# 1. WSTĘP

## 1.1. Charakterystyka badanego zagadnienia

System zdrowotny w Stanach Zjednoczonych jest dosyć złożony. W USA obowiązuje rynkowy model opieki zdrowotnej, w którym tylko 40 proc. wydatków finansowana jest przez władze federalne lub stanowe, a aż 60 proc. – przez sektor prywatny: towarzystwa ubezpieczeniowe i bezpośrednie wpłaty od pacjentów. W Stanach Zjednoczonych ubezpieczenie zdrowotne oferowane jest przez kilka tysięcy agencji ubezpieczeniowych. Proponują one rożne plany ubezpieczeniowe o zróżnicowanym poziomie cen. Wszystkie plany są opłacane składką, zazwyczaj miesięczną. Większość z nich wymaga także opłaty z góry określonej kwoty za wizytę u lekarza (współpłacenie) lub ustalenia procentowego rozdziału wydatków na usługę medyczną pomiędzy pacjentem a firmą ubezpieczeniową (współubezpieczeni). Niektóre usługi wymagają obu powyższych opłat. Ponadto każdego roku większość osób musi zapłacić tak zwaną franszyzę redukcyjną czyli pewną kwotę, zanim ubezpieczenie zacznie pokrywać koszty leczenia.

Pozostałe ubezpieczenia zdrowotne to ubezpieczenia indywidualne, wykupywane głownie przez: osoby samo- zatrudniające się, studentów, emerytów, którzy nie mogą jeszcze skorzystać z programu Medicare, ponieważ mają poniżej 65 lat, bezrobotnych, osoby zmieniające pracę, osoby zatrudnione w mniejszych firmach, w których pracodawca nie oferuje ubezpieczenia zdrowotnego. Ten typ ubezpieczenia wykupuje niecałe 10% Amerykanów.

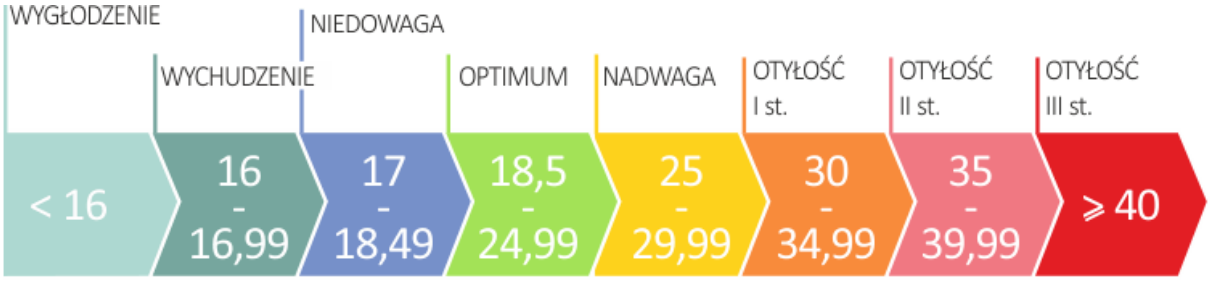
## 1.2. Problem i cel badawczy badanego zagadnienia

Problemem badawczym badania jest pytanie: Jakie czynniki wpływają na wysokość kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne danego beneficjenta palącego i niepalącego, w zależności od wybranych czynników dla danej populacji? Celem badania jest prognoza kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne w USA zależności od wybranych czynników dla danej populacji beneficjentów palących oraz niepalących.

## 1.3. Przygotowanie i prezentacja zbioru danych

Analiza predykcji wysokości stawki ubezpieczeniowej została przeprowadzona na zbiorze 1339 jednostek. W zbiorze danych nie ma braków. Baza danych zawiera następujące charakterystyki:

- **Age**: Wiek głównego beneficjenta (18 – 64 lata) – w latach
- **Sex**: Płeć kontrahenta ubezpieczeniowego (male – mężczyzna, female – kobieta)
- **Bmi**: Wskaźnik masy ciała, który zapewnia zrozumienie ciała – wagi, która jest stosunkowo wysoka lub niskia w stosunku do wzrostu. Obiektywny wskaźnik masy ciała (kg/m$^2$) przy użyciu stosunku wzrostu do masy ciała. Skala wskaźnika bmi znajduje się poniżej:




- **Children:** Liczba osób na utrzymaniu beneficjenta
- **Smoker:** Czy beneficjent jest palaczem (yes – tak; no – nie)
- **Region:** obszar mieszkalny beneficjenta w USA, północny wschód, południowy wschód, południowy zachód lub północny zachód.
- **Charges:** opłaty - indywidualne koszty leczenia rozliczane przez ubezpieczenie zdrowotne

### 1.4. Wybór zmiennych do badania

Zbiór wszystkich jednostek, liczący 1338 osób, został podzielony na dwie grupy – beneficjenci palący i niepalący. Do budowy wszystkich modeli dla grupy beneficjentów palących wybrano z całego zbioru tylko osoby palące - 274 jednostki. Do budowy wszystkich modeli dla grupy beneficjentów niepalących wybrano z całego również 274 jednostki. Przed badaniem wszystkim 1064 osobą niepalącym została nadana funkcja los(), która wylosowała 274 jednostki.

## 2. ETAPY BUDOWY MODELI DLA PALĄCYCH BENEFICJENTÓW

- Każdy wykres oraz tabela zostały wygenerowane w programie R Studio (źródło).
- W testach statystycznych przyjmujemy poziom istotności alfa = 0.01.
- W projekcie zastosowano kryterium dopasowania modelu wg. kryterium informacyjnego Akaike.

**2.1. W pierwszym etapie dokonano budowy, estymacji, a kolejno weryfikacji otrzymanych modeli liniowych objaśniających koszty leczenia rozliczane przez ubezpieczenie zdrowotne dla beneficjentów palących. Estymacji dokonano klasyczną metodą MNK - metoda najmniejszych kwadratów.**

### Niezbędne pakiety:

```
library("car") # funkcja vif()
library("ggplot2") # wykresy - funkcja ggplot()
library("lmtest") # testy diagnostyczne modeli lm
library("pscl") #pseudo-R2 funkcja pR2()
library("pROC") #funkcje roc, auc
```

### Wczytanie danych i podstawowe statystyki opisowe dla poszczególnych zmiennych zostały zaprezentowane w tabeli 1:

*Tabela 1*

```
dane1 <-  read.table("palacy.csv", header = TRUE, sep = ";",dec=",")
dane1$sex<-as.factor(dane1$sex)
dane1$region<-as.factor(dane1$region)
summary(dane1)
```

```
##       age            sex           bmi           children          region
##  Min.   :18.00   female:115   Min.   :17.20   Min.   :0.000   northeast:67
##  1st Qu.:27.00    male  :159   1st Qu.:26.08   1st Qu.:0.000    northwest:58
##  Median :38.00                Median :30.45   Median :1.000   southeast:91
##  Mean   :38.51                Mean   :30.71   Mean   :1.113   southwest:58
##  3rd Qu.:49.00                3rd Qu.:35.20   3rd Qu.:2.000
##  Max.   :64.00                Max.   :52.58   Max.   :5.000

charges
##  Min.   :12829
##  1st Qu.:20826
##  Median :34456
##  Mean   :32050
##  3rd Qu.:41019
##  Max.   :63770
```

Zmienną objaśnianą są koszty leczenia rozliczane przez ubezpieczenie zdrowotne (zmienna charges). Na początku stworzono modele 1-5 objaśniające koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależności od każdej ze zmiennych objaśniających ze z osobna.

## Model liniowy 1 ze zmienną *age*

*Tabela 2*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku beneficjenta palącego
m1 <- lm(charges  ~ age, data = dane1)
summary(m1)
```

```
## Call:
## lm(formula = charges ~ age, data = dane1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16072 -11137   5764   8592  28815
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20294.13    1913.40  10.606  < 2e-16 ***
## age           305.24      46.73   6.532 3.18e-10 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 10750 on 272 degrees of freedom
## Multiple R-squared:  0.1356, Adjusted R-squared:  0.1324
## F-statistic: 42.67 on 1 and 272 DF,  p-value: 3.181e-10
```

**Postać modelu:** charges = 20294.13 + 305.24 * *age*

Miary dopasowania z tabeli 2:

- Odchylenie standardowe reszt: Se = 10750
- Współczynnik determinacji $R^2$ = 0.1356
- $R^2$ skorygowany = 0.1324

**Wniosek:** Model 1 wyjaśnia 13,56% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

## Model liniowy 2 ze zmienną *sex*

*Tabela 3*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od płci beneficjenta palącego
m2 <- lm(charges  ~ sex, data = dane1)
summary(m2)
```

```
## Call:
## lm(formula = charges ~ sex, data = dane1)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -20213 -11153   1907   9139  33091
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    30679       1073  28.600   <2e-16 ***
## sexmale         2363       1408   1.678   0.0945 .
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 11500 on 272 degrees of freedom
## Multiple R-squared:  0.01025,    Adjusted R-squared:  0.006608
## F-statistic: 2.816 on 1 and 272 DF,  p-value: 0.09448
```

**Postać modelu:** charges = 30679 + 2363 * *sex*

Miary dopasowania z tabeli 3:

