

# Choroby układu krążenia

Paweł Strzałecki

2023-02-03

Źródło danych: <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>

Zmienna	Opis
age	- wiek (w dniach)
height	- wysokość (w cm)
weight	- waga (w kg)
gender	- płeć (1 - kobieta, 2 - mężczyzna)
ap_hi	- skurczowe ciśnienie krwi
ap_lo	- rozkurczowe ciśnienie krwi
cholesterol	- cholesterol (1 - norma, 2 - powyżej normy, 3 - dużo powyżej normy)
gluc	- glukoza (1 - norma, 2 - powyżej normy, 3 - dużo powyżej normy)
smoke	- 0 - pacjent niepalący, 1 - pacjent palący
alco	- 0 - pacjent nie spożywa alkoholu, 1 - pacjent spożywa alkohol
active	- 0 - pacjent nie jest aktywny fizycznie, 1 - pacjent jest aktywny fizycznie
cardio	- 0 - pacjent nie ma chorób układu krążenia, 1 - pacjent ma choroby układu krążenia

- age, height, weight, gender - informacje faktyczne
- ap\_hi, ap\_lo, cholesterol, gluc - rezultat badań medycznych
- smoke, alco, active - informacje subiektywne pacjenta

## CZEŚĆ I

**HIPOTEZA:** Ciśnienie skurczowe krwi, a choroby układu krążenia. Czy zdrowy styl życia redukuje ciśnienie i zmniejsza efekt chorób układu krążenia?

```
library('dplyr')

## 
## Dołączanie pakietu: 'dplyr'

## Następujące obiekty zostały zakryte z 'package:stats':
## 
##     filter, lag

## Następujące obiekty zostały zakryte z 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```

library('ggplot2')
library('GGally')

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

library('statmod')

## Warning: pakiet 'statmod' został zbudowany w wersji R 4.2.3

library('ResourceSelection')

## Warning: pakiet 'ResourceSelection' został zbudowany w wersji R 4.2.3

## ResourceSelection 0.3-5 2019-07-22

```

## Wczytanie i eksploracja danych

```

data = read.csv('cardio_train.csv', sep=';')

head(data)

##   id   age gender height weight ap_hi ap_lo cholesterol gluc smoke alco active
## 1  0 18393     2    168     62    110     80             1     1     0     0      1
## 2  1 20228     1    156     85    140     90             3     1     0     0      1
## 3  2 18857     1    165     64    130     70             3     1     0     0      0
## 4  3 17623     2    169     82    150    100             1     1     0     0      1
## 5  4 17474     1    156     56    100     60             1     1     0     0      0
## 6  8 21914     1    151     67    120     80             2     2     0     0      0
##   cardio
## 1      0
## 2      1
## 3      1
## 4      1
## 5      0
## 6      0

summary(data)

```

	id	age	gender	height
##	Min. : 0	Min. :10798	Min. :1.00	Min. : 55.0
##	1st Qu.:25007	1st Qu.:17664	1st Qu.:1.00	1st Qu.:159.0
##	Median :50002	Median :19703	Median :1.00	Median :165.0
##	Mean :49972	Mean :19469	Mean :1.35	Mean :164.4
##	3rd Qu.:74889	3rd Qu.:21327	3rd Qu.:2.00	3rd Qu.:170.0
##	Max. :99999	Max. :23713	Max. :2.00	Max. :250.0
##	weight	ap_hi	ap_lo	cholesterol

```

##   Min.    : 10.00    Min.    :-150.00    Min.    :-70.00    Min.    :1.000
## 1st Qu.: 65.00    1st Qu.: 120.00    1st Qu.:  80.00    1st Qu.:1.000
## Median : 72.00    Median : 120.00    Median :  80.00    Median :1.000
## Mean   : 74.21    Mean   : 128.80    Mean   :  96.63    Mean   :1.367
## 3rd Qu.: 82.00    3rd Qu.: 140.00    3rd Qu.:  90.00    3rd Qu.:2.000
## Max.   :200.00    Max.   :16020.00   Max.   :11000.00   Max.   :3.000
##          gluc           smoke          alco          active
##   Min.    :1.000    Min.    :0.00000    Min.    :0.00000    Min.    :0.0000
## 1st Qu.:1.000    1st Qu.:0.00000    1st Qu.:0.00000    1st Qu.:1.0000
## Median :1.000    Median :0.00000    Median :0.00000    Median :1.0000
## Mean   :1.226    Mean   :0.08813    Mean   :0.05377    Mean   :0.8037
## 3rd Qu.:1.000    3rd Qu.:0.00000    3rd Qu.:0.00000    3rd Qu.:1.0000
## Max.   :3.000    Max.   :1.00000    Max.   :1.00000    Max.   :1.0000
##          cardio
##   Min.    :0.0000
## 1st Qu.:0.0000
## Median :0.0000
## Mean   :0.4997
## 3rd Qu.:1.0000
## Max.   :1.0000

```

Wiek pacjenta jest podany w dniach. Dla lepszej czytelności danych zamienimy dni na lata.

```
data['age'] = round(data['age']/365)
```

Niektóre kolumny zawierają nierealne wartości, trzeba zatem wyczyścić dane.

```

data = filter(data, height >= 140,
              weight >= 40,
              ap_hi >= 90 & ap_hi <= 150,
              ap_lo >= 60 & ap_lo <= 100)

```

```
summary(data)
```

```

##      id          age        gender       height
##   Min.    : 0    Min.    :30.00    Min.    :1.000    Min.    :140.0
## 1st Qu.:24994  1st Qu.:48.00    1st Qu.:1.000    1st Qu.:159.0
## Median :50058  Median :54.00    Median :1.000    Median :165.0
## Mean   :49978  Mean   :53.17    Mean   :1.346    Mean   :164.5
## 3rd Qu.:74895  3rd Qu.:58.00    3rd Qu.:2.000    3rd Qu.:170.0
## Max.   :99999  Max.   :65.00    Max.   :2.000    Max.   :250.0
##          weight      ap_hi      ap_lo      cholesterol
##   Min.    :40.00    Min.    :90.0    Min.    :60.00    Min.    :1.000
## 1st Qu.:65.00    1st Qu.:120.0   1st Qu.:80.00    1st Qu.:1.000
## Median :71.00    Median :120.0   Median :80.00    Median :1.000
## Mean   :73.63    Mean   :123.8   Mean   :80.31    Mean   :1.348
## 3rd Qu.:81.00    3rd Qu.:130.0   3rd Qu.:80.00    3rd Qu.:1.000
## Max.   :200.00   Max.   :150.0   Max.   :100.00   Max.   :3.000
##          gluc           smoke          alco          active
##   Min.    :1.000    Min.    :0.00000    Min.    :0.00000    Min.    :0.0000
## 1st Qu.:1.000    1st Qu.:0.00000    1st Qu.:0.00000    1st Qu.:1.0000
## Median :1.000    Median :0.00000    Median :0.00000    Median :1.0000

```

```

##   Mean    :1.219    Mean    :0.08679    Mean    :0.05191    Mean    :0.8033
## 3rd Qu.:1.000    3rd Qu.:0.00000    3rd Qu.:0.00000    3rd Qu.:1.0000
## Max.    :3.000    Max.    :1.00000    Max.    :1.00000    Max.    :1.0000
##      cardio
##  Min.    :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean    :0.4682
##  3rd Qu.:1.0000
##  Max.    :1.0000

data = data %>%
  mutate(cholesterol = factor(cholesterol, levels = c('norm' = 1, 'a_norm' = 2, 'wa_norm' = 3), labels =
  mutate(gluc = factor(gluc, levels = c('norm' = 1, 'a_norm' = 2, 'wa_norm' = 3), labels = c('norm', 'a_norm', 'wa_norm'))
  mutate(smoke = factor(smoke, levels = c('non smoker' = 0, 'smoker' = 1), labels = c('non smoker', 'smoker'))
  mutate(alco = factor(alco, levels = c('non alc' = 0, 'alc' = 1), labels = c('non alc', 'alc')))) %>%
  mutate(active = factor(active, levels = c('non active' = 0, 'active' = 1), labels = c('non active', 'active'))
  mutate(gender = factor(gender, levels = c('female' = 1, 'male' = 2), labels = c('female', 'male')))) %>%
  mutate(cardio = factor(cardio, levels = c('healthy' = 0, 'sick' = 1), labels = c('healthy', 'sick'))))

head(data)

##   id age gender height weight ap_hi ap_lo cholesterol   gluc   smoke   alco
## 1  0  50 male     168     62    110     80      norm  norm non smoker non alc
## 2  1  55 female   156     85    140     90     aw_norm  norm non smoker non alc
## 3  2  52 female   165     64    130     70     aw_norm  norm non smoker non alc
## 4  3  48 male     169     82    150    100      norm  norm non smoker non alc
## 5  4  48 female   156     56    100     60      norm  norm non smoker non alc
## 6  8  60 female   151     67    120     80      a_norm a_norm non smoker non alc
##      active cardio
## 1      active healthy
## 2      active sick
## 3 non active sick
## 4      active sick
## 5 non active healthy
## 6 non active healthy

```

Kolumna z id pacjenta nie będzie potrzebna w dalszej analizie postawionej hipotezy, dlatego zostanie ona usunięta.

```
data = select(data, -id)
```

```
summary(data)
```

```

##       age        gender      height      weight
##  Min.   :30.00   female:41562   Min.   :140.0   Min.   : 40.00
##  1st Qu.:48.00   male  :22026   1st Qu.:159.0   1st Qu.: 65.00
##  Median :54.00
##  Mean   :53.17
##  3rd Qu.:58.00
##  Max.   :65.00
##      ap_hi        ap_lo      cholesterol      gluc

```

```

##  Min.   : 90.0   Min.   : 60.00   norm    :48443   norm    :54426
##  1st Qu.:120.0   1st Qu.: 80.00   a_norm : 8190   a_norm : 4388
##  Median :120.0   Median : 80.00   aw_norm: 6955   wa_norm: 4774
##  Mean    :123.8   Mean    : 80.31
##  3rd Qu.:130.0   3rd Qu.: 80.00
##  Max.    :150.0   Max.    :100.00
##              smoke          alco          active        cardio
##  non smoker:58069   non alc:60287   non active:12509   healthy:33814
##  smoker     : 5519   alc     : 3301   active    :51079   sick    :29774
##
## 
## 
## 
##
```

- Grupa badanych pacjentów jest w przedziale wiekowym 30-65 lat.
- Większość pacjentów jest płci żeńskiej.
- Wzrost pacjentów to przedział od 140cm do 250cm wzrostu.
- Waga pacjentów wynosi od 40 do 200kg.
- Większość pacjentów ma cholesterol w normie, ok. 8200 osób ma poziom cholesterolu powyżej normy, a ok. 7000 osób dużo powyżej normy.
- Większość pacjentów ma poziom glukozy w normie, ok. 4400 osób powyżej normy, a ok. 4800 osób dużo powyżej normy.
- Ok. 58000 pacjentów to osoby nie palące, 5500 pacjentów to osoby palące.
- Grupa pacjentów nie pijących alkoholu liczy 60300 osób. 3300 pacjentów spożywa alkohol.
- Ok. 51100 pacjentów jest aktywna fizycznie. Osób nieaktywnych fizycznie jest 12500.
- 34000 pacjentów nie posiada choroby układu krążenia, natomiast osób chorych jest ok. 30000.

## Analiza poszczególnych zmiennych i ich wpływu na ciśnienie skurczowe krwi

```

plots = ggpairs(data)
ggsave('wykresy.jpeg', plots)
```

```

## Saving 6.5 x 4.5 in image
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

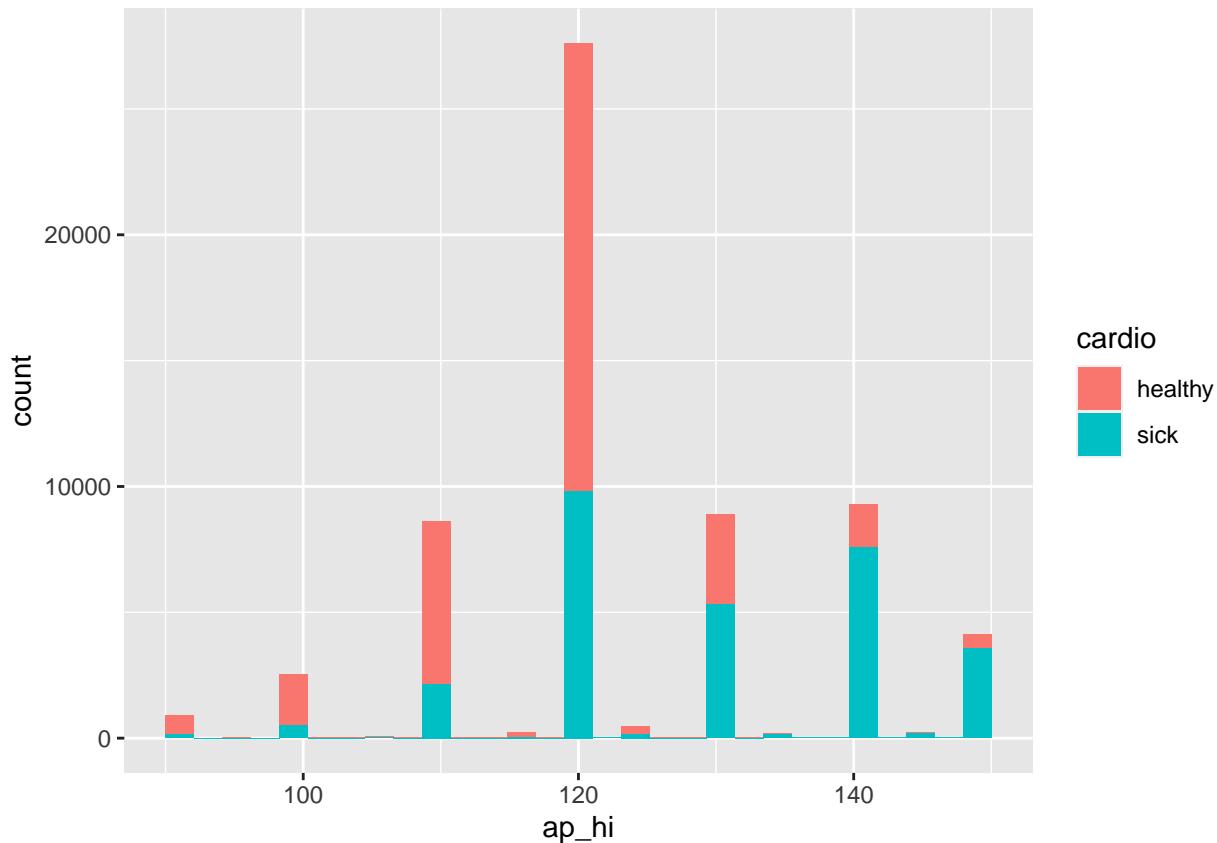
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Występuje duża korelacja między ciśnieniem skurczowym, a rozkurczowym krwi ~ 0.7.

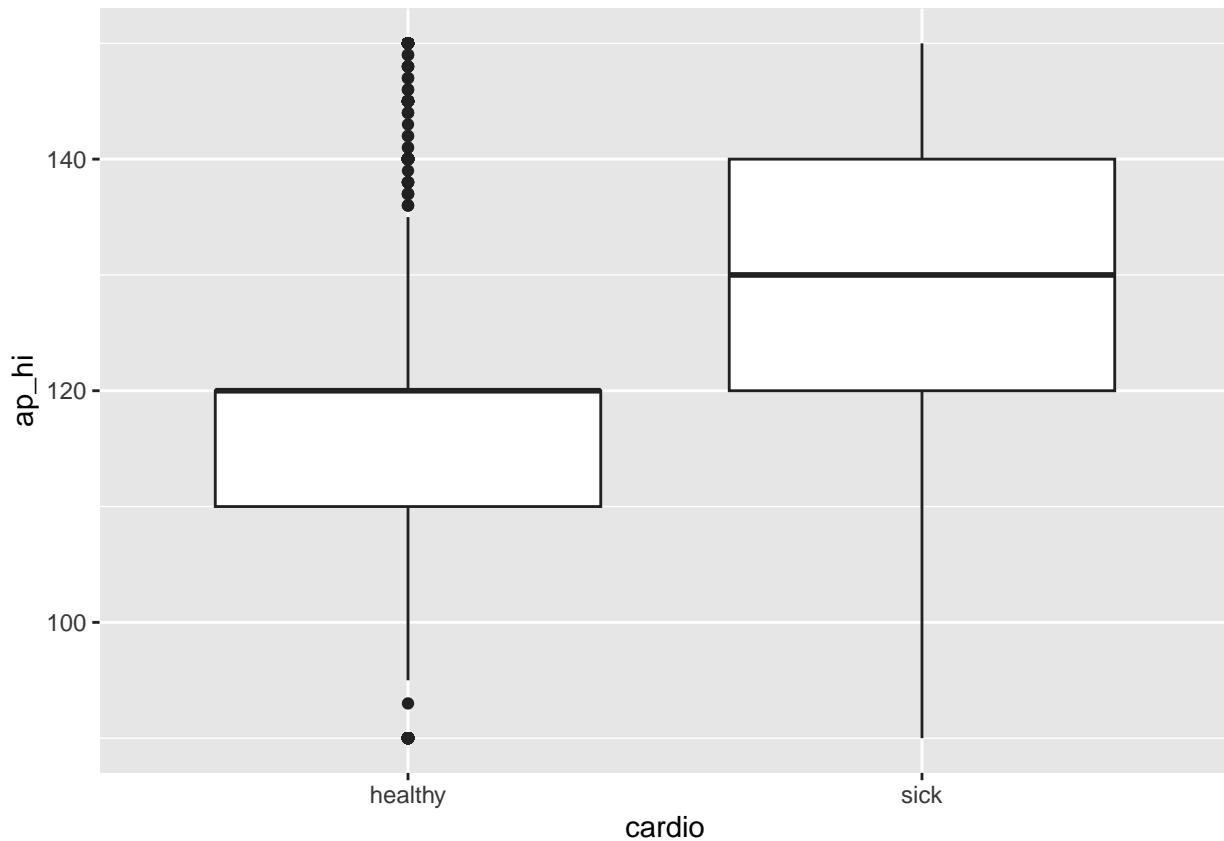
## Cardio

```
ggplot(data, aes(x = ap_hi, fill = cardio)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(data, aes(x = cardio, y = ap_hi)) + geom_boxplot()
```

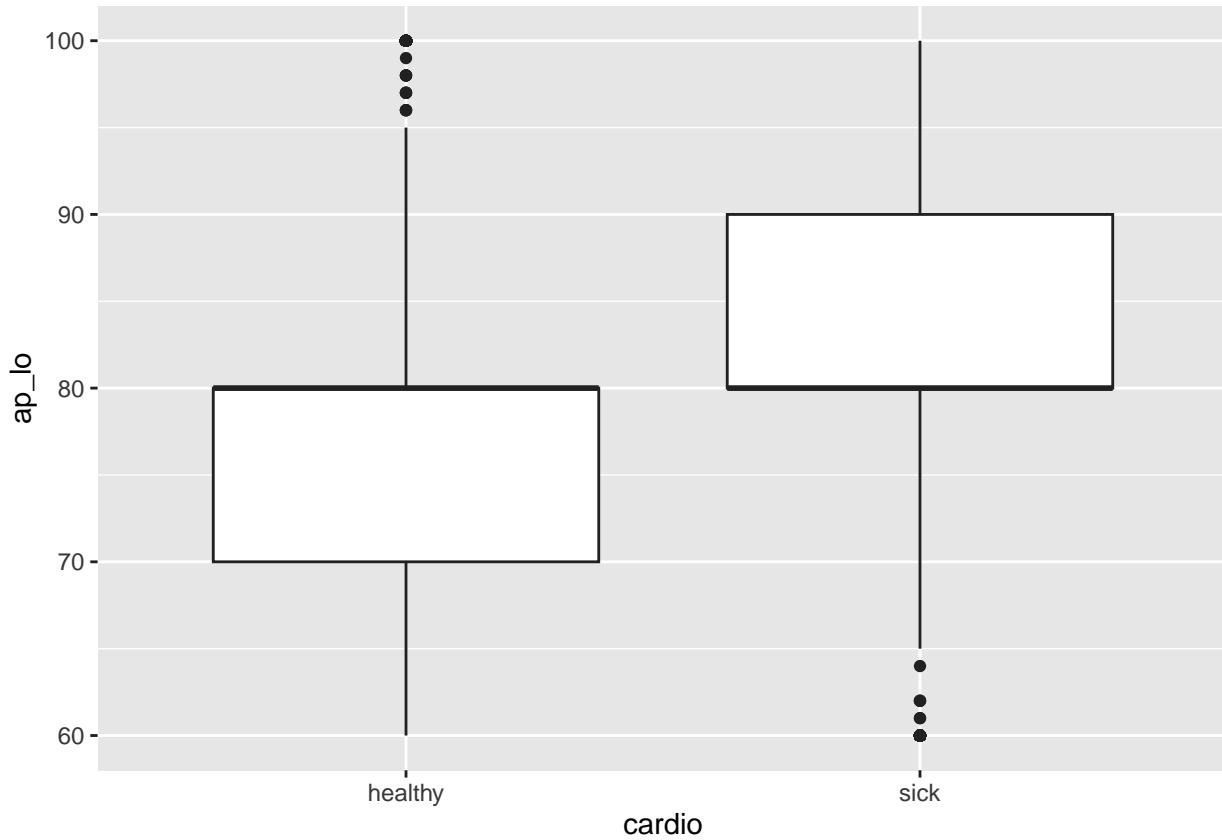


```
summary(lm(ap_hi ~ cardio, data = data))
```

```
##
## Call:
## lm(formula = ap_hi ~ cardio, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -39.560 -8.815  1.185  10.440  31.185 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 118.81508   0.06318 1880.5 <2e-16 ***
## cardio>sick 10.74541    0.09233   116.4 <2e-16 ***
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.62 on 63586 degrees of freedom
## Multiple R-squared:  0.1756, Adjusted R-squared:  0.1756 
## F-statistic: 1.354e+04 on 1 and 63586 DF,  p-value: < 2.2e-16
```

Osoby mające choroby układu krążenia mają wyższe ciśnienie skurczowe krwi o 10 mm/Hg od osób zdrowych.

```
ggplot(data, aes(x = cardio, y = ap_lo)) + geom_boxplot()
```



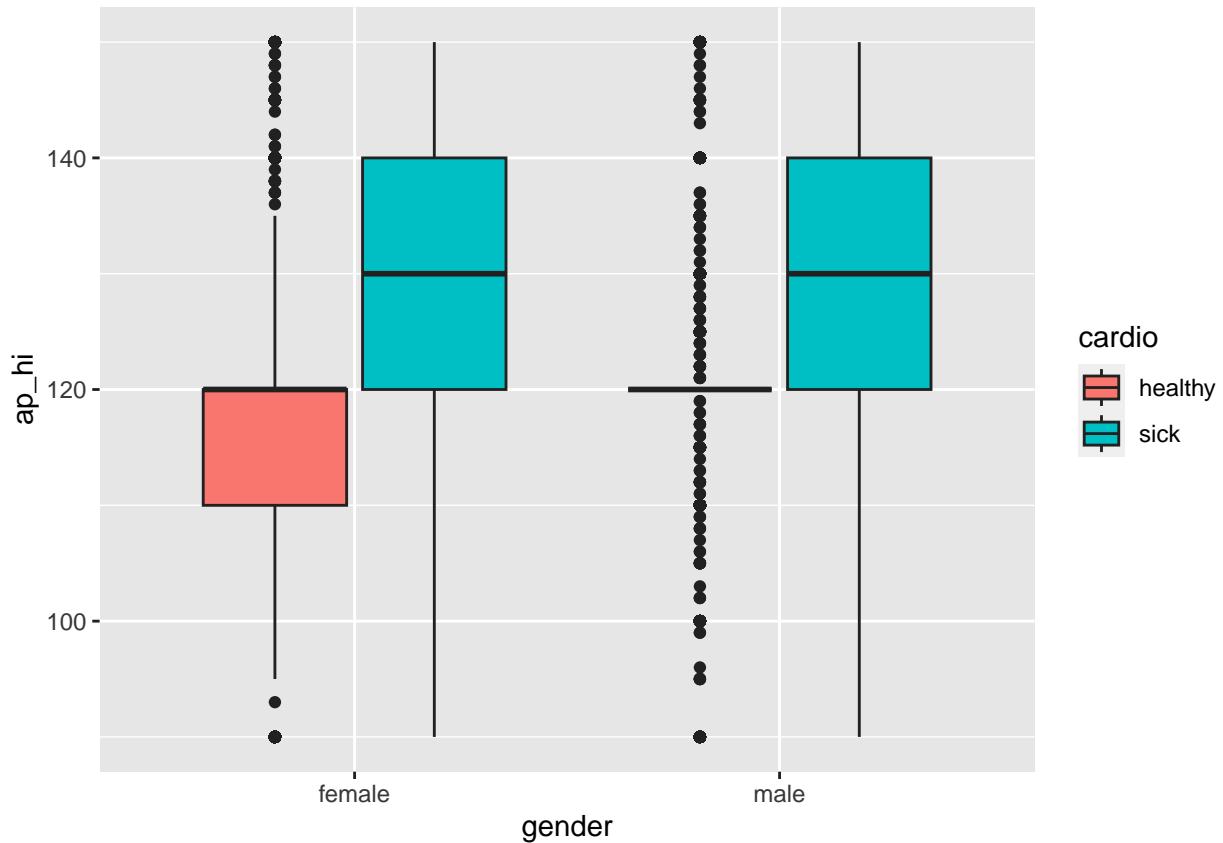
```
summary(lm(ap_lo ~ cardio, data = data))
```

```
##  
## Call:  
## lm(formula = ap_lo ~ cardio, data = data)  
##  
## Residuals:  
##     Min      1Q  Median      3Q     Max  
## -23.08  -3.08   2.13   2.13  22.13  
##  
## Coefficients:  
##             Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 77.87002   0.04260 1828.1 <2e-16 ***  
## cardiosick   5.21025   0.06225   83.7 <2e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 7.833 on 63586 degrees of freedom  
## Multiple R-squared:  0.09924,    Adjusted R-squared:  0.09923  
## F-statistic: 7006 on 1 and 63586 DF,  p-value: < 2.2e-16
```

Osoby mające choroby układu krążenia mają wyższe ciśnienie rozkurczowe krwi o 5 mm/Hg od osób zdrowych.

## Gender

```
ggplot(data, aes(x = gender, y = ap_hi, fill = cardio)) + geom_boxplot()
```

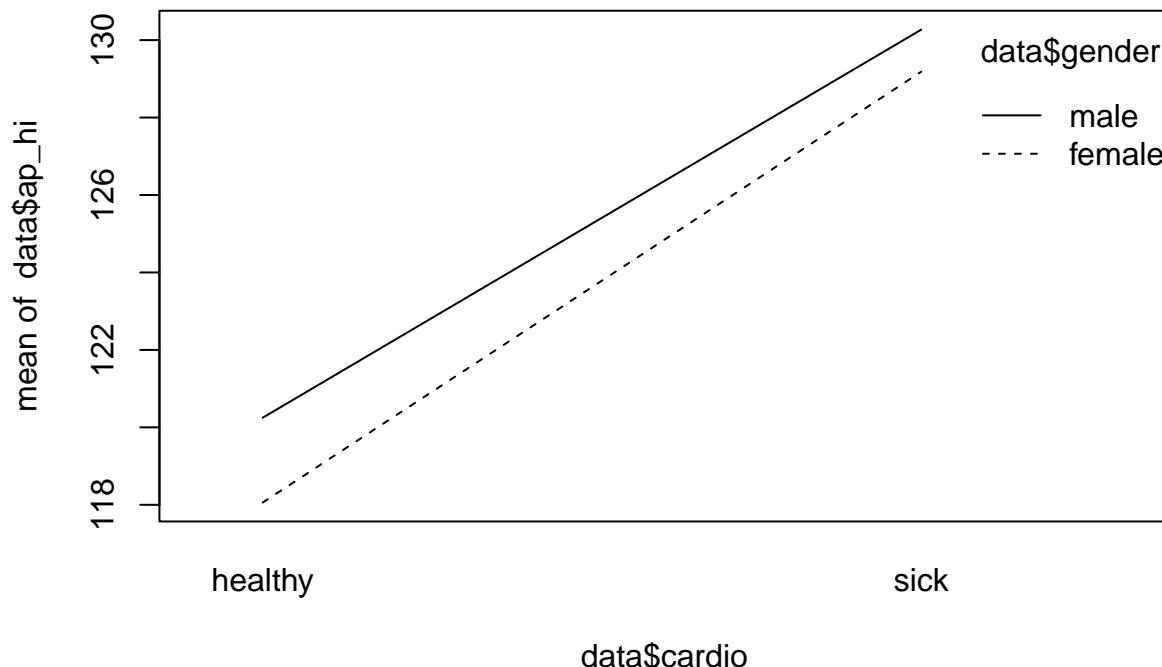


```
summary(lm(ap_hi ~ gender, data = data))
```

```
##
## Call:
## lm(formula = ap_hi ~ gender, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -34.955  -4.955  -3.259   6.741  26.741 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 123.25913    0.06264 1967.73 <2e-16 ***
## gendermale    1.69551    0.10643   15.93 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.77 on 63586 degrees of freedom
## Multiple R-squared:  0.003975, Adjusted R-squared:  0.00396 
## F-statistic: 253.8 on 1 and 63586 DF, p-value: < 2.2e-16
```

Średnie ciśnienie skurczowe krwi różni się w zależności od płci. Średnie ciśnienie skurczowe krwi u kobiet wynosi 123 mm/Hg, natomiast u mężczyzn jest to wartość większa o ok. 1.7 mm/Hg. Testy statystyczne wskazują na dużą istotność tej różnicy, zatem w dalszej analizie trzeba uwzględnić wyniki w zależności od płci.

```
interaction.plot(x.factor = data$cardio,
                 trace.factor = data$gender,
                 response = data$ap_hi)
```



Brak interakcji.

```
summary(lm(ap_hi ~ cardio + gender, data = data))

##
## Call:
## lm(formula = ap_hi ~ cardio + gender, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -40.652  -8.979   1.021   9.348  31.763 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 118.23723   0.07132 1657.76   <2e-16 ***
## cardio.sick 10.74217   0.09212 116.61   <2e-16 ***
```

```

## gendermale    1.67261    0.09660   17.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.59 on 63585 degrees of freedom
## Multiple R-squared:  0.1795, Adjusted R-squared:  0.1794
## F-statistic:  6953 on 2 and 63585 DF,  p-value: < 2.2e-16

```

Zdrowa kobieta ma średnio ciśnienie skurczowe krwi na poziomie 118 mm/Hg. Chora kobieta natomiast ma już ciśnienie krwi ok. 129 mm/Hg.

Zdrowy mężczyzna ma średnio ciśnieni skurczowe krwi na poziomie 120 mm/Hg. Chory mężczyzna natomiast ma ciśnienie krwi ok. 131 mm/Hg.

### Active

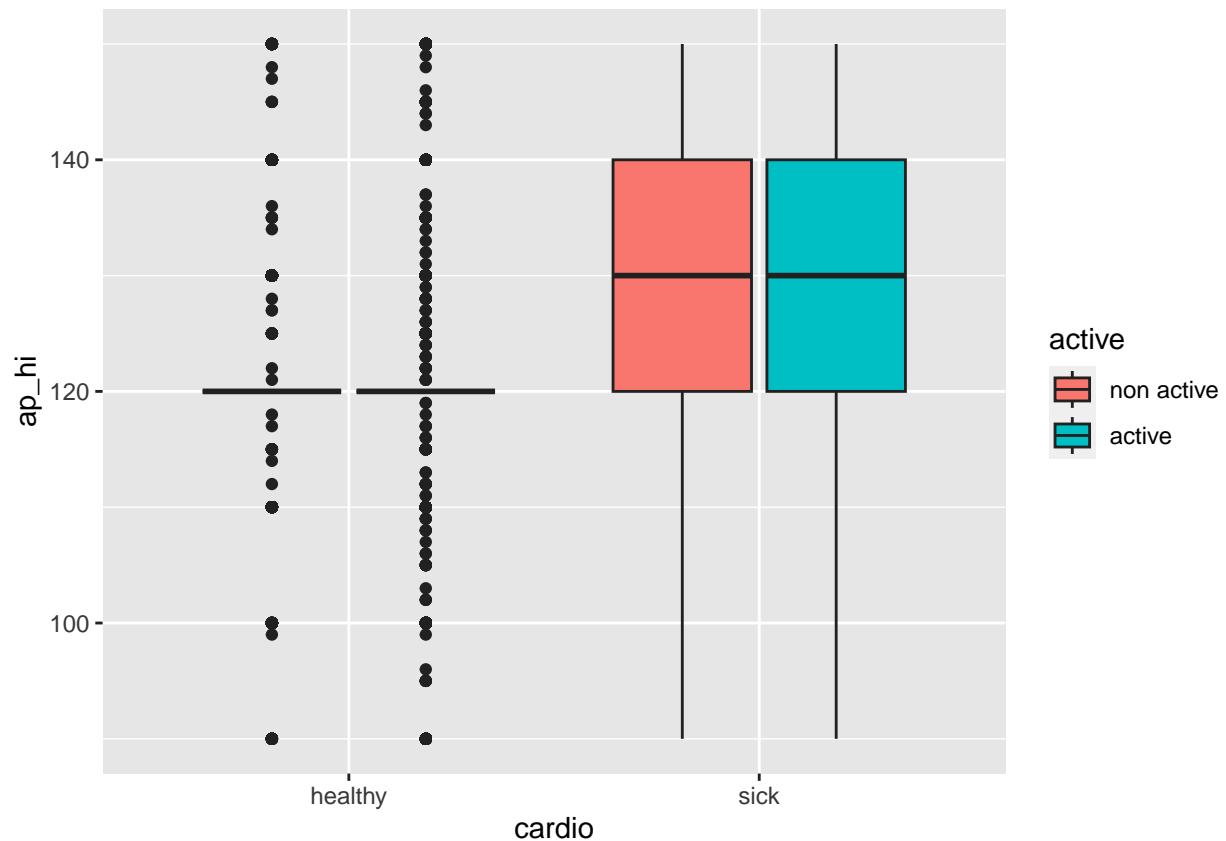
Mówią się, że prowadzenie zdrowego stylu życia pomaga zredukować ryzyko chorób. Można zatem sprawdzić jak różnią się wyniki pacjentów, którzy byli aktywni fizycznie.

```

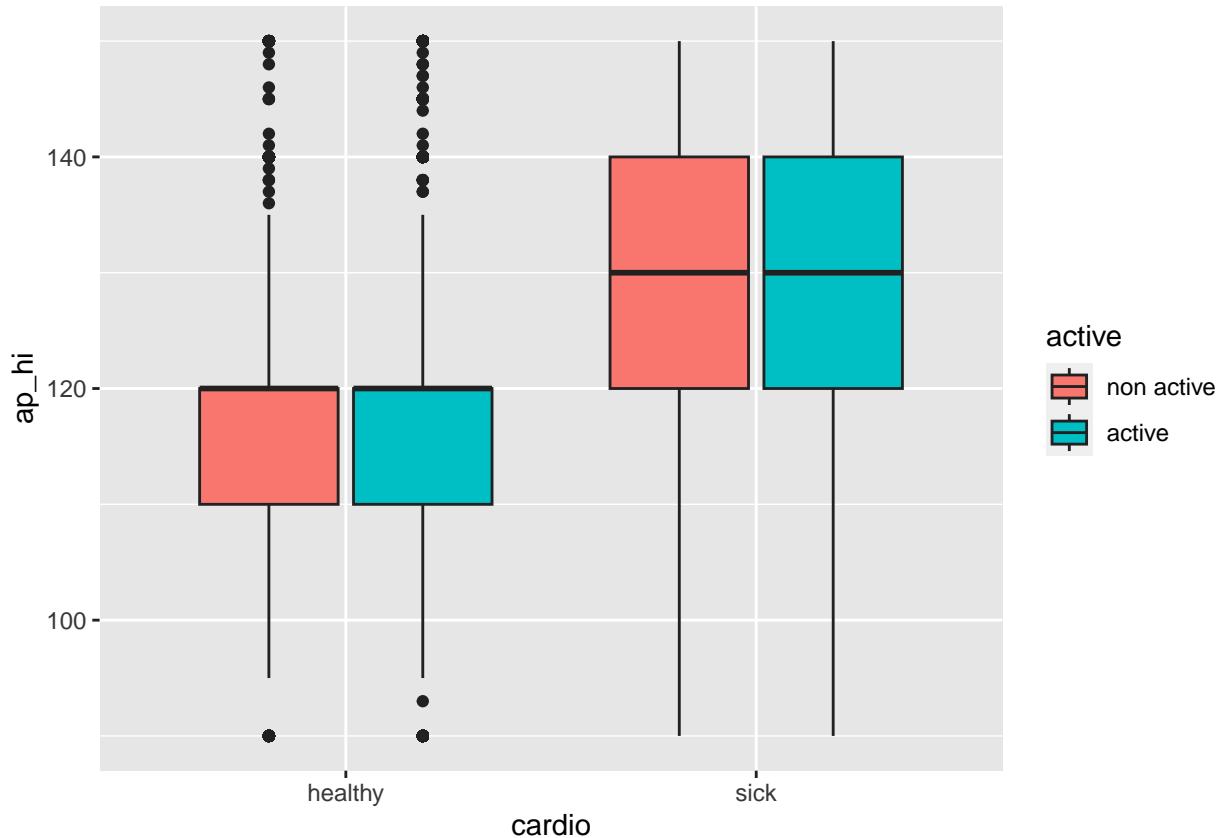
data_male = filter(data, gender == 'male')
data_female = filter(data, gender == 'female')

ggplot(data_male, aes(x = cardio, y = ap_hi, fill = active)) + geom_boxplot()

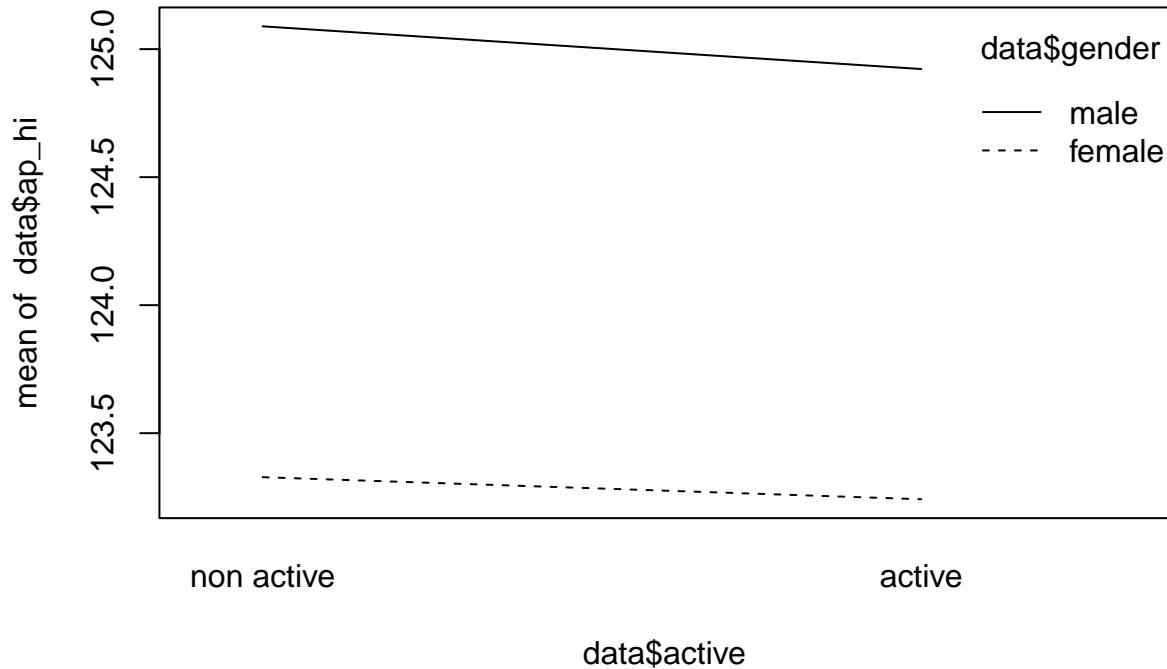
```



```
ggplot(data_female, aes(x = cardio, y = ap_hi, fill = active)) + geom_boxplot()
```



```
interaction.plot(x.factor = data$active,
                  trace.factor = data$gender,
                  response = data$ap_hi)
```



Brak interakcji.

