
extractAIC(model.zlozony.aic)
```

```
## [1]    6.00 1058.63
```

Wnioski i obserwacje:

Znaleźliśmy najlepiej dopasowany model do naszych danych. Jest to model.zlozony.aic.

Wykresy diagnostyczne najlepszego modelu

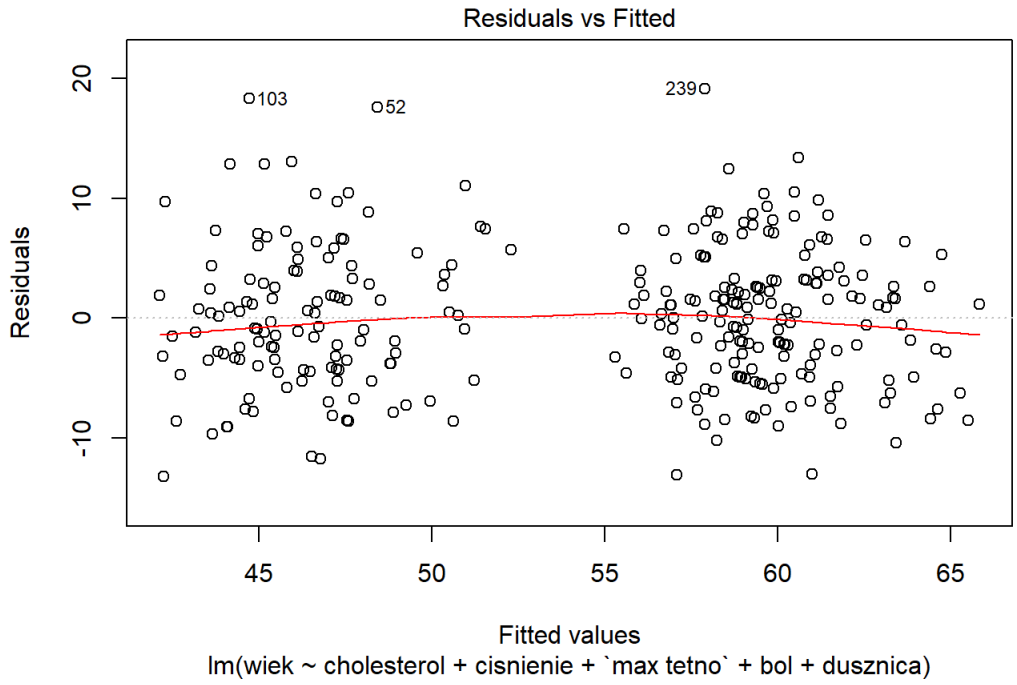
Pierwszy wykres przedstawia reszty wykreślone względem wartości dopasowanych. Punkty powinny być równomiernie rozrzucone względem osi `x`. Czerwona linia (wygładzenie) wskazuje na lekkie niedopasowane modelu dla mniejszych wartości.

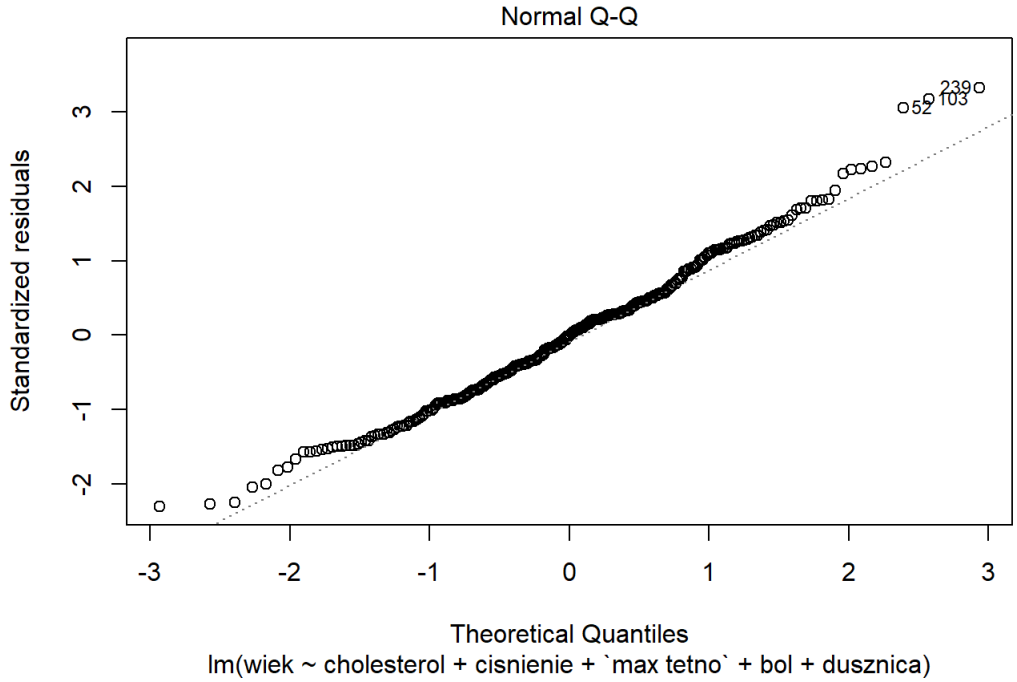
Drugim wykresem jest wykres kwantyl-kwantyl standaryzowanych reszt. Reszty standaryzowane, są to reszty podzielone przez odchylenie standardowe: $$t_i = \frac{e_i}{s \sqrt{1 - h_{ii}} }$$ Wykres kwantyl-kwantyl powinien przypominać linię prostą. Tutaj widzimy pewne niedopasowanie w przypadku bardziej ekstremalnych obserwacji.