- Odchylenie standardowe reszt: Se = 11500
- Współczynnik determinacji $R^2$ = 0.01025
- $R^2$ skorygowany = 0.006608

**Wniosek:** Model 2 wyjaśnia 1,03% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

## Model liniowy 3 ze zmienną *bmi*

*Tabela 4*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od bmi beneficjenta palącego
m3 <- lm(charges  ~ bmi, data = dane1)
summary(m3)
```

```
## Call:
## lm(formula = charges ~ bmi, data = dane1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19768.0  -4487.9     34.4   3263.9  31055.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13186.58    2052.88  -6.423 5.93e-10 ***
## bmi           1473.11      65.48  22.496  < 2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
```

```
## Residual standard error: 6837 on 272 degrees of freedom
## Multiple R-squared:  0.6504, Adjusted R-squared:  0.6491
## F-statistic: 506.1 on 1 and 272 DF,  p-value: < 2.2e-16
```

**Postać modelu:** charges = -13186.58 + 1473.11 * *bmi*

Miary dopasowania z tabeli 4:

- Odchylenie standardowe reszt: Se = 6837
- Współczynnik determinacji $R^2$ = 0.6504
- $R^2$ skorygowany= 0.6491

**Wniosek:** Model 3 wyjaśnia 65,04% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

## Model liniowy 4 ze zmienną *children*

*Tabela 5*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od liczby osób na utrzymaniu beneficjenta palącego
m4 <- lm(charges  ~ children, data = dane1)
summary(m4)

## Call:
## lm(formula = charges ~ children, data = dane1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19539 -11408   2793   9173  32119
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31651.1      969.5  32.646   <2e-16 ***
## children       358.5      604.4   0.593    0.554
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 11560 on 272 degrees of freedom
## Multiple R-squared:  0.001292,   Adjusted R-squared:  -0.00238
## F-statistic: 0.3519 on 1 and 272 DF,  p-value: 0.5535
```

**Postać modelu:** charges = 31651.1 + 358.5 * *children*

Miary dopasowania z tabeli 5:

- Odchylenie standardowe reszt: Se = 11560
- Współczynnik determinacji $R^2$ = 0.001292
- $R^2$ skorygowany = -0.00238

**Wniosek:** Model 4 wyjaśnia 0,13% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

## Model liniowy 5 ze zmienną *region*

*Tabela 6*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od obszaru zamieszkania w USA beneficjenta palącego
m5 <- lm(charges  ~ region, data = dane1)
summary(m5)
```

```
## Call:
## lm(formula = charges ~ region, data = dane1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18425 -10181   2154   9281  29829
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      29673.5     1392.4  21.311  < 2e-16 ***
## regionnorthwest    518.5     2044.1   0.254  0.79997
## regionsoutheast   5171.5     1834.8   2.819  0.00518 **
## regionsouthwest   2595.5     2044.1   1.270  0.20527
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 11400 on 270 degrees of freedom
## Multiple R-squared:  0.03554,    Adjusted R-squared:  0.02482
## F-statistic: 3.316 on 3 and 270 DF,  p-value: 0.02046
```

**Postać modelu:** charges = 29673.5 + 518.5 * `regionnorthwest` *+ 5171.5 * `regionsoutheast`* + 2595.5 * *`regionsouthwest`*

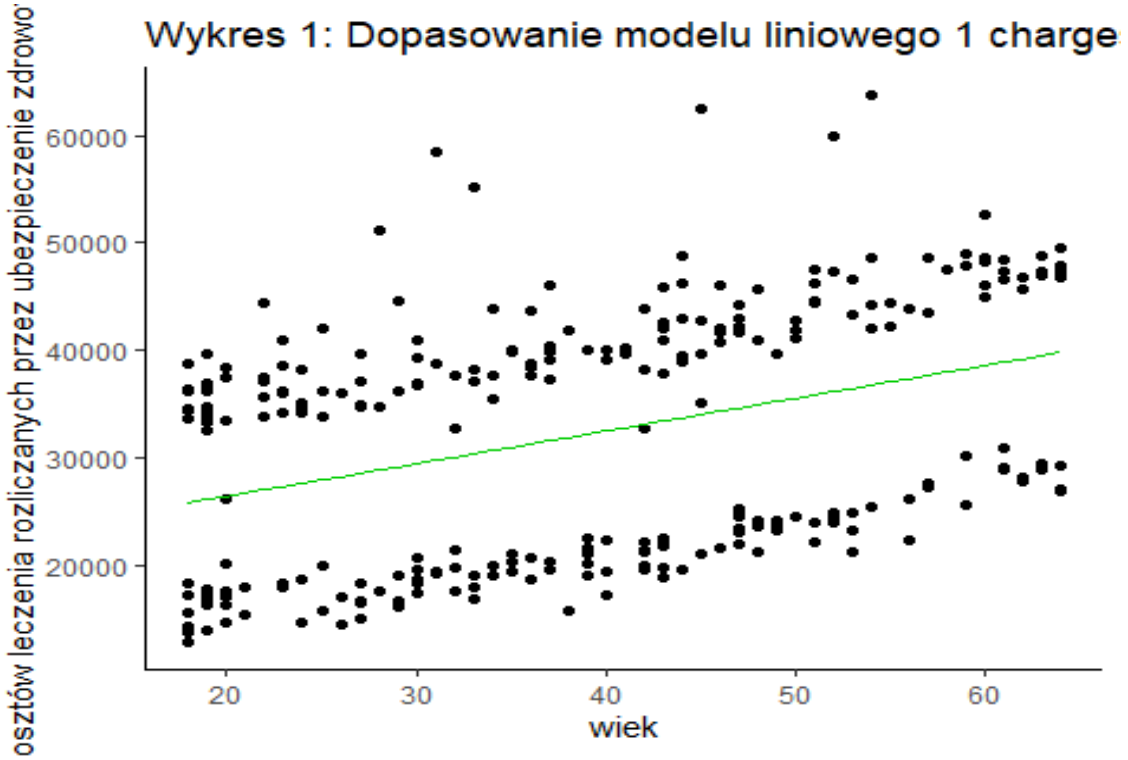
Miary dopasowania z tabeli 6:

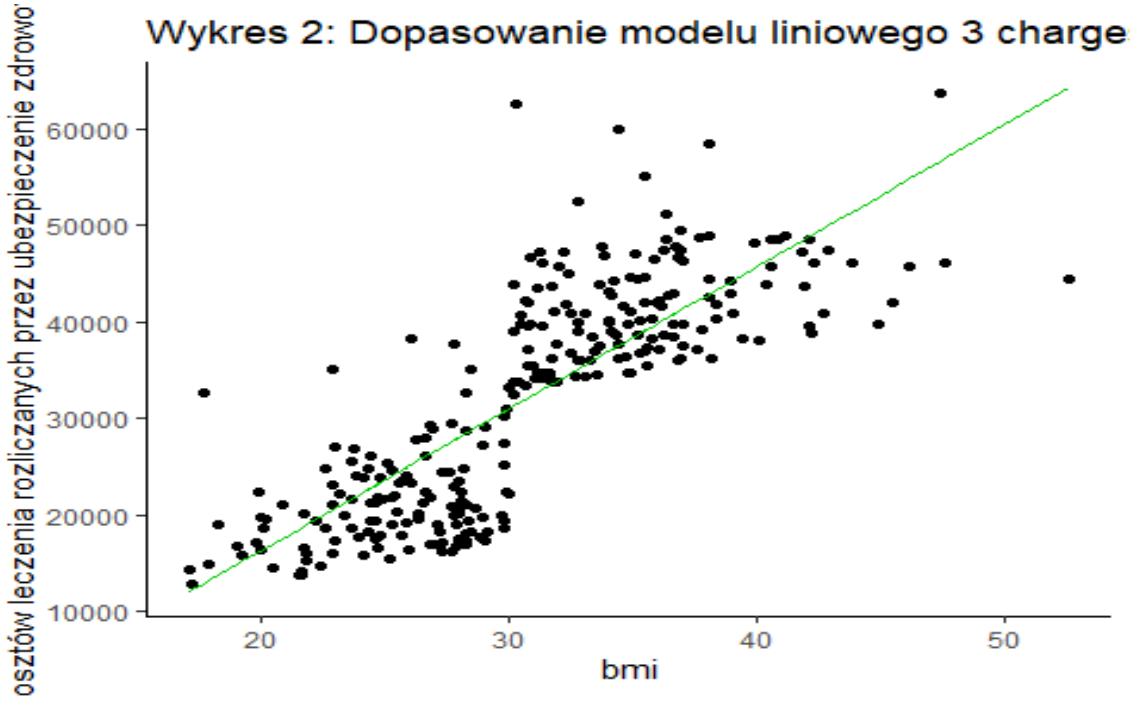
- Odchylenie standardowe reszt: Se = 11400
- Współczynnik determinacji $R^2$ = 0.03554
- $R^2$ skorygowany = 0.02482

**Wniosek:** Model 5 wyjaśnia 3,55% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Model liniowy m2 oraz m4 wyjaśniają bardzo mały procent kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne, natomiast modele m1, m5, a zwłaszcza m3 wyjaśniają w większym stopniu badane zjawisko.

**Ponieważ modele 1 i 3 mają tylko jedną zmienną objaśniającą można pokazać je na wykresie:**



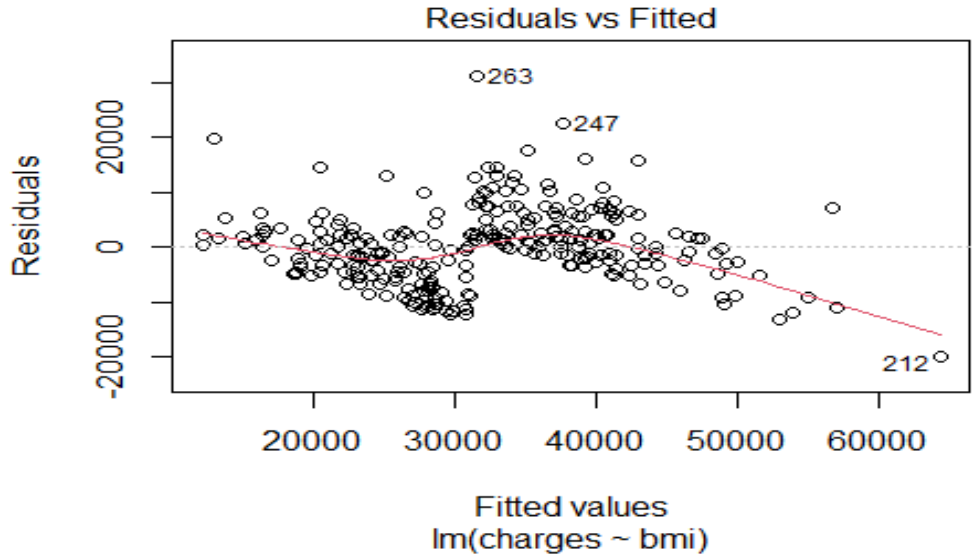




**Wniosek:** Analizując powyższy wykres 1 można zauważyć, że dopasowanie modelu nie jest dobre. Jednostki zdecydowanie odstają od linii prostej. Analizując wykres 2 można zauważyć, że jednostki są zdecydowanie bardziej zbliżone do linii prostej i dopasowanie modelu jest lepsze niż w przypadku modelu 1. Główną zmienną objaśniającą w budowaniu modelu kształtującego koszty leczenia rozliczane przez ubezpieczenie zdrowotne będzie zmienna *bmi*.

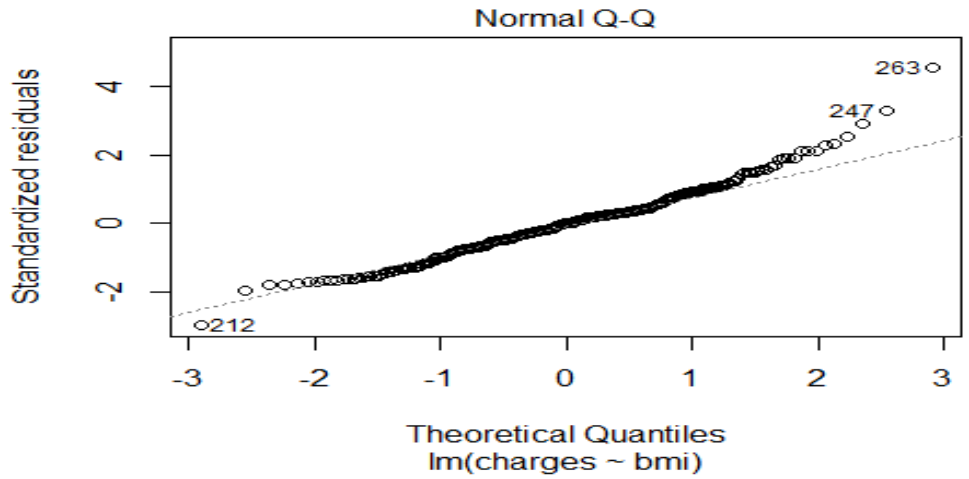
**W celu lepszego zbadania modelu 3 wykonano dla niego wykresy diagnostyczne:**