```
summary(lm(ap_hi ~ cardio + gender + active, data = data))

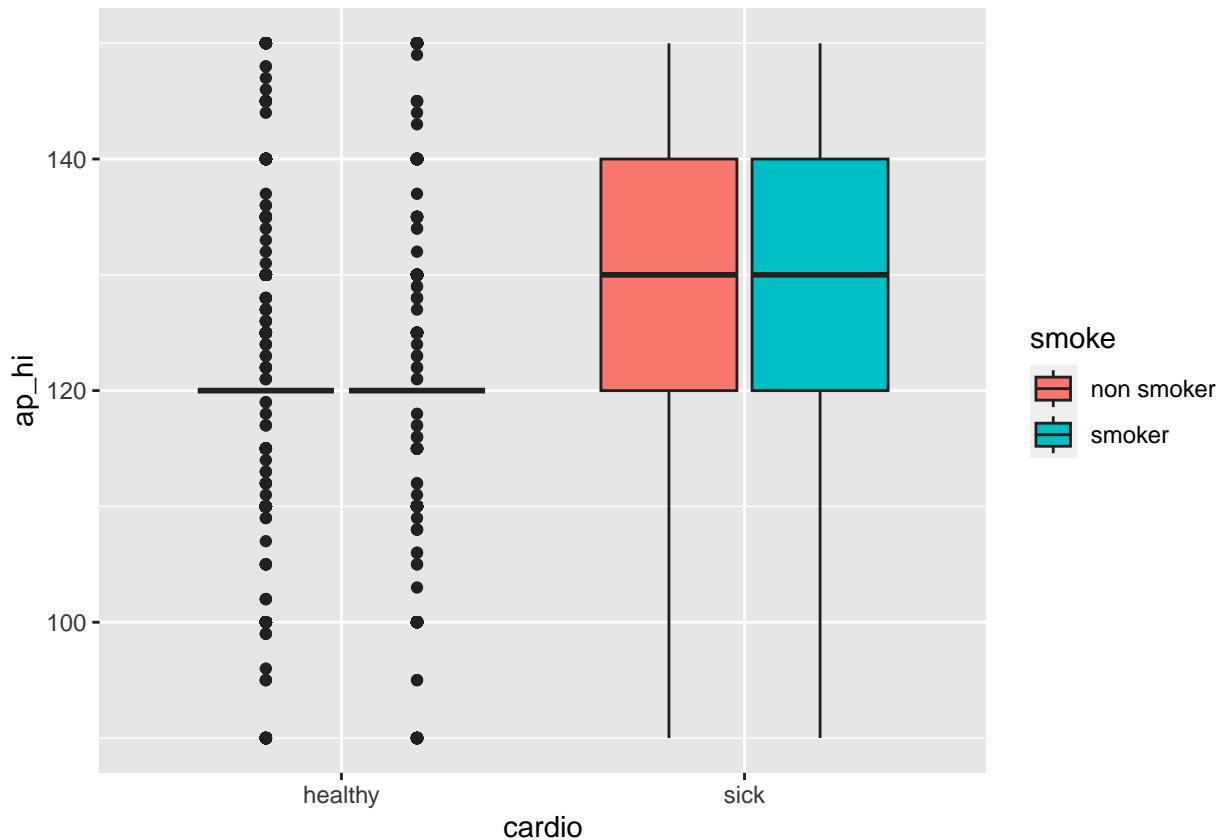
##
## Call:
## lm(formula = ap_hi ~ cardio + gender + active, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -40.748  -9.077   0.923   9.252  32.133 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 117.86724   0.11846 995.008 < 2e-16 ***
## cardio      10.75729   0.09219 116.689 < 2e-16 ***
## gendermale   1.67056   0.09659  17.295 < 2e-16 ***
## activeactive  0.45266   0.11572   3.912 9.18e-05 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.59 on 63584 degrees of freedom
## Multiple R-squared:  0.1797, Adjusted R-squared:  0.1796 
## F-statistic: 4642 on 3 and 63584 DF,  p-value: < 2.2e-16
```

Aktywność fizyczna wpływa istotnie na obniżenie ciśnienia skurczowego krwi. Z powyższego wykresu możemy

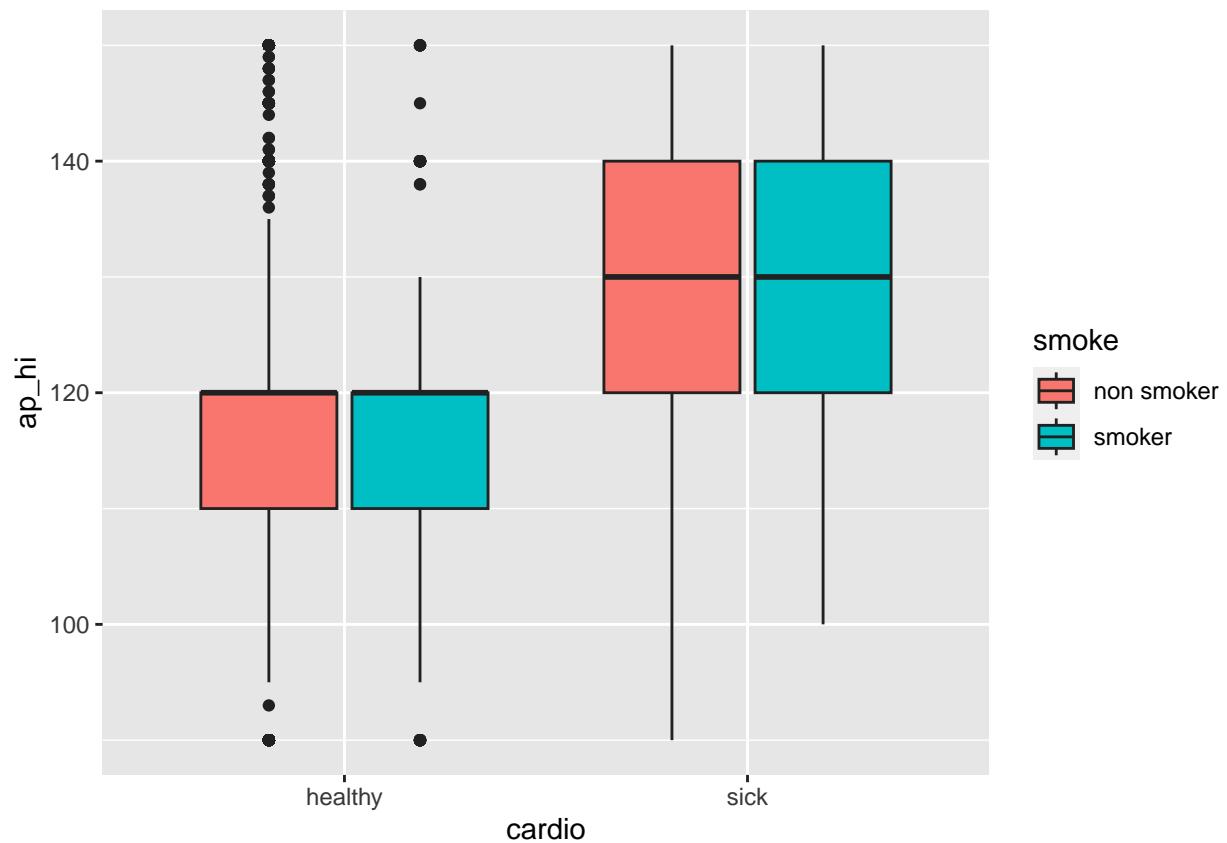
odczytać, że aktywność fizyczna w większym stopniu redukuje ciśnienie skurczowe krwi wśród mężczyzn, niż wśród kobiet.

### Smoke

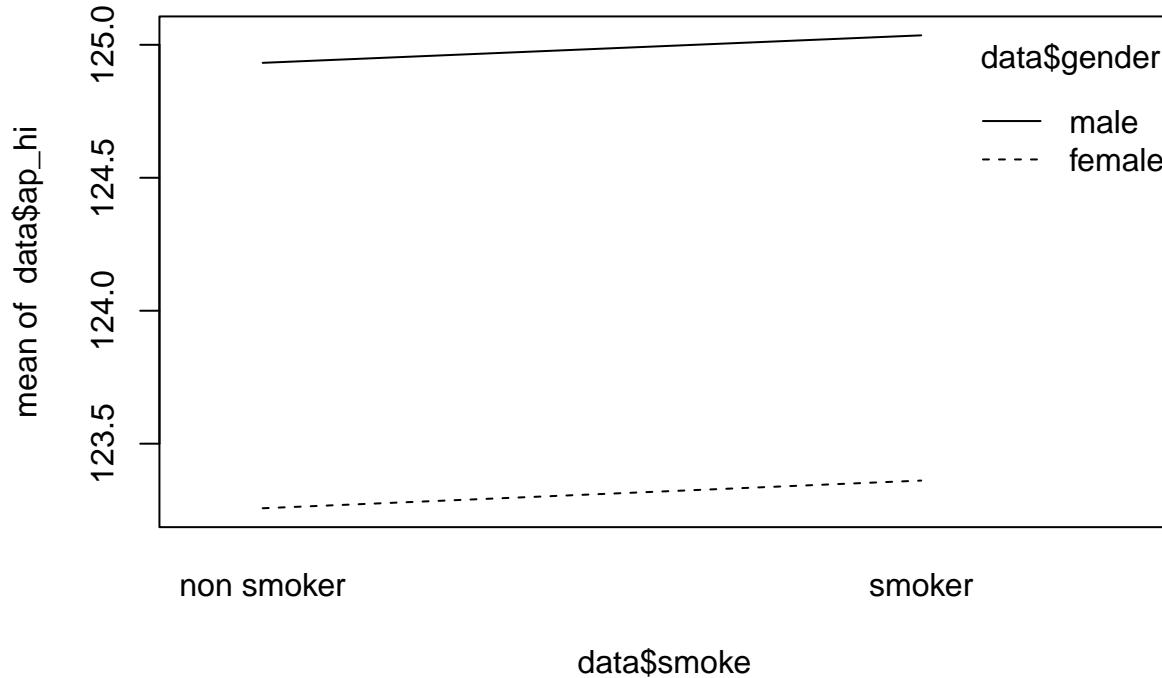
```
ggplot(data_male, aes(x = cardio, y = ap_hi, fill = smoke)) + geom_boxplot()
```



```
ggplot(data_female, aes(x = cardio, y = ap_hi, fill = smoke)) + geom_boxplot()
```



```
interaction.plot(x.factor = data$smoke,
                 trace.factor = data$gender,
                 response = data$ap_hi)
```



Brak interakcji.

```
summary(lm(ap_hi ~ cardio + gender + smoke, data = data))

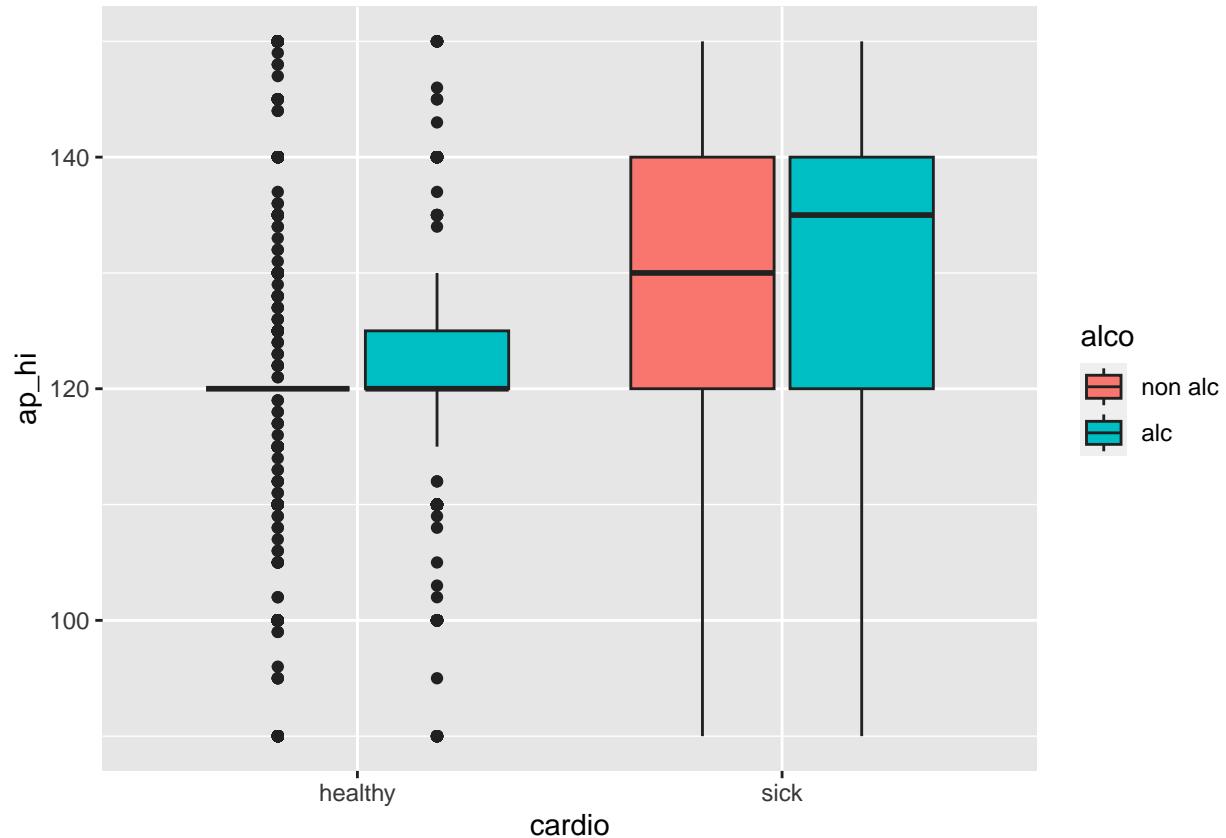
##
## Call:
## lm(formula = ap_hi ~ cardio + gender + smoke, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -41.124  -8.973   1.027   9.474  31.777 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 118.22299   0.07144 1654.923 < 2e-16 ***
## cardio      10.74991   0.09214 116.674 < 2e-16 ***
## gendermale  1.55358   0.10259  15.143 < 2e-16 ***
## smokesmoker  0.59733   0.17343   3.444 0.000573 *** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.59 on 63584 degrees of freedom
## Multiple R-squared:  0.1796, Adjusted R-squared:  0.1796 
## F-statistic: 4640 on 3 and 63584 DF,  p-value: < 2.2e-16
```

Efekt palenia jest istotny statystycznie. Osoby palące mają większe ciśnienie skurczowe krwi o ok. 0.6

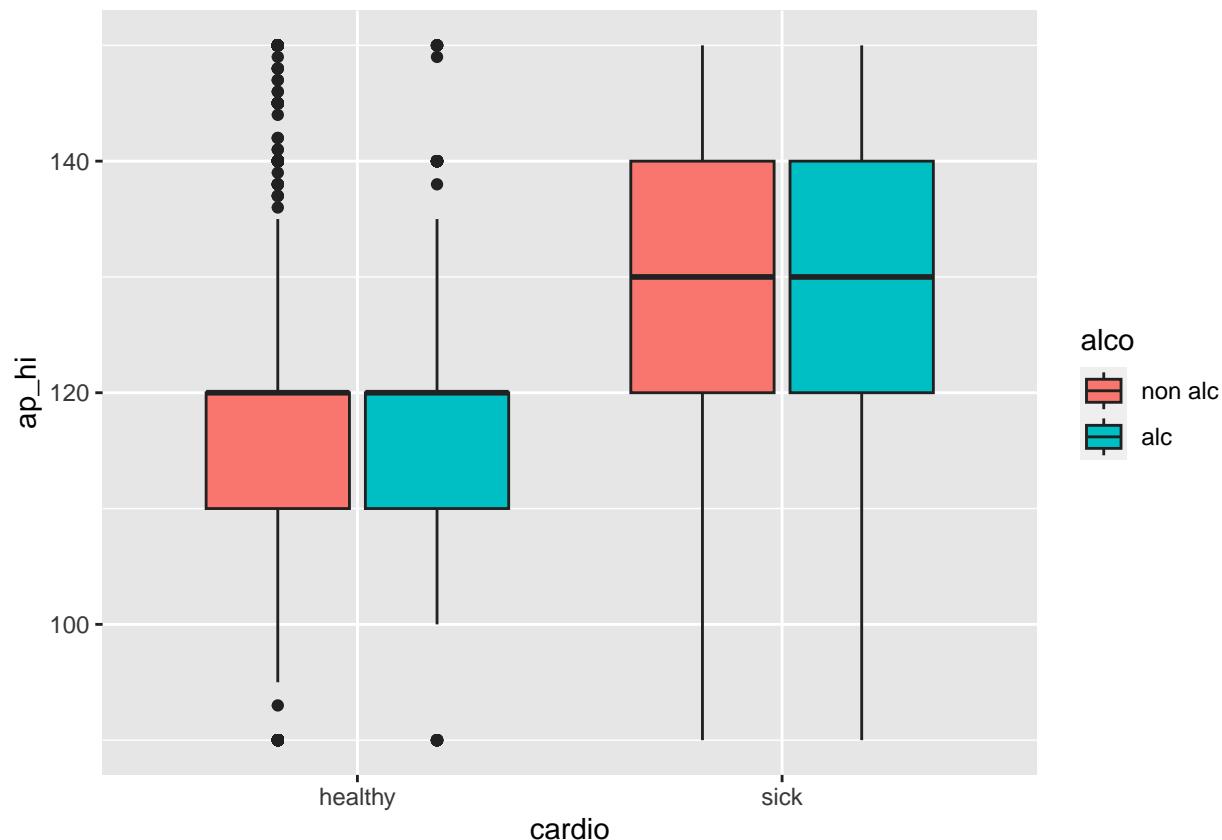
mm/Hg.

### Alco

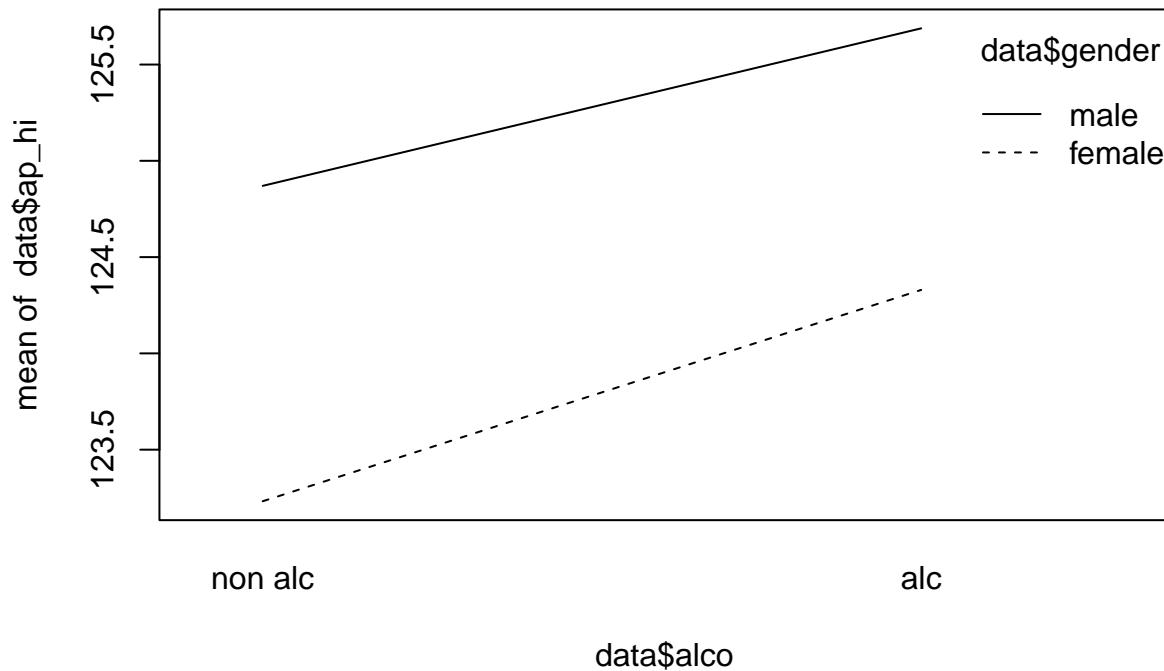
```
ggplot(data_male, aes(x = cardio, y = ap_hi, fill = alco)) + geom_boxplot()
```



```
ggplot(data_female, aes(x = cardio, y = ap_hi, fill = alco)) + geom_boxplot()
```



```
interaction.plot(x.factor = data$alco,
                 trace.factor = data$gender,
                 response = data$ap_hi)
```



Brak interakcji.

```
summary(lm(ap_hi ~ cardio + gender + alco, data = data))

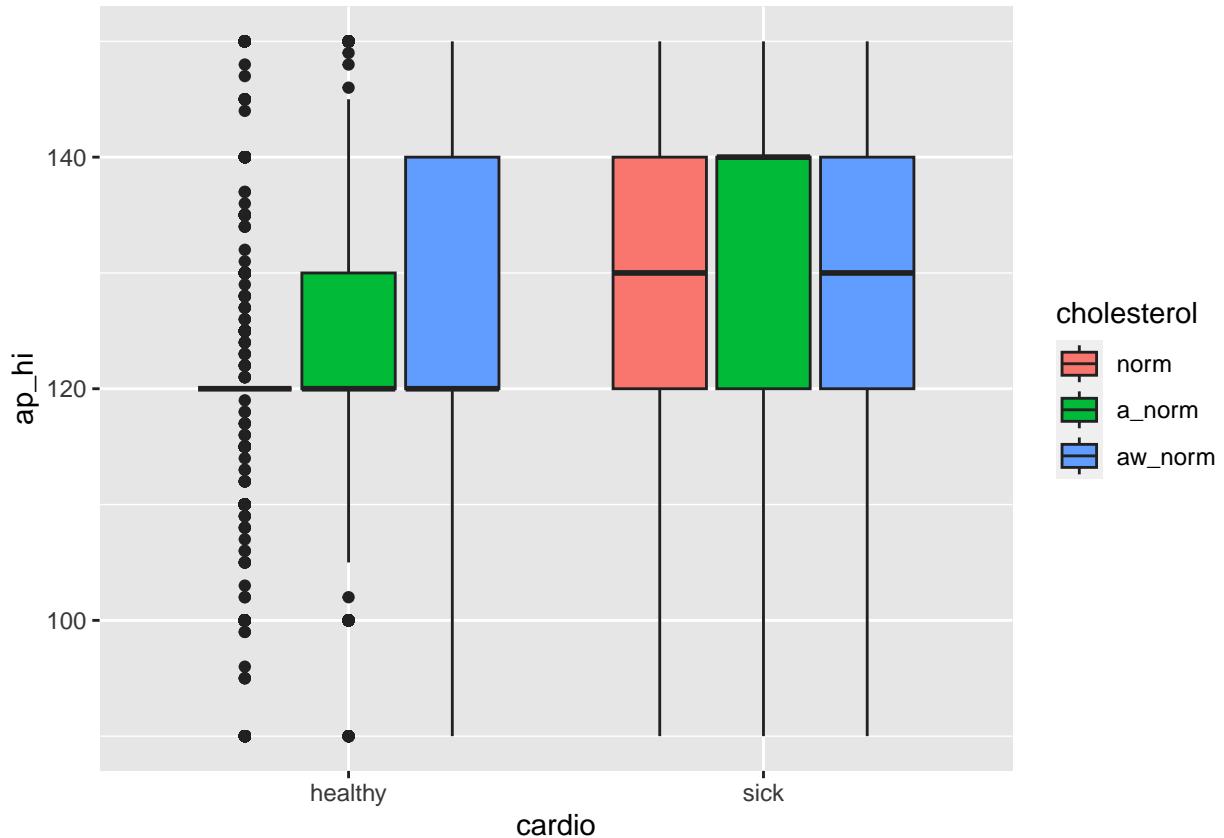
##
## Call:
## lm(formula = ap_hi ~ cardio + gender + alco, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -41.789  -8.953   1.047   9.474  31.798 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 118.20242   0.07154 1652.293 < 2e-16 ***
## cardio      10.75015   0.09210  116.721 < 2e-16 ***
## gendermale   1.57292   0.09799   16.052 < 2e-16 ***
## alcoalc     1.26352   0.21019    6.011 1.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.59 on 63584 degrees of freedom
## Multiple R-squared:  0.1799, Adjusted R-squared:  0.1799 
## F-statistic: 4650 on 3 and 63584 DF,  p-value: < 2.2e-16
```

Spożywanie alkoholu ma bardzo istotny wpływ na ciśnienie skurczowe krwi. Osoby spożywające duże ilości

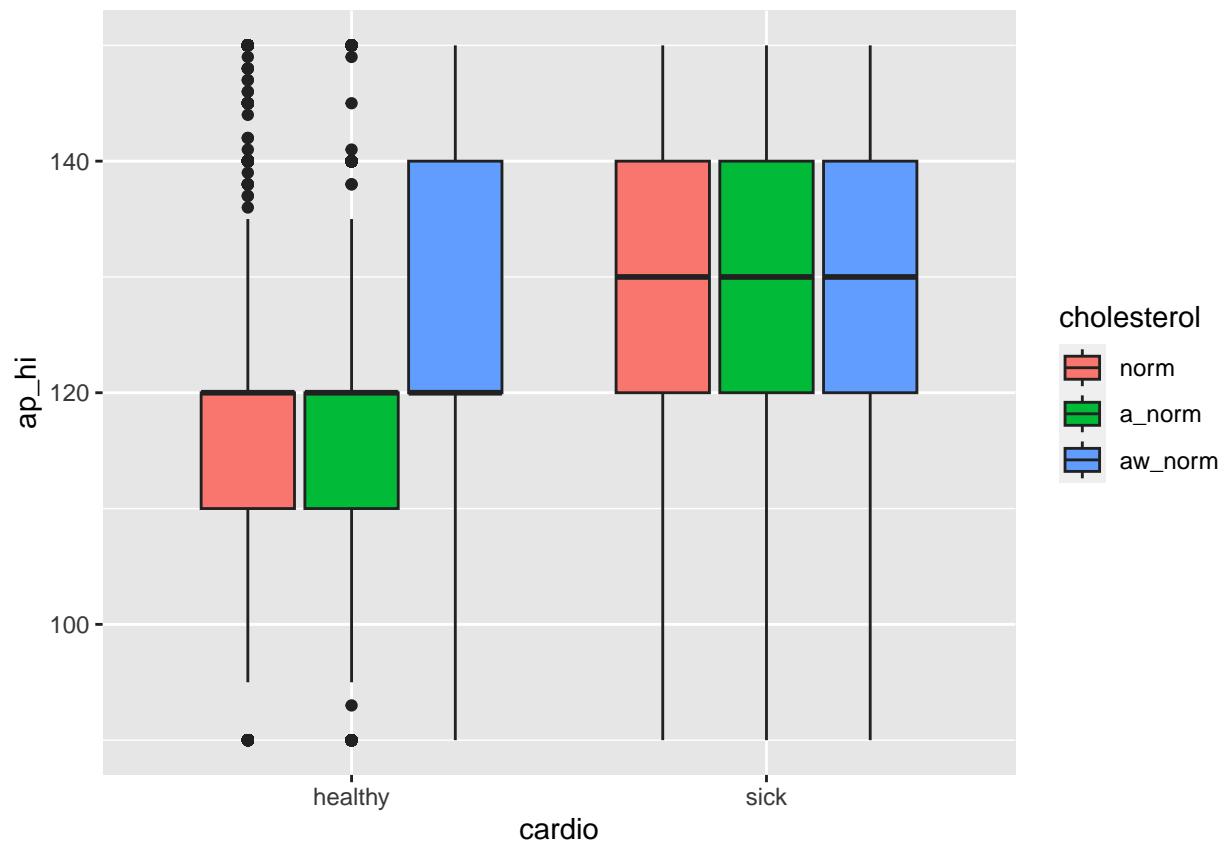
alkoholu mają ciśnienie skurczowe krwi większe o aż 1.3 mm/Hg.

## Cholesterol

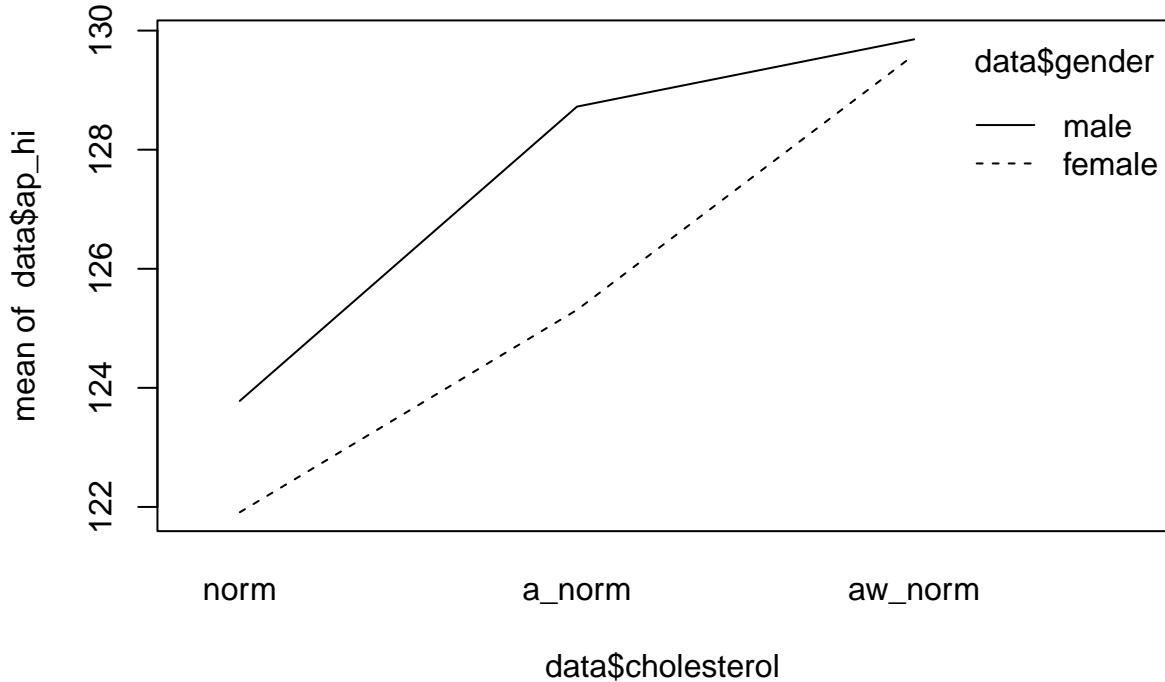
```
ggplot(data_male, aes(x = cardio, y = ap_hi, fill = cholesterol)) + geom_boxplot()
```



```
ggplot(data_female, aes(x = cardio, y = ap_hi, fill = cholesterol)) + geom_boxplot()
```



```
interaction.plot(x.factor = data$cholesterol,
                 trace.factor = data$gender,
                 response = data$ap_hi)
```



```
anova(lm(ap_hi ~ cardio + gender + cholesterol, data = data),
      lm(ap_hi ~ cardio + gender * cholesterol, data = data))
```

```
## Analysis of Variance Table
##
## Model 1: ap_hi ~ cardio + gender + cholesterol
## Model 2: ap_hi ~ cardio + gender * cholesterol
##   Res.Df   RSS Df Sum of Sq    F    Pr(>F)
## 1  63583 8437438
## 2  63581 8432533  2     4905.3 18.493 9.354e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Zachodzi interakcja.

```
summary(lm(ap_hi ~ cardio + gender * cholesterol, data = data))
```

```
##
## Call:
## lm(formula = ap_hi ~ cardio + gender * cholesterol, data = data)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -42.839 -7.918  0.463  7.937 32.236
```

```

## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           117.76442   0.07567 1556.275 < 2e-16 ***
## cardiosick            10.15395   0.09388 108.153 < 2e-16 ***
## gendermale             1.77302   0.10923 16.232 < 2e-16 ***
## cholesterola_norm     1.98486   0.16804 11.812 < 2e-16 ***
## cholesterolaw_norm    4.14499   0.18105 22.894 < 2e-16 ***
## gendermale:cholesterol_norm 1.16287   0.29390  3.957 7.61e-05 ***
## gendermale:cholesterolaw_norm -1.30136   0.31911 -4.078 4.55e-05 ***
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 11.52 on 63581 degrees of freedom
## Multiple R-squared:  0.19, Adjusted R-squared:  0.19 
## F-statistic:  2486 on 6 and 63581 DF, p-value: < 2.2e-16

```

Z racji interakcji, wpływ cholesterolu na ciśnienie skurczowe krwi jest inny dla kobiet i inny dla mężczyzn.

Kobiety z poziomem cholesterolu powyżej normy mają ciśnienie skurczowe krwi wyższe o 2 mm/Hg, natomiast kobiety z cholesterololem dużo powyżej normy mają średnio ciśnienie skurczowe krwi wyższe aż o 4 mm/Hg.

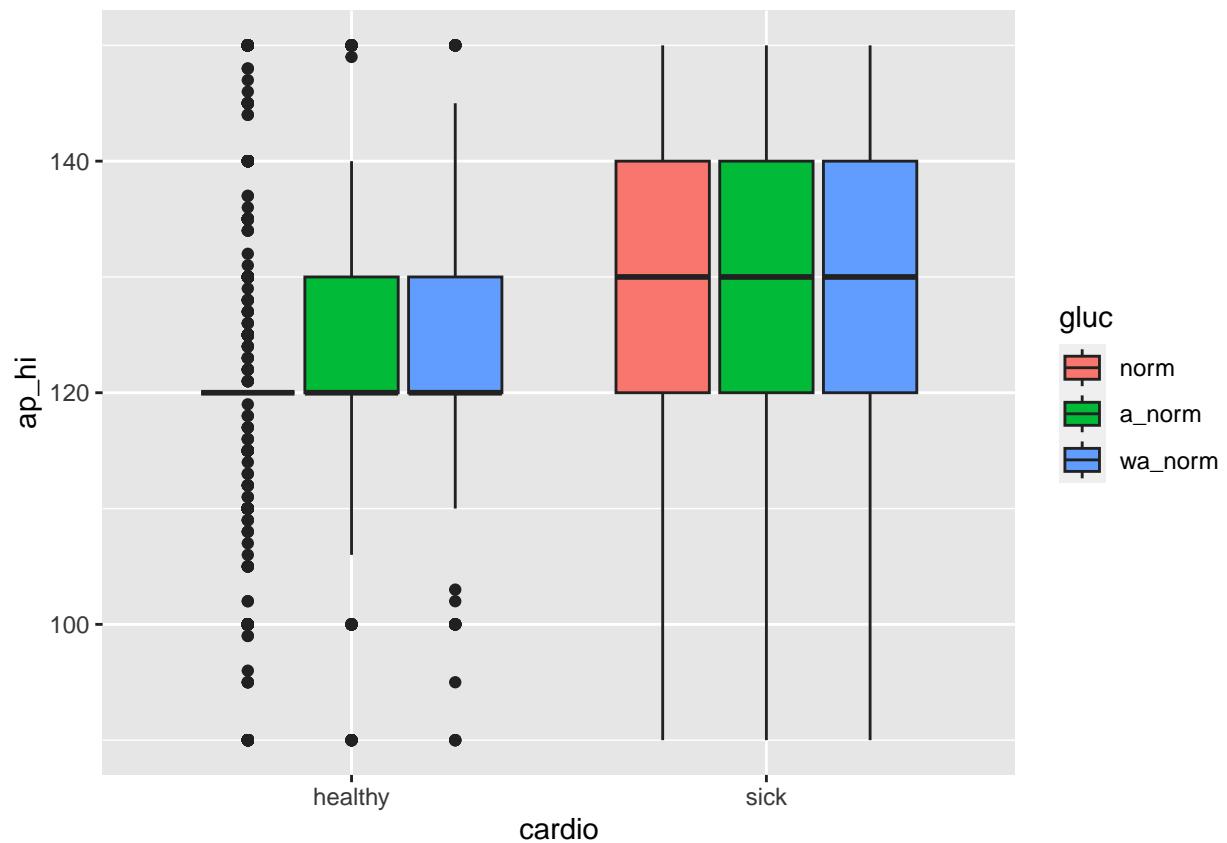
Sytuacja wygląda inaczej w przypadku mężczyzn.

Mężczyźni z poziomem cholesterolu powyżej normy mają ciśnienie skurczowe krwi wyższe o ok. 3 mm/Hg, natomiast mężczyźni z cholesterololem dużo powyżej normy mają średnio ciśnienie skurczowe krwi wyższe o ok. 4.6 mm/Hg.

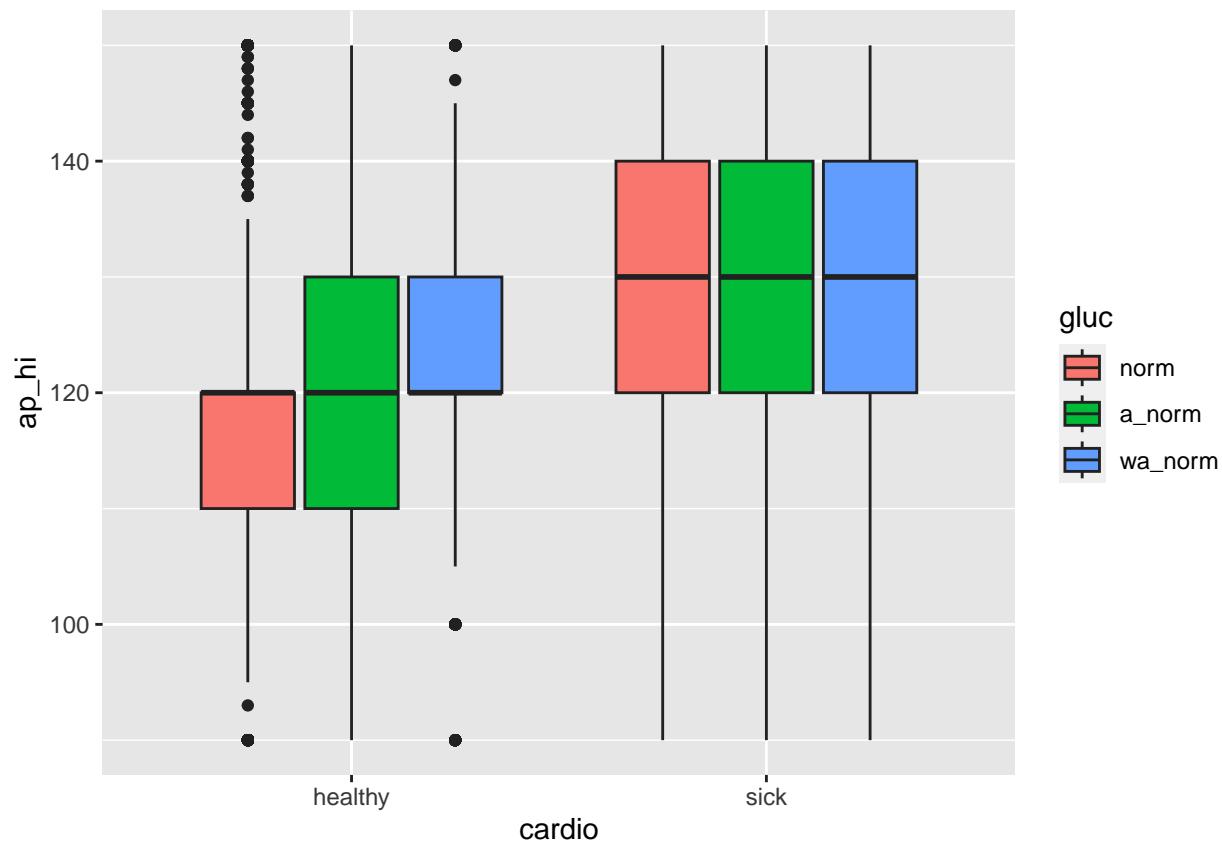
Widać znaczną różnicę w oddziaływaniu cholesterolu na ciśnienie krwi wśród kobiet, a wśród mężczyzn. Przyrost u kobiet między wysokim poziomem cholesterolu, a bardzo wysokim wyniósł 2 mm/Hg, natomiast przyrost w grupie mężczyzn wyniósł 1.6 mm/Hg. Powyższy wykres idealnie to obrazuje.