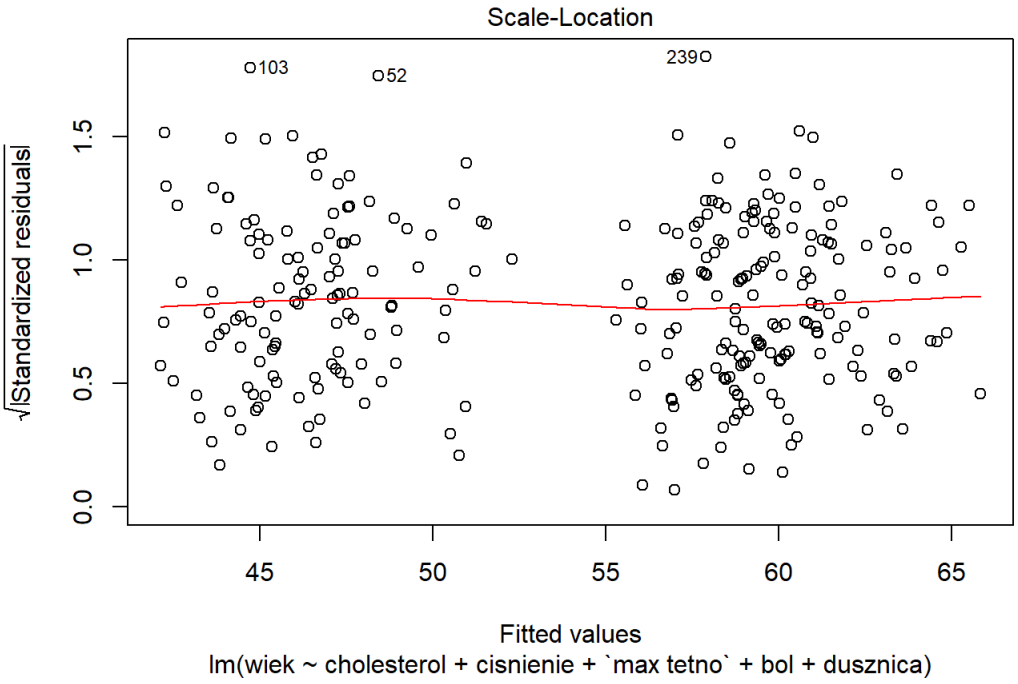
Trzeci wykres jest wygodny do badania homoskedastyczności (czerwona krzywa powinna być jak najbardziej pozioma).

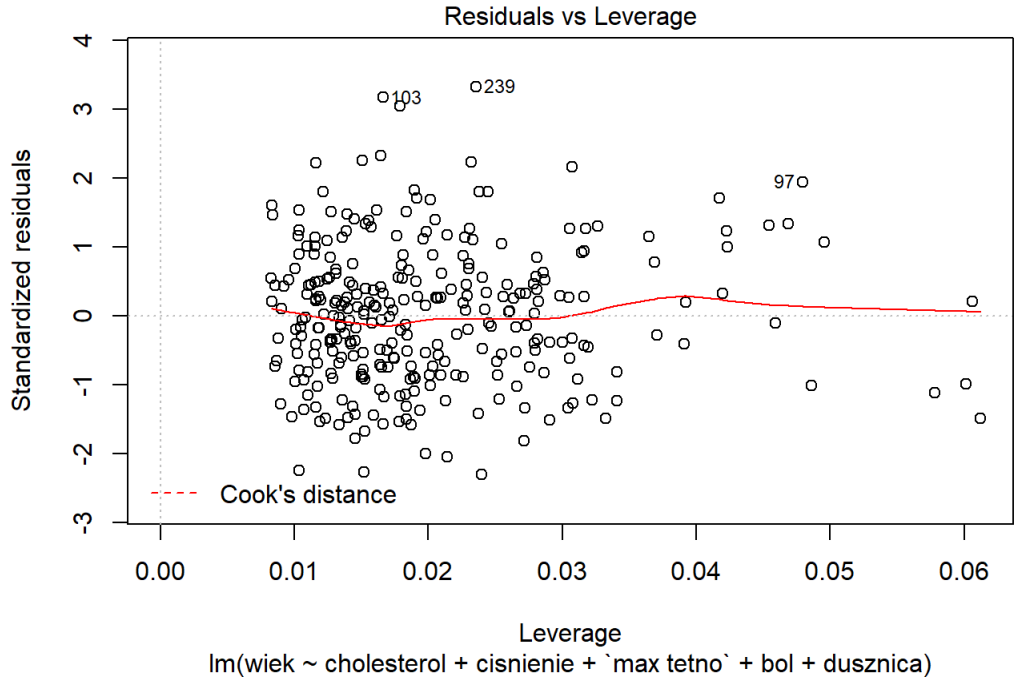
Czwarty wykres służy do badania, które obserwacje są wpływowe. Uwidoczniona jest na nim odległość Cooka, która bada różnicę predykcji modelu wyjściowego z modelem po usunięciu obserwacji. Można udowodnić, że wyraża się ona wzorem $$C_i = e_i \left(\frac{n-p}{p} \cdot\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2} $$

```
plot(model.zlozony.aic)
```






Scale-Location
103
52
239
√|Standardized residuals|
0.0
0.5
1.0
1.5
45
50
55
60
65
Fitted values
lm(wiek ~ cholesterol + cisnienie + `max tetno` + bol + dusznica)


Residuals vs Leverage
103
239
97
Cook's distance
Standardized residuals
-3
-2
-1
0
1
2
3
4
0.00
0.01
0.02
0.03
0.04
0.05
0.06
Leverage
lm(wiek ~ cholesterol + cisnienie + `max tetno` + bol + dusznica)