```
plot(m3, which = 1:3)
```

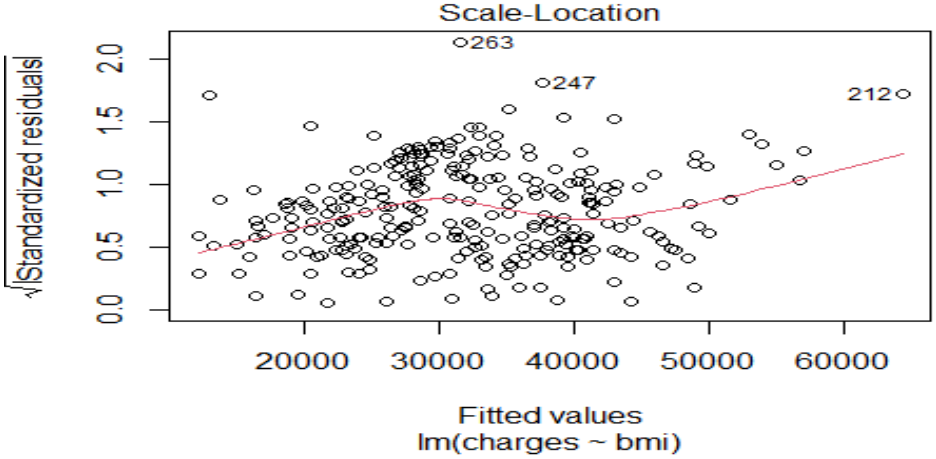
*Wykres 3*




*Wykres 4*




*Wykres 5*



**Wniosek:** Na podstawie wykresów 3, 4 i 5 można stwierdzić, że wariancja (rozrzut) reszt nie jest równomierne rozmieszczony wzdłuż linii poziomej, na poziomie = 0. Rozkład reszt nie jest normalny, można zauważyć obserwacje nietypowe o dużych resztach. Wartość średnia reszt jest bliska 0, co wskazuje właściwą postać funkcyjną modelu, choć występuje kilka jednostek odstających.

Interpretacja testów istotności parametrów **modelu 1** (test F i test t)

**test F:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- F= 42.67
- p-value: 3.181e-10
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 6.53
- p-value: 3.18e-10
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wiek statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Interpretacja testów istotności parametrów **modelu 3** (test F i test t)

**test F:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- F= 506.1
- p-value: 5.93e-10
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 22.50
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Bmi statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Interpretacja testów istotności parametrów **modelu 5** (test F i test t)

**test F:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- F= 3.316
- p-value: 0.02046
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 21.311
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Region statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

## Model liniowy 6 ze zmiennymi *age* i *bmi*

*Tabela 7*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku i bmi beneficjenta palącego
m6 <- lm(charges  ~ age + bmi , data = dane1)
summary(m6)
```

```
## Call:
## lm(formula = charges ~ age + bmi, data = dane1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14604.4  -4315.1   -240.5   3638.0  29316.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22367.45    1931.86  -11.58   <2e-16 ***
## age            266.29      25.06   10.63   <2e-16 ***
## bmi           1438.09      55.22   26.05   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 5754 on 271 degrees of freedom
## Multiple R-squared:  0.7532, Adjusted R-squared:  0.7514
## F-statistic: 413.6 on 2 and 271 DF,  p-value: < 2.2e-16
```

**Postać modelu:** charges = -22367.45 + 266.29 * `age` + 1438.09 * *`bmi`*

Miary dopasowania z tabeli 7:

- Odchylenie standardowe reszt: Se = 5754
- Współczynnik determinacji $R^2$ = 0.7532
- $R^2$ skorygowany = 0.7514

Model wyjaśnia 75,54% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Interpretacja testów istotności parametrów **modelu 6** (test F i test t)

**test F:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- F= 413.6
- p-value: < 2.2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t:**

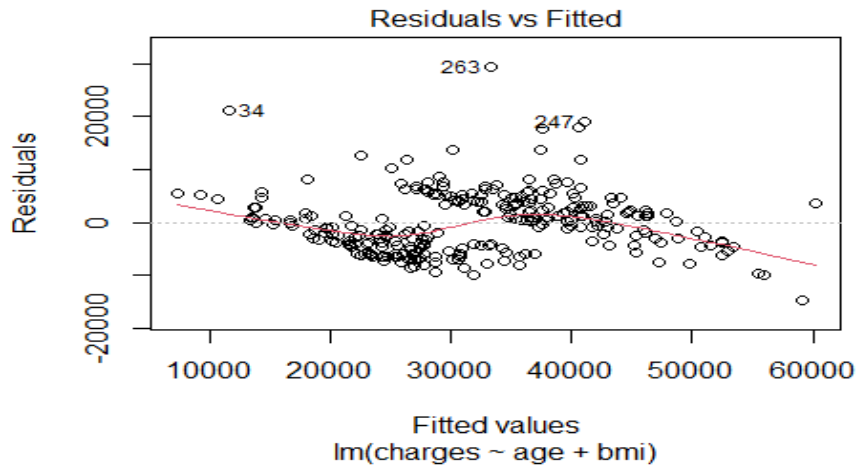
- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -11.58
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wszystkie parametry modelu istotnie różnią się od 0 na wszystkich poziomach istotności.