## Gluc

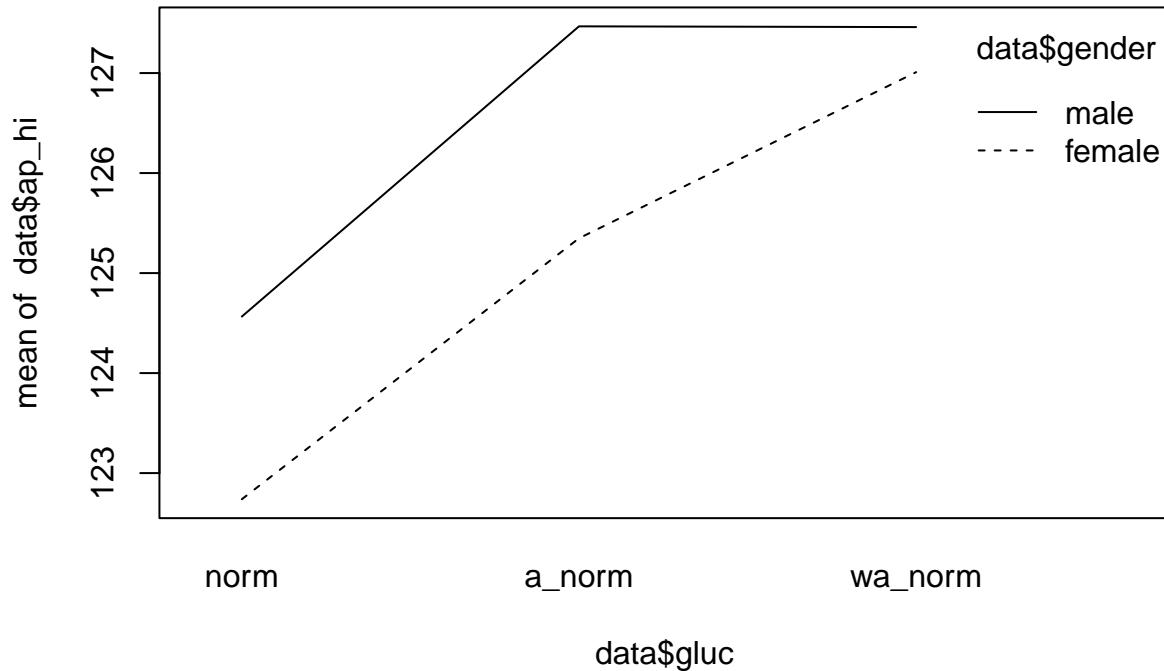
```
ggplot(data_male, aes(x = cardio, y = ap_hi, fill = gluc)) + geom_boxplot()
```



```
ggplot(data_female, aes(x = cardio, y = ap_hi, fill = gluc)) + geom_boxplot()
```



```
interaction.plot(x.factor = data$gluc,
                 trace.factor = data$gender,
                 response = data$ap_hi)
```



Zachodzi interakcja.

```
summary(lm(ap_hi ~ cardio + gender * gluc, data = data))
```

```
##
## Call:
## lm(formula = ap_hi ~ cardio + gender * gluc, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -42.111 -8.600   0.263   9.649  32.013 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 117.98681  0.07413 1591.704 < 2e-16 ***
## cardio      10.61309  0.09234 114.940 < 2e-16 ***
## gendermale  1.75068  0.10396  16.840 < 2e-16 ***
## gluca_norm  1.61982  0.22305   7.262 3.85e-13 ***
## glucwa_norm 2.49418  0.21163  11.786 < 2e-16 ***
## gendermale:gluca_norm  0.08529  0.38432    0.222    0.824  
## gendermale:glucwa_norm -0.73379  0.37609   -1.951    0.051 .  
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.57 on 63581 degrees of freedom
## Multiple R-squared:  0.1825, Adjusted R-squared:  0.1824
```

```
## F-statistic: 2365 on 6 and 63581 DF, p-value: < 2.2e-16
```

Wpływ poziomu glukozy powyżej normy jest taki sam wśród mężczyzn jak i kobiet. Znaczne różnice pojawiają się jeżeli poziom glukozy jest dużo powyżej normy. Wtedy wpływ poziomu glukozy jest zależny od płci.

Kobieta z poziomem glukozy we krwi powyżej normy ma średnio większe ciśnienie skurczowe krwi o 1.6 mm/Hg, natomiast gdy poziom glukozy jest dużo powyżej normy to ciśnienie skurczowe krwi jest większe o aż 2.5 mm/Hg.

Mężczyzna tak samo jak kobieta, ma średnio większe ciśnienie skurczowe krwi o 1.6 mm/Hg w przypadku glukozy powyżej normy, natomiast gdy poziom glukozy we krwi dużo powyżej normy to ciśnienie skurczowe krwi jest średnio większe o 1.8 mm/Hg.

Jak widać, wśród mężczyzn to czy pacjent ma poziom glukozy powyżej normy czy dużo powyżej normy nie ma aż tak dużego znaczenia, natomiast wśród kobiet ta różnica jest znacznie większa.

## BMI

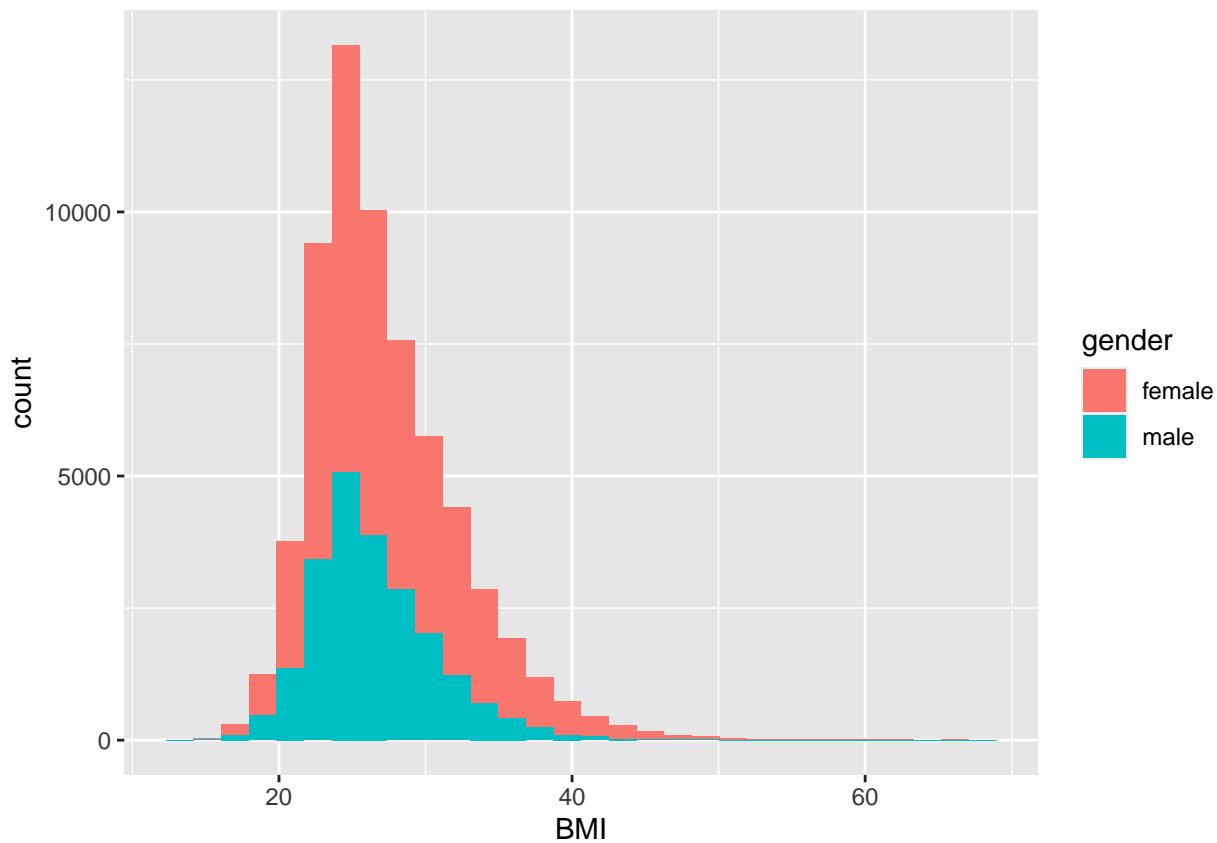
Mając dostęp do wagi oraz wzoru pacjentów można obliczyć wskaźnik BMI i sprawdzić wpływ wychudzenia lub otyłości na ciśnienie skurczowe krwi.

```
data$BMI = data$weight/(data$height/100)^2
summary(data$BMI)

##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##    13.52    23.81    26.17    27.24    29.76    68.31

ggplot(data, aes(x = BMI, fill = gender)) + geom_histogram()

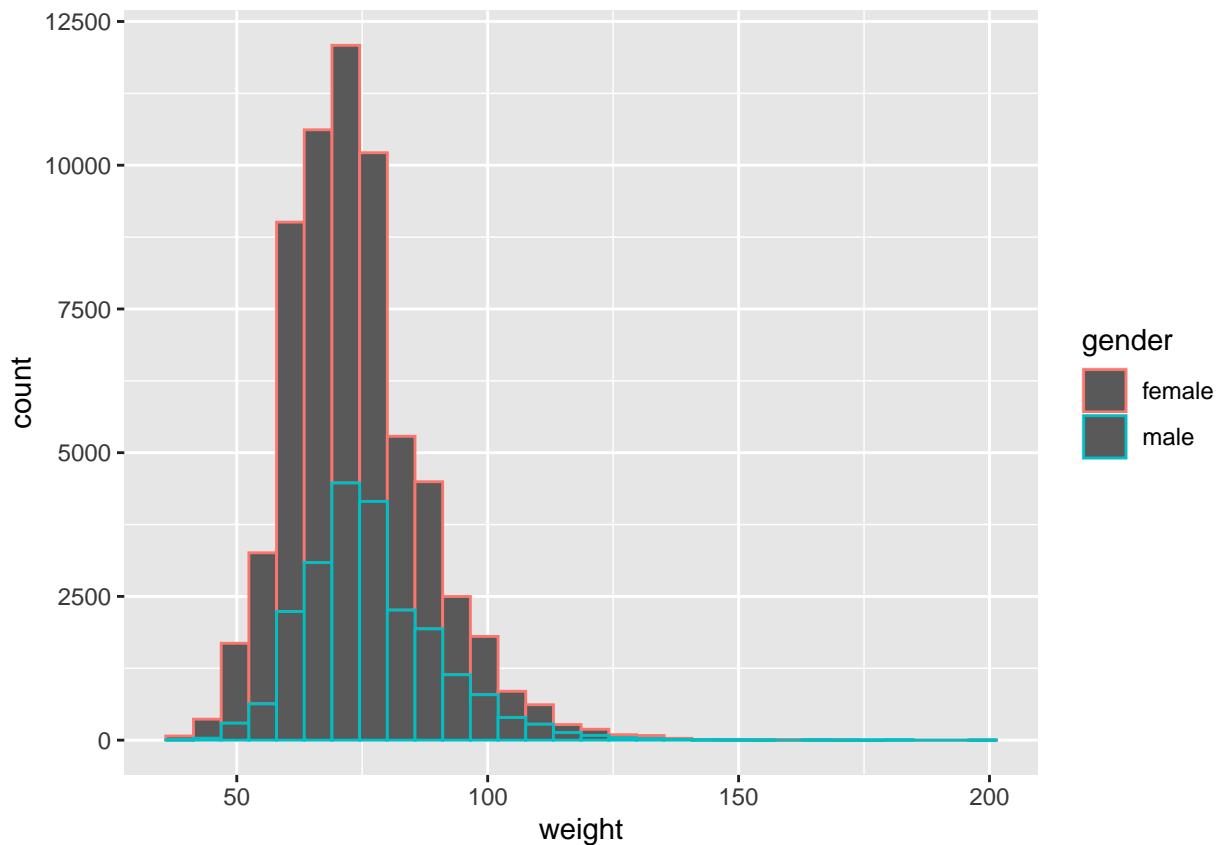
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Wskaźnik BMI opiera się o wzrost i wagę pacjenta. Z racji, że kobiety są niższe i lżejsze od mężczyzn, może nie będzie trzeba uwzględniać zmiennej 'gender' przy budowie modelu.

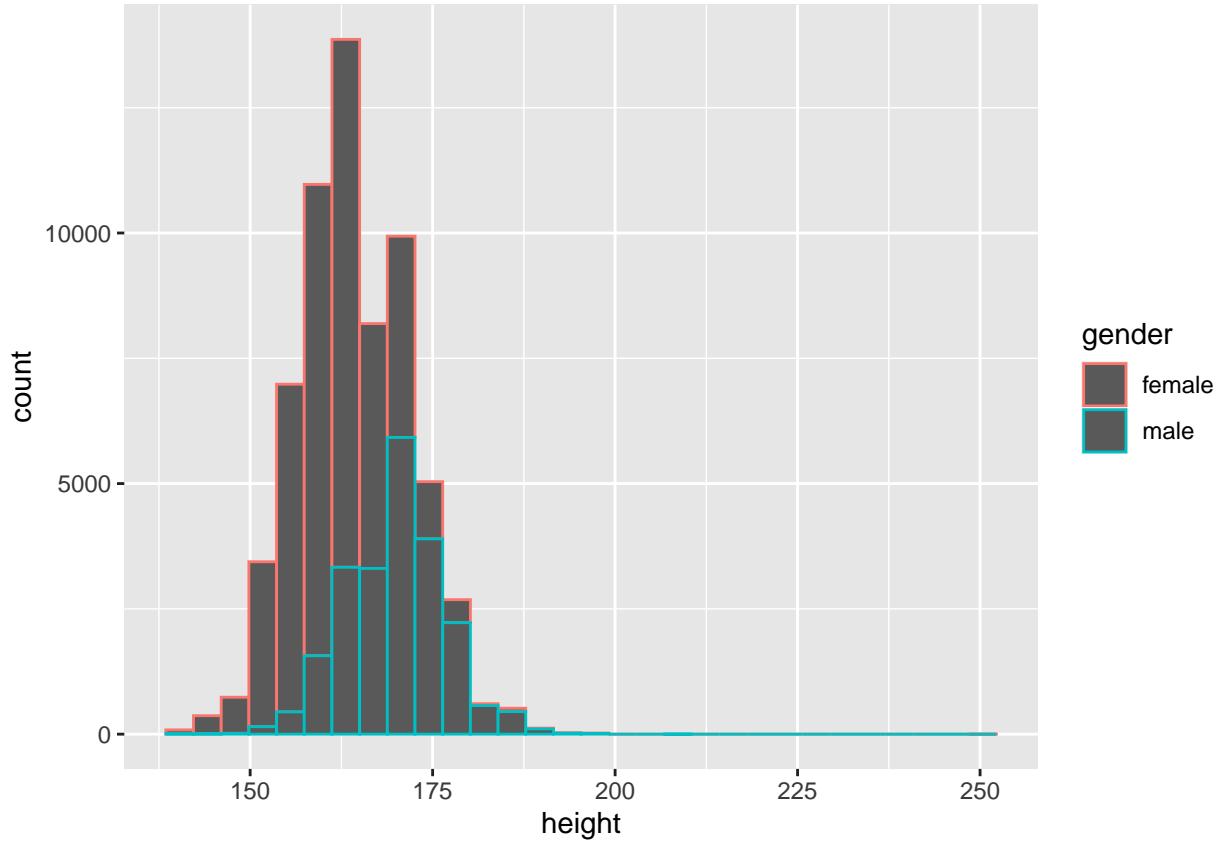
```
ggplot(data, aes(x = weight, color = gender)) + geom_histogram()
```

```
## 'stat_bin()' using 'bins = 30'. Pick better value with 'binwidth'.
```



```
ggplot(data, aes(x = height, color = gender)) + geom_histogram()
```

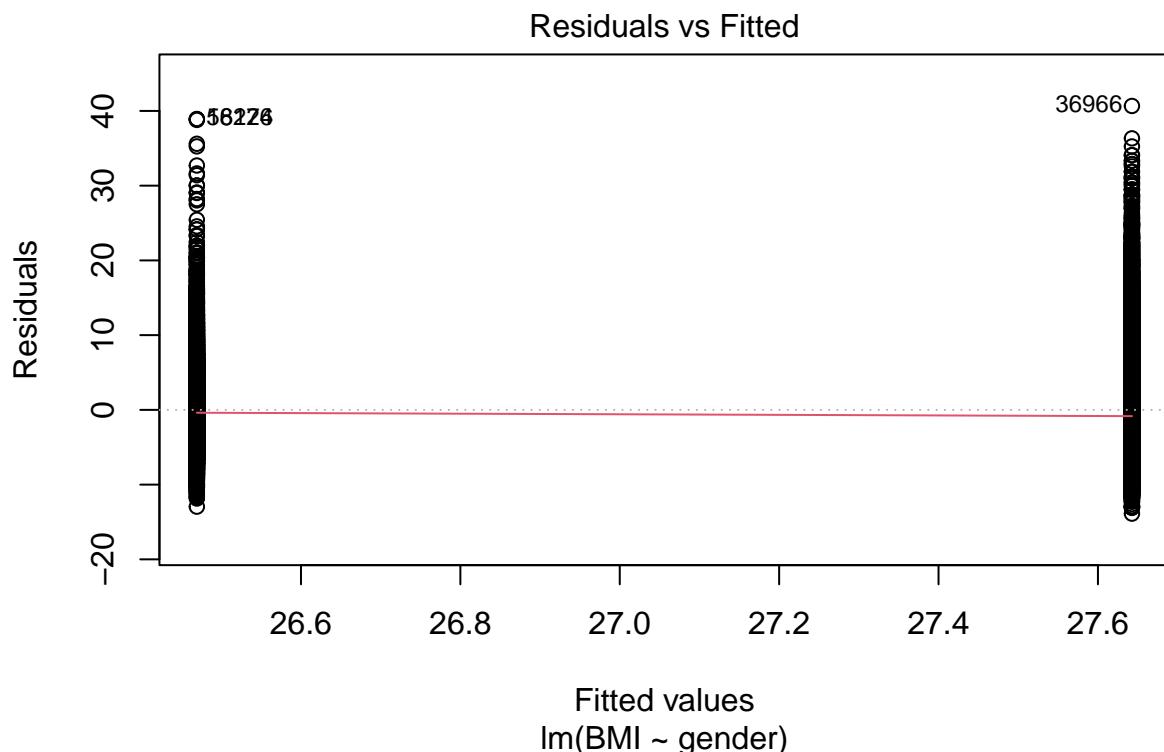
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

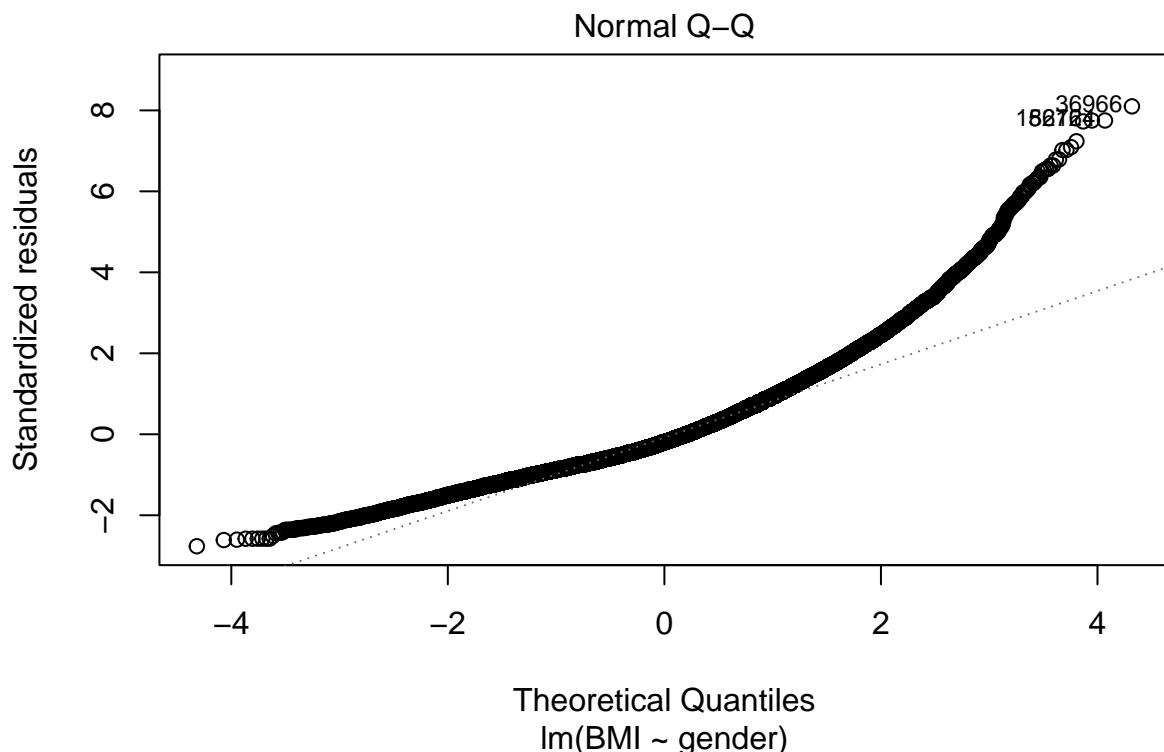


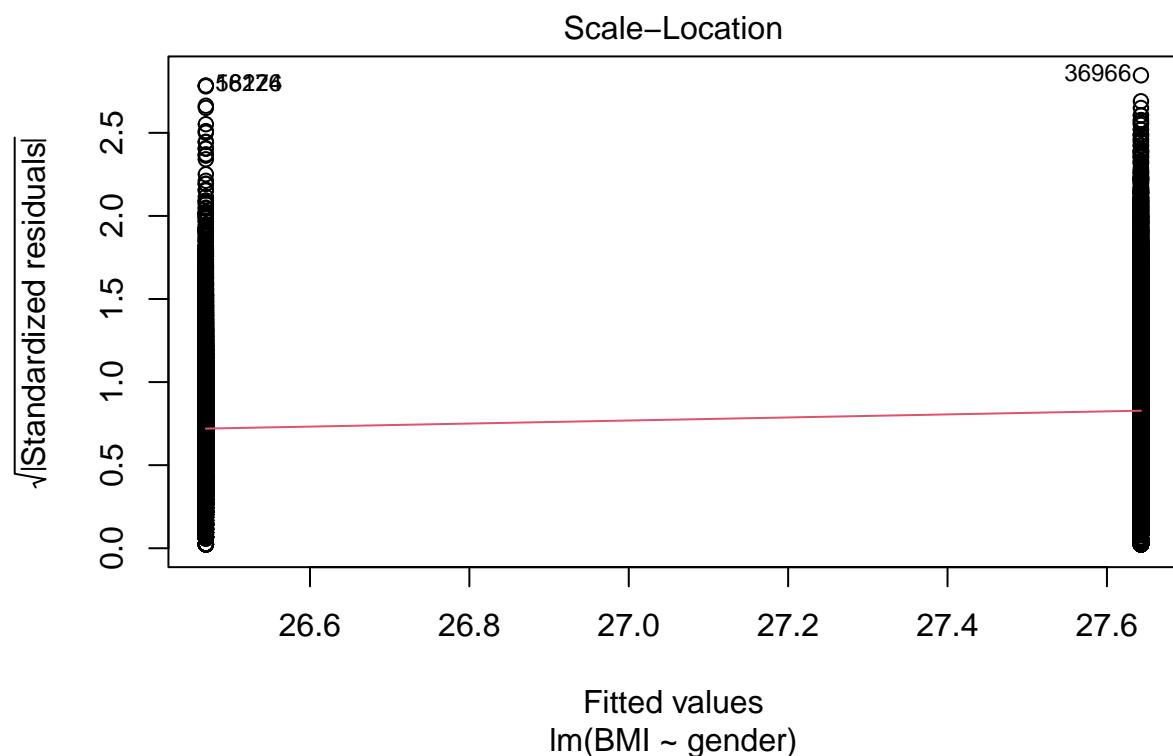
```
summary(lm(BMI ~ gender, data = data))
```

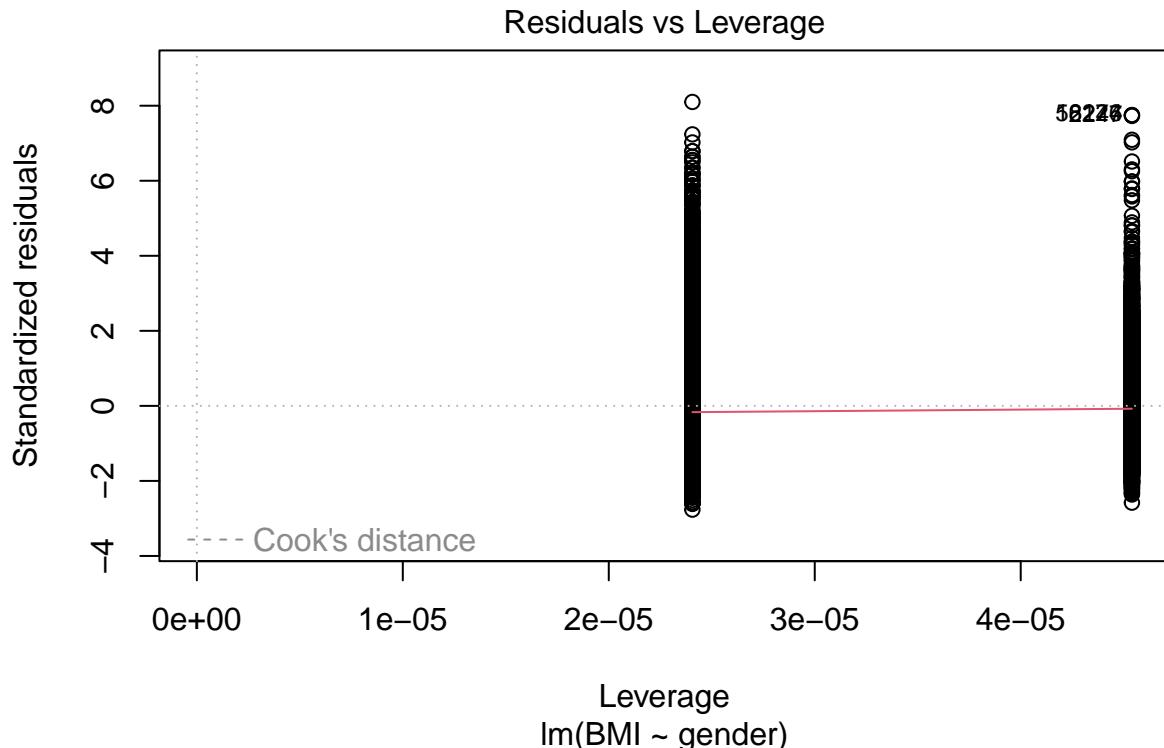
```
##
## Call:
## lm(formula = BMI ~ gender, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -13.883  -3.479  -0.970   2.655  40.666 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 27.64269   0.02463 1122.44 <2e-16 ***
## gendermale -1.17336   0.04184 -28.04  <2e-16 ***  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.021 on 63586 degrees of freedom
## Multiple R-squared:  0.01221,    Adjusted R-squared:  0.0122 
## F-statistic: 786.3 on 1 and 63586 DF,  p-value: < 2.2e-16
```

```
plot(lm(BMI ~ gender, data = data))
```



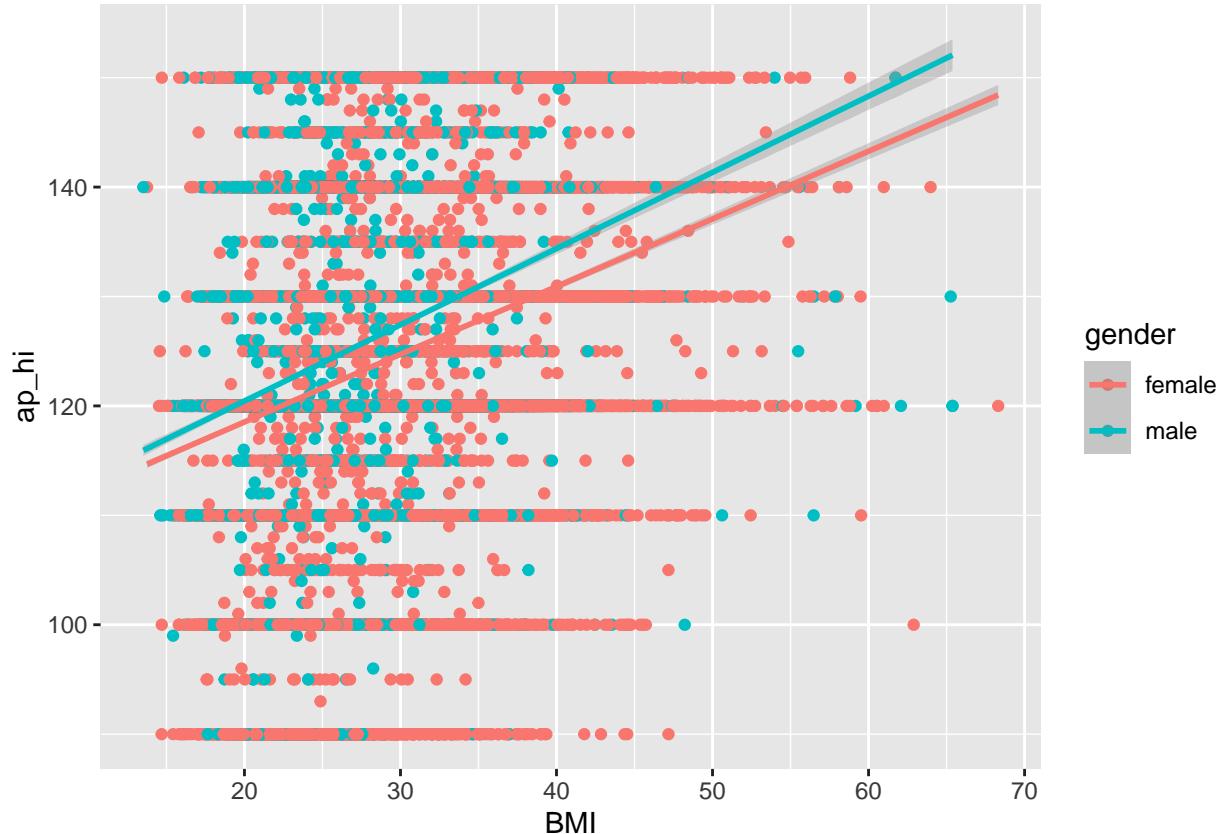






Zróżnicowanie w BMI względem płci jest istotne statystycznie, natomiast trudno będzie dobrze opisać płeć pacjenta za pomocą samego BMI. Przy budowie modelu będzie trzeba mimo wszystko uwzględnić płeć.

```
ggplot(data = data, aes(x = BMI, y = ap_hi, group = gender, colour = gender)) + geom_point() + geom_smooth()
## `geom_smooth()` using formula = 'y ~ x'
```



Brak interakcji - wpływ BMI na skurczowe ciśnienie krwi jest bardzo podobny zarówno wśród kobiet jak i mężczyzn.

```
summary(lm(ap_hi ~ cardio + gender + BMI, data = data))

##
## Call:
## lm(formula = ap_hi ~ cardio + gender + BMI, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -45.039  -7.848   0.636   7.517  37.328 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.059e+02  2.544e-01  416.35 <2e-16 ***
## cardio      9.907e+00  9.183e-02  107.88 <2e-16 ***
## gendermale  2.215e+00  9.534e-02   23.23 <2e-16 ***
## BMI         4.604e-01  9.127e-03   50.44 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.37 on 63584 degrees of freedom
## Multiple R-squared:  0.211, Adjusted R-squared:  0.211 
## F-statistic: 5669 on 3 and 63584 DF,  p-value: < 2.2e-16
```

Współczynnik BMI jest istotny statystycznie. Jest on dodatni, zatem wraz ze wzrostem BMI wzrasta skurczowe ciśnienie krwi.

## WNIOSKI

- Aktywność fizyczna obniża istotnie ciśnienie skurczowe krwi.
- Spożywanie alkoholu bardzo mocno wpływa na ciśnienie skurczowe krwi.
- Efekt palenia ma wpływ na ciśnienie skurczowe krwi.
- Poziom cholesterolu wpływa na ciśnienie skurczowe krwi. Jego wpływ jest również zależny od płci.
- Poziom glukozy powyżej normy ma wpływ na ciśnienie skurczowe krwi. Wpływ poziomu glukozy jest zależny od płci.
- Im większe BMI tym większe skurczowe ciśnienie krwi.

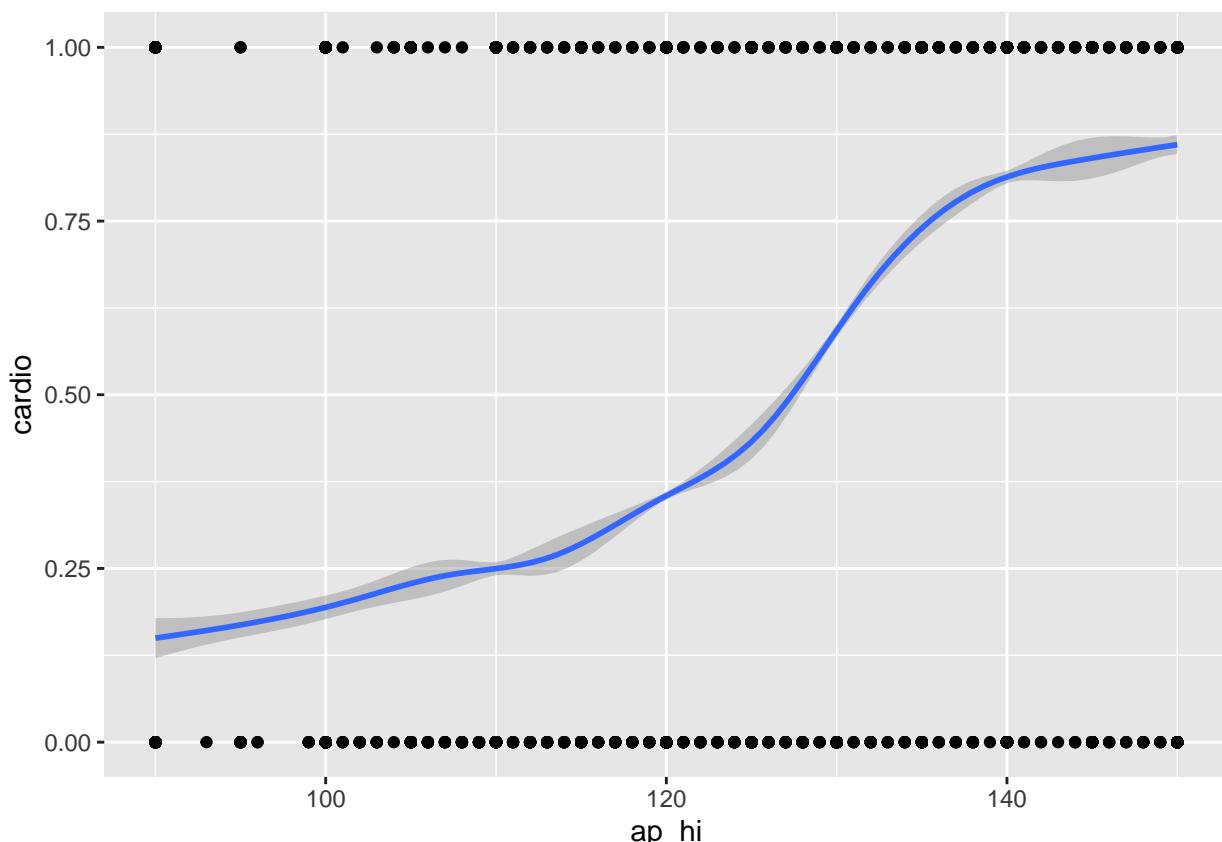
## CZEŚĆ II

**HIPOTEZA:** Czy wysokie skurczowe ciśnienie krwi zwiększa szanse na posiadanie choroby układu krążenia?

```
data = data %>% mutate_at('cardio', as.numeric)
data$cardio = data$cardio - 1

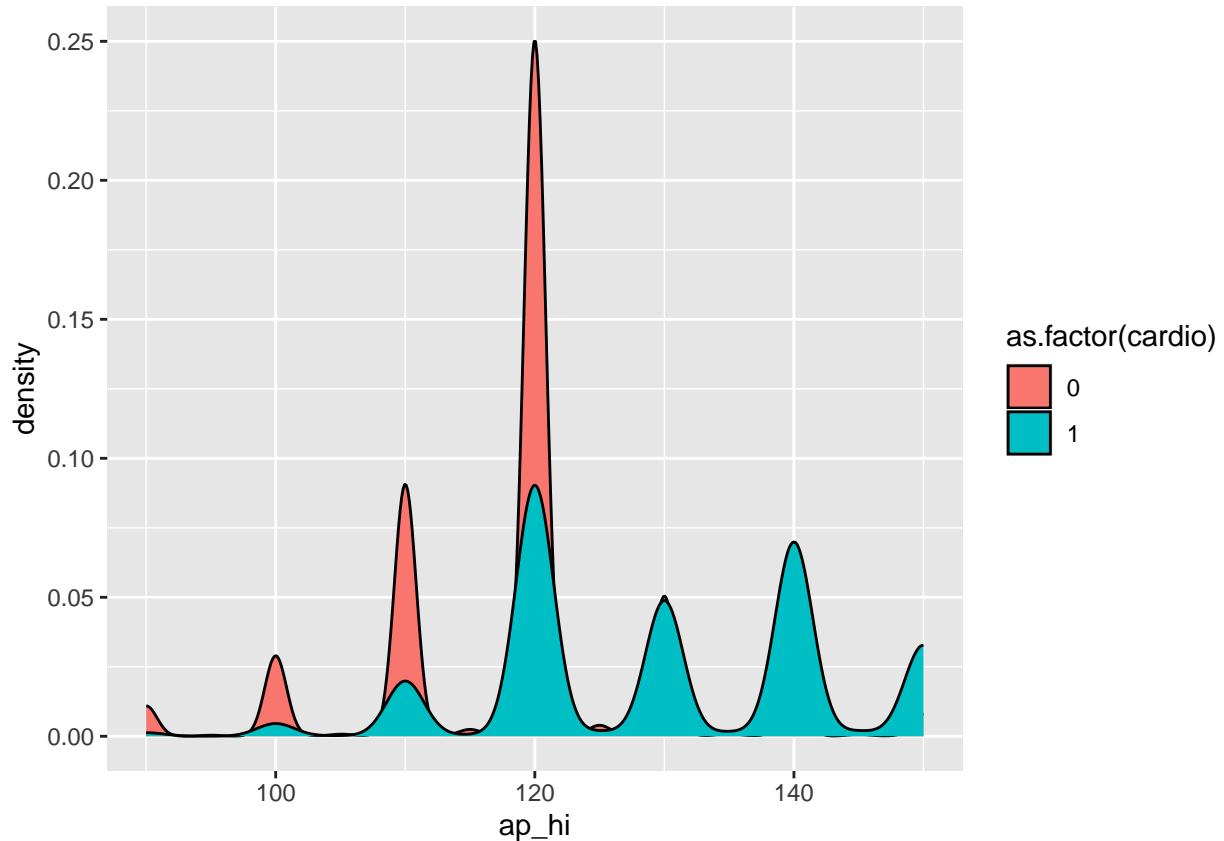
ggplot(data = data, aes(x = ap_hi, y = cardio)) + geom_point() + geom_smooth()

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Wraz ze wzrostem ciśnienia skurczowego krwi rośnie prawdopodobieństwo, że dana osoba ma chorobę układu krążenia.

```
ggplot(data = data, aes(x = ap_hi, fill = as.factor(cardio))) + geom_density()
```



```
xtabs(~cardio + ap_hi, data = data)
```

```
##      ap_hi
## cardio  90   93   95   96   99   100  101  102  103  104  105  106
##   0    778     1   26     2     4  2052     3     8     7     1    48     9
##   1    139     0     2     0     0   496     1     0     1     5    16     1
##      ap_hi
## cardio 107  108  109  110  111  112  113  114  115  116  117  118
##   0      6     8     9  6434     6    15    11     8   167     6    17    10
##   1      1     1     0  2154     3     5     3     3    49     3     3     4
##      ap_hi
## cardio 119  120  121  122  123  124  125  126  127  128  129  130
##   0      9 17773     7    13    12    15  267    11    15    24     4  3580
##   1      4  9808     5     3    10     2  169     5     7    10     2  5294
##      ap_hi
## cardio 131  132  133  134  135  136  137  138  139  140  141  142
##   0      3    10     3     6    65     3     5     8     1  1705     2     2
##   1      6     4     3     8   141    11     3    10     9  7569    18     5
##      ap_hi
## cardio 143  144  145  146  147  148  149  150
```

```

##      0      1      3     46      3      3      6      3    570
##      1     10      6    177      5      9     10      6   3555

```

Osób zdrowych jest znacznie więcej w przedziale ciśnienia skurczowego krwi na poziomie < 130 mm/Hg, natomiast osób chorych jest znacznie więcej powyżej ciśnienia skurczowego 130 mm/Hg.

```

model_zero = glm(cardio ~ ap_hi, family = 'binomial', data = data)
summary(model_zero)

```

```

##
## Call:
## glm(formula = cardio ~ ap_hi, family = "binomial", data = data)
##
## Deviance Residuals:
##       Min        1Q    Median        3Q       Max
## -2.0387 -0.9954 -0.4951  1.0321  2.4044
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.000e+01  1.028e-01 -97.30  <2e-16 ***
## ap_hi        7.963e-02  8.271e-04   96.28  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895 on 63587 degrees of freedom
## Residual deviance: 75647 on 63586 degrees of freedom
## AIC: 75651
##
## Number of Fisher Scoring iterations: 3

```

Wartości współczynników modelu wskazują na to, że wraz ze wzrostem ciśnienia skurczowego krwi wzrasta prawdopodobieństwo, że osoba ma chorobę układu krążenia.