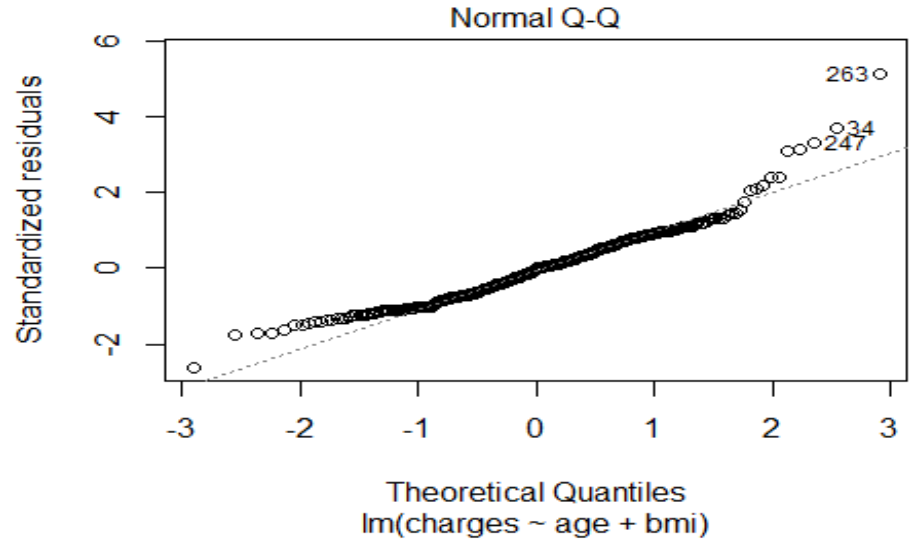
**W celu lepszego zbadania modelu 6 wykonano dla niego wykresy diagnostyczne oraz obserwację jednostek wpływowych**

```
plot(m6, which = 1:5)
```

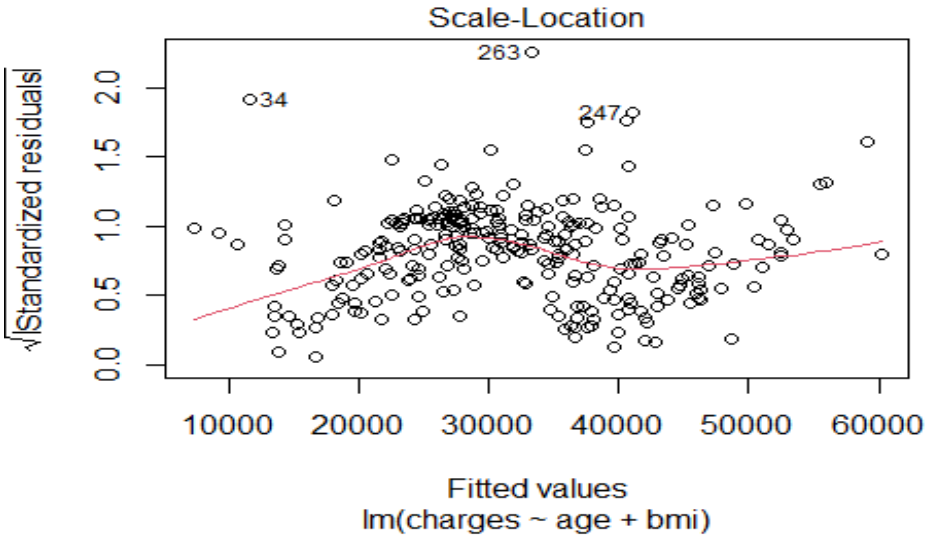
*Wykres 6*



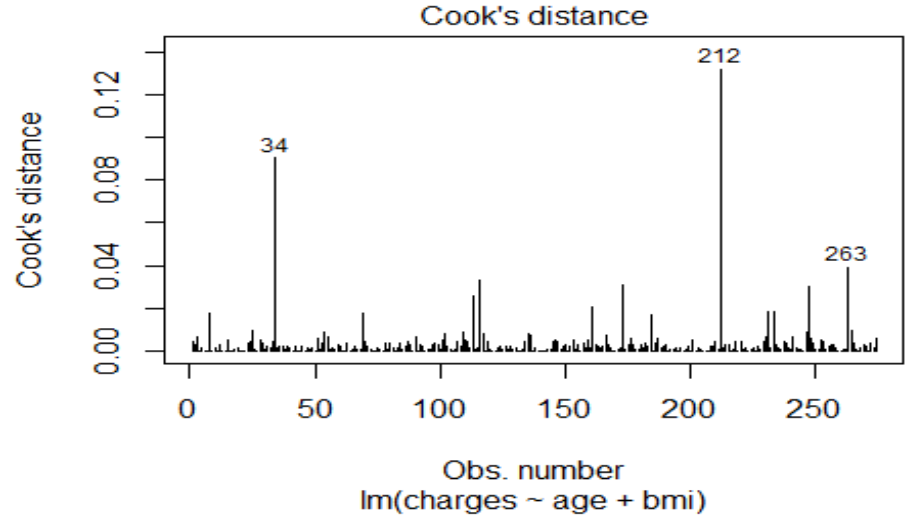
*Wykres 7*



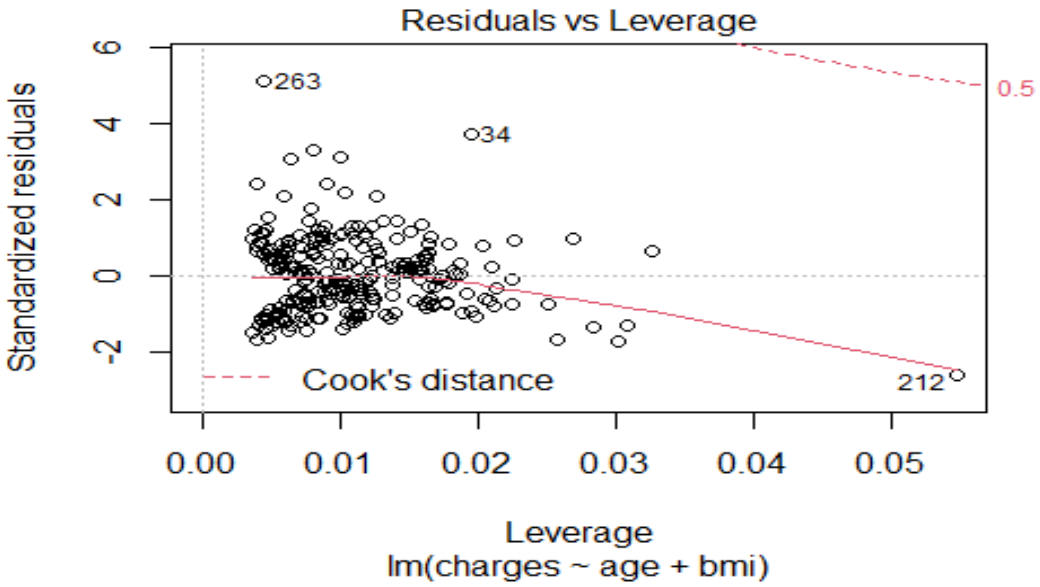

*Wykres 8*




*Wykres 9*



*Wykres 10*




**Wniosek:** Dzięki wykresom 6-10 można stwierdzić, że wariancja (rozrzut) reszt nie jest równomierne rozmieszczony wzdłuż linii poziomej, na poziomie=0. Rozkład reszt nie jest normalny, można zauważyć obserwacje nietypowe o dużych resztach. Wartość średnia reszt jest bliska 0, co wskazuje właściwą postać funkcyjną modelu, choć występuje kilka jednostek odstających. Za pomocą miar: odległość Cooka oraz wskaźnik wpływu leverage (dźwignia) dokonano oceny wpływu poszczególnych obserwacji na parametry strukturalne modelu 6, które zostały kolejno usunięte i zapisane jako *dane2*.

## Usunięcie jednostek odstających z modelu 6

```
dane2<-dane1[-c(34,212,263,113,116,173,247,233,69,161,184,231,8,54,13
5,3,90,264,25,159,246,109,187,229,87,218,248),]
```

**Następnie wykonano testy statystyczne dla modelu 6**

### Test normalności Shapiro-Wilka dla reszt modelu 6

*Tabela 8*

```
shapiro.test(m6$residuals)

##  Shapiro-Wilk normality test
##
## data:  m6$residuals
## W = 0.9454, p-value = 1.45e-08
```

**Wniosek:**

Ponieważ p-value < 0.01 (tabela 8) to odrzucamy $H_0$ na korzyść $H_1$. Reszty modelu nie mają rozkładu normalnego.

## Test Breuscha-Pagana jednorodności wariancji reszt modelu 6

*Tabela 9*

```
bptest(m6)
```

```
##  studentized Breusch-Pagan test
##
## data:  m6
## BP = 0.43093, df = 2, p-value = 0.8062
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 9) to nie ma podstaw do odrzucenia $H_0$ o jednorodności wariancji reszt.

## Test Durbina-Watsona niezależności reszt modelu 6

*Tabela 10*

```
dwtest(m6, order.by = ~age, data = dane1)
```

```
##  Durbin-Watson test
##
## data:  m6
## DW = 2.0887, p-value = 0.7514
## alternative hypothesis: true autocorrelation is greater than 0
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 10) to nie ma podstaw do odrzucenia $H_0$, mówiącej o niezależności reszt.

## Test rainbow na liniowość modelu modelu 6

*Tabela 11*

```
raintest(m6)
```

```
##  Rainbow test
##
## data:  m6
## Rain = 1.2276, df1 = 137, df2 = 134, p-value = 0.1172
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 11) to nie ma podstaw do odrzucenia $H_0$, mówiącej o liniowości modelu.

## Model liniowy 7, który jest modelem 6 po usunięciu jednostek odstających (ze zmiennymi *age* i *bmi*)

*Tabela 12*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku i bmi beneficjenta palącego po wyeliminowaniu jednostek
odsających
m7 <- lm(charges  ~ age + bmi , data = dane2)
summary(m7)
```

```
## Call:
## lm(formula = charges ~ age + bmi, data = dane2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9136.4 -3299.5  -155.9  3156.9 14260.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28320.58    1683.83  -16.82   <2e-16 ***
## age            242.90      20.16   12.05   <2e-16 ***
## bmi           1646.64      50.57   32.56   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
##
## Residual standard error: 4374 on 244 degrees of freedom
## Multiple R-squared:  0.8407, Adjusted R-squared:  0.8394
## F-statistic: 643.9 on 2 and 244 DF,  p-value: < 2.2e-16
```

**Wnioski:**

**Postać modelu:** charges = -28320.58 + 242.90 * age *+ 1646.64 * bmi*

**Interpretacja modelu 7 na podstawie tabeli 12:**

Wyraz wolny $\beta_0$ = -28320.58 w tym modelu nie ma interpretacji, ponieważ badanie nie obejmowało osób w wieku 0 lat (niemowląt).

Współczynnik przy zmiennej “age” $\beta_1$ = 242.90:

- Jeżeli wiek beneficjenta palącego wzrośnie o 1 rok, to koszty leczenia rozliczane przez ubezpieczenie zdrowotne wzrosną średnio o 242.90 $ (ceteris paribus).
- Jeżeli wiek beneficjenta palącego wzrośnie o 10 lat, to koszty leczenia rozliczane przez ubezpieczenie zdrowotne wzrosną średnio o 2429.0 $ (ceteris paribus).

Współczynnik przy zmiennej “bmi” $\beta_2$ = 1646.64:

- Jeżeli bmi beneficjenta palącego wzrośnie o 1 jednostkę miary, to koszty leczenia rozliczane przez ubezpieczenie zdrowotne wzrosną średnio o 1646.64 $ (ceteris paribus).

Miary dopasowania z tabeli 12:

- Odchylenie standardowe reszt: Se = 4374
- Współczynnik determinacji $R^2$ = 0.8407
- $R^2$ skorygowany= 0.8394

Model wyjaśnia 84,07% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Interpretacja testów istotności parametrów **modelu 7** (test F i test t)

**test F:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- F= 643.9
- p-value: < 2.2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t:**

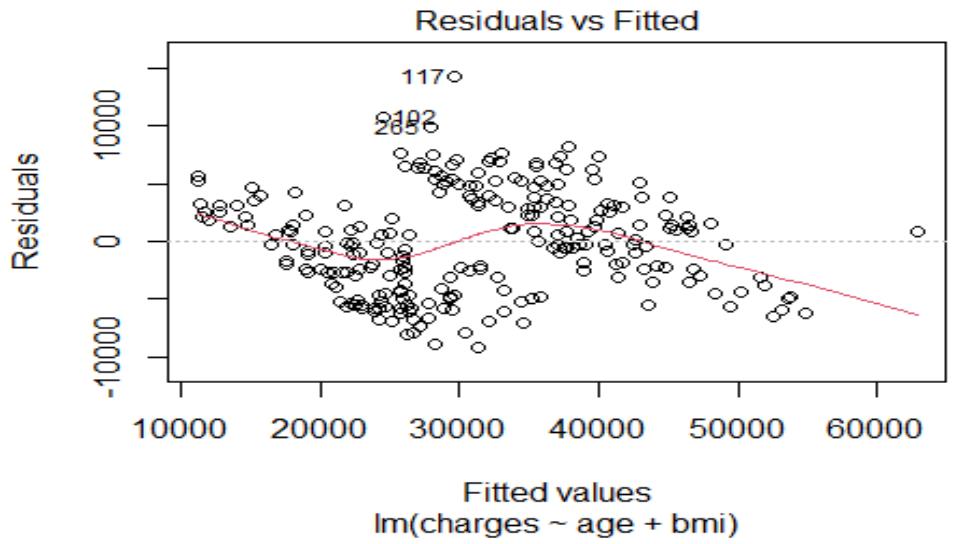
- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 643.9
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wszystkie parametry modelu istotnie różnią się od 0 na wszystkich poziomach istotności.

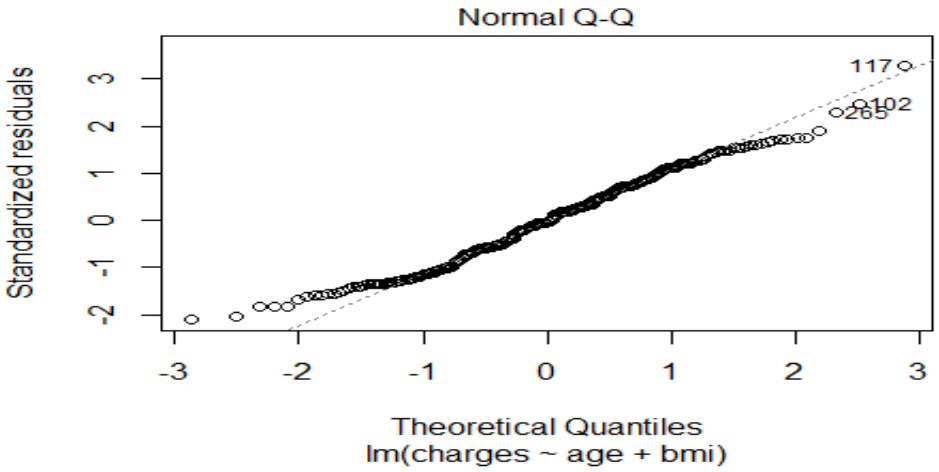
**Wykresy diagnostyczne modelu liniowego 7:**

```
plot(m7, which = 1:3)
```

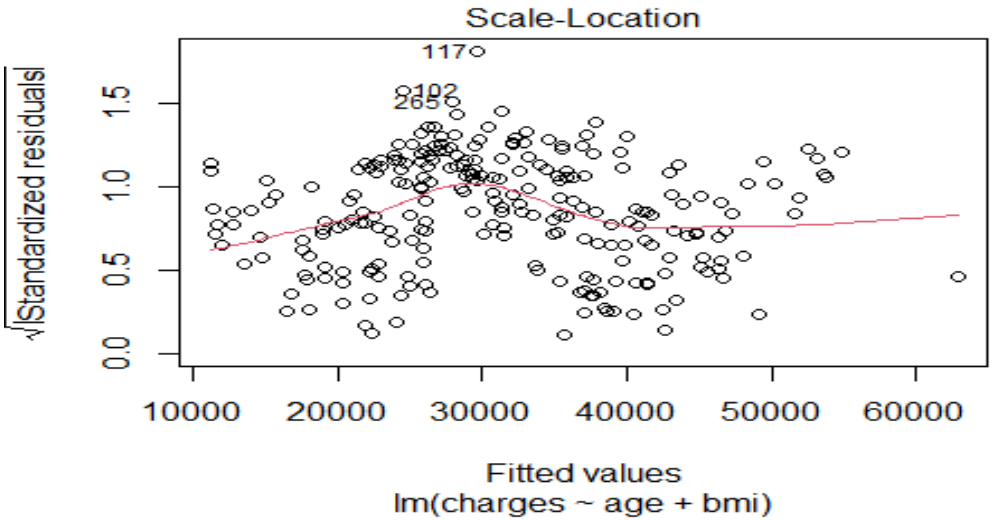
*Wykres 11*



*Wykres 12*




*Wykres 13*




**Wniosek:** Na podstawie wykresów 11,12 i 13 można stwierdzić, że wariancja (rozrzut) reszt nie jest równomierne rozmieszczony wzdłuż linii poziomej, na poziomie=0. Rozkład reszt nie jest normalny, można zauważyć obserwacje nietypowe o dużych resztach. Wartość średnia reszt jest bliska 0, co wskazuje właściwą postać funkcyjną modelu, choć występuje kilka jednostek odstających.

## Testy statystyczne dla modelu 7

### Test normalności Shapiro-Wilka dla reszt modelu 7

*Tabela 13*

```
shapiro.test(m7$residuals)

##  Shapiro-Wilk normality test
##
## data:  m7$residuals
## W = 0.98285, p-value = 0.004479
```

**Wniosek:**

Ponieważ p-value < 0.01 (tabela 13) to odrzucamy $H_0$ na korzyść $H_1$. Reszty modelu nie mają rozkładu normalnego.