Reszta dewiancyjna wskazuje na duże niedopasowanie modelu.

```
a = exp(-10 + 0.08 * 120)
```

Szansa, że osoba mająca ciśnienie skurczowe krwi na poziomie 120 mm/Hg ma chorobę układu krążenia wynosi 0.67

```
a/(1+a)
```

```
## [1] 0.4013123
```

Prawdopodobieństwo, że osoba mająca ciśnienie skurczowe krwi na poziomie 120 mm/Hg ma chorobę układu krążenia wynosi 0.4

```
exp(0.08)
```

```
## [1] 1.083287
```

Jeśli ciśnienie skurczowe krwi wzrośnie o 1 mm/Hg to szansa na to, że osoba ma chorobę układu krążenia wzrasta 1.08 razy

```
b = exp(-10 + 0.08 * 140)
```

Szansa, że osoba mająca ciśnienie skurczowe krwi na poziomie 140 mm/Hg ma chorobę układu krążenia wynosi 3.32

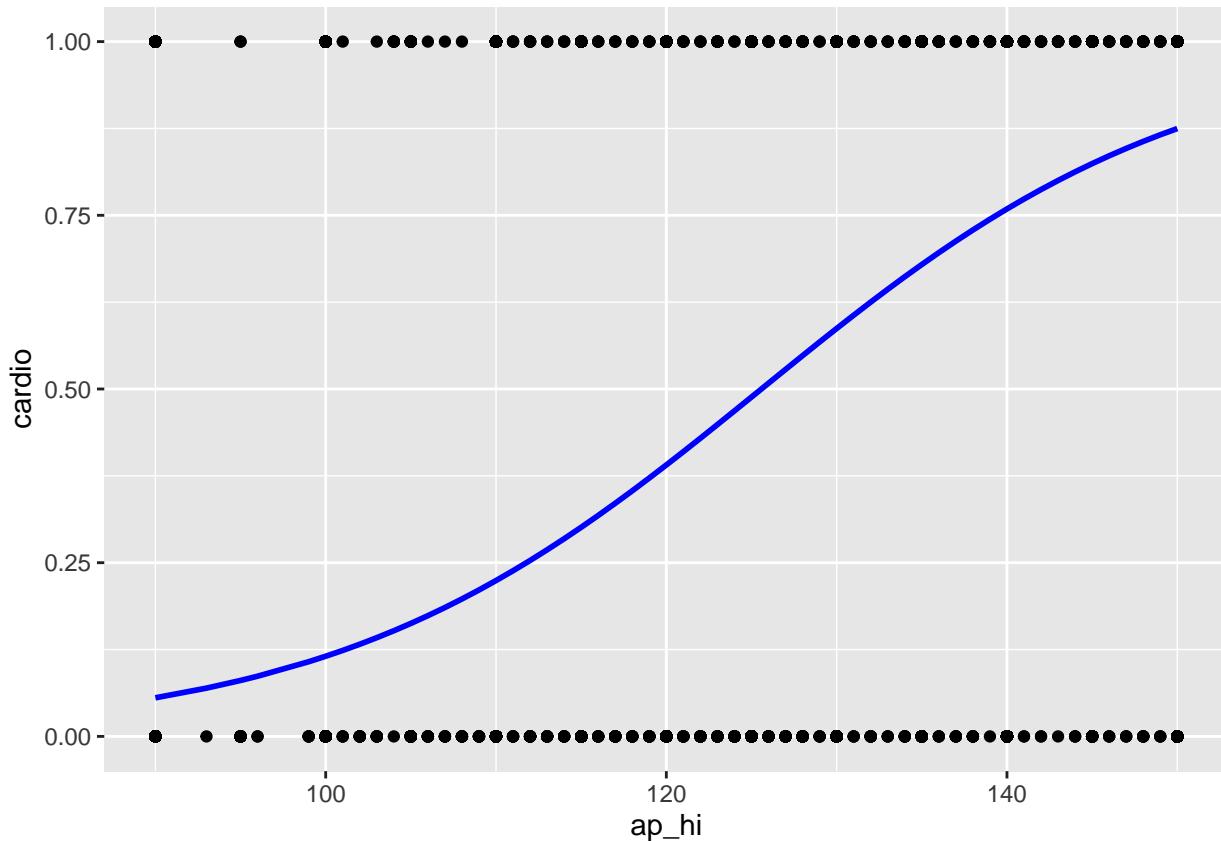
```
b/(b+1)
```

```
## [1] 0.7685248
```

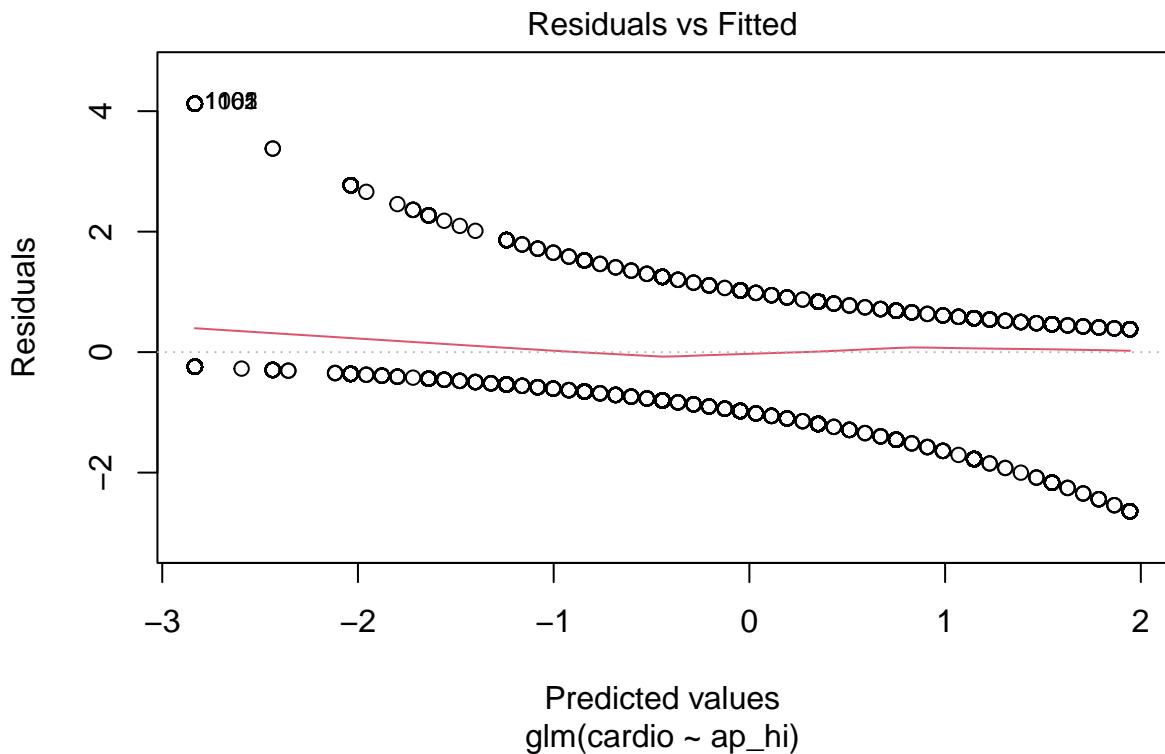
Prawdopodobieństwo, że osoba mająca ciśnienie skurczowe krwi na poziomie 140 mm/Hg ma chorobę układu krążenia wynosi 0.77

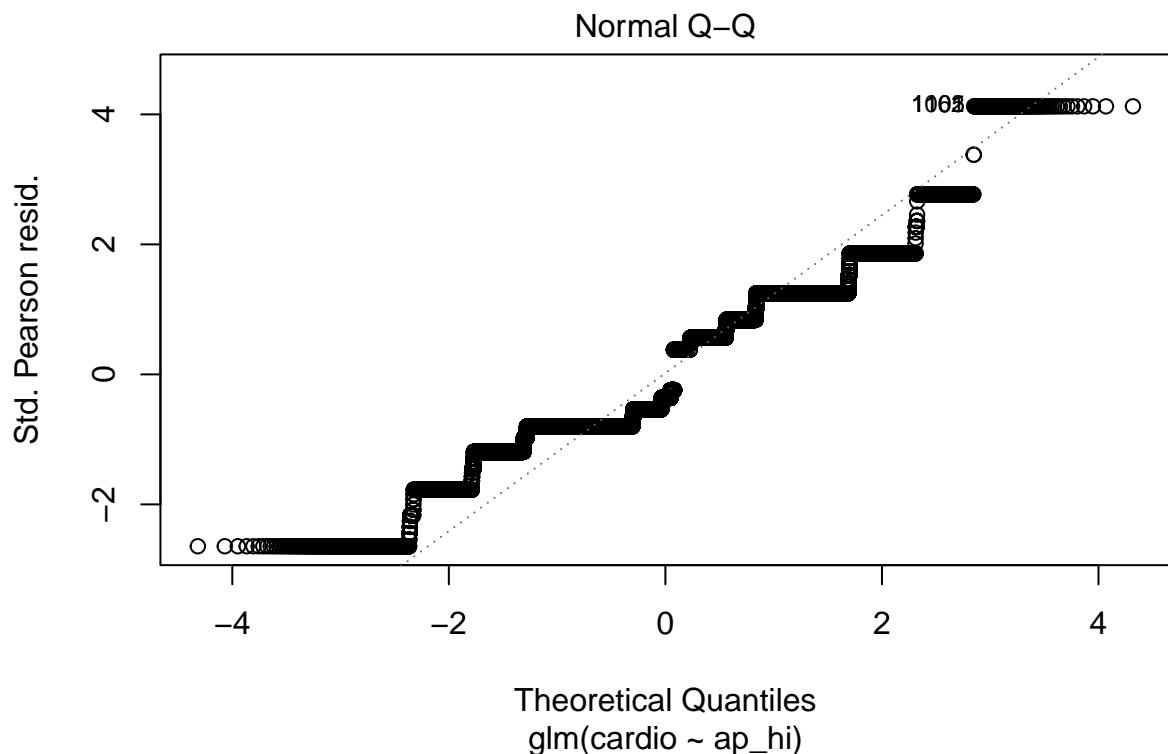
```
preds_plot = function(model){
  ggplot(data = data, aes(x = model$model$ap_hi)) +
    geom_point(aes(y = model$model$cardio)) +
    geom_line(aes(y = model$fitted), colour = 'blue', linewidth = 1) +
    labs(x = 'ap_hi', y = 'cardio')
}
```

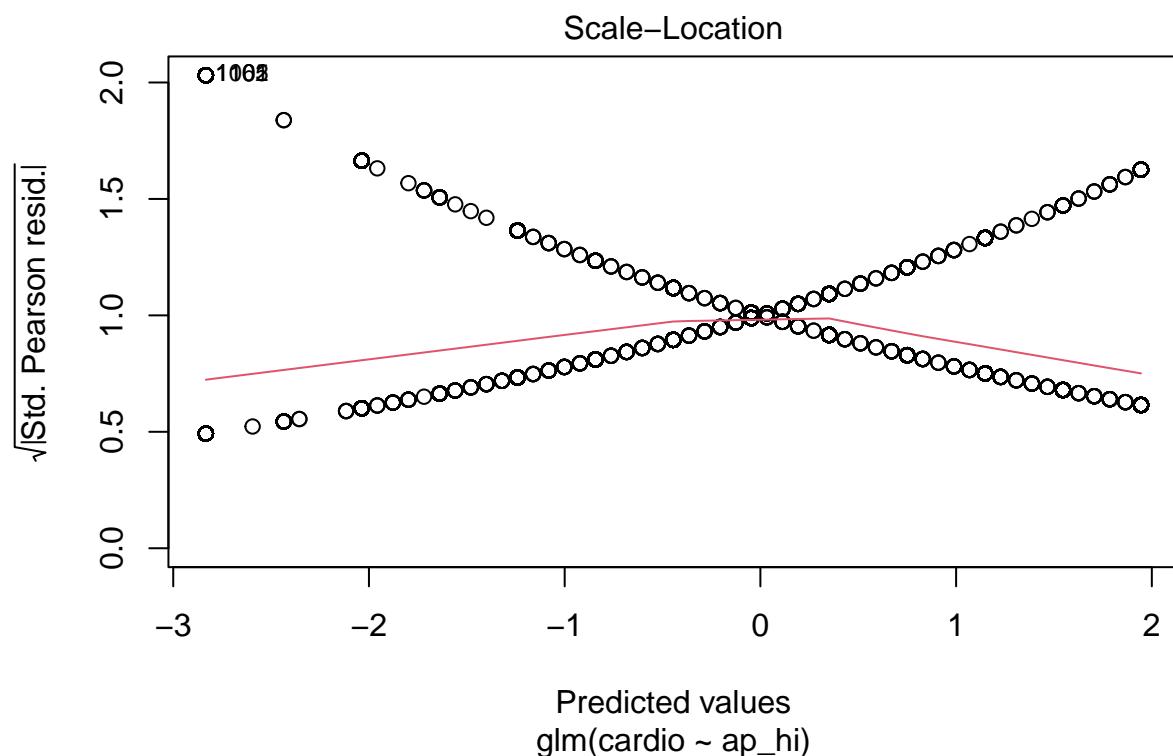
```
preds_plot(model_zero)
```

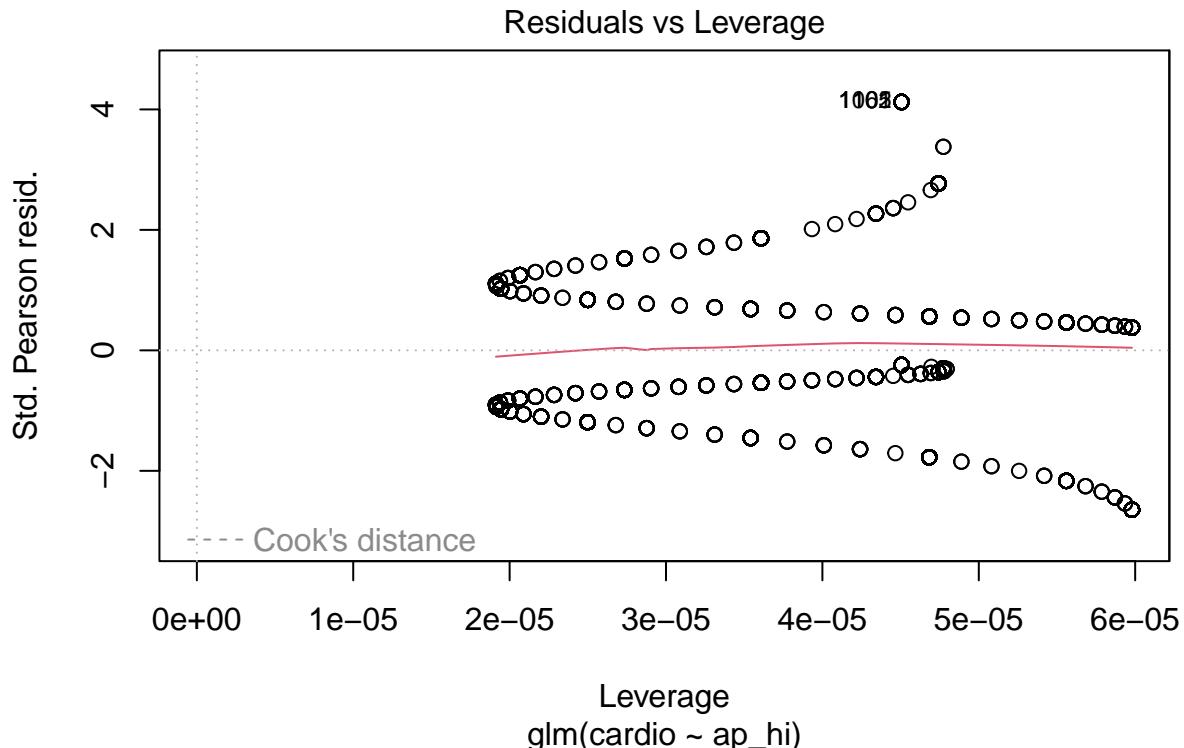


```
plot(model_zero)
```









Ostatni wykres diagnostyczny wskazuje na brak wartości odstających.

```

qres_plot = function(model){
  qres = statmod::qresid(model)
  pred = fitted(model)

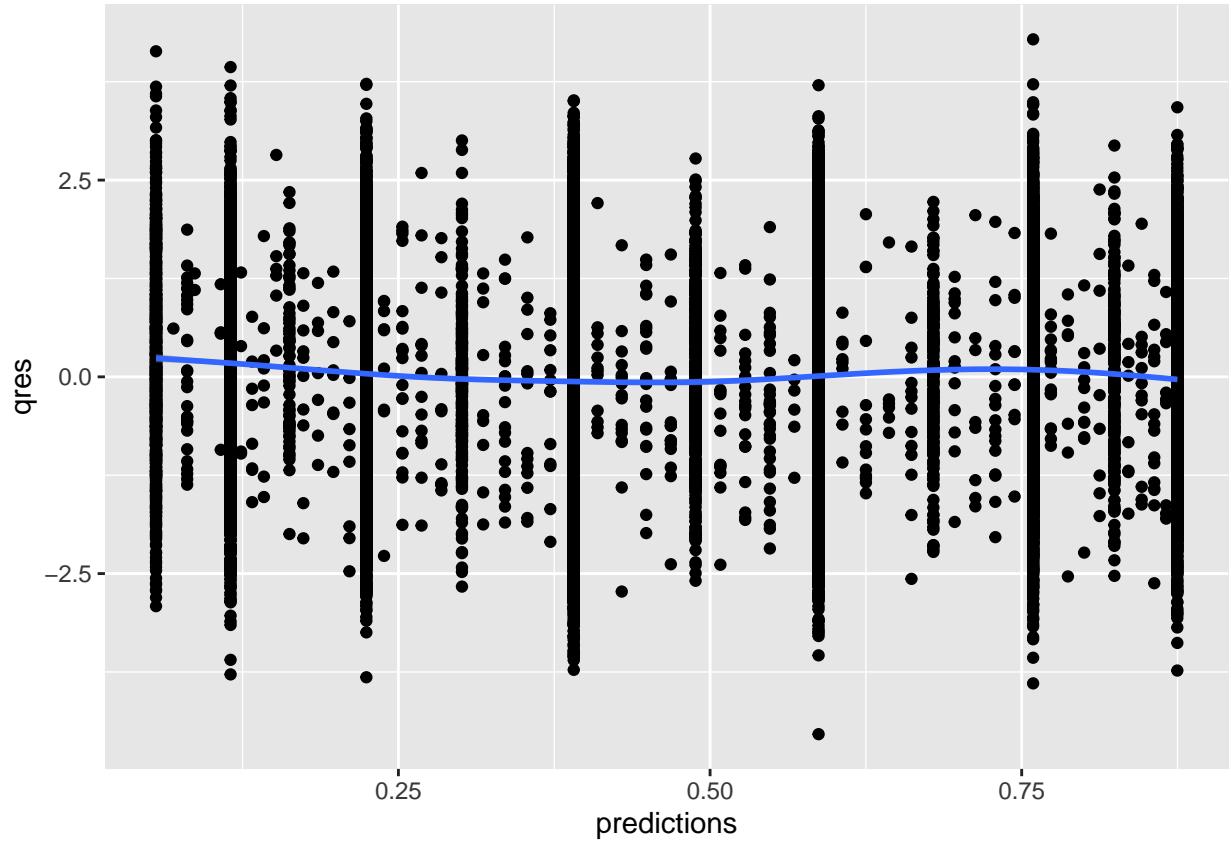
  df = data.frame(qres = qres, predictions = pred)

  ggplot(data = df, aes(x = predictions, y = qres)) + geom_point() + geom_smooth()
}

qres_plot(model_zero)

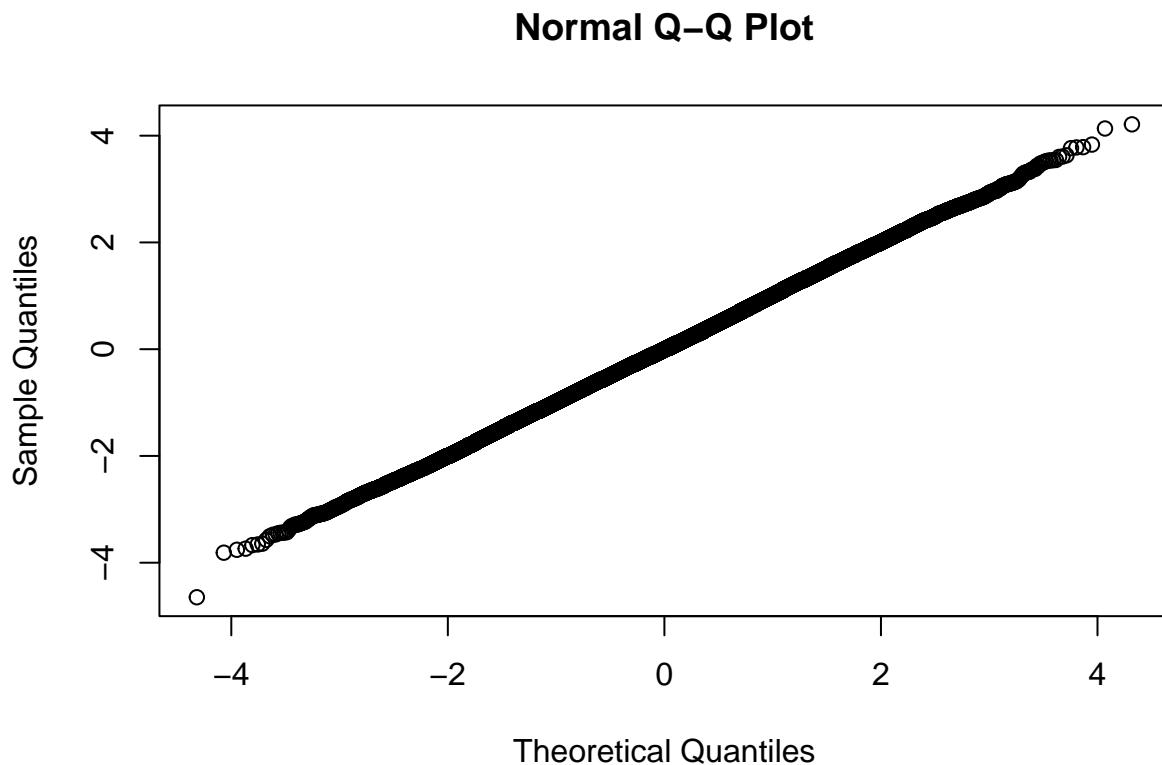
## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

```



Wykres reszt kwantylowych oscyluje w 0, natomiast widać niepożądane wzrosty wartości (skala wykresu jest dość duża).

```
qqnorm(qresid(model_zero))
```



Reszty pochodzą z rozkładu normalnego.

```
hoslem.test(x = model_zero$model$cardio, y = fitted(model_zero))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
##  data:  model_zero$model$cardio, fitted(model_zero)
##  X-squared = 456.21, df = 8, p-value < 2.2e-16
```

Test Hosmera-Lemeshowa wskazuje na niedopasowanie modelu.

Wartości ciśnienia skurczowego krwi tworzą rozkłady normalne. Teoria matematyczna sugeruje zastosowanie predyktorów  $x$  i  $x^2$ .

```
model_1 = glm(cardio ~ ap_hi + I(ap_hi^2), family = 'binomial', data = data)
summary(model_1)
```

```
##
## Call:
## glm(formula = cardio ~ ap_hi + I(ap_hi^2), family = "binomial",
##      data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.000000 -0.999999 -0.999999 -0.999999  1.000000
```

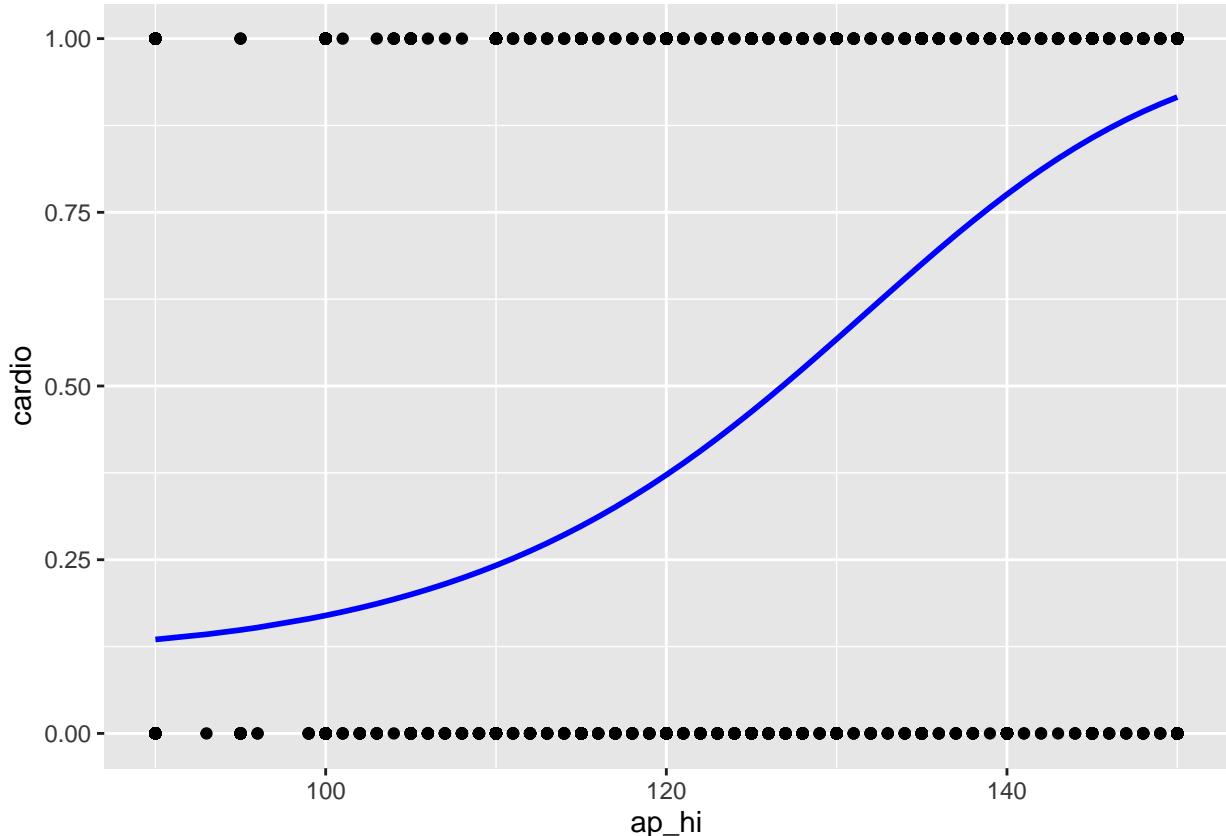
```

## -2.2247 -0.9647 -0.6101  1.0645  2.0006
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.610e+00 7.518e-01  4.802 1.57e-06 ***
## ap_hi      -1.396e-01  1.212e-02 -11.516 < 2e-16 ***
## I(ap_hi^2) 8.763e-04  4.872e-05 17.985 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895  on 63587  degrees of freedom
## Residual deviance: 75340  on 63585  degrees of freedom
## AIC: 75346
##
## Number of Fisher Scoring iterations: 4

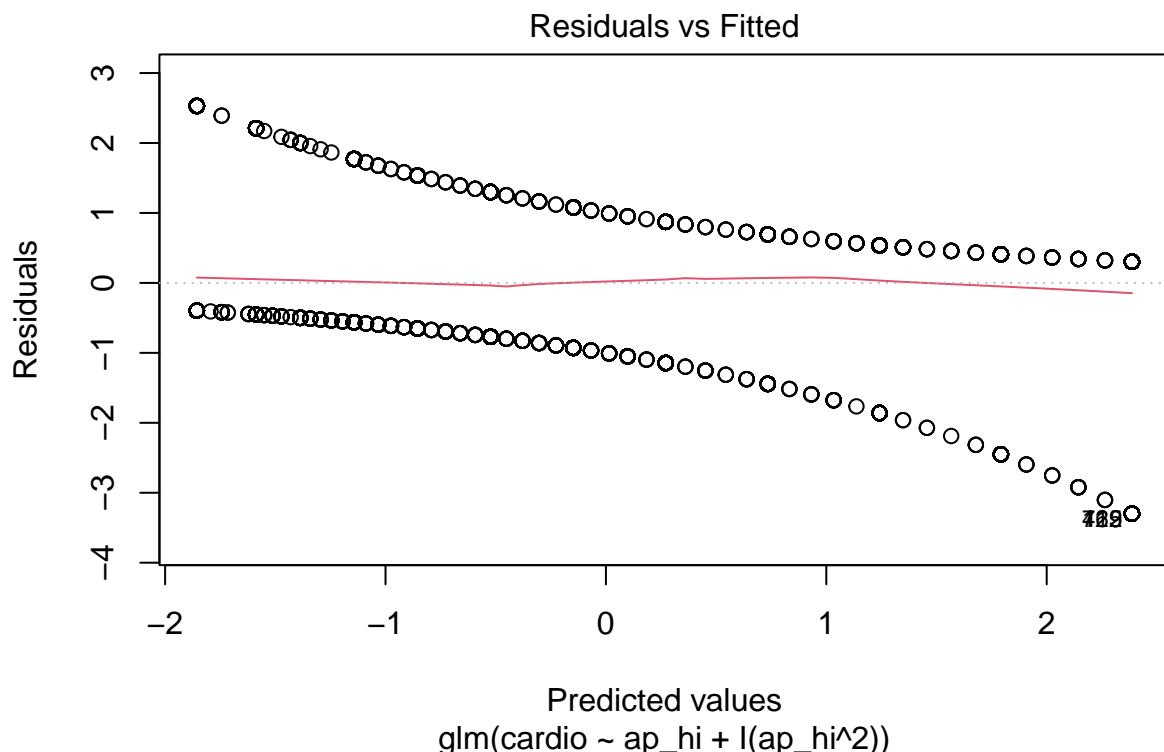
```

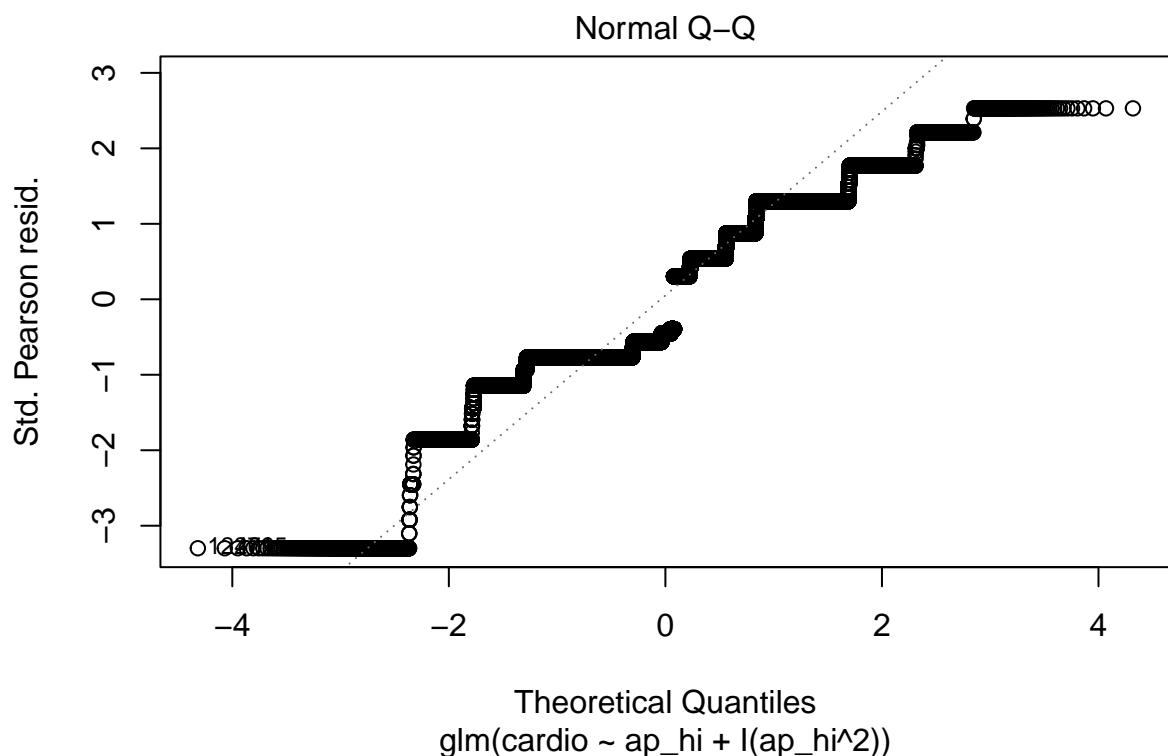
Wszystkie współczynniki są istotne statystycznie. Uwzględniając dodatkowo zmienną  $ap\_hi^2$  udało się zmniejszyć resztę dewiancyjną. Również kryterium AIC wskazuje na lepsze dopasowanie modelu.

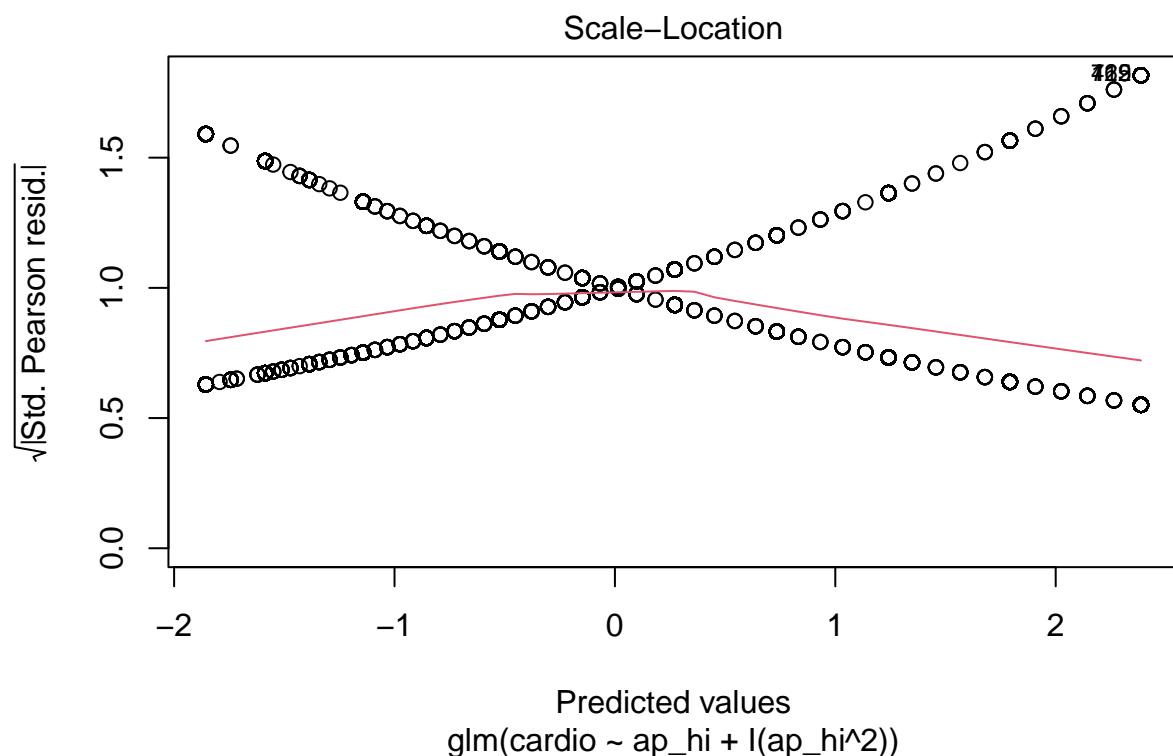
```
preds_plot(model_1)
```

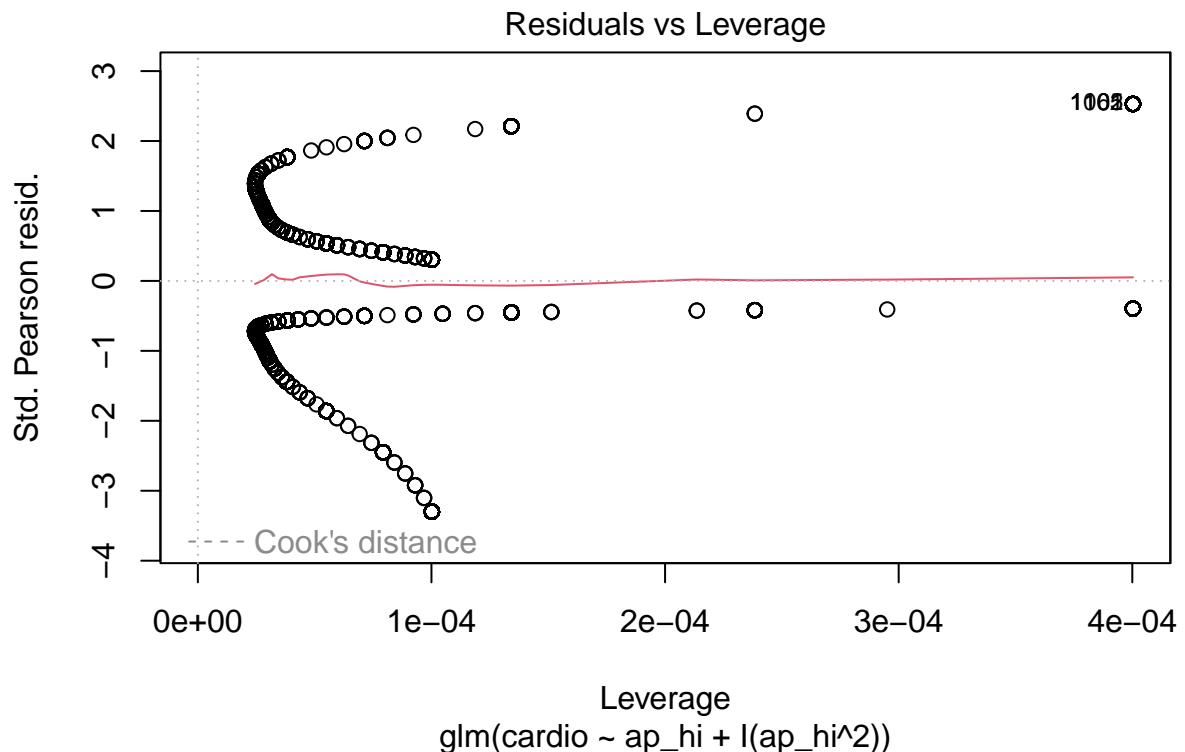


```
plot(model_1)
```



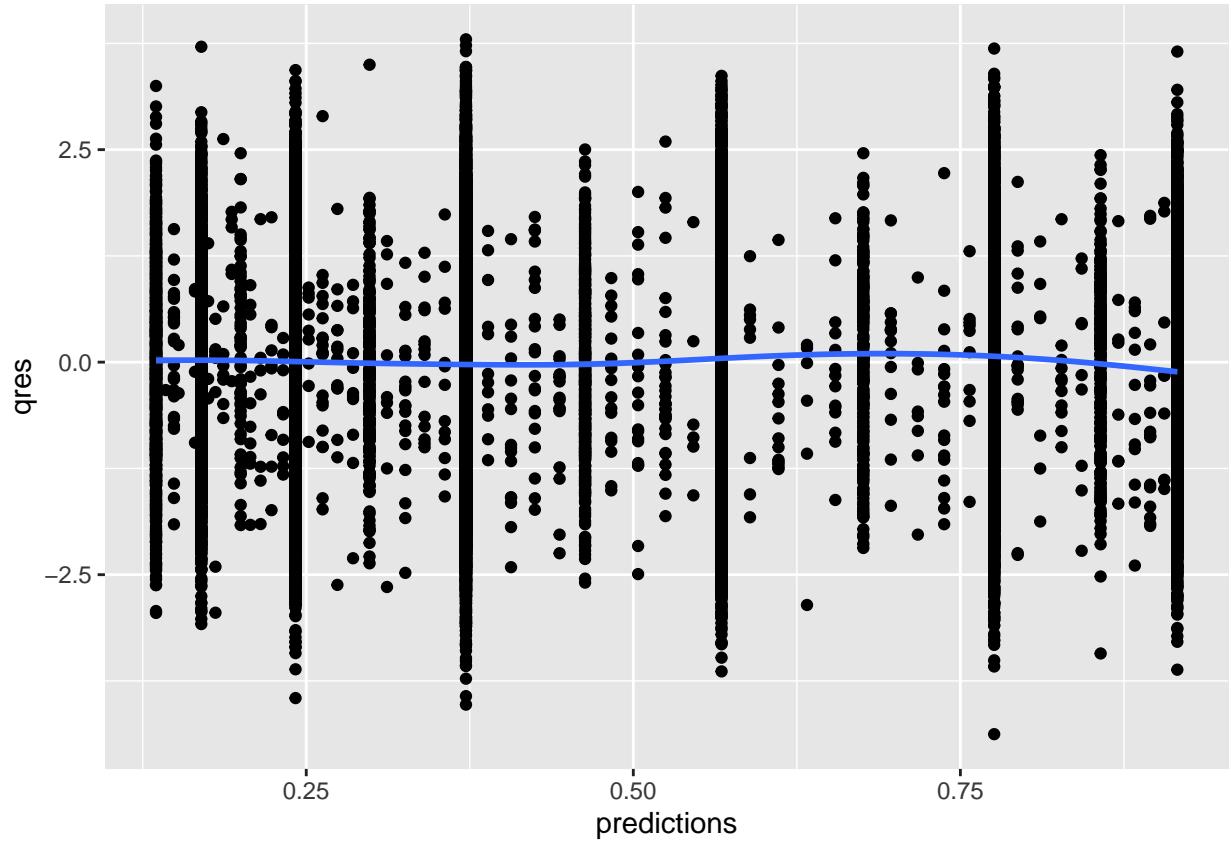






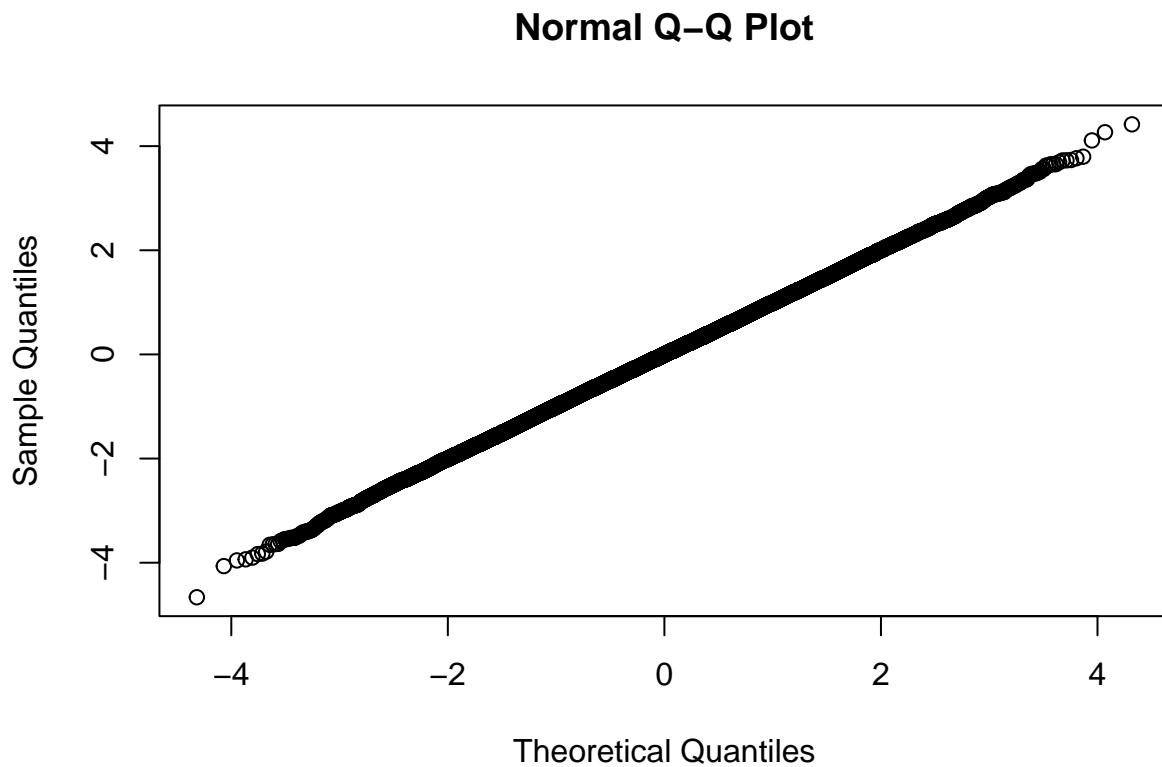
```
qres_plot(model_1)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Po uwzględnieniu dodatkowej zmiennej, wykres reszt kwantylowych znacznie się poprawił.

```
qqnorm(statmod::qresid(model_1))
```



Reszty dalej pochodzą z rozkładu normalnego.

```
hoslem.test(x = model_1$model$cardio, y = fitted(model_1))
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  model_1$model$cardio, fitted(model_1)
## X-squared = 312.55, df = 8, p-value < 2.2e-16
```

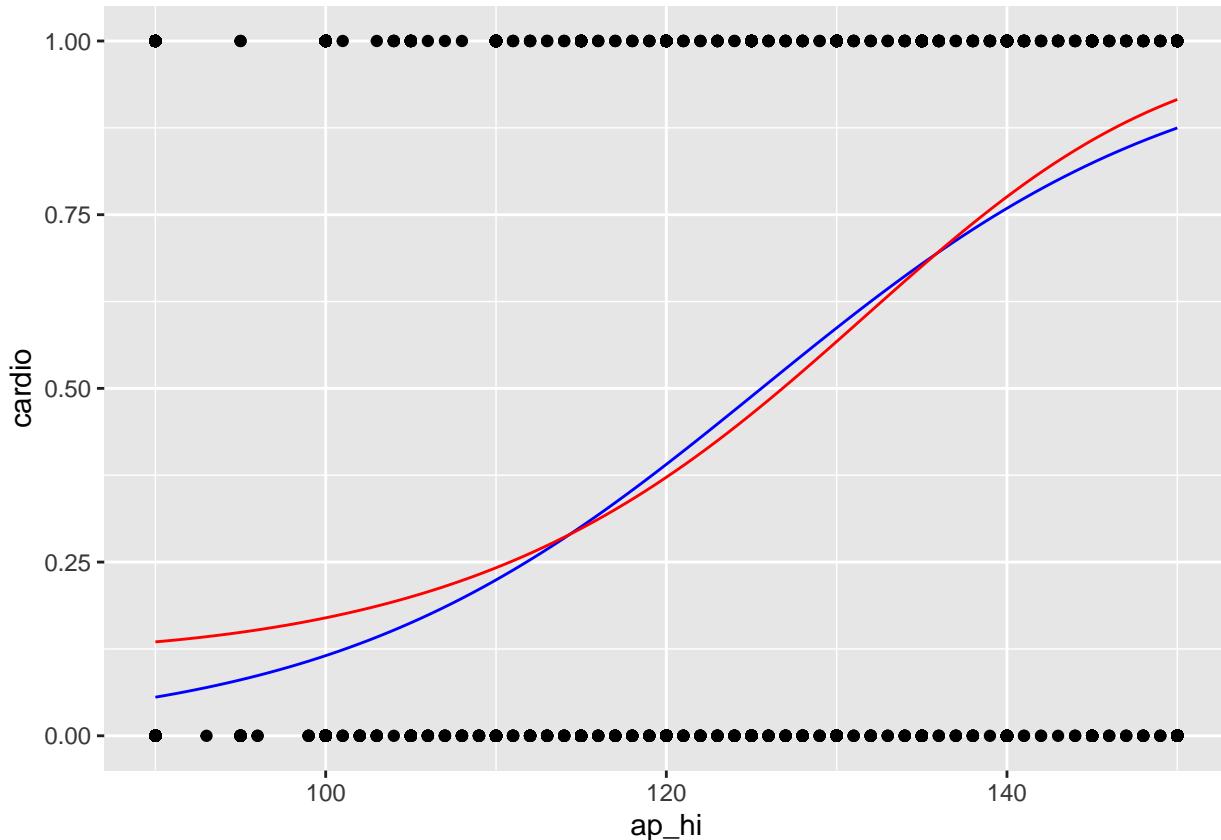
Mimo uwzględnienia dodatkowej zmiennej, test wskazuje na dalsze niedopasowanie modelu.

```
anova(model_zero, model_1, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: cardio ~ ap_hi
## Model 2: cardio ~ ap_hi + I(ap_hi^2)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      63586     75647
## 2      63585     75340  1      307.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test Chi<sup>2</sup> wskazuje na istotność zmiennej ap\_hi<sup>2</sup>.

```
ggplot(data, aes(x = ap_hi, y = cardio)) +
  geom_point() +
  stat_function(fun = function(x) predict(object = model_zero, newdata = data.frame(ap_hi = x), type = "response"))
  stat_function(fun = function(x) predict(object = model_1, newdata = data.frame(ap_hi = x), type = "response"))
```



Podane wartości ciśnienia krwi pacjenta są wartościami średnimi z kilku pomiarów. Celem zatem będzie jak najlepsze oszacowanie ciśnienia skurczowego krwi. Z wcześniejszej analizy wiemy jaki wpływ na ciśnienie skurczowe krwi mają poszczególne czynniki.

Z analizy w części I wynika, że występuje istotne zróżnicowanie ciśnienia skurczowego krwi względem płci. Warto zatem uwzględnić w modelu tę zmienną.

```
model_2 = glm(cardio ~ ap_hi + I(ap_hi^2) + gender, family = 'binomial',
              data = data)
```

```
summary(model_2)
```

```
##
## Call:
## glm(formula = cardio ~ ap_hi + I(ap_hi^2) + gender, family = "binomial",
##       data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.2395  -0.9786  -0.6166   1.0491   2.0383
##
```

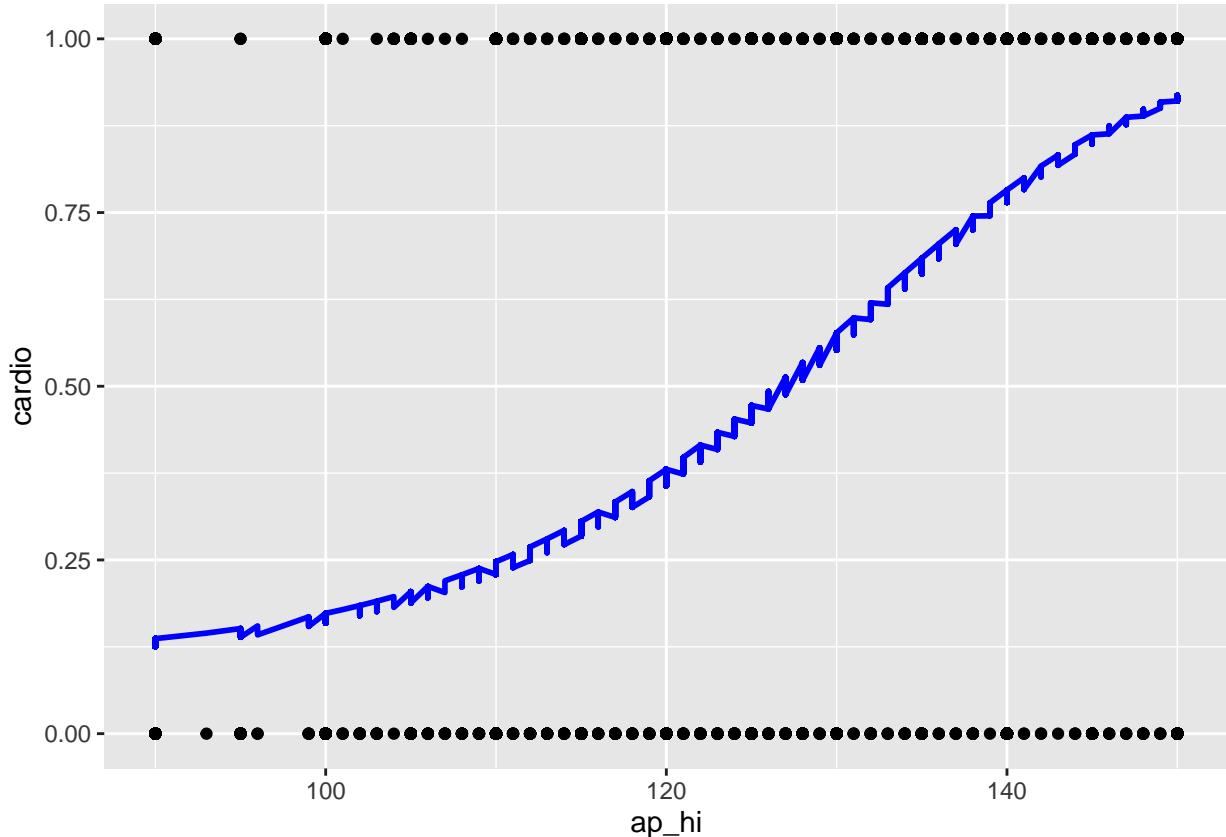
```

## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.427e+00 7.526e-01  4.554 5.27e-06 ***
## ap_hi       -1.363e-01 1.214e-02 -11.233 < 2e-16 ***
## I(ap_hi^2)  8.642e-04 4.877e-05 17.720 < 2e-16 ***
## gendermale -1.013e-01 1.853e-02 -5.469 4.53e-08 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895 on 63587 degrees of freedom
## Residual deviance: 75310 on 63584 degrees of freedom
## AIC: 75318
##
## Number of Fisher Scoring iterations: 4

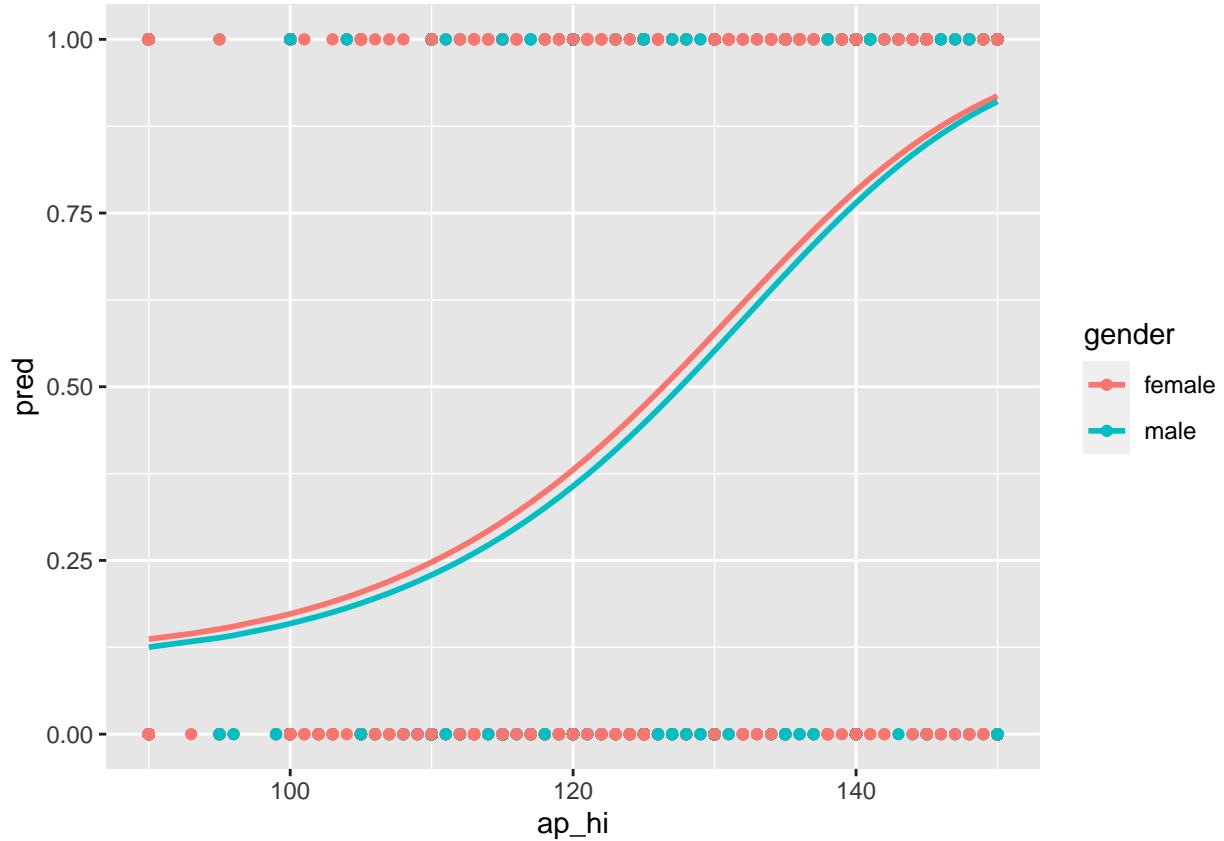
```

Wszystkie współczynniki są istotne statystycznie. Dewiancja resztowa zmalała.

```
preds_plot(model_2)
```

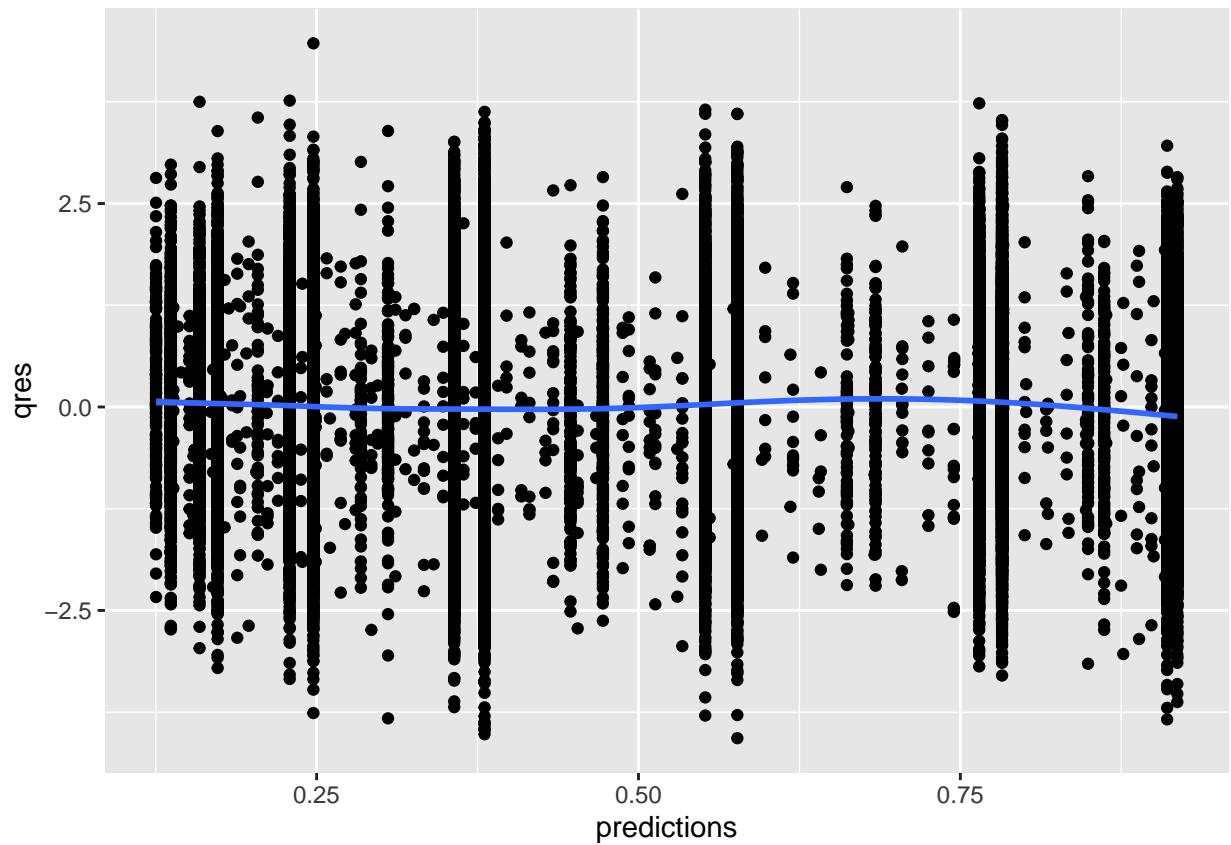


```
ggplot(data = data, aes(x = model_2$ap_hi, y = model_2$fitted, colour = gender)) + geom_point(aes
```



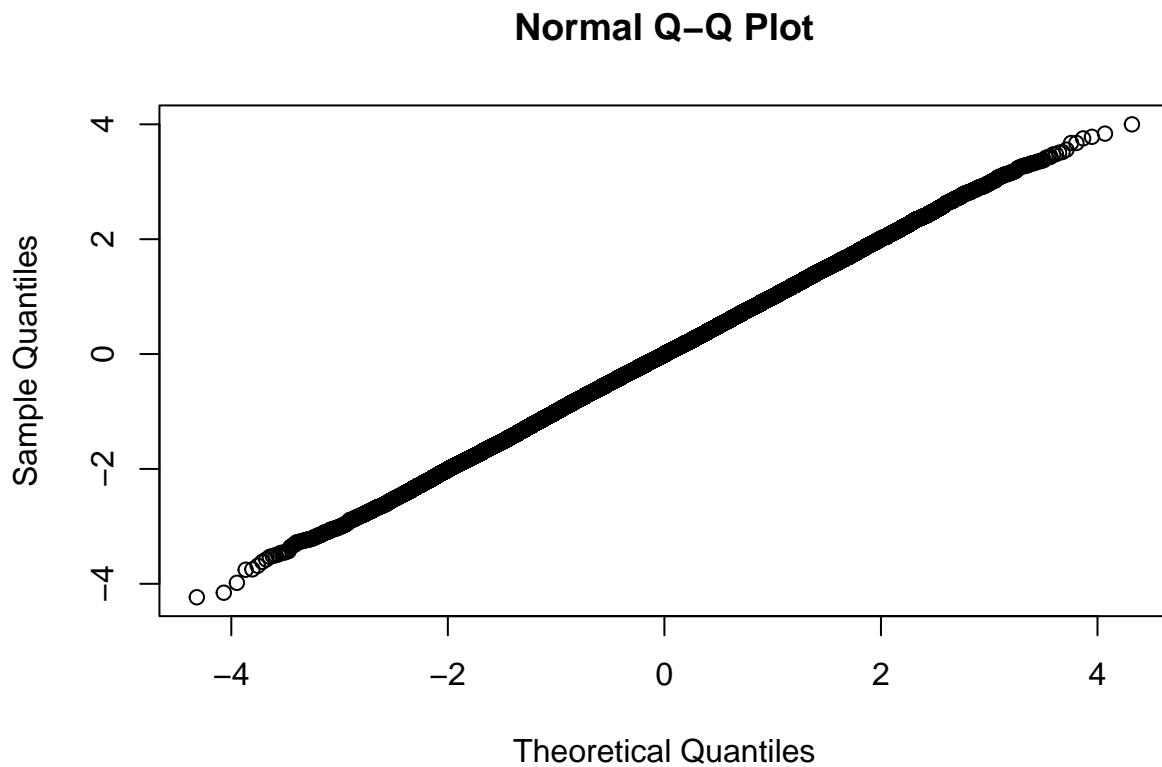
```
qres_plot(model_2)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



Wykres reszt kwantylowych delikatnie się wypłaszczył.

```
qqnorm(statmod::qresid(model_2))
```



Reszty kwantylowe dalej pochodzą z rozkładu normalnego.

```
anova(model_1, model_2, test = 'Chisq')

## Analysis of Deviance Table
##
## Model 1: cardio ~ ap_hi + I(ap_hi^2)
## Model 2: cardio ~ ap_hi + I(ap_hi^2) + gender
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      63585     75340
## 2      63584     75310  1    29.961 4.407e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test Chi<sup>2</sup> wskazuje na istotność włączenia zmiennej gender do modelu.

```
hoslem.test(x = model_2$model$cardio, y = fitted(model_2))
```

```
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: model_2$model$cardio, fitted(model_2)
## X-squared = 309.74, df = 8, p-value < 2.2e-16
```

Test Hosmera-Lemeshowa dalej wskazuje na niedopasowanie modelu.

```

model_3 = glm(cardio ~ ap_hi + I(ap_hi^2) + gender * cholesterol, family = 'binomial', data = data)

summary(model_3)

## 
## Call:
## glm(formula = cardio ~ ap_hi + I(ap_hi^2) + gender * cholesterol,
##      family = "binomial", data = data)
## 
## Deviance Residuals:
##       Min      1Q   Median      3Q      Max
## -2.5764 -0.9180 -0.5886  0.9796  2.0617
## 
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                3.094e+00  7.495e-01  4.127 3.67e-05 ***
## ap_hi                   -1.297e-01  1.208e-02 -10.736 < 2e-16 ***
## I(ap_hi^2)                 8.211e-04  4.853e-05 16.921 < 2e-16 ***
## gendermale                -6.987e-02  2.106e-02 -3.317 0.000908 ***
## cholesterol_norm            3.753e-01  3.283e-02 11.431 < 2e-16 ***
## cholesterolaw_norm          1.168e+00  3.806e-02 30.696 < 2e-16 ***
## gendermale:cholesterol_norm 5.570e-02  5.770e-02  0.965 0.334404
## gendermale:cholesterolaw_norm -7.273e-02  6.722e-02 -1.082 0.279287
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
## Null deviance: 87895  on 63587  degrees of freedom
## Residual deviance: 73774  on 63580  degrees of freedom
## AIC: 73790
## 
## Number of Fisher Scoring iterations: 4

```

Nie wszystkie współczynniki są istotne statystycznie. We wcześniejszej analizie wyszło, że efekt poziomu cholesterolu jest istotny statystycznie w każdej z norm. Mogło się to zmienić ze względu na dodatkową zmienną  $ap\_hi^2$ . Poziom cholesterolu jest powiązany z ciśnieniem skurczowym krwi, dlatego dodanie  $ap\_hi^2$  mogło zmniejszyć istotność interakcji.

```

model_4 = glm(cardio ~ ap_hi + I(ap_hi^2) + gender + cholesterol, family = 'binomial', data = data)

summary(model_4)

## 
## Call:
## glm(formula = cardio ~ ap_hi + I(ap_hi^2) + gender + cholesterol,
##      family = "binomial", data = data)
## 
## Deviance Residuals:
##       Min      1Q   Median      3Q      Max
## -2.5680 -0.9179 -0.5883  0.9741  2.0623
## 
```

```

## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)            3.079e+00  7.495e-01   4.108 3.99e-05 ***
## ap_hi                 -1.295e-01  1.208e-02  -10.719 < 2e-16 ***
## I(ap_hi^2)             8.204e-04  4.853e-05   16.906 < 2e-16 ***
## gendermale            -6.939e-02  1.878e-02  -3.695 0.00022 ***
## cholesterola_norm     3.930e-01  2.705e-02   14.529 < 2e-16 ***
## cholesterolaw_norm    1.146e+00  3.139e-02   36.488 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895  on 63587  degrees of freedom
## Residual deviance: 73776  on 63582  degrees of freedom
## AIC: 73788
##
## Number of Fisher Scoring iterations: 4

anova(model_4, model_3, test = 'Chisq')

```

```

## Analysis of Deviance Table
##
## Model 1: cardio ~ ap_hi + I(ap_hi^2) + gender + cholesterol
## Model 2: cardio ~ ap_hi + I(ap_hi^2) + gender * cholesterol
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      63582     73776
## 2      63580     73774  2    2.3666   0.3063

```

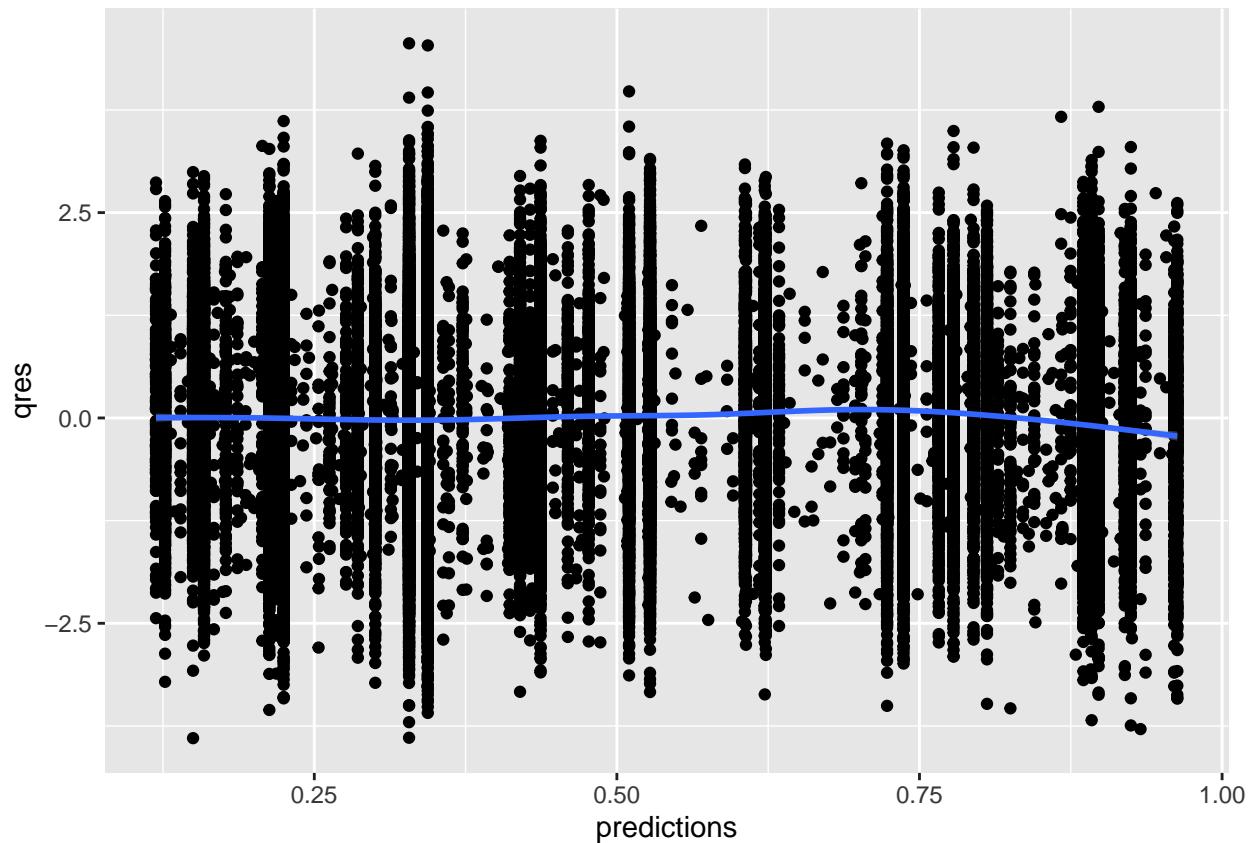
Interakcja zmiennej cholesterol i gender nie jest istotna statystycznie.

```

qres_plot(model_4)

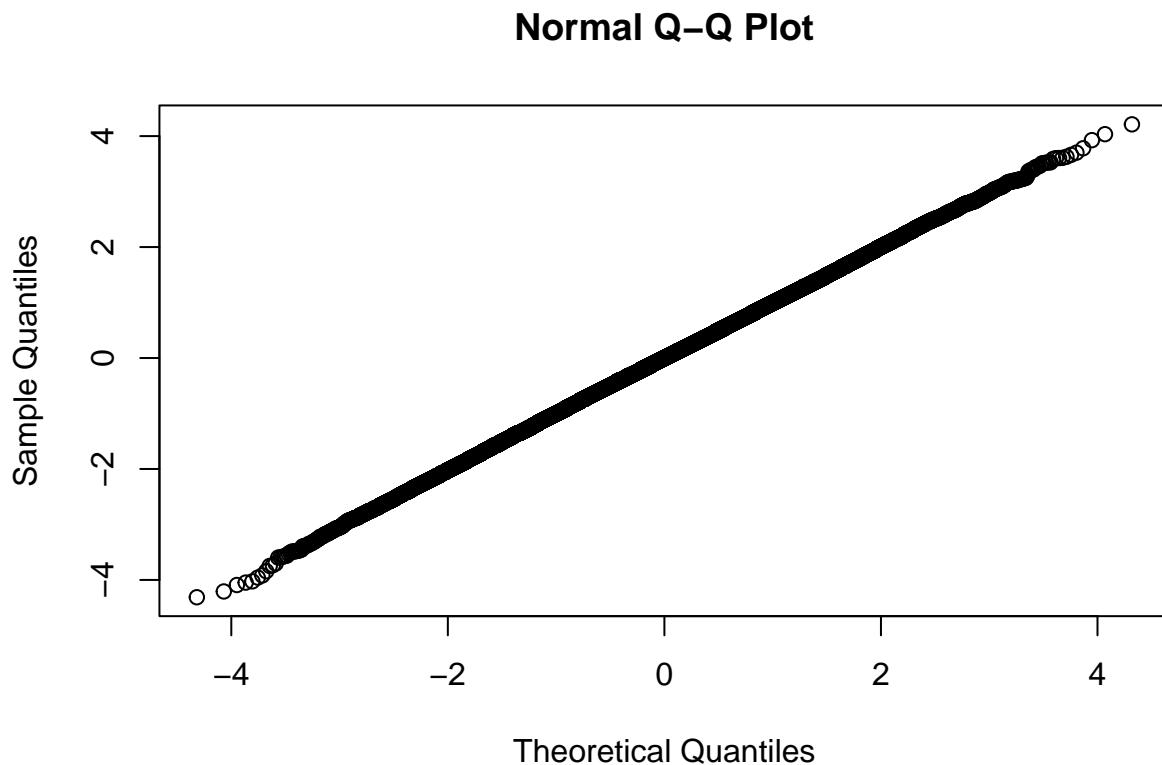
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

```



Wykres reszt kwantylowych troszkę się pogorszył.

```
qqnorm(statmod::qresid(model_4))
```



Reszty kwantylowe pochodzą z rozkładu normalnego.

```
hoslem.test(x = model_4$y, y = model_4$fitted)

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: model_4$y, model_4$fitted
## X-squared = 412.63, df = 8, p-value < 2.2e-16
```

Test Hosmera-Lemeshowa dalej wskazuje na niedopasowanie modelu.

```
anova(model_2, model_4, test = 'Chisq')

## Analysis of Deviance Table
##
## Model 1: cardio ~ ap_hi + I(ap_hi^2) + gender
## Model 2: cardio ~ ap_hi + I(ap_hi^2) + gender + cholesterol
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      63584     75310
## 2      63582     73776  2    1533.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Uwzględnienie zmiennej cholesterol w modelu jest istotne.

```

model_5 = glm(cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol, family = 'binomial', data = data)

summary(model_5)

## 
## Call:
## glm(formula = cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol,
##      family = "binomial", data = data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.6491 -0.9197 -0.5887  0.9756  2.2270
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.999e+00 7.499e-01 3.999 6.37e-05 ***
## ap_hi                   -1.281e-01 1.209e-02 -10.601 < 2e-16 ***
## I(ap_hi^2)                 8.150e-04 4.855e-05 16.786 < 2e-16 ***
## gendermale                -6.793e-02 2.016e-02 -3.369 0.000754 ***
## gluca_norm                 4.908e-02 4.526e-02  1.084 0.278173
## glucwa_norm                -2.558e-01 4.770e-02 -5.362 8.25e-08 ***
## cholesterol_norm            3.806e-01 2.814e-02 13.523 < 2e-16 ***
## cholesterol_low_norm        1.278e+00 3.683e-02 34.710 < 2e-16 ***
## gendermale:gluca_norm     1.029e-01 7.572e-02  1.358 0.174351
## gendermale:glucwa_norm   -1.365e-01 7.489e-02 -1.823 0.068282 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895  on 63587  degrees of freedom
## Residual deviance: 73708  on 63578  degrees of freedom
## AIC: 73728
##
## Number of Fisher Scoring iterations: 4

```

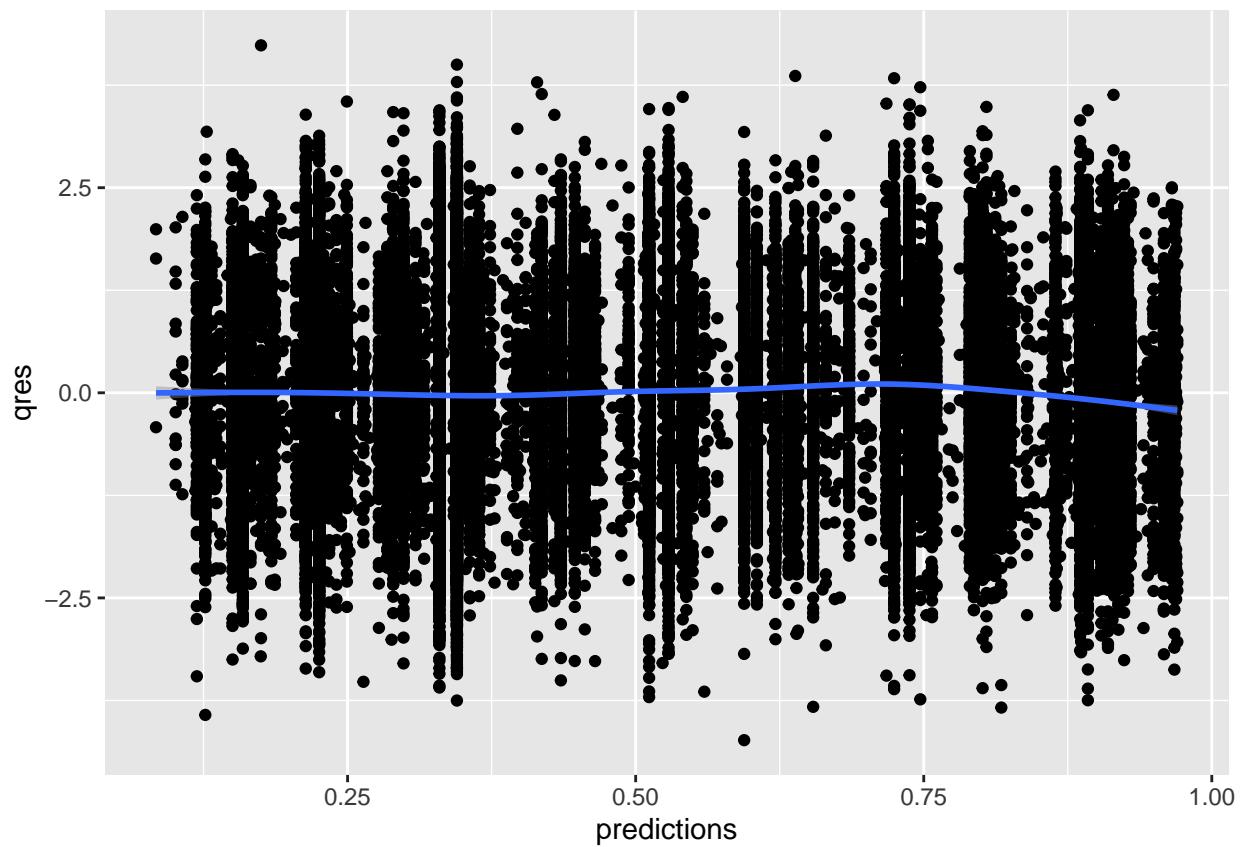
Przy uwzględnieniu innych zmiennych, interakcja zmiennej gluc z gender zmieniła się. Zachodzi ona tylko w przypadku poziomu glukozy we krwi dużo powyżej normy, mimo wszystko nie można uprościć modelu pomijając interakcje tak jak w przypadku zmiennej cholesterol.