### Test Breuscha-Pagana jednorodności wariancji reszt modelu 7

*Tabela 14*

```
bptest(m7)
```

```
##  studentized Breusch-Pagan test
##
## data:  m7
## BP = 2.6391, df = 2, p-value = 0.2673
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 14) to nie ma podstaw do odrzucenia $H_0$ o jednorodności wariancji reszt.

### Test Durbina-Watsona niezależności reszt modelu 7

*Tabela 15*

```
dwtest(m7, order.by = ~age, data = dane2)
```

```
##  Durbin-Watson test
##
## data:  m7
## DW = 1.9833, p-value = 0.4244
## alternative hypothesis: true autocorrelation is greater than 0
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 15) to nie ma podstaw do odrzucenia $H_0$, mówiącej o niezależności reszt.

### Test rainbow na liniowość modelu modelu 7

*Tabela 16*

```
raintest(m7)
```

```
##  Rainbow test
##
## data:  m7
## Rain = 0.8157, df1 = 124, df2 = 120, p-value = 0.8693
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 16) to nie ma podstaw do odrzucenia $H_0$, mówiącej o liniowości modelu.

## Model liniowy 8, który jest modelem 6 (ze zmiennymi *age* i *bmi*) oraz z dodatkową zmienną *region*

*Tabela 17*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku, bmi i regionu zamieszania beneficjenta palącego
m8 <- lm(charges  ~ age + bmi + region , data = dane1)
summary(m8)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + region, data = dane1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14237.0  -4232.1    -67.8   3422.6  30358.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -22627.67    1988.80 -11.378   <2e-16 ***
## age                268.83      25.10  10.709   <2e-16 ***
## bmi               1471.07      57.68  25.503   <2e-16 ***
## regionnorthwest   -620.40    1029.98  -0.602   0.5475
## regionsoutheast  -1897.24     960.16  -1.976   0.0492 *
## regionsouthwest   -419.77    1040.48  -0.403   0.6869
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 5738 on 268 degrees of freedom
## Multiple R-squared:  0.7574, Adjusted R-squared:  0.7528
## F-statistic: 167.3 on 5 and 268 DF,  p-value: < 2.2e-16
```

**Wnioski:**

**Postać modelu:** charges = - 22627.67 + 268.83 * age + *1471.07* * *bmi* - 620.40 * regionnorthwest - *1897.24* * *regionsoutheast* - 419.77 * *regionsouthwest*

Miary dopasowania z tabeli 17:

- Odchylenie standardowe reszt: Se = 5738
- Współczynnik determinacji $R^2 = 0.7574$
- $R^2$ skorygowany = 0.7528

Model wyjaśnia 75,74% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Zmienna *region* samodzielnie wyjaśniała zaledwie 3,55% natomiast w porównaniu z innymi zmiennymi nie wnosi dużo do modelu.

Interpretacja testów istotności parametrów **modelu 8** (test F i test t)

**test F:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- F= 167.3
- p-value: < 2.2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t:**

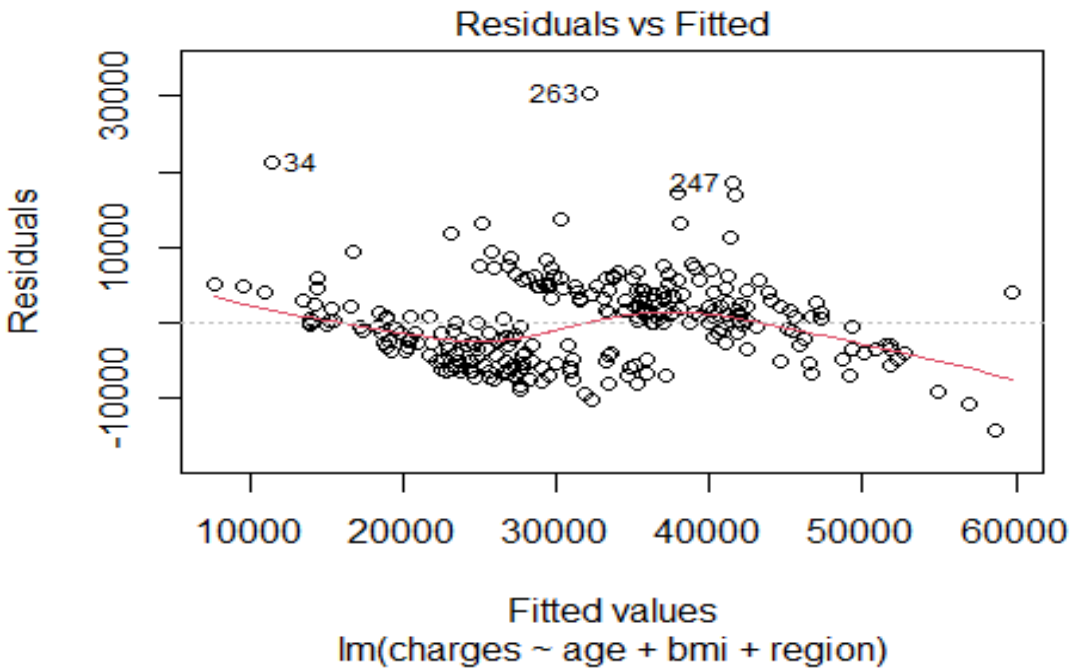
- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -11.378
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wszystkie parametry modelu istotnie różnią się od 0 na wszystkich poziomach istotności, poza parametrami dotyczącymi regionu gdzie tylko `regionsoutheast` jest istotny od 0 na poziomie istotności 0.01 i 0.05, a `regionnorthwest` i `regionsouthwest` są istotne na poziomie istotności 0.1 i 1.

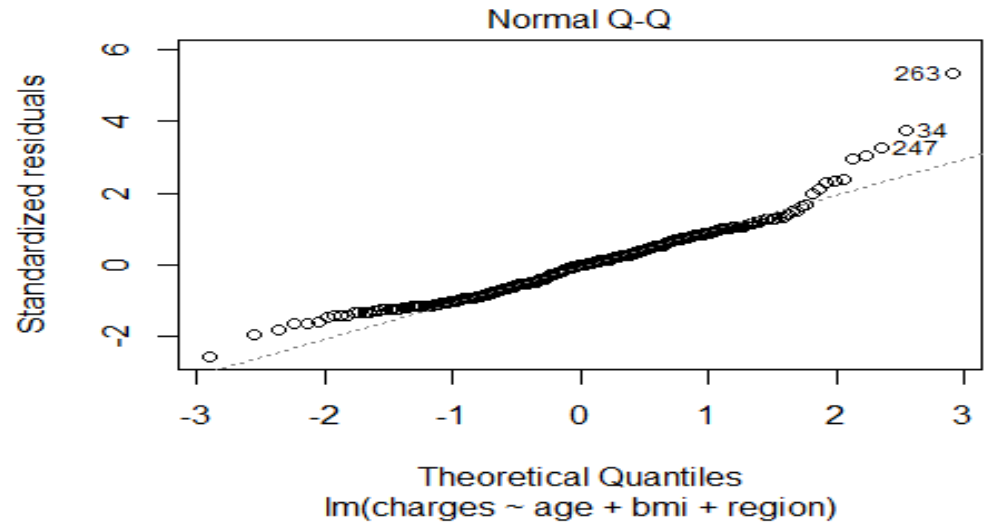
**Wykresy diagnostyczne modelu liniowego 8 oraz obserwacja jednostek wpływowych:**

```
plot(m8, which = 1:5)
```