```

qres_plot(model_5)

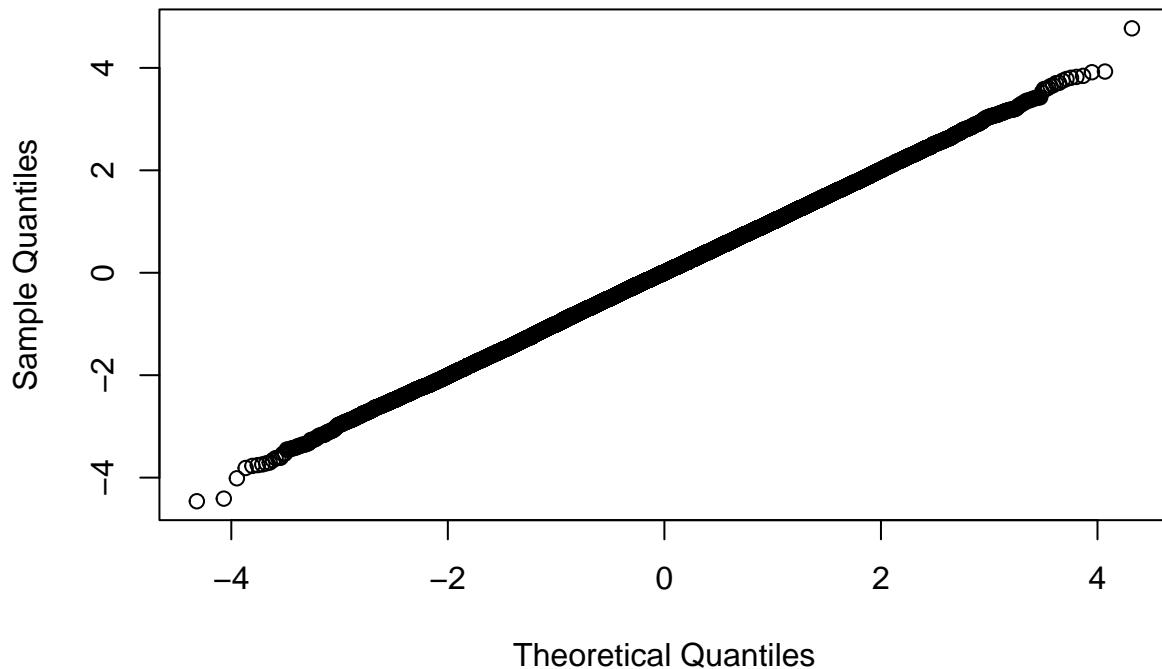
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

```



```
qqnorm(statmod::qresid(model_5))
```

## Normal Q-Q Plot



```
hoslem.test(x = model_5$y, y = model_5$fitted)
```

```
##  
## Hosmer and Lemeshow goodness of fit (GOF) test  
##  
## data: model_5$y, model_5$fitted  
## X-squared = 428.78, df = 8, p-value < 2.2e-16
```

```
anova(model_4, model_5, test = 'Chisq')
```

```
## Analysis of Deviance Table  
##  
## Model 1: cardio ~ ap_hi + I(ap_hi^2) + gender + cholesterol  
## Model 2: cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol  
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1      63582     73776  
## 2      63578     73708  4    68.225 5.379e-14 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_6 = glm(cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol + active, family = 'binomial',  
summary(model_6)
```

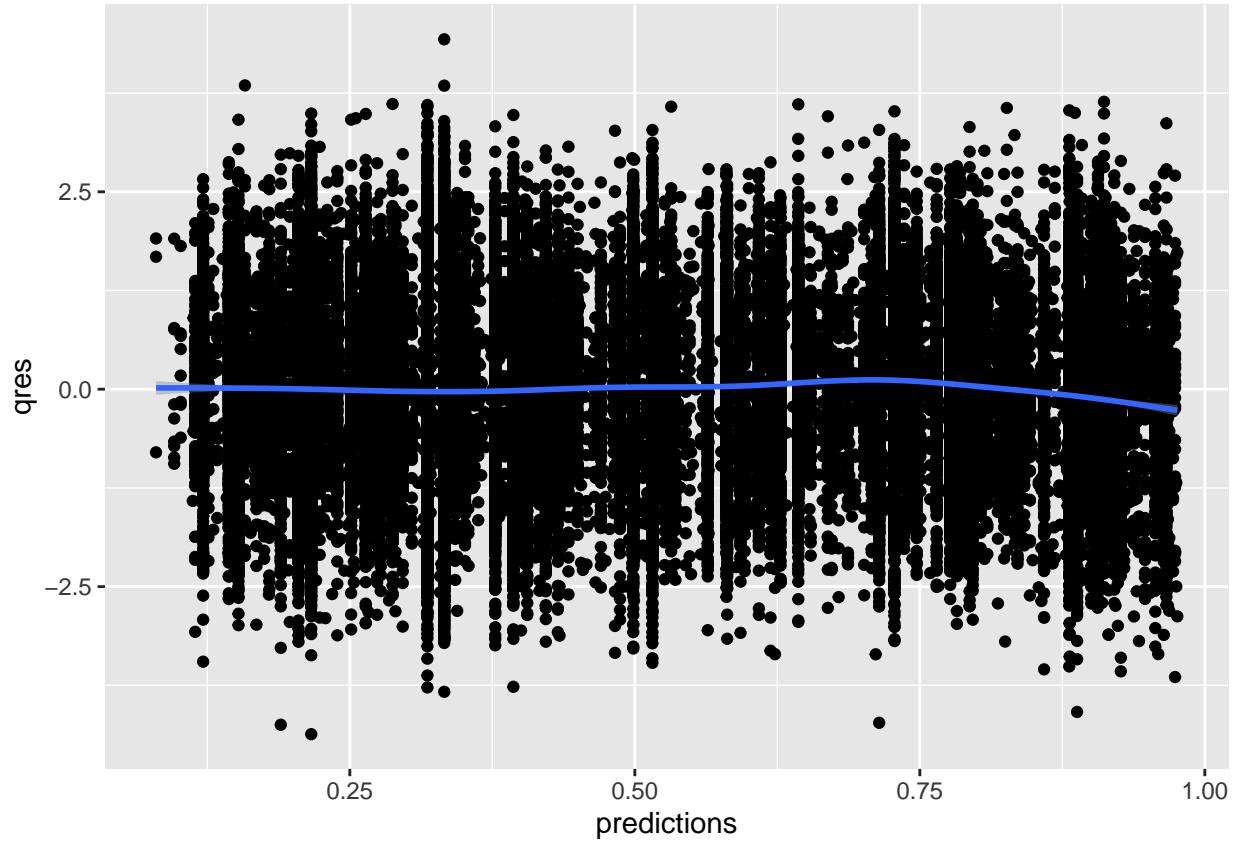
```

## 
## Call:
## glm(formula = cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol +
##      active, family = "binomial", data = data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.7290 -0.8997 -0.5750  0.9648  2.2475
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)            3.316e+00 7.503e-01  4.420 9.86e-06 ***
## ap_hi                  -1.300e-01 1.209e-02 -10.754 < 2e-16 ***
## I(ap_hi^2)              8.228e-04 4.855e-05 16.946 < 2e-16 ***
## gendermale             -6.713e-02 2.019e-02 -3.326 0.000882 ***
## gluca_norm              4.307e-02 4.531e-02  0.951 0.341841
## glucwa_norm             -2.632e-01 4.777e-02 -5.511 3.58e-08 ***
## cholesterol_norm        3.807e-01 2.818e-02 13.508 < 2e-16 ***
## cholesterolaw_norm      1.286e+00 3.685e-02 34.892 < 2e-16 ***
## activeactive            -2.627e-01 2.225e-02 -11.810 < 2e-16 ***
## gendermale:gluca_norm   1.050e-01 7.581e-02  1.385 0.165922
## gendermale:glucwa_norm -1.321e-01 7.494e-02 -1.762 0.078007 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895  on 63587  degrees of freedom
## Residual deviance: 73569  on 63577  degrees of freedom
## AIC: 73591
##
## Number of Fisher Scoring iterations: 4

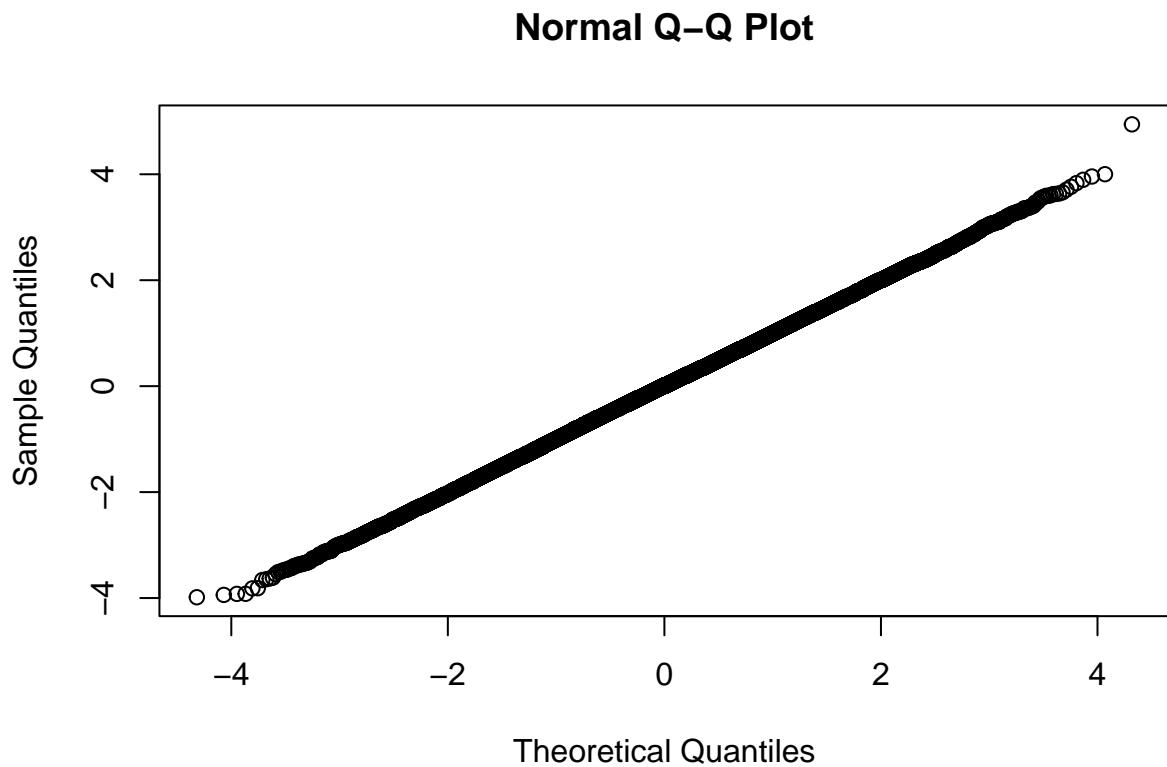
qres_plot(model_6)

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

```



```
qqnorm(statmod::qresid(model_6))
```



```

hoslem.test(x = model_6$y, y = model_6$fitted)

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: model_6$y, model_6$fitted
## X-squared = 417.48, df = 8, p-value < 2.2e-16

model_7 = glm(cardio ~ ap_hi + I(ap_hi^2) + gender * gluc +
              cholesterol + active + smoke, family = 'binomial', data = data)

summary(model_7)

##
## Call:
## glm(formula = cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol +
##       active + smoke, family = "binomial", data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.7582   -0.9009   -0.5766    0.9748    2.2377
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)

```

```

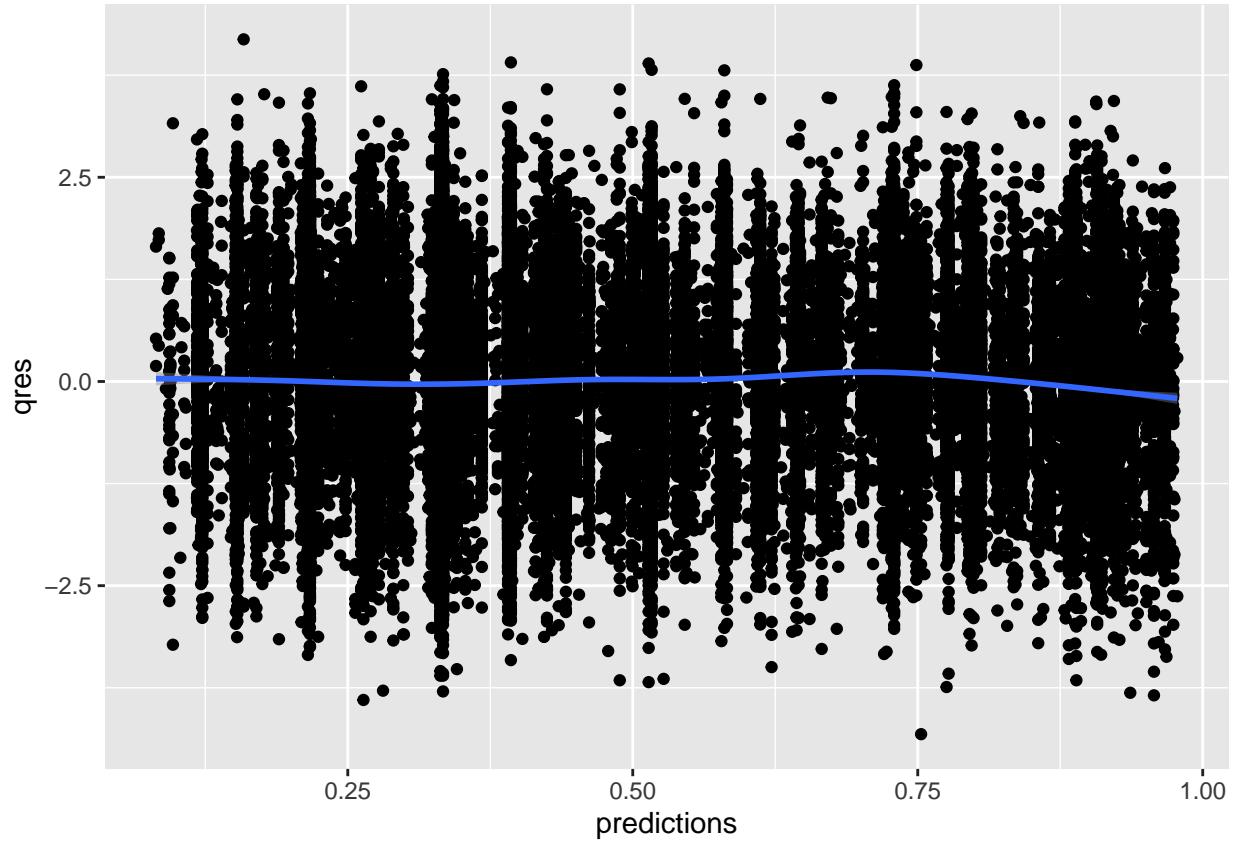
## (Intercept) 3.434e+00 7.511e-01 4.572 4.82e-06 ***
## ap_hi      -1.320e-01 1.210e-02 -10.909 < 2e-16 ***
## I(ap_hi^2) 8.314e-04 4.861e-05 17.102 < 2e-16 ***
## gendermale -1.049e-02 2.123e-02 -0.494 0.6212
## gluca_norm 4.208e-02 4.534e-02 0.928 0.3534
## glucwa_norm -2.657e-01 4.780e-02 -5.558 2.73e-08 ***
## cholesterol_norm 3.888e-01 2.822e-02 13.775 < 2e-16 ***
## cholesterolaw_norm 1.292e+00 3.687e-02 35.027 < 2e-16 ***
## activeactive -2.578e-01 2.226e-02 -11.580 < 2e-16 ***
## smokesmoker -2.917e-01 3.426e-02 -8.512 < 2e-16 ***
## gendermale:gluca_norm 1.181e-01 7.588e-02 1.557 0.1195
## gendermale:glucwa_norm -1.403e-01 7.504e-02 -1.870 0.0614 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895 on 63587 degrees of freedom
## Residual deviance: 73496 on 63576 degrees of freedom
## AIC: 73520
##
## Number of Fisher Scoring iterations: 4

```

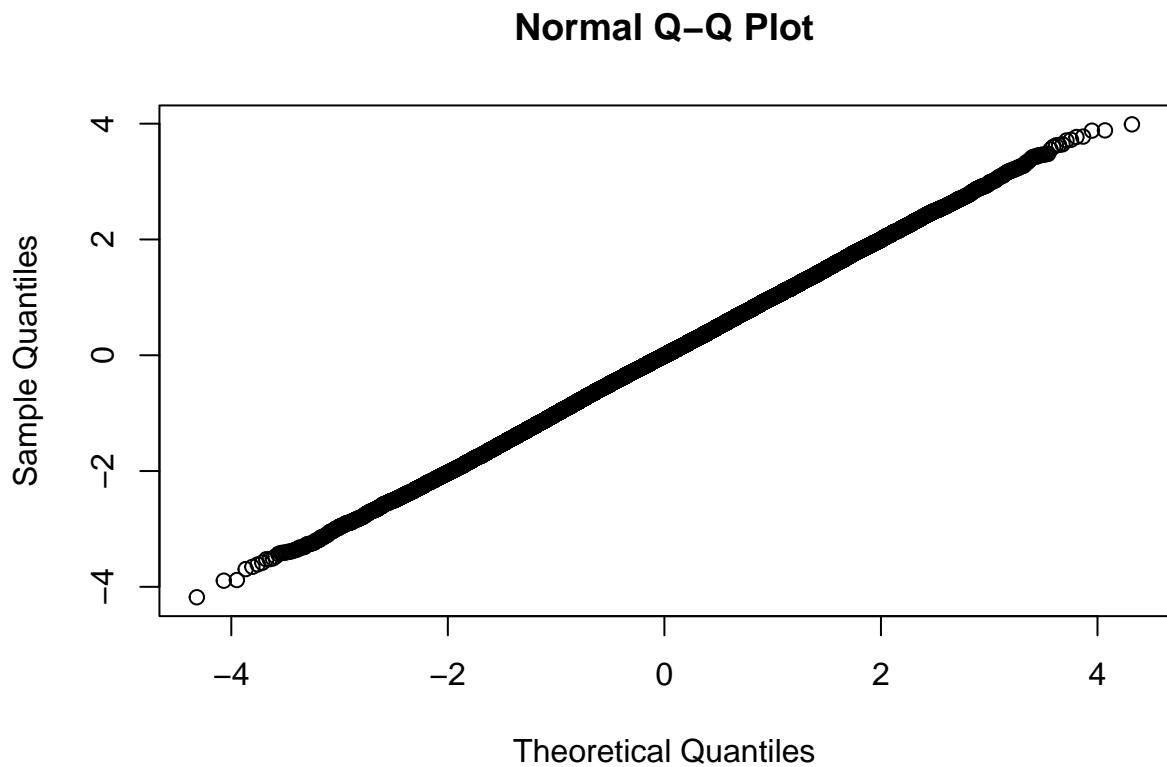
Po uwzględnieniu jeszcze zmiennej smoke, zmienna gendermale straciła istotność statystyczną. Nie można uprościć modelu usuwając zmienną gender, ponieważ zachodzi istotna statystycznie interakcja między gender, a gluc.

```
qres_plot(model_7)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
qqnorm(statmod::qresid(model_7))
```



```

hoslem.test(x = model_7$y, y = model_7$fitted)

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: model_7$y, model_7$fitted
## X-squared = 435.17, df = 8, p-value < 2.2e-16

model_8 = glm(cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol +
               active + smoke + alco, family = 'binomial', data = data)

summary(model_8)

##
## Call:
## glm(formula = cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol +
##       active + smoke + alco, family = "binomial", data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.7651   -0.9022   -0.5774    0.9708    2.2213
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## 
```

```

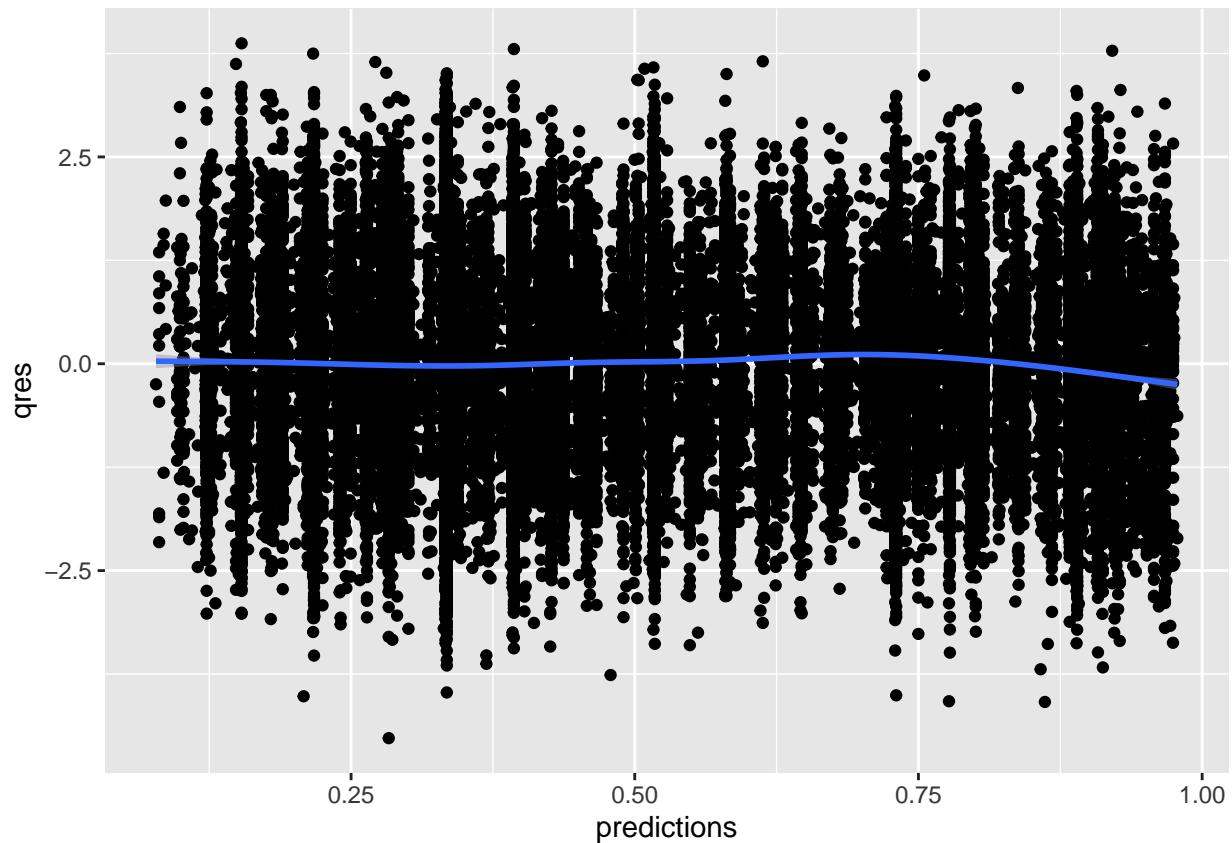
## (Intercept) 3.453e+00 7.513e-01 4.597 4.29e-06 ***
## ap_hi -1.324e-01 1.210e-02 -10.935 < 2e-16 ***
## I(ap_hi^2) 8.332e-04 4.863e-05 17.133 < 2e-16 ***
## gendermale -4.752e-03 2.126e-02 -0.223 0.8232
## gluca_norm 4.350e-02 4.536e-02 0.959 0.3375
## glucwa_norm -2.676e-01 4.782e-02 -5.597 2.18e-08 ***
## cholesterol_norm 3.932e-01 2.825e-02 13.919 < 2e-16 ***
## cholesterol_low_norm 1.295e+00 3.688e-02 35.114 < 2e-16 ***
## activeactive -2.558e-01 2.227e-02 -11.488 < 2e-16 ***
## smokesmoker -2.354e-01 3.596e-02 -6.545 5.94e-11 ***
## alcoalc -2.244e-01 4.394e-02 -5.107 3.28e-07 ***
## gendermale:gluca_norm 1.218e-01 7.590e-02 1.605 0.1085
## gendermale:glucwa_norm -1.394e-01 7.506e-02 -1.858 0.0632 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895 on 63587 degrees of freedom
## Residual deviance: 73470 on 63575 degrees of freedom
## AIC: 73496
##
## Number of Fisher Scoring iterations: 4

```

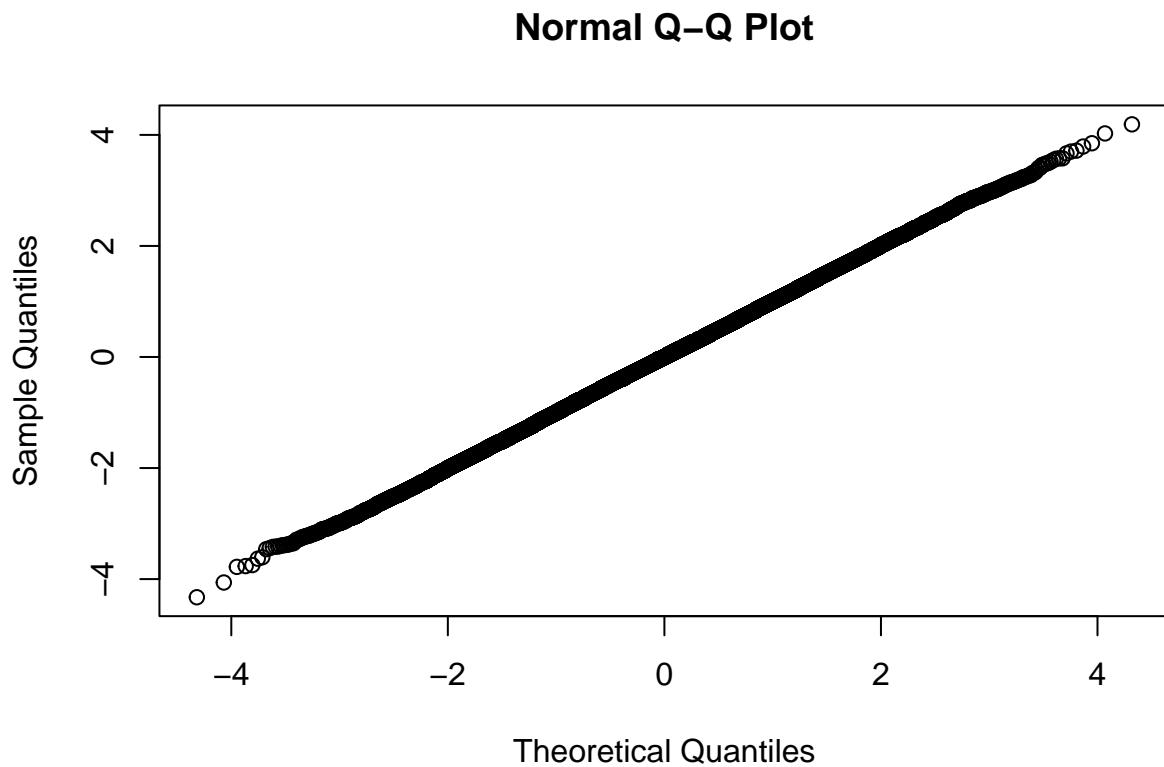
Po uwzględnieniu zmiennych active, smoke oraz alco, interakcja zmiennej gluc i gender jest większa. Gendermale:gluca\_norm jest na pograniczu istotności statystycznej.

```
qres_plot(model_8)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
qqnorm(statmod::qresid(model_8))
```



```

hoslem.test(x = model_8$y, y = model_8$fitted)

##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: model_8$y, model_8$fitted
## X-squared = 449.21, df = 8, p-value < 2.2e-16

model_9 = glm(cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol +
               active + smoke + alco + BMI, family = 'binomial', data = data)

summary(model_9)

##
## Call:
## glm(formula = cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol +
##       active + smoke + alco + BMI, family = "binomial", data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.8585   -0.8969   -0.5948    0.9625    2.2165
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)

```

```

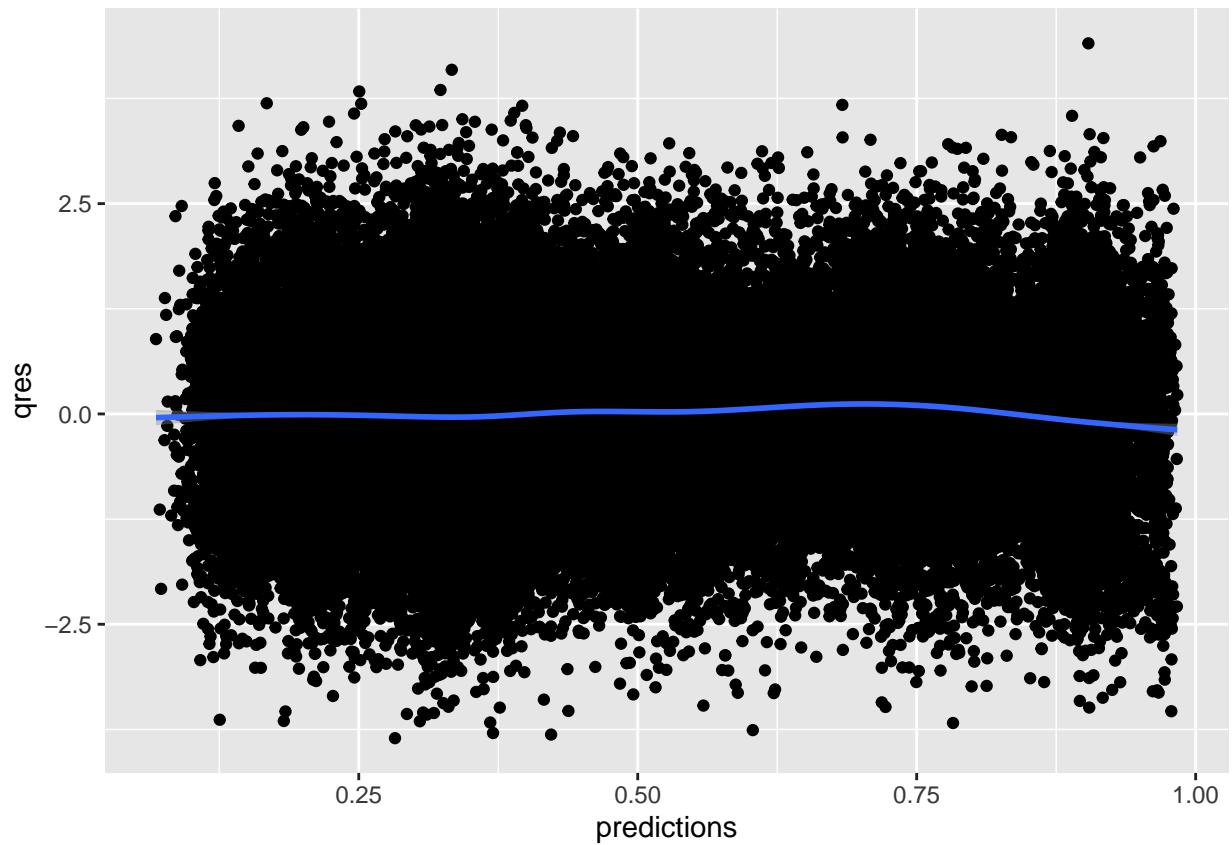
## (Intercept) 2.9143214 0.7500873 3.885 0.000102 ***
## ap_hi -0.1348055 0.0120752 -11.164 < 2e-16 ***
## I(ap_hi^2) 0.0008324 0.0000485 17.162 < 2e-16 ***
## gendermale 0.0343477 0.0214330 1.603 0.109032
## gluca_norm 0.0010958 0.0455758 0.024 0.980818
## glucwa_norm -0.2739774 0.0479298 -5.716 1.09e-08 ***
## cholesterola_norm 0.3684809 0.0283286 13.007 < 2e-16 ***
## cholesterolaw_norm 1.2534376 0.0370724 33.811 < 2e-16 ***
## activeactive -0.2539239 0.0223239 -11.375 < 2e-16 ***
## smokesmoker -0.2328470 0.0359959 -6.469 9.88e-11 ***
## alcoalc -0.2452043 0.0440282 -5.569 2.56e-08 ***
## BMI 0.0312038 0.0018898 16.512 < 2e-16 ***
## gendermale:gluca_norm 0.1208951 0.0760122 1.590 0.111729
## gendermale:glucwa_norm -0.1428698 0.0750978 -1.902 0.057113 .
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895 on 63587 degrees of freedom
## Residual deviance: 73195 on 63574 degrees of freedom
## AIC: 73223
##
## Number of Fisher Scoring iterations: 4

```

Po uwzględnieniu zmiennej BMI, gendermale jest na pograniczu istotności statystycznej.

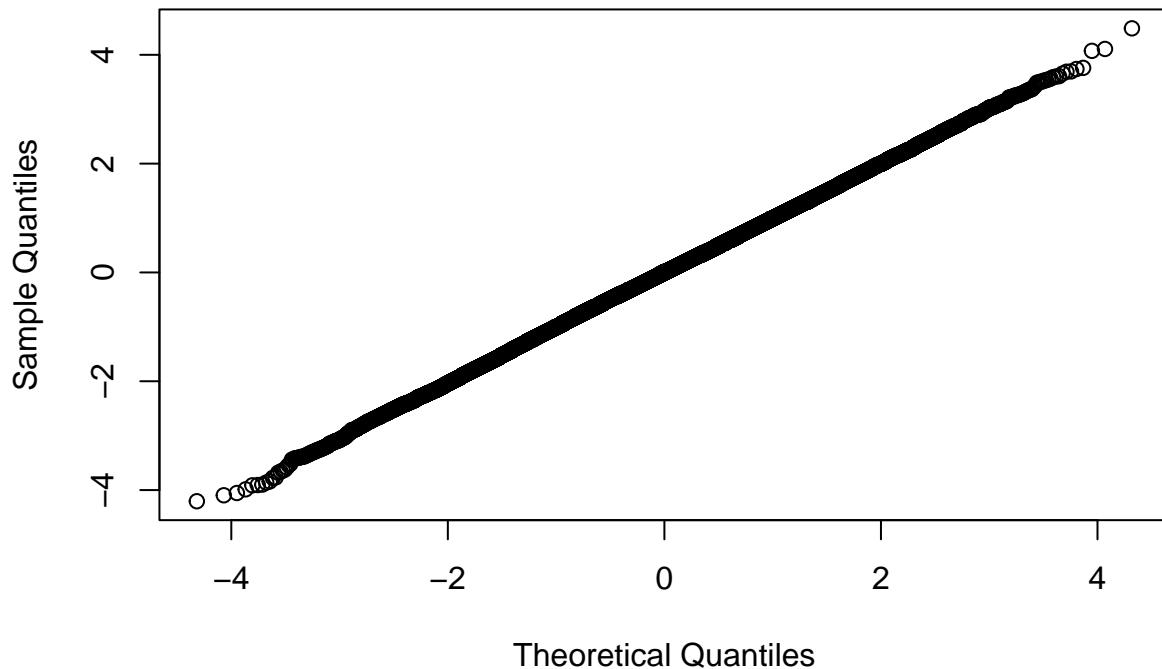
```
qres_plot(model_9)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
qqnorm(statmod::qresid(model_9))
```

## Normal Q-Q Plot



```
hoslem.test(x = model_9$y, y = model_9$fitted)
```

```
##  
## Hosmer and Lemeshow goodness of fit (GOF) test  
##  
## data: model_9$y, model_9$fitted  
## X-squared = 401.58, df = 8, p-value < 2.2e-16
```

```
model_10 = glm(cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol +  
active + smoke + alco + BMI + I(BMI^2), family = 'binomial',  
data = data)
```

```
summary(model_10)
```

```
##  
## Call:  
## glm(formula = cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol +  
##       active + smoke + alco + BMI + I(BMI^2), family = "binomial",  
##       data = data)  
##  
## Deviance Residuals:  
##      Min        1Q    Median        3Q       Max  
## -2.7806  -0.9033  -0.5928   0.9600   2.2446  
##
```

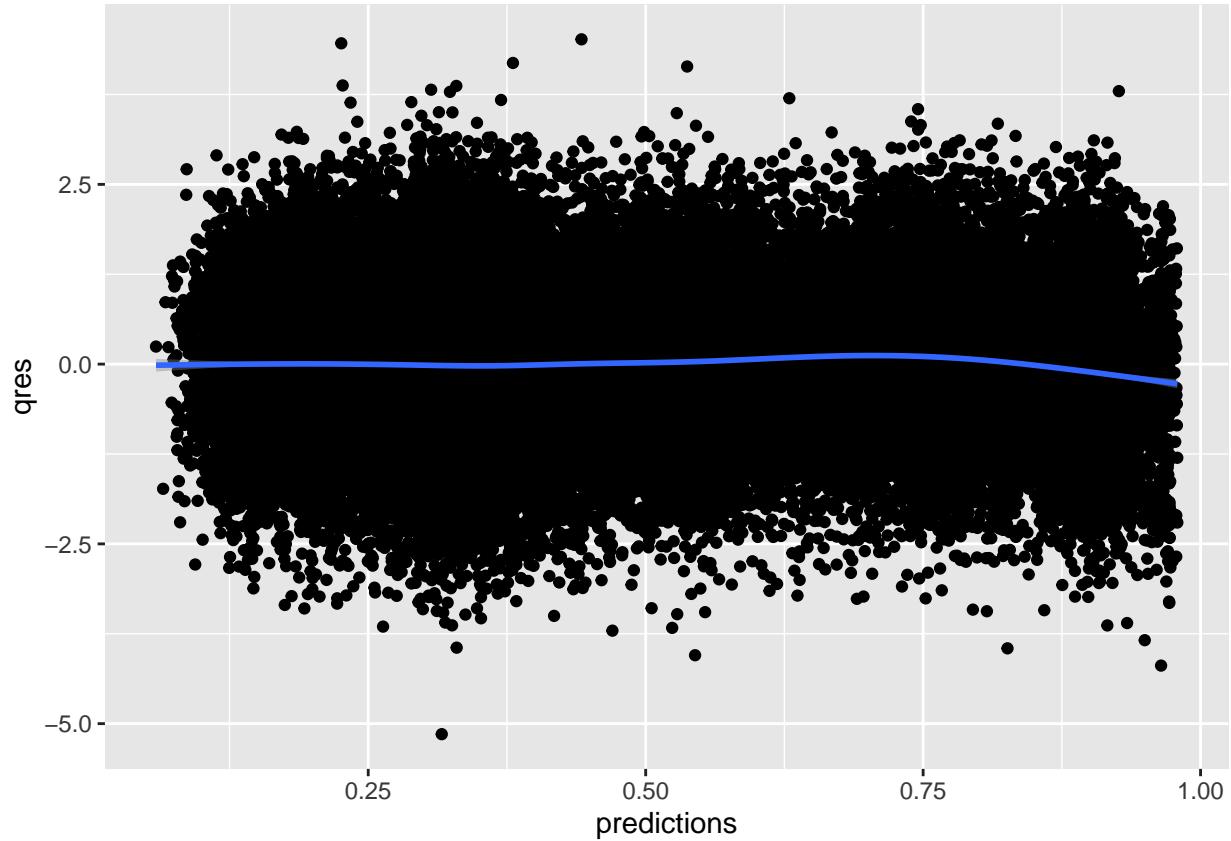
```

## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)           1.751e+00 7.667e-01  2.284  0.0224 *
## ap_hi            -1.374e-01 1.209e-02 -11.363 < 2e-16 ***
## I(ap_hi^2)          8.415e-04 4.855e-05 17.332 < 2e-16 ***
## gendermale          3.109e-02 2.146e-02   1.449  0.1475
## gluca_norm          2.374e-03 4.552e-02   0.052  0.9584
## glucwa_norm         -2.711e-01 4.790e-02  -5.660 1.52e-08 ***
## cholesterola_norm   3.665e-01 2.834e-02 12.935 < 2e-16 ***
## cholesterolaw_norm  1.252e+00 3.707e-02 33.763 < 2e-16 ***
## activeactive        -2.541e-01 2.233e-02 -11.379 < 2e-16 ***
## smokesmoker         -2.309e-01 3.603e-02  -6.409 1.47e-10 ***
## alcoalc             -2.469e-01 4.404e-02  -5.605 2.08e-08 ***
## BMI                 1.233e-01 1.250e-02   9.864 < 2e-16 ***
## I(BMI^2)            -1.523e-03 2.041e-04  -7.465 8.32e-14 ***
## gendermale:gluca_norm 1.130e-01 7.603e-02   1.486  0.1373
## gendermale:glucwa_norm -1.497e-01 7.511e-02  -1.993  0.0463 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895 on 63587 degrees of freedom
## Residual deviance: 73139 on 63573 degrees of freedom
## AIC: 73169
##
## Number of Fisher Scoring iterations: 4

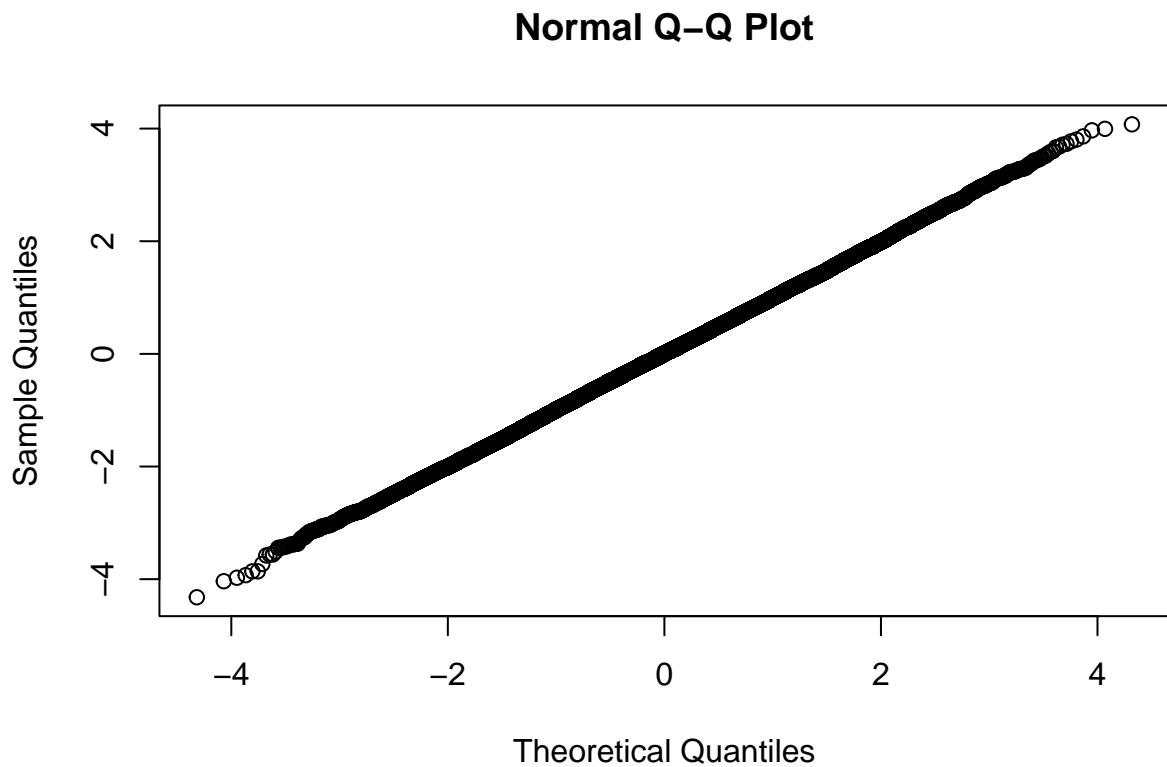
qres_plot(model_10)

## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

```



```
qqnorm(statmod::qresid(model_10))
```



```
hoslem.test(x = model_10$y, y = model_10$fitted)
```

```
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: model_10$y, model_10$fitted
## X-squared = 412.81, df = 8, p-value < 2.2e-16
```

## WNIOSKI

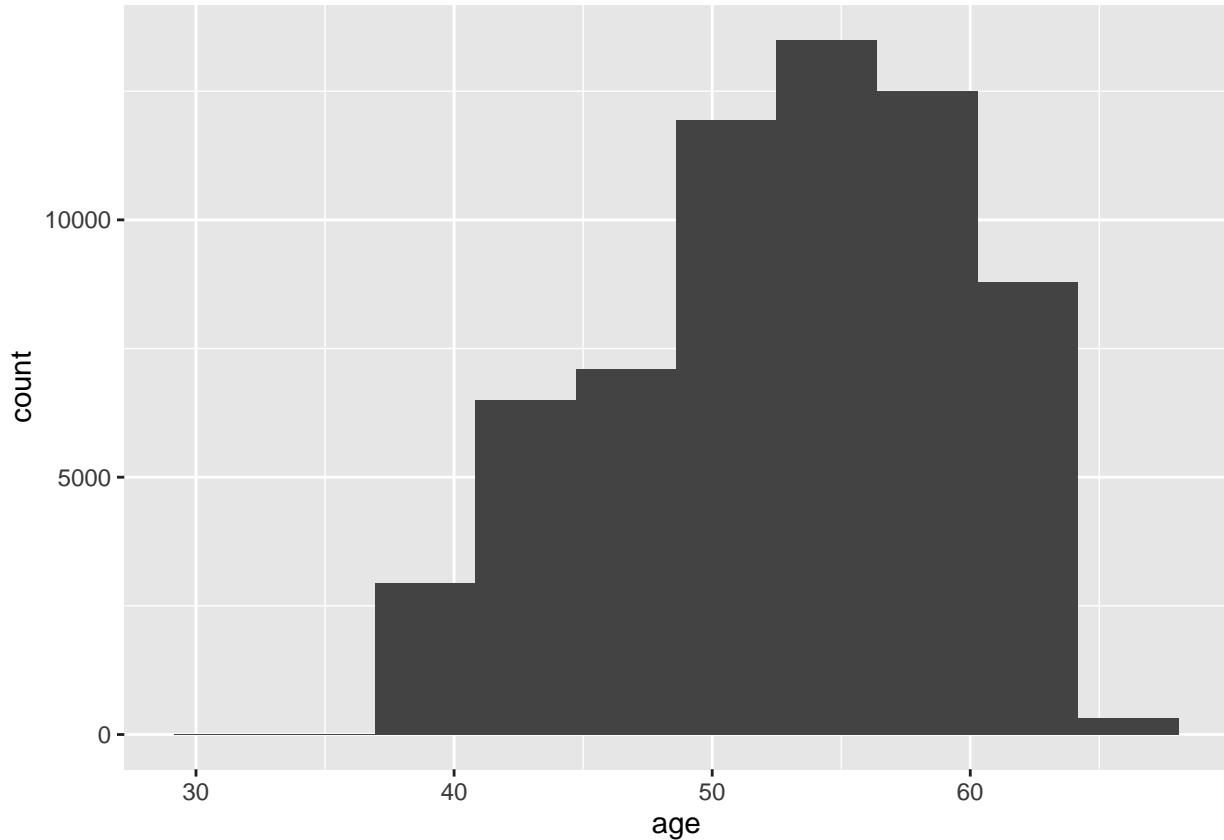
- Występuje problem z nadmierną dyspersją
- Interakcja zmiennej cholesterol i gender nie jest istotna statystycznie
- Uwzględnienie predyktorów x i  $x^2$  zmiennych ciągłych o rozkładzie normalnym daje lepsze efekty

## Część III

W analizie nie zostały uwzględnione jeszcze dwie zmienne: ap\_lo - ciśnienie rozkurczowe krwi, age - wiek. Logika podpowiada, że starsze osoby są bardziej narażone na choroby układu krążenia, zatem warto przeanalizować i uwzględnić zmienną age.

**HIPOTEZA:** Czy osoby starsze mają większą szansę posiadania choroby układu krążenia?

```
ggplot(data = data, aes(x = age)) + geom_histogram(bins = 10)
```



```
model.zero = glm(cardio ~ age, family = 'binomial', data = data)
summary(model.zero)
```

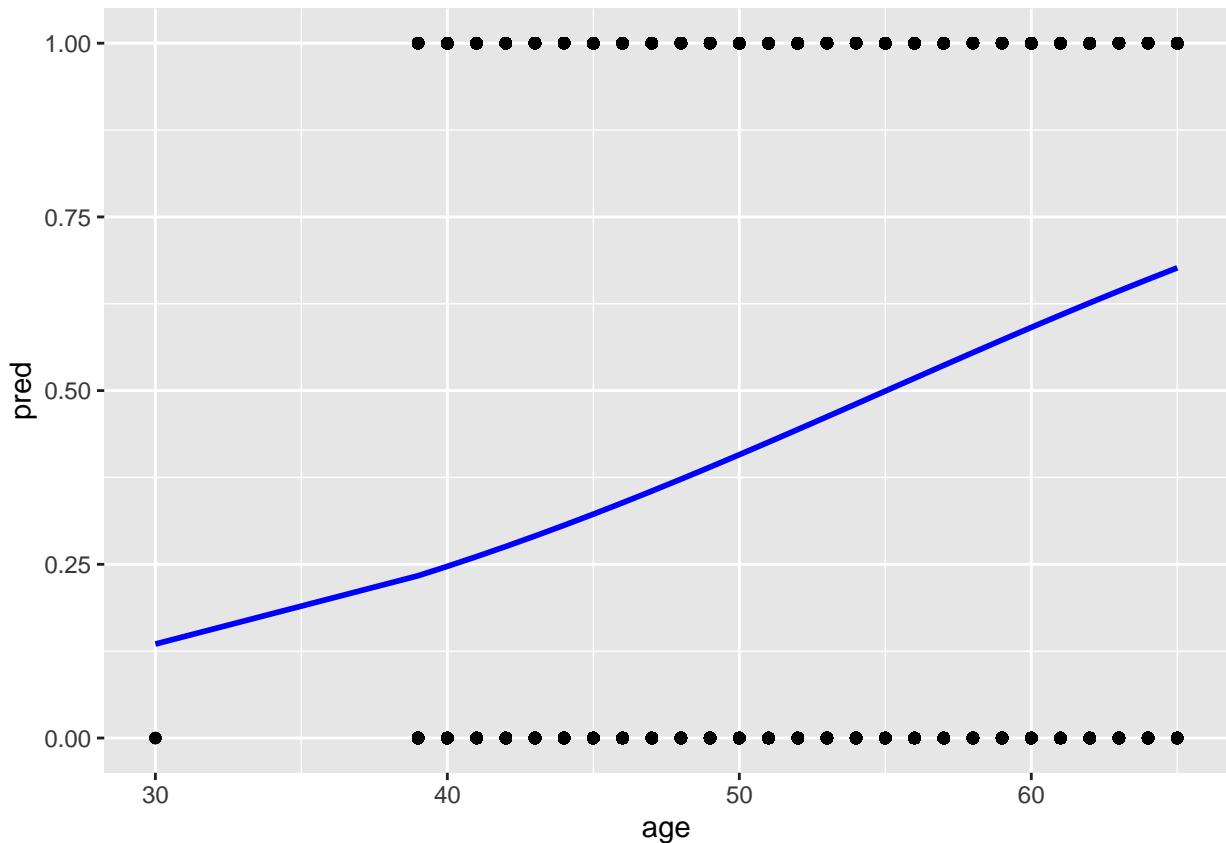
```
##
## Call:
## glm(formula = cardio ~ age, family = "binomial", data = data)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -1.5027   -1.0835   -0.7534    1.1472    1.7054
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.078385   0.067720  -60.22   <2e-16 ***
## age          0.074106   0.001257   58.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895  on 63587  degrees of freedom
## Residual deviance: 84174  on 63586  degrees of freedom
## AIC: 84178
```

```

## 
## Number of Fisher Scoring iterations: 4

ggplot(data = data, aes(x = model.zero$model$age, y = model.zero$fitted)) +
  geom_point(aes(y = cardio)) + geom_line(linewidth = 1, colour = 'blue') +
  labs(x = 'age', y = 'pred')

```



```

p = exp(-4.08 + 40 * 0.074)
1 - (1/(1+p))

```

```

## [1] 0.2460113

```

Prawdopodobieństwo, że osoba mająca 40 lat ma chorobę układu krążenia wynosi 0.25

```

q = exp(-4.08 + 60 * 0.074)
1 - (1/(1+q))

```

```

## [1] 0.5890404

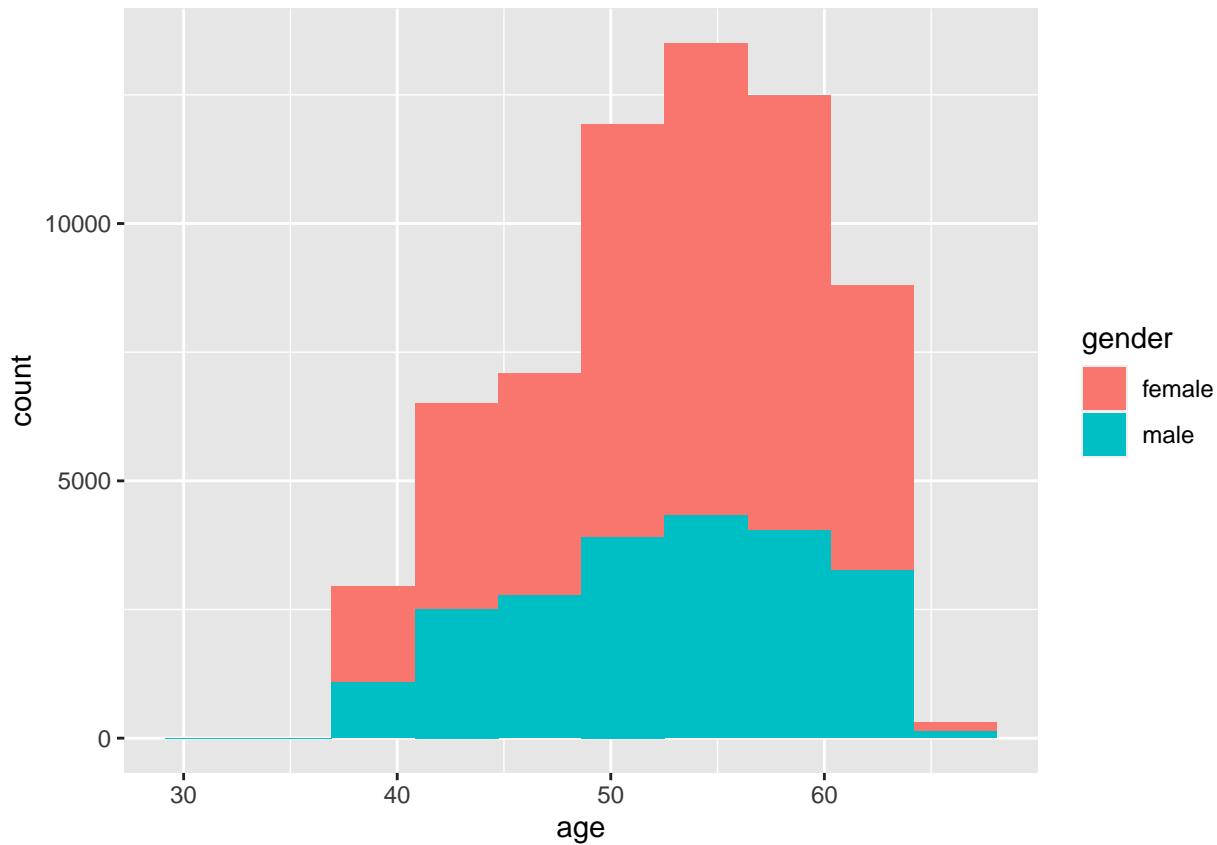
```

Prawdopodobieństwo, że osoba mająca 60 lat ma chorobę układu krążenia wynosi 0.59

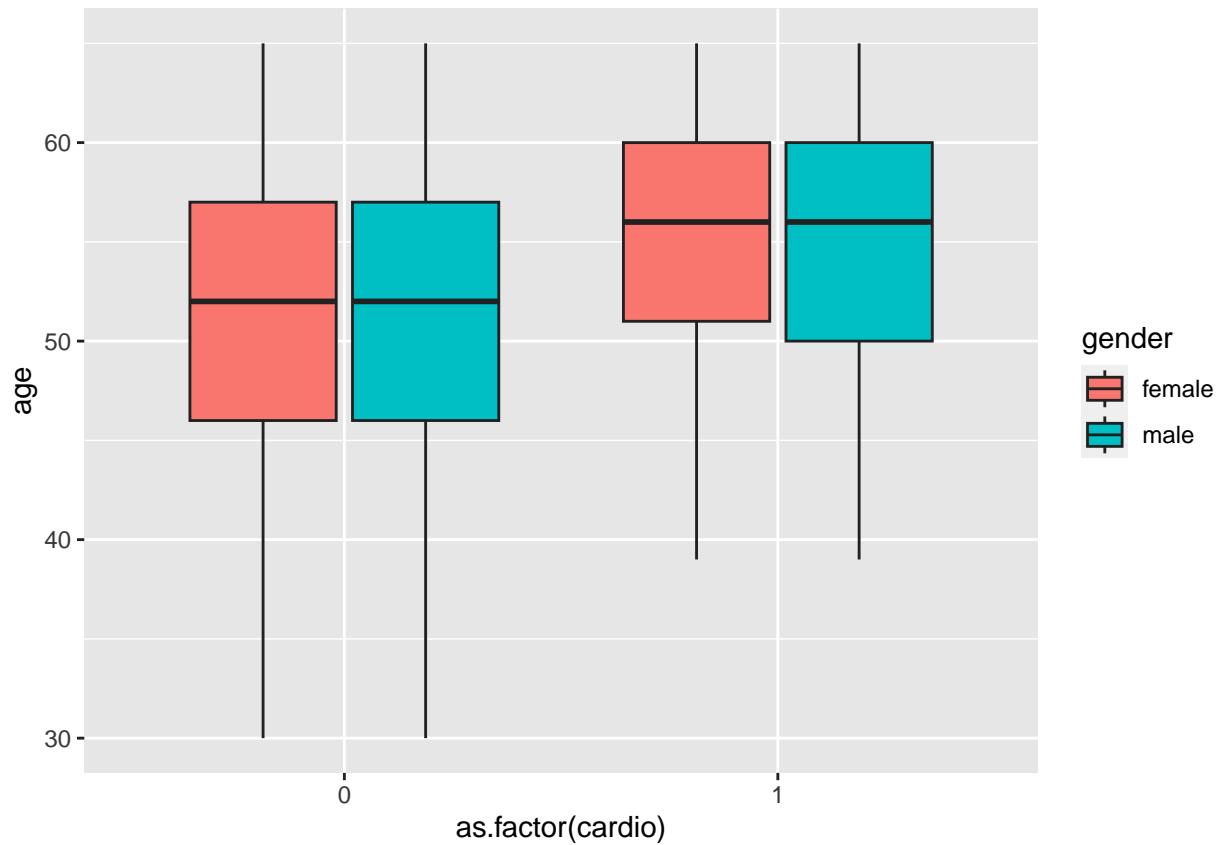
```

ggplot(data = data, aes(x = age, fill = gender)) + geom_histogram(bins = 10)

```

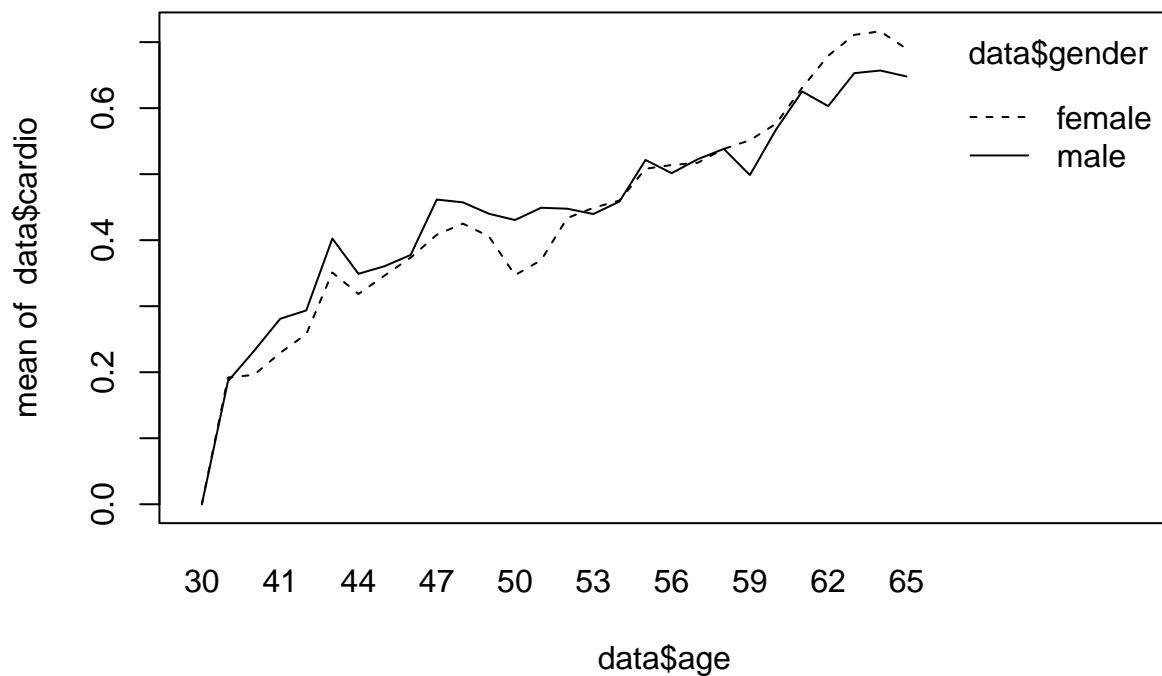


```
ggplot(data = data, aes(x = as.factor(cardio), y = age, fill = gender)) + geom_boxplot()
```



Osoby chore mają średnio większy wiek od osób zdrowych, niezależnie od płci.

```
interaction.plot(x.factor = data$age,
                  trace.factor = data$gender,
                  response = data$cardio)
```

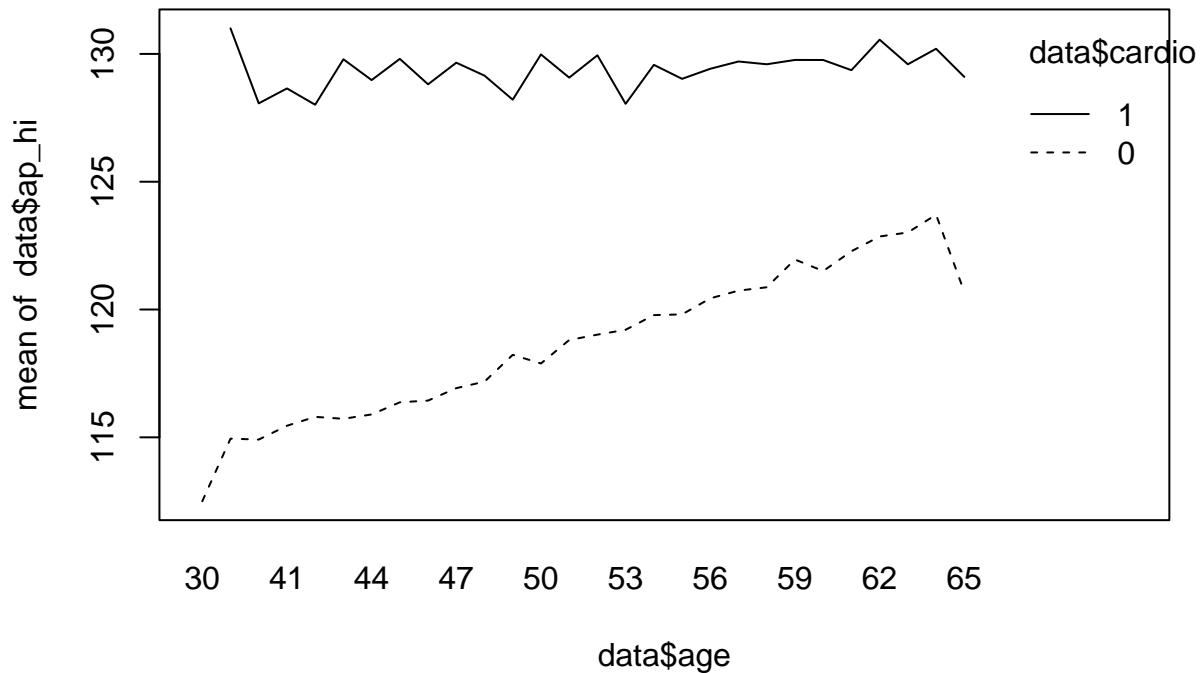


Brak interakcji między zmienną gender i age.

Prawdopodobieństwo posiadania choroby wzrasta w podobnym tempie zarówno u kobiet jak i u mężczyzn względem wzrostu wieku.

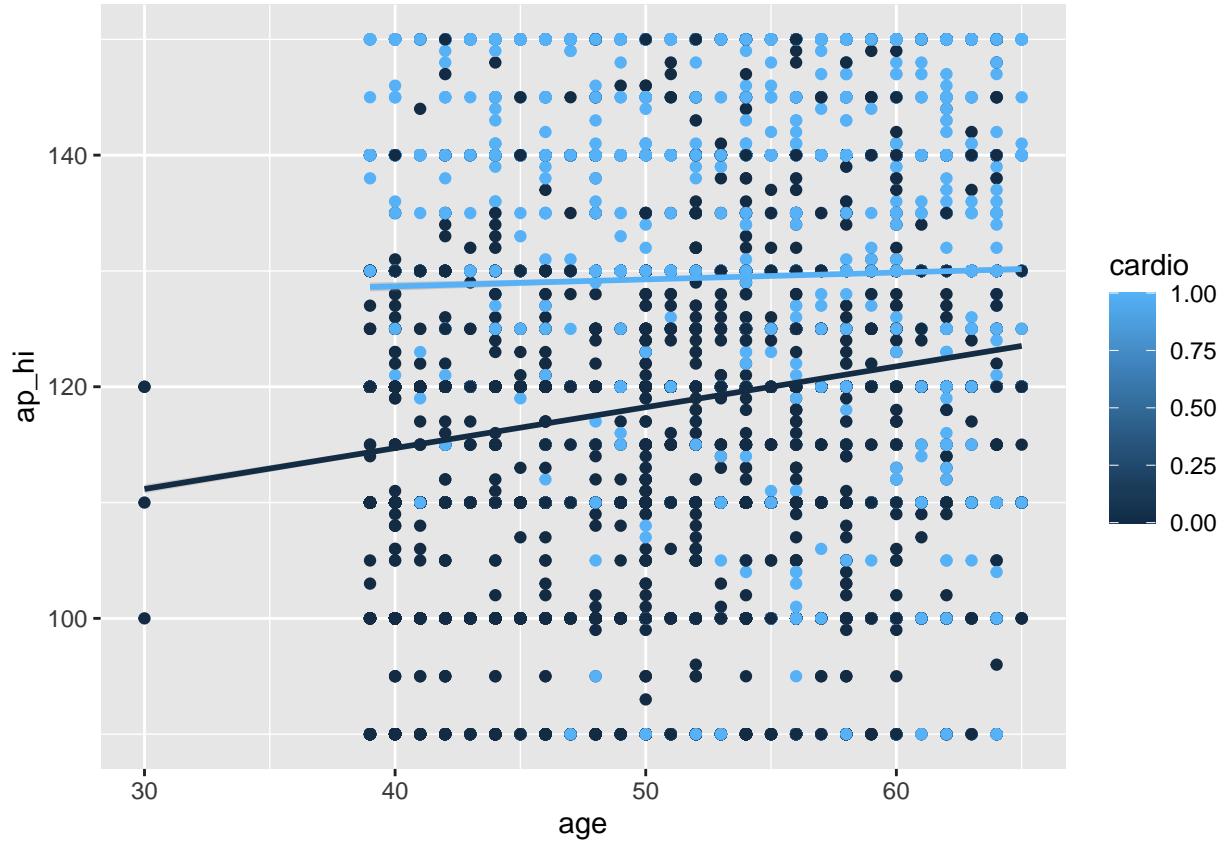
Wraz z wiekiem wzrasta ciśnienie krwi. Można zatem sprawdzić jaki wpływ ma wzrost wieku na ciśnienie krwi w obu grupach.

```
interaction.plot(x.factor = data$age,
                 trace.factor = data$cardio,
                 response = data$ap_hi)
```



```
ggplot(data = data, aes(x = age, y = ap_hi, group = cardio, colour = cardio)) + geom_point() + geom_smooth()
```

```
## 'geom_smooth()' using formula = 'y ~ x'
```



```
anova(glm(cardio ~ ap_hi + age, family = 'binomial', data = data),
      glm(cardio ~ ap_hi * age, family = 'binomial', data = data),
      test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: cardio ~ ap_hi + age
## Model 2: cardio ~ ap_hi * age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     63585    73823
## 2     63584    73445  1    377.66 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Osoby chore mają stale wysokie ciśnienie - wraz z wiekiem ich ciśnienie skurczowe krwi praktycznie nie rośnie. U osób zdrowych ciśnienie skurczowe krwi rośnie wraz z wiekiem.

```
model_11 = glm(cardio ~ ap_hi * age + gender, family = 'binomial',
                data = data)
summary(model_11)
```

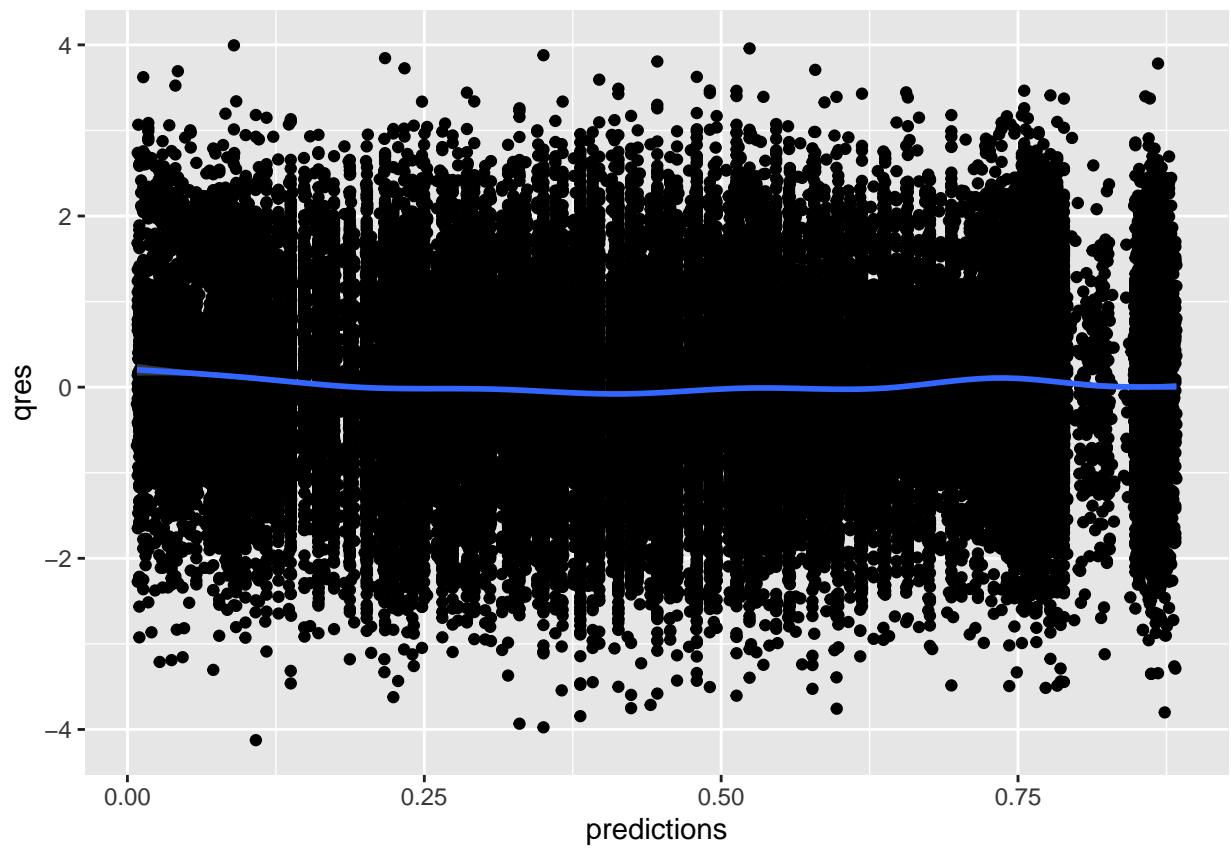
```
##
## Call:
## glm(formula = cardio ~ ap_hi * age + gender, family = "binomial",
```

```

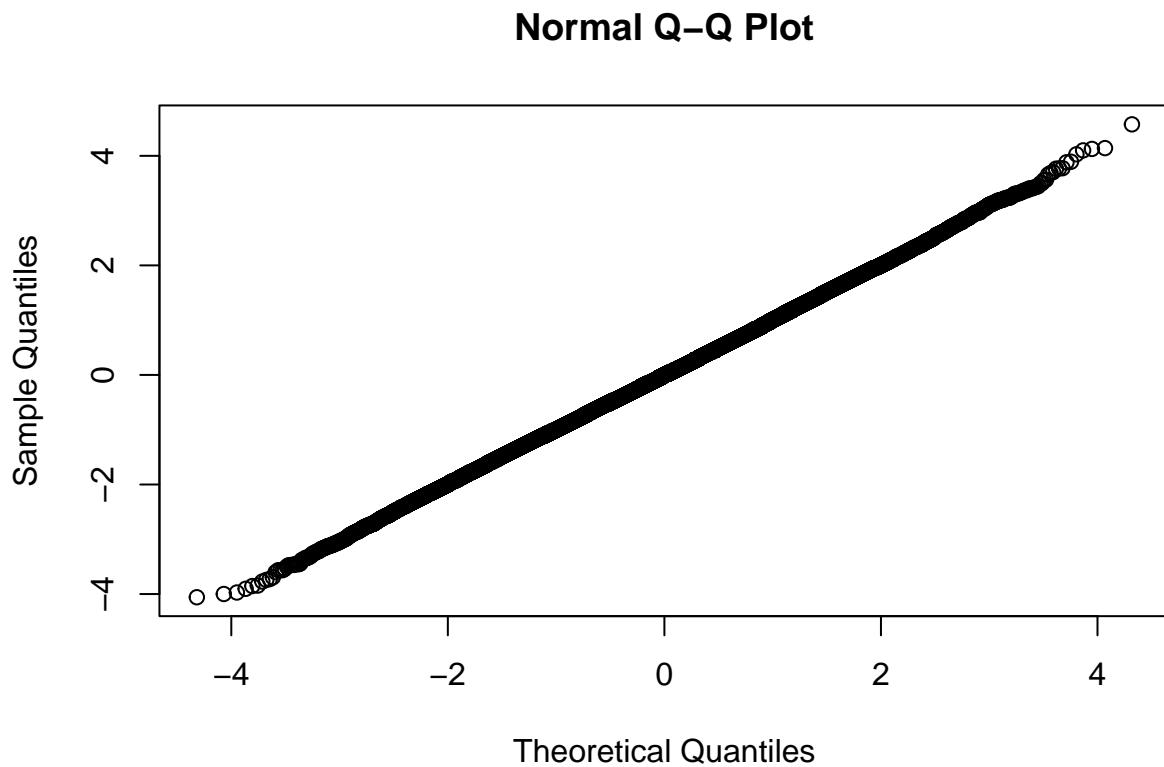
##      data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.0695  -0.9801  -0.4159   0.9801   3.0793
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.923e+01 8.854e-01 -33.010 < 2e-16 ***
## ap_hi        2.105e-01 7.121e-03 29.563 < 2e-16 ***
## age          3.682e-01 1.621e-02 22.719 < 2e-16 ***
## gendermale  -8.993e-02 1.875e-02 -4.797 1.61e-06 ***
## ap_hi:age   -2.510e-03 1.301e-04 -19.290 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895 on 63587 degrees of freedom
## Residual deviance: 73422 on 63583 degrees of freedom
## AIC: 73432
##
## Number of Fisher Scoring iterations: 4
```

```
qres_plot(model_11)
```

```
## 'geom_smooth()' using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
qqnorm(statmod::qresid(model_11))
```



Reszty kwantylowe pochodzą z rozkładu normalnego. Widoczne są delikatne ogony na krańcach.

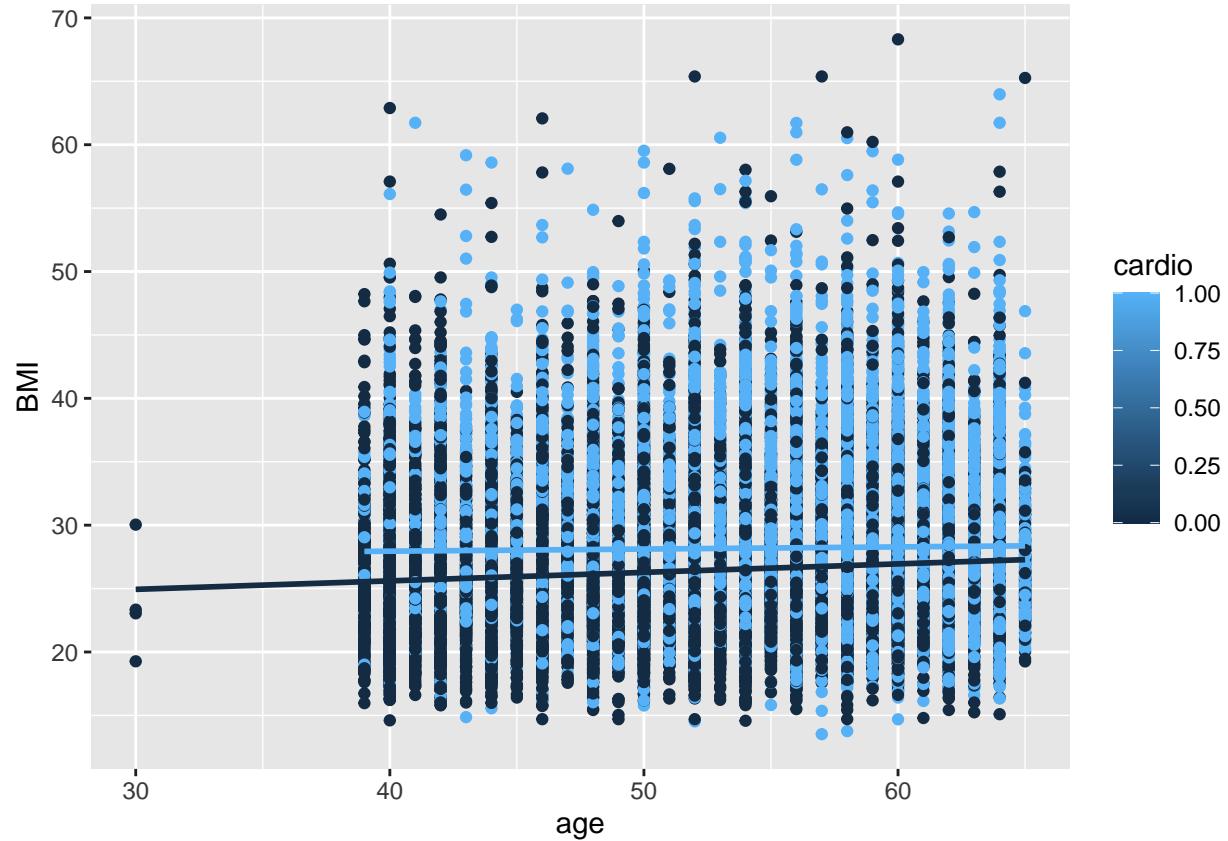
```
hoslem.test(x = model_11$y, y = model_11$fitted)
```

```
##  
## Hosmer and Lemeshow goodness of fit (GOF) test  
##  
## data: model_11$y, model_11$fitted  
## X-squared = 546.74, df = 8, p-value < 2.2e-16
```

Test Hosmera-Lemeshowa wskazuje na niedopasowanie modelu.