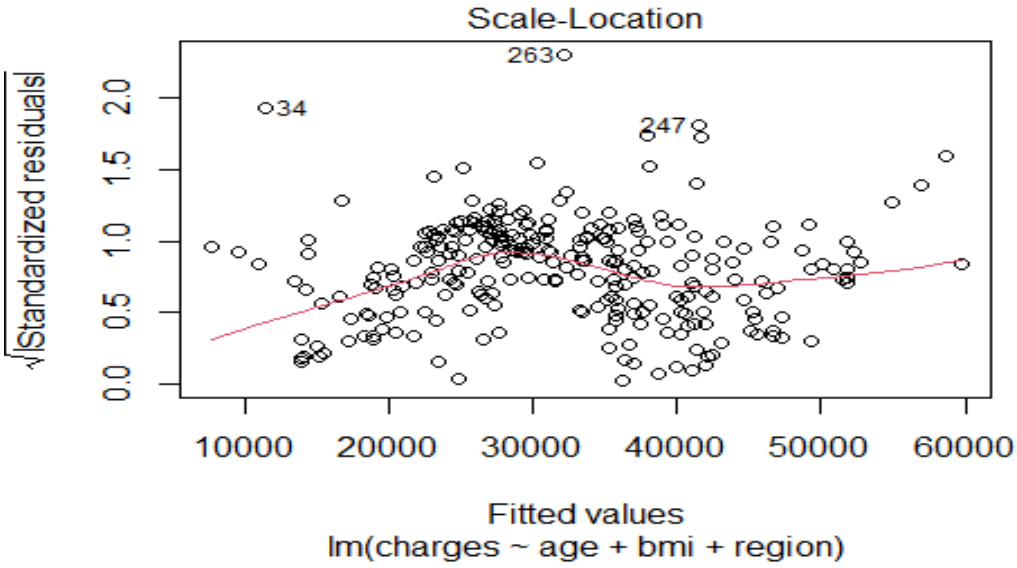
*Wykres 14*



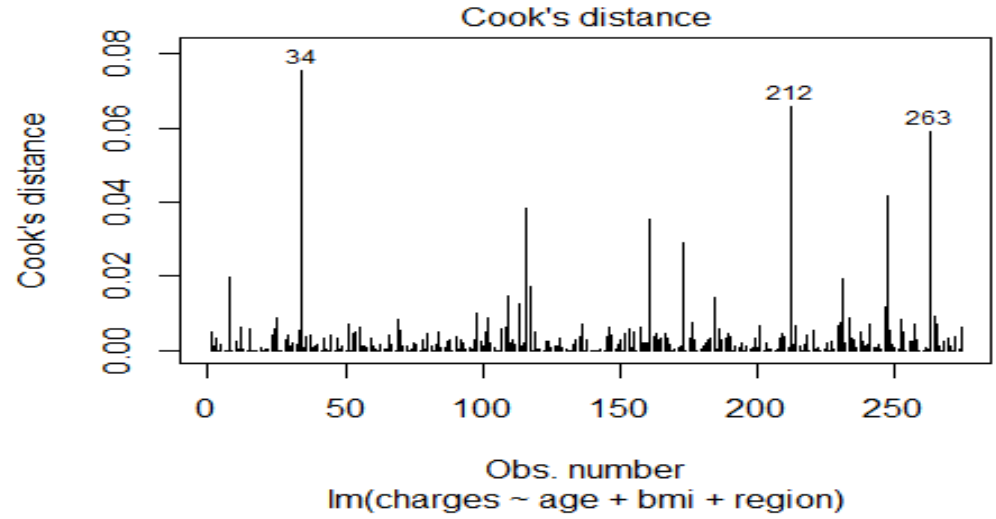
*Wykres 15*



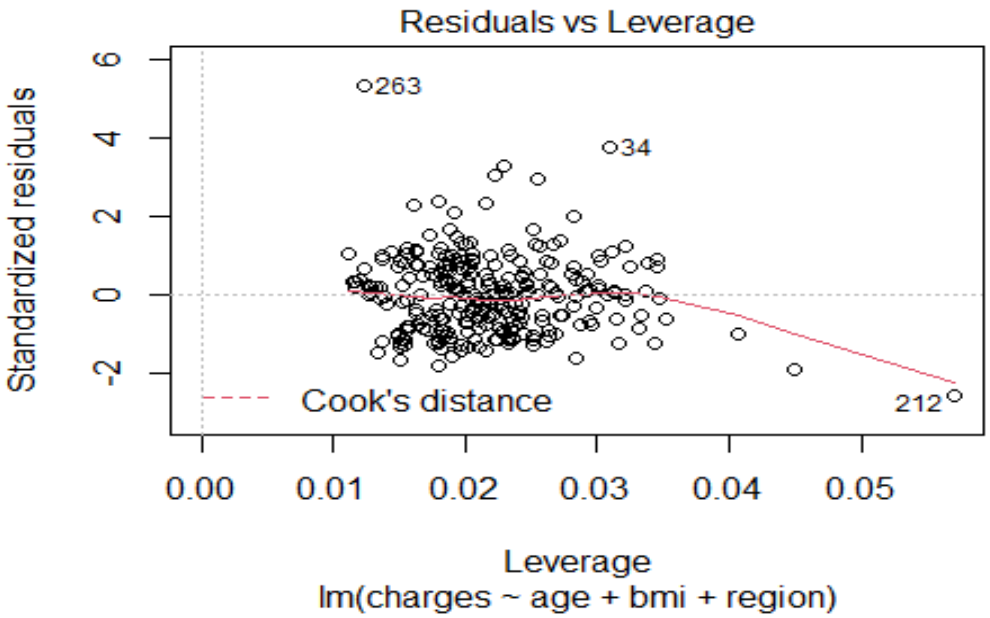

*Wykres 16*




*Wykres 17*



*Wykres 18*




**Wniosek:** Za pomocą wykresów 14-18 można zauważyć, że wariancja (rozrzut) reszt nie jest równomierne rozmieszczony wzdłuż linii poziomej, na poziomie=0. Rozkład reszt nie jest normalny, można zauważyć obserwacje nietypowe o dużych resztach. Wartość średnia reszt jest bliska 0, co wskazuje właściwą postać funkcyjną modelu, choć występuje kilka jednostek odstających. Za pomocą miar: odległość Cooka oraz wskaźnik wpływu leverage (dźwignia) dokonano oceny wpływu poszczególnych obserwacji na parametry strukturalne modelu 8, które zostały kolejno usunięte i zapisane jako *dane3*.

## Usunięcie jednostek odstających z modelu 8

```
dane3<-dane1[-c(34,212,263,247,173,116,161,117,231,8),]
```

### Testy statystyczne dla modelu 8

#### Test normalności Shapiro-Wilka dla reszt modelu 8

*Tabela 18*

```
shapiro.test(m8$residuals)
```

```
##  Shapiro-Wilk normality test
##
## data:  m8$residuals
## W = 0.94519, p-value = 1.38e-08
```

**Wniosek:**

Ponieważ p-value < 0.01 (tabela 18) to odrzucamy $H_0$ na korzyść $H_1$. Reszty modelu nie mają rozkładu normalnego.

### Test Breuscha-Pagana jednorodności wariancji reszt modelu 8

*Tabela 19*

```
bptest(m8)
```

```
##  studentized Breusch-Pagan test
##
## data:  m8
## BP = 2.9181, df = 5, p-value = 0.7126
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 19) to nie ma podstaw do odrzucenia $H_0$ o jednorodności wariancji reszt.

### Test Durbina-Watsona niezależności reszt modelu 8

*Tabela 20*

```
dwtest(m8, order.by = ~age, data = dane1)
```

```
##  Durbin-Watson test
##
## data:  m8
## DW = 2.0955, p-value = 0.7702
## alternative hypothesis: true autocorrelation is greater than 0
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 20) to nie ma podstaw do odrzucenia $H_0$, mówiącej o niezależności reszt.

### Test rainbow na liniowość modelu modelu 8

Tabela 21

```
raintest(m8)
```

```
##  Rainbow test
##
## data:  m8
## Rain = 1.357, df1 = 137, df2 = 131, p-value = 0.03951
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 21) to nie ma podstaw do odrzucenia $H_0$, mówiącej o liniowości modelu.

## Model liniowy 9, który jest modelem 8 po usunięciu jednostek odstających (ze zmiennymi *age*, *bmi* i *region*)

*Tabela 22*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku, bmi i regionu zamieszkania beneficjenta palącego po wy
eliminowaniu jednostek odstających
m9 <- lm(charges  ~ age + bmi +region , data = dane3)
summary(m9)
```

```
## Call:
## lm(formula = charges ~ age + bmi + region, data = dane3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9917.6 -3943.1   159.7  3402.5 13559.3
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -23562.68    1665.23 -14.150   <2e-16 ***
## age                261.28      20.68  12.636   <2e-16 ***
## bmi               1505.19      49.47  30.426   <2e-16 ***
## regionnorthwest  -1689.41     856.23  -1.973   0.0496 *
## regionsoutheast  -1965.03     789.04  -2.490   0.0134 *
## regionsouthwest   -505.08     859.74  -0.587   0.5574
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 4661 on 258 degrees of freedom
## Multiple R-squared:  0.8248, Adjusted R-squared:  0.8214
## F-statistic:   243 on 5 and 258 DF,  p-value: < 2.2e-16
```

**Wnioski:**

**Postać modelu:** charges = -23562.68 + 261.28 * age *+ 1505.19* * *bmi* - 1689.41 * regionnorthwest - 1965.*034* * *regionsoutheast* - 505.08 * *regionsouthwest*

**Interpretacja modelu 9 na podstawie tabeli 22**

Wyraz wolny $\beta_0$ = -23562.68 w tym modelu nie ma interpretacji, ponieważ badanie nie obejmowało osób w wieku 0 lat (niemowląt).

Współczynnik przy zmiennej “age” $\beta_1$ = 261.28:

- Jeżeli wiek beneficjenta palącego wzrośnie o 1 rok, to koszty leczenia rozliczane przez ubezpieczenie zdrowotne wzrosną średnio o 261.28 $ (ceteris paribus).
- Jeżeli wiek beneficjenta palącego wzrośnie o 10 lat, to koszty leczenia rozliczane przez ubezpieczenie zdrowotne wzrosną średnio o 2612.8 $ (ceteris paribus).

Współczynnik przy zmiennej “bmi” $\beta_2$ = 1505.19:

- Jeżeli bmi beneficjenta palącego wzrośnie o 1 jednostkę miary, to koszty leczenia rozliczane przez ubezpieczenie zdrowotne wzrosną średnio o 1505.19 $ (ceteris paribus).

Współczynnik przy zmiennej “regionnorthwest” $\beta_3$ = -1689.41:

- Beneficjenci palący mieszkający w rejonie northwest mają koszty leczenia rozliczane przez ubezpieczenie zdrowotne niższe średnio o 1689.41 $ (ceteris paribus) w stosunku do tych zamieszkujących region northeast.

Współczynnik przy zmiennej “regionsoutheast” $\beta_4$ = -1965.03:

- Beneficjenci palący mieszkający w rejonie southeast mają koszty leczenia rozliczane przez ubezpieczenie zdrowotne niższse średnio o 1965.03 $ (ceteris paribus) w stosunku do tych zamieszkujących region northeast.