Zmienna BMI we wcześniejszych analizach okazała się bardzo istotna w oszacowaniu prawdopodobieństwa posiadania choroby układu krążenia u pacjenta. Wraz z wiekiem zazwyczaj waga ciała rośnie, co za tym idzie, BMI również rośnie. Można zatem sprawdzić powiązanie zmiennych age i BMI.

```
ggplot(data = data, aes(x = age, y = BMI, group = cardio, colour = cardio)) + geom_point() + geom_smooth()  
  
## 'geom_smooth()' using formula = 'y ~ x'
```



Brak interakcji. Tempo wzrostu BMI wraz z wiekiem jest bardzo podobne w przypadku osób zdrowych i chorych, natomiast średnia wartość BMI u osób zdrowych i chorych różni się - osoby chore mają większe BMI.

```
model_12 = glm(cardio ~ ap_hi * age + gender + BMI, family = 'binomial',
                data = data)
summary(model_12)
```

```
##
## Call:
## glm(formula = cardio ~ ap_hi * age + gender + BMI, family = "binomial",
##      data = data)
##
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.3926   -0.9585   -0.4009    0.9542    3.0789
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -29.520861   0.884711 -33.368 <2e-16 ***
## ap_hi        0.205356   0.007116  28.860 <2e-16 ***
## age          0.362787   0.016191  22.406 <2e-16 ***
## gendermale   -0.045229   0.018932  -2.389  0.0169 *
## BMI          0.035221   0.001874  18.798 <2e-16 ***
## ap_hi:age    -0.002473   0.000130 -19.022 <2e-16 ***
## ---
```

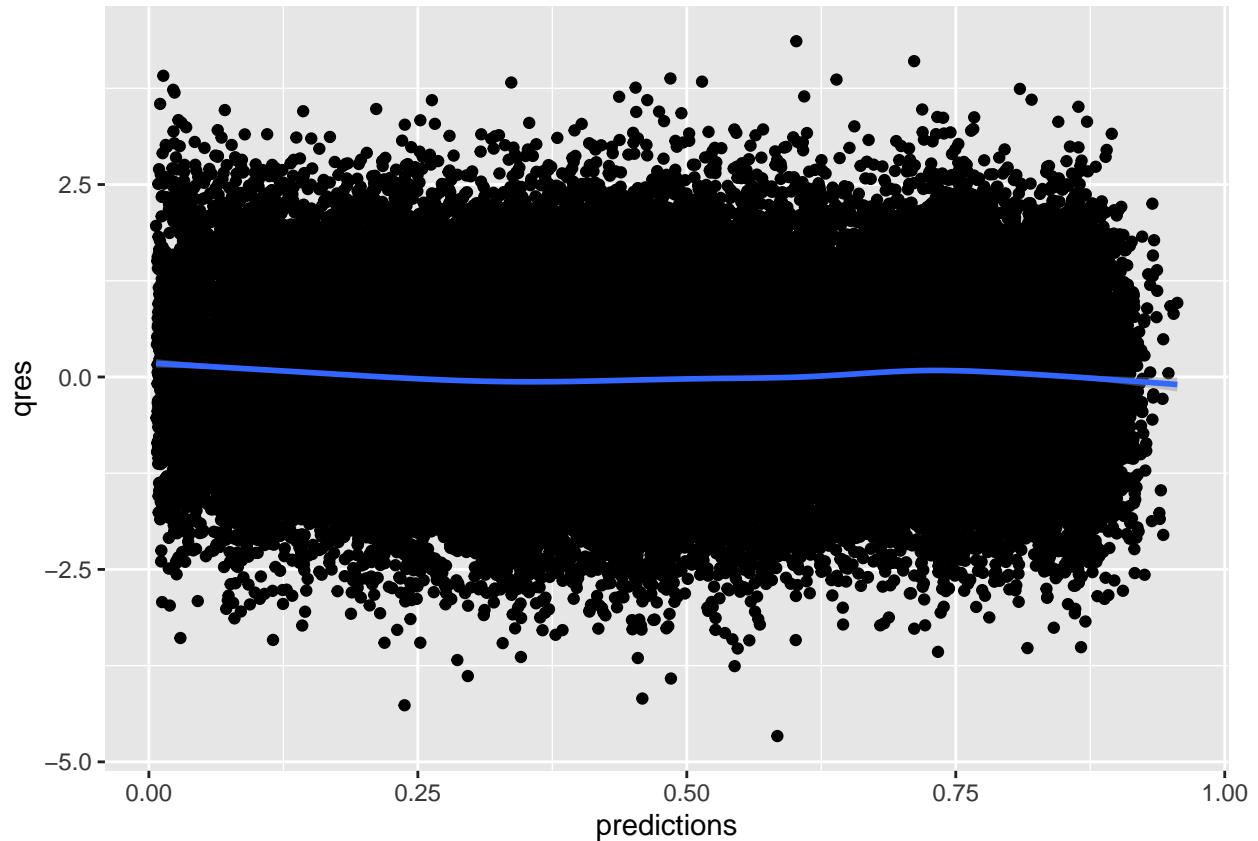
```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 87895  on 63587  degrees of freedom
## Residual deviance: 73063  on 63582  degrees of freedom
## AIC: 73075
##
## Number of Fisher Scoring iterations: 4

```

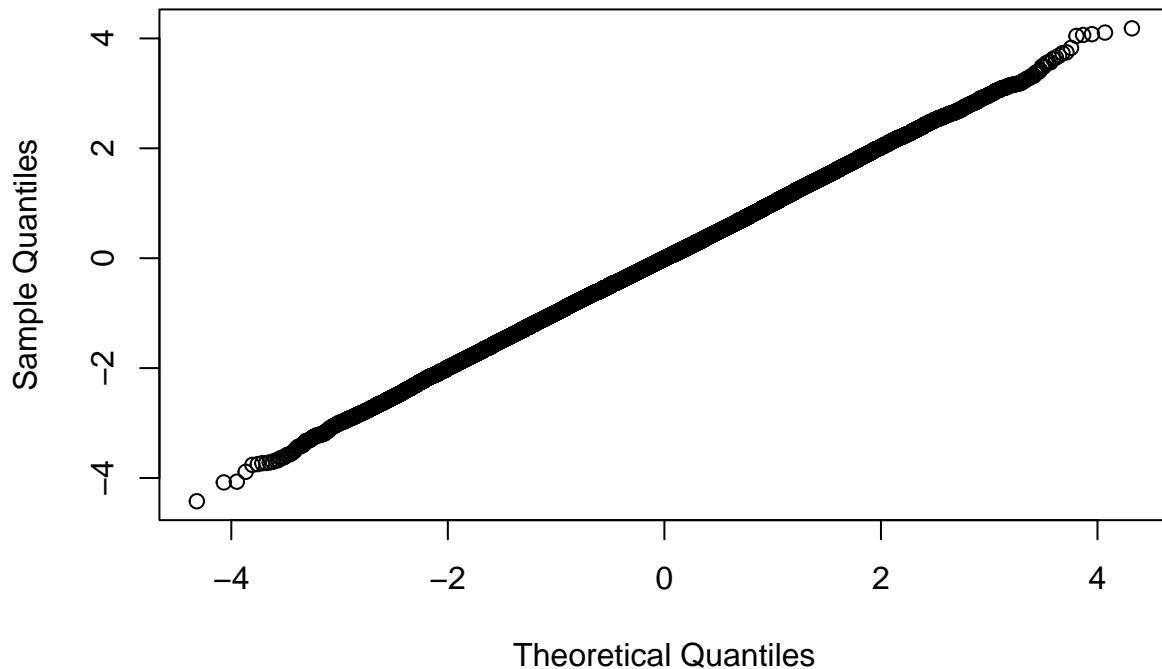
```
qres_plot(model_12)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
qqnorm(statmod::qresid(model_12))
```

## Normal Q-Q Plot



```
hoslem.test(x = model_12$y, y = model_12$fitted)
```

```
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: model_12$y, model_12$fitted
## X-squared = 437.88, df = 8, p-value < 2.2e-16
```

Tak jak wcześniej, do predyktorów ap\_hi oraz BMI można dodać ich odpowiedniki  $ap\_hi^2$  oraz  $BMI^2$ .

```
model_13 = glm(cardio ~ (ap_hi + I(ap_hi^2)) * age + gender + BMI + I(BMI^2),
                family = 'binomial', data = data)
summary(model_13)
```

```
##
## Call:
## glm(formula = cardio ~ (ap_hi + I(ap_hi^2)) * age + gender +
##      BMI + I(BMI^2), family = "binomial", data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.5815   -0.9297   -0.4765    0.9817    2.4478
##
## Coefficients:
```

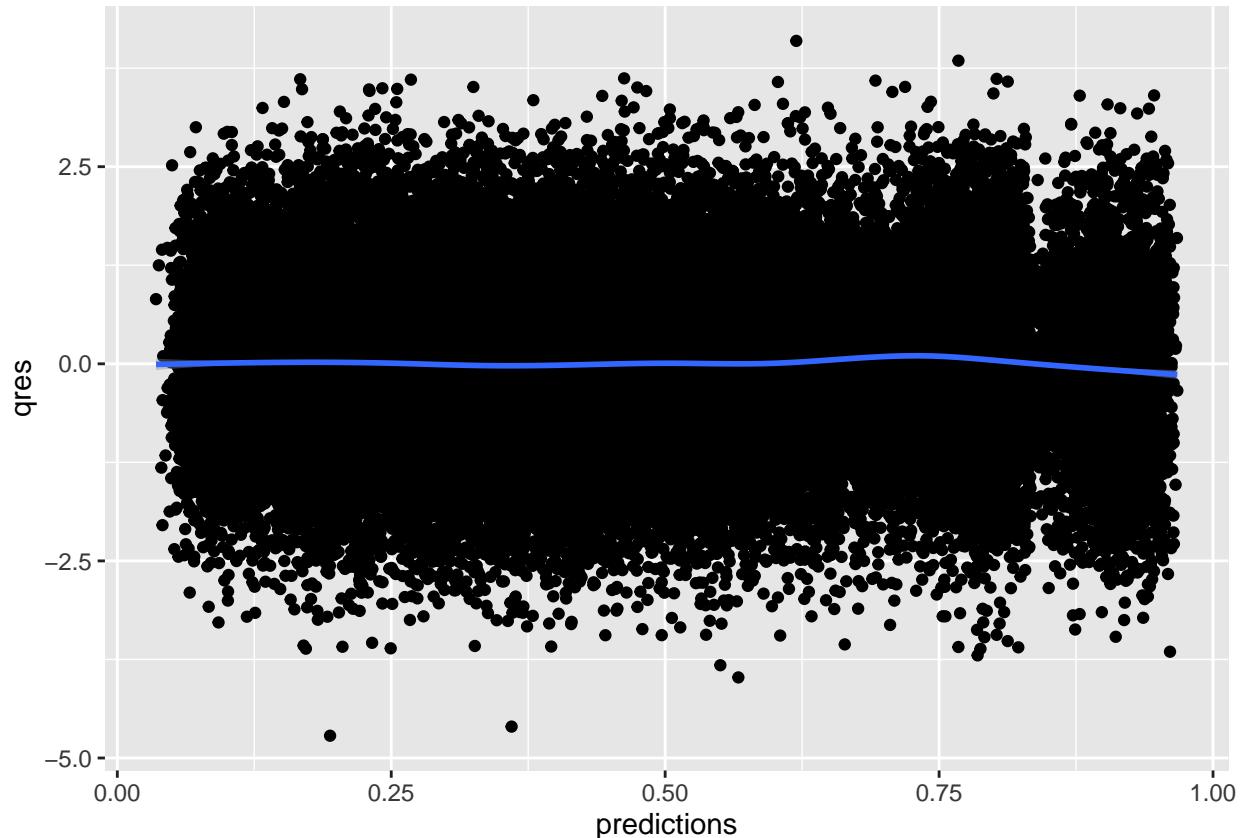
```

##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.610e+01  6.186e+00  5.836 5.35e-09 ***
## ap_hi      -8.873e-01  1.002e-01 -8.854 < 2e-16 ***
## I(ap_hi^2) 4.435e-03  4.048e-04 10.956 < 2e-16 ***
## age        -5.665e-01  1.149e-01 -4.930 8.22e-07 ***
## gendermale -3.274e-02  1.914e-02 -1.711  0.0871 .
## BMI         1.165e-01  1.249e-02  9.326 < 2e-16 ***
## I(BMI^2)   -1.344e-03  2.035e-04 -6.606 3.95e-11 ***
## ap_hi:age   1.280e-02  1.857e-03  6.895 5.40e-12 ***
## I(ap_hi^2):age -6.230e-05 7.477e-06 -8.333 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895 on 63587 degrees of freedom
## Residual deviance: 72511 on 63579 degrees of freedom
## AIC: 72529
##
## Number of Fisher Scoring iterations: 4

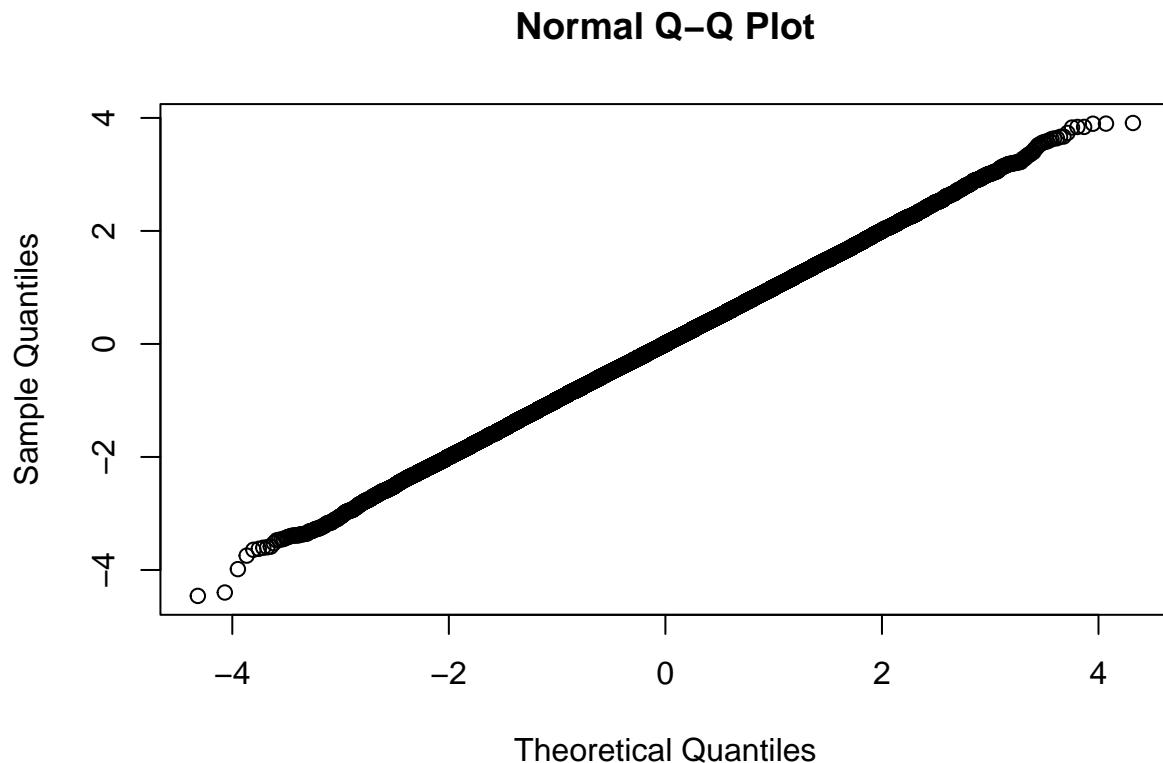
```

```
qres_plot(model_13)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
qqnorm(statmod::qresid(model_13))
```



```
hoslem.test(x = model_13$y, y = model_13$fitted)
```

```
##  
## Hosmer and Lemeshow goodness of fit (GOF) test  
##  
## data: model_13$y, model_13$fitted  
## X-squared = 173.39, df = 8, p-value < 2.2e-16
```

Test wskazuje na niedopasowanie modelu.

```
model_14 = glm(cardio ~ (ap_hi + I(ap_hi^2)) * age + gender * gluc +  
    cholesterol + BMI + I(BMI^2) + active + smoke + alco,  
    family = 'binomial', data = data)  
summary(model_14)
```

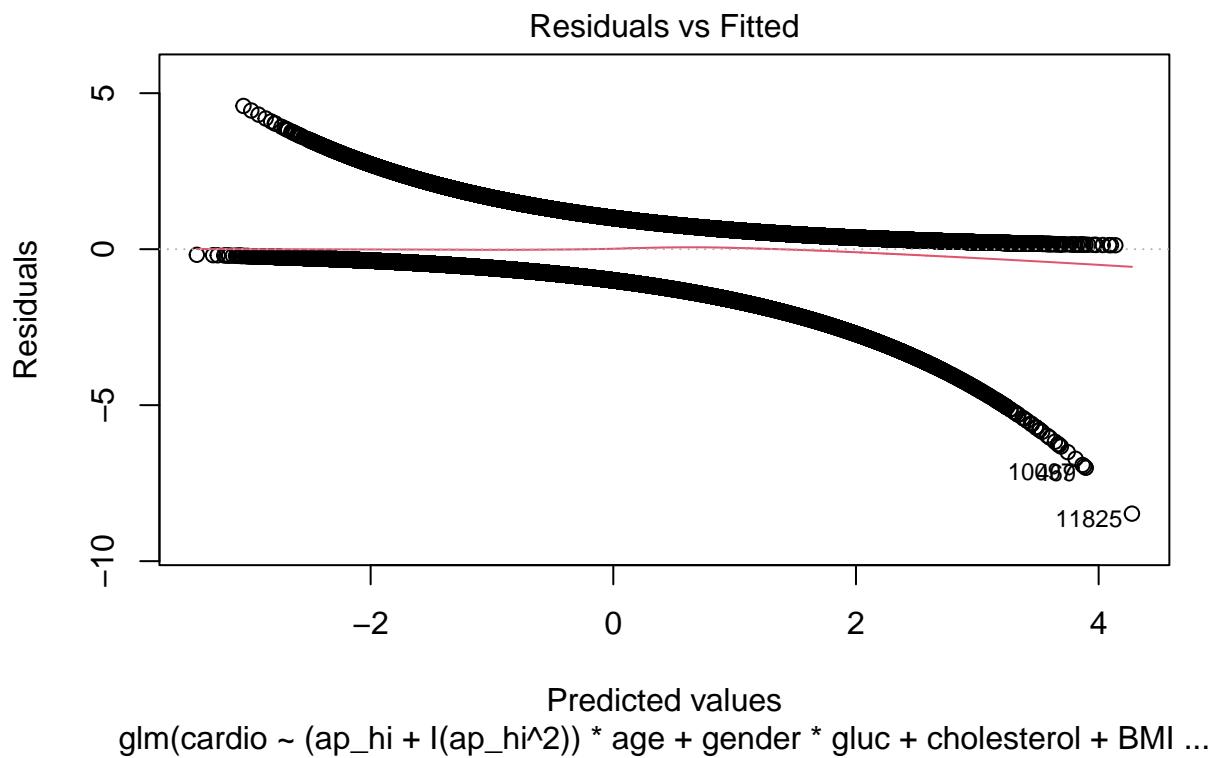
```
##  
## Call:  
## glm(formula = cardio ~ (ap_hi + I(ap_hi^2)) * age + gender *  
##   gluc + cholesterol + BMI + I(BMI^2) + active + smoke + alco,  
##   family = "binomial", data = data)  
##  
## Deviance Residuals:
```

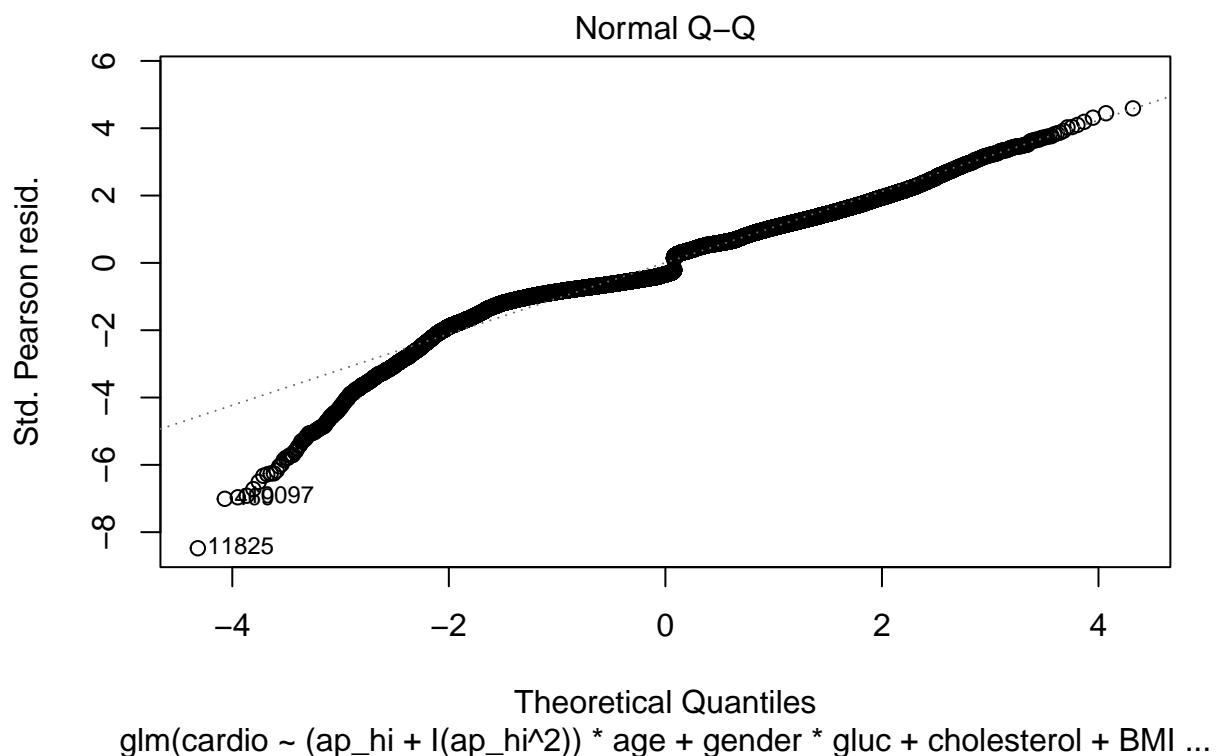
```

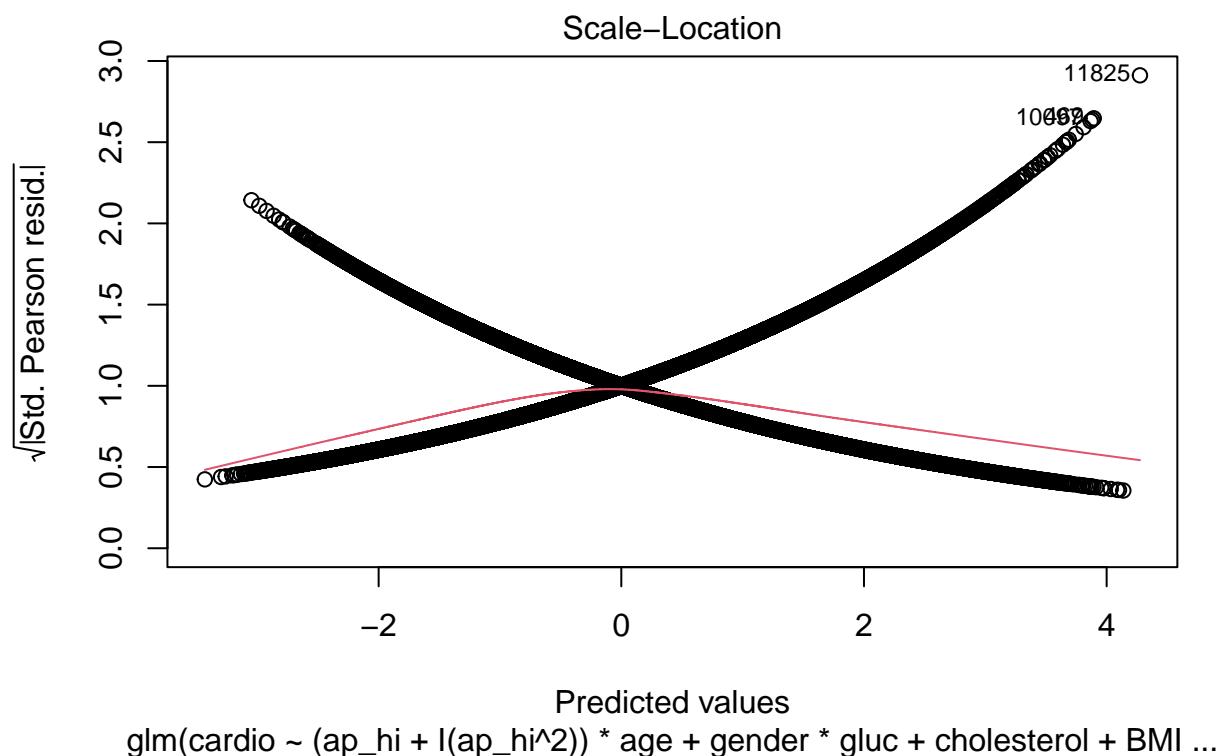
##      Min       1Q     Median      3Q      Max
## -2.9286 -0.9022 -0.4695  0.9134  2.4877
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)            3.652e+01  6.175e+00  5.914 3.34e-09 ***
## ap_hi                 -8.850e-01  1.000e-01 -8.849 < 2e-16 ***
## I(ap_hi^2)              4.424e-03  4.039e-04 10.953 < 2e-16 ***
## age                  -5.742e-01  1.148e-01 -5.002 5.66e-07 ***
## gendermale             4.244e-02  2.186e-02  1.941  0.0522 .
## gluca_norm              6.289e-03  4.632e-02  0.136  0.8920
## glucwa_norm             -3.058e-01  4.798e-02 -6.375 1.83e-10 ***
## cholesterola_norm        3.471e-01  2.876e-02 12.071 < 2e-16 ***
## cholesterolaw_norm       1.144e+00  3.731e-02 30.654 < 2e-16 ***
## BMI                   1.082e-01  1.266e-02  8.548 < 2e-16 ***
## I(BMI^2)                -1.311e-03  2.065e-04 -6.349 2.16e-10 ***
## activeactive            -2.528e-01  2.274e-02 -11.116 < 2e-16 ***
## smokesmoker             -1.916e-01  3.672e-02 -5.220 1.79e-07 ***
## alcoalc                -2.217e-01  4.482e-02 -4.947 7.53e-07 ***
## ap_hi:age               1.288e-02  1.854e-03  6.946 3.75e-12 ***
## I(ap_hi^2):age          -6.274e-05  7.463e-06 -8.406 < 2e-16 ***
## gendermale:gluca_norm   9.311e-02  7.733e-02  1.204  0.2285
## gendermale:glucwa_norm -1.492e-01  7.527e-02 -1.982  0.0474 *
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 87895 on 63587 degrees of freedom
## Residual deviance: 71162 on 63570 degrees of freedom
## AIC: 71198
##
## Number of Fisher Scoring iterations: 4

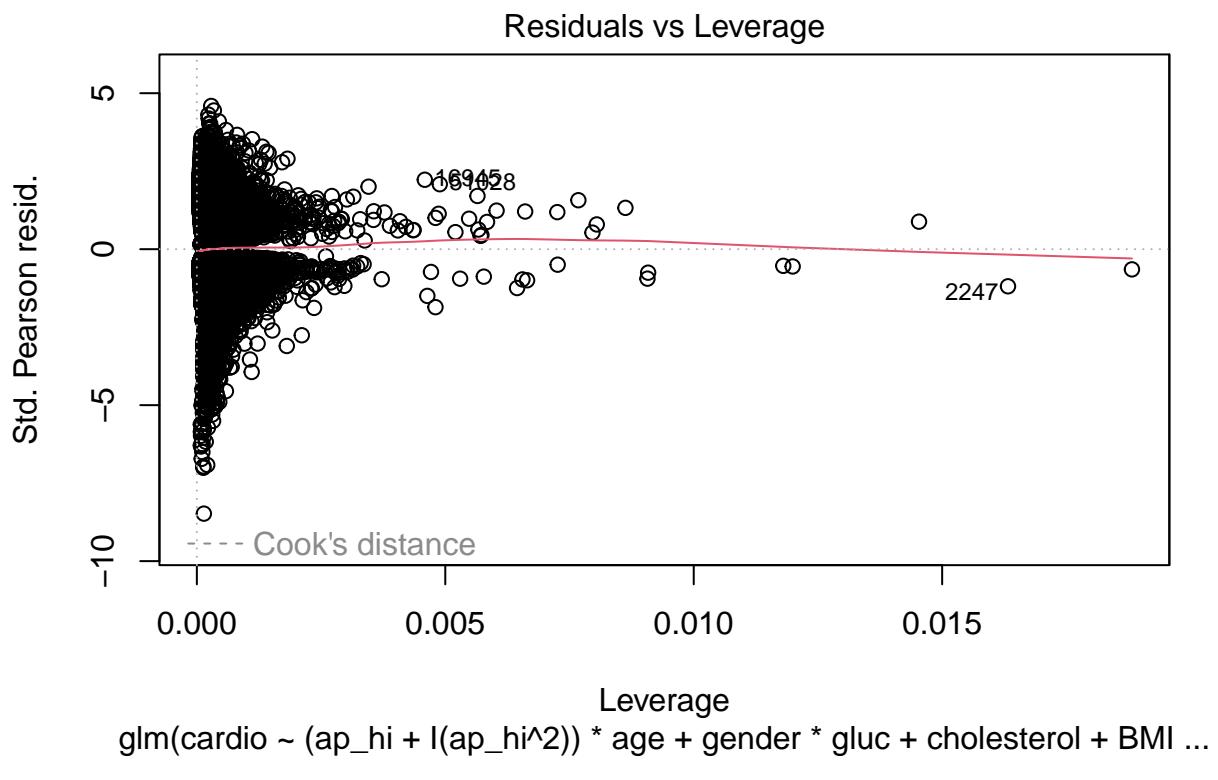
```

```
plot(model_14)
```



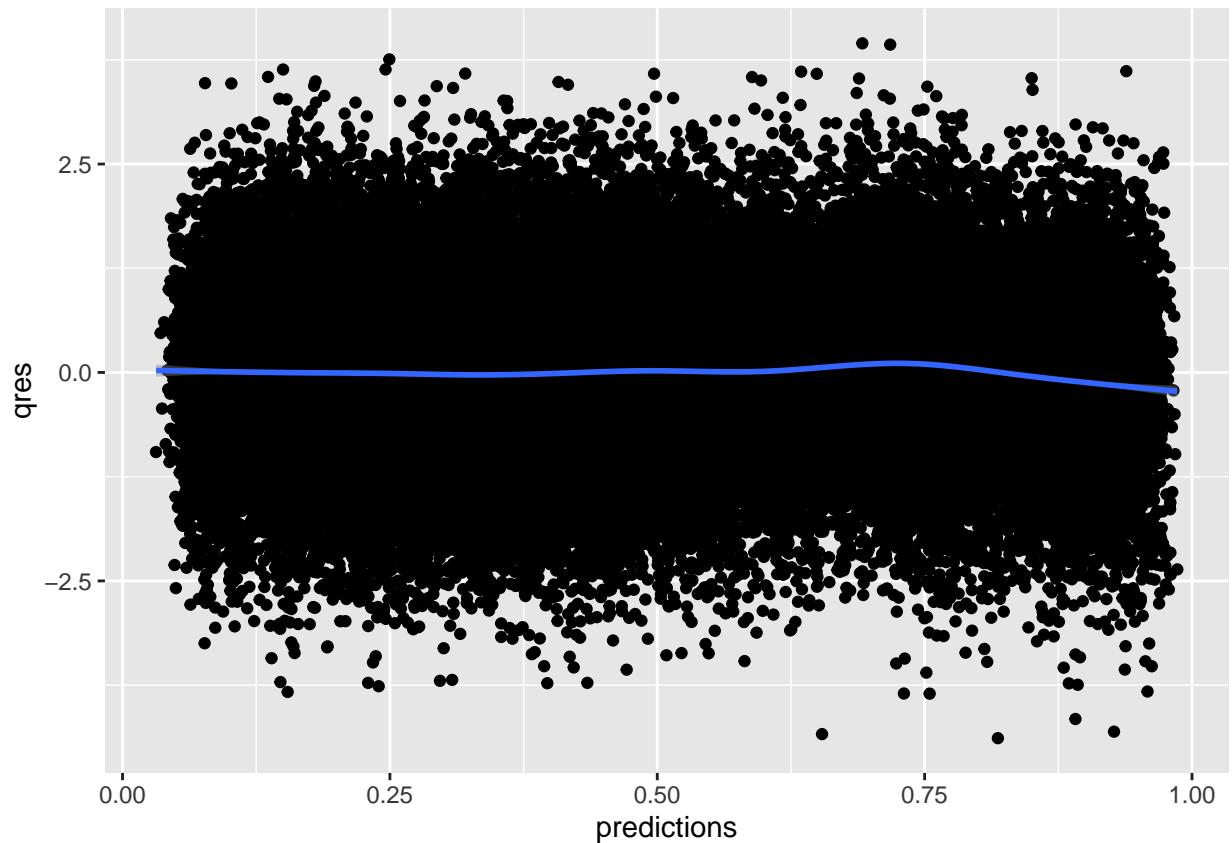




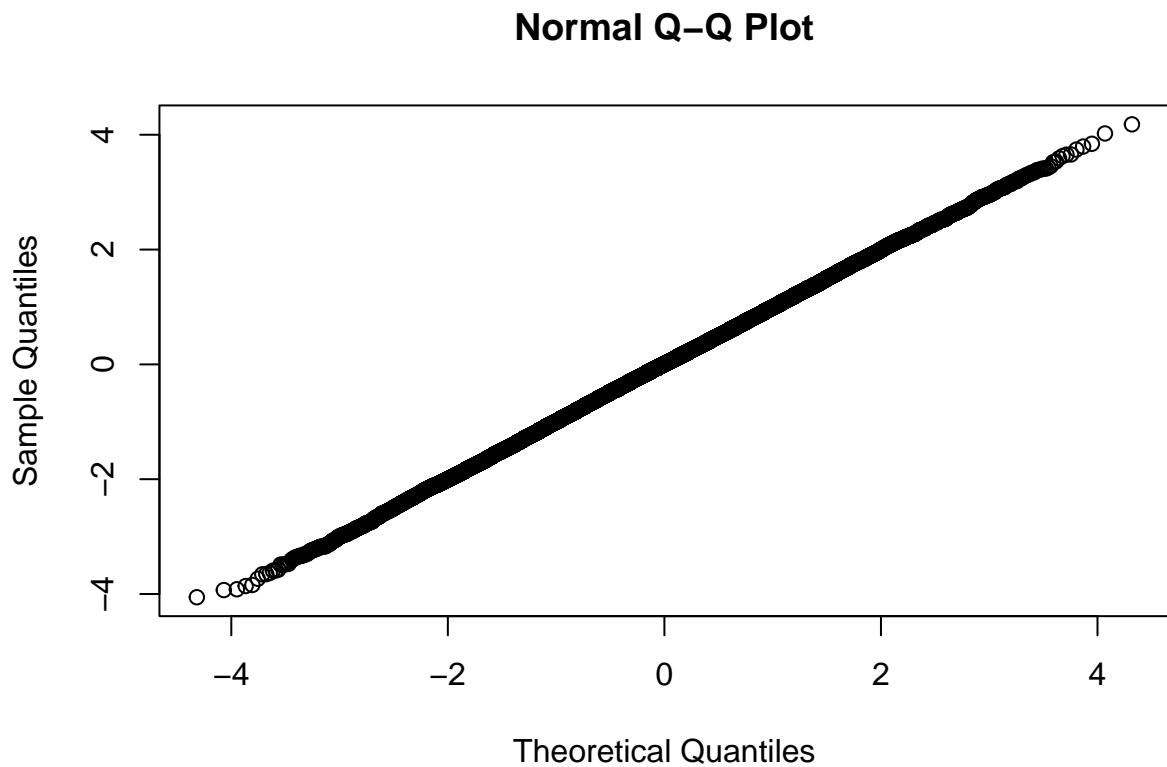


```
qres_plot(model_14)
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



```
qqnorm(statmod::qresid(model_14))
```



```
hoslem.test(x = model_14$y, y = model_14$fitted)
```

```
##
## Hosmer and Lemeshow goodness of fit (GOF) test
##
## data: model_14$y, model_14$fitted
## X-squared = 328.62, df = 8, p-value < 2.2e-16
```

Model jest niedopasowany.

```
anova(model_10, model_14, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: cardio ~ ap_hi + I(ap_hi^2) + gender * gluc + cholesterol + active +
##           smoke + alco + BMI + I(BMI^2)
## Model 2: cardio ~ (ap_hi + I(ap_hi^2)) * age + gender * gluc + cholesterol +
##           BMI + I(BMI^2) + active + smoke + alco
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1     63573      73139
## 2     63570      71162  3    1976.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Uwzględnienie zmiennej age w modelu jest bardzo istotne.

Wszystkie współczynniki modelu są istotne statystycznie. Brak wartości odstających, reszty oscylują w zerze i pochodzą z rozkładu normalnego. Mimo wszystko dewiancja resztowa jest znacznie za duża - test Hosmera-Lemeshowa wskazuje na niedopasowanie modelu. Tak jak wcześniej występuje problem z nadmierną dyspersją. Można zatem spróbować zmienić rodzinę rozkładu z binomial na quasibinomial, wówczas zmieni się parametr dyspersji.

```
model_update = update(model_14, family = quasibinomial)
summary(model_update)
```

```
## 
## Call:
## glm(formula = cardio ~ (ap_hi + I(ap_hi^2)) * age + gender *
##       gluc + cholesterol + BMI + I(BMI^2) + active + smoke + alco,
##       family = quasibinomial, data = data)
## 
## Deviance Residuals:
##      Min        1Q     Median        3Q       Max
## -2.9286  -0.9022  -0.4695   0.9134   2.4877
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            3.652e+01  6.288e+00  5.808 6.37e-09 ***
## ap_hi                 -8.850e-01  1.018e-01 -8.690 < 2e-16 ***
## I(ap_hi^2)              4.424e-03  4.113e-04 10.756 < 2e-16 ***
## age                  -5.742e-01  1.169e-01 -4.912 9.03e-07 ***
## gendermale             4.244e-02  2.227e-02  1.906  0.0566 .  
## gluca_norm              6.289e-03  4.717e-02  0.133  0.8939
## glucwa_norm             -3.058e-01  4.885e-02 -6.260 3.88e-10 ***
## cholesterol_norm         3.471e-01  2.929e-02 11.853 < 2e-16 ***
## cholesterolaw_norm       1.144e+00  3.800e-02 30.103 < 2e-16 ***
## BMI                   1.082e-01  1.289e-02  8.394 < 2e-16 ***
## I(BMI^2)                -1.311e-03  2.103e-04 -6.235 4.55e-10 ***
## activeactive            -2.528e-01  2.316e-02 -10.916 < 2e-16 ***
## smokesmoker             -1.916e-01  3.739e-02 -5.126 2.97e-07 ***
## alcoalc                -2.217e-01  4.564e-02 -4.858 1.19e-06 ***
## ap_hi:age               1.288e-02  1.888e-03  6.821 9.10e-12 ***
## I(ap_hi^2):age           -6.274e-05  7.600e-06 -8.255 < 2e-16 ***
## gendermale:gluca_norm    9.311e-02  7.875e-02  1.182  0.2370
## gendermale:glucwa_norm   -1.492e-01  7.664e-02 -1.947  0.0516 .
## ---                     
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for quasibinomial family taken to be 1.036996)
## 
## Null deviance: 87895  on 63587  degrees of freedom
## Residual deviance: 71162  on 63570  degrees of freedom
## AIC: NA
## 
## Number of Fisher Scoring iterations: 4
```

Odchylenia standardowe współczynników delikatnie wzrosły. Parametr dyspersji zmienił się z 1 na zaledwie 1.037.

```

anova(model_update, test = 'F')

## Analysis of Deviance Table
##
## Model: quasibinomial, link: logit
##
## Response: cardio
##
## Terms added sequentially (first to last)
##
##
##          Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
## NULL             63587     87895
## ap_hi            1   12247.4    63586    75647 11810.4551 < 2.2e-16 ***
## I(ap_hi^2)       1     307.4    63585    75340   296.4382 < 2.2e-16 ***
## age              1   1911.4    63584    73429   1843.1963 < 2.2e-16 ***
## gender           1     12.5    63583    73416   12.0169 0.0005276 ***
## gluc              2     109.3    63581    73307   52.6948 < 2.2e-16 ***
## cholesterol      2    1153.6    63579    72153   556.2369 < 2.2e-16 ***
## BMI               1     236.4    63578    71917   227.9630 < 2.2e-16 ***
## I(BMI^2)          1     40.5    63577    71876   39.0888 4.075e-10 ***
## active            1    131.1    63576    71745   126.3854 < 2.2e-16 ***
## smoke             1     46.2    63575    71699   44.5807 2.461e-11 ***
## alco              1     24.4    63574    71675   23.5056 1.248e-06 ***
## ap_hi:age         1    439.0    63573    71236   423.3317 < 2.2e-16 ***
## I(ap_hi^2):age   1     67.9    63572    71168   65.5061 5.895e-16 ***
## gender:gluc       2      5.8    63570    71162   2.7738 0.0624306 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Wszystkie współczynniki modelu i ich powiązania są istotne statystycznie.

## WNIOSKI

- Dalej występuje problem z nadmierną dyspersją
- Osoby chore są z reguły starsze
- U osób chorych ciśnienie skurczowe krwi praktycznie nie rośnie z wiekiem
- U osób zdrowych ciśnienie skurczowe krwi rośnie wraz z wiekiem
- Wiek jest bardzo istotny w określeniu p.p posiadania choroby układu krążenia

Informacje o aktywności, spożywaniu alkoholu i paleniu są subiektywną oceną pacjenta. Skala 0-1 w przypadku aktywności może być zbyt mała aby dobrze określić wpływ aktywności fizycznej na ciśnienie krwi. Również wyniki samego ciśnienia skurczowego krwi są mocno zaokrąglone - wartości w większości zmieniają się co 5 mm/Hg. Wszystko to może właśnie spowodować niedopasowanie modelu. Również nieuwzględnienie zmiennej ap\_lo - ciśnienie rozkurczowe krwi - może się do tego przyczynić. Mimo dużej korelacji ap\_lo z ap\_hi ~ 0.7, pewnie warto by było się jej bliżej przyjrzeć i ewentualnie uwzględnić w modelu.