Współczynnik przy zmiennej “regionsoutheast” $\beta_5$ = -505.08:

- Beneficjenci palący mieszkający w rejonie southeast mają koszty leczenia rozliczane przez ubezpieczenie zdrowotne niższe średnio o 505.08 $ (ceteris paribus) w stosunku do tych zamieszkujących region northeast.

Miary dopasowania z tabeli 22:

- Odchylenie standardowe reszt: Se= 4661
- Współczynnik determinacji $R^2$ = 0.8248
- $R^2$ skorygowany= 0.8214

Model wyjaśnia 82,48% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Interpretacja testów istotności parametrów **modelu 9** (test F i test t)

**test F:**

- $H_0$= $\beta i = 0$
- $H_1$= $\beta i \neq 0$
- F= 243
- p-value: < 2.2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t:**

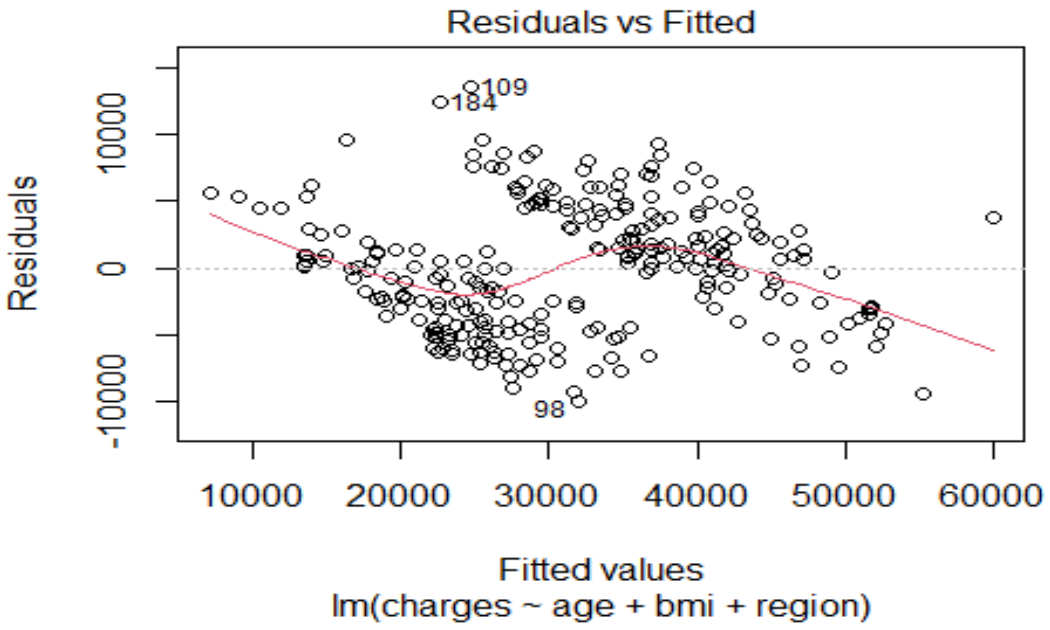
- $H_0$= $\beta i = 0$
- $H_1$= $\beta i \neq 0$
- t= -14.15
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wszystkie parametry modelu istotnie różnią się od 0 na wszystkich poziomach istotności, poza parametrami dotyczącymi regionu gdzie `regionsoutheast` i `regionnorthwest` są istotne od 0 na poziomie istotności 0.01 i 0.05, a`regionsouthwest` jest istoty na poziomie istotności 0.1 i 1.

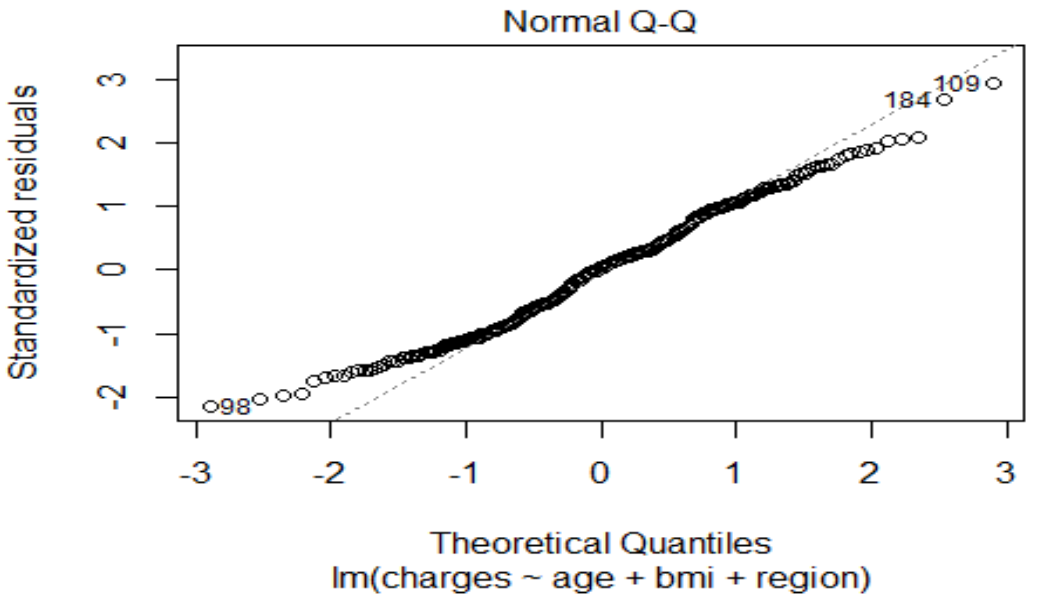
**Wykresy diagnostyczne modelu liniowego 9:**

```
plot(m9, which = 1:3)
```

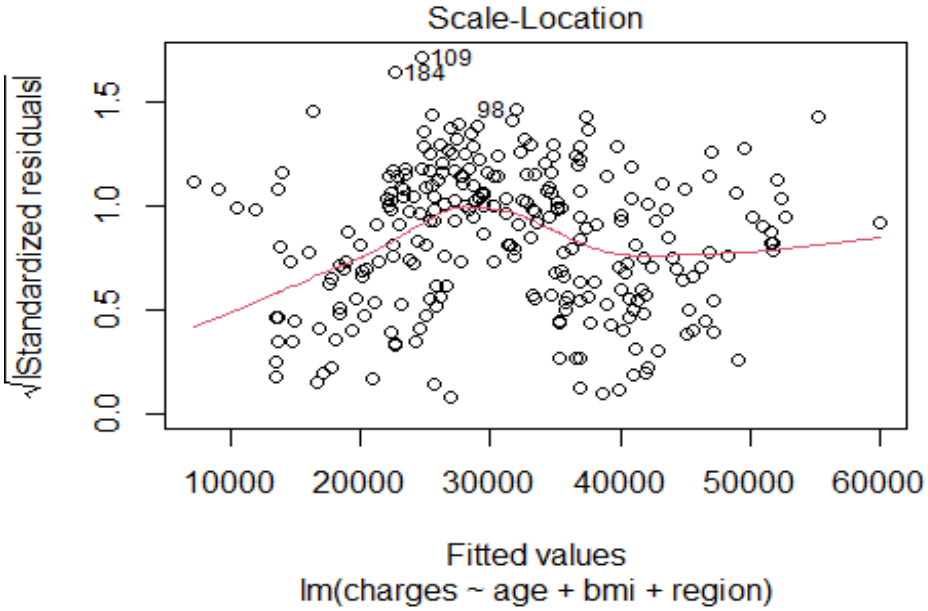
*Wykres 19*




*Wykres 20*



*Wykres 21*




**Wniosek:** Na podstawie wykresów 19, 20 i 21 można stwierdzić, że wariancja (rozrzut) reszt nie jest równomierne rozmieszczony wzdłuż linii poziomej, na poziomie=0. Rozkład reszt jest normalny. Wartość średnia reszt jest bliska 0, co wskazuje właściwą postać funkcyjną modelu, choć występuje kilka jednostek odstających.

## Testy statystyczne dla modelu 9

### Test normalności Shapiro-Wilka dla reszt modelu 9

*Tabela 23*

```
shapiro.test(m9$residuals)
```

```
##  Shapiro-Wilk normality test
##
## data:  m9$residuals
## W = 0.9861, p-value = 0.01165
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 23) to nie ma podstaw do odrzucenia $H_0$ na rzecz $H_1$. Reszty modelu mają rozkład normalny.

### Test Breuscha-Pagana jednorodności wariancji reszt modelu 9

*Tabela 24*

```
bptest(m9)
```

```
##  studentized Breusch-Pagan test
##
## data:  m9
## BP = 1.8537, df = 5, p-value = 0.869
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 24) to nie ma podstaw do odrzucenia $H_0$ o jednorodności wariancji reszt.

### Test Durbina-Watsona niezależności reszt modelu 9

*Tabela 25*

```
dwtest(m9, order.by = ~age, data = dane3)
```

```
##  Durbin-Watson test
##
## data:  m9
## DW = 1.9782, p-value = 0.4071
## alternative hypothesis: true autocorrelation is greater than 0
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 25) to nie ma podstaw do odrzucenia $H_0$, mówiącej o niezależności reszt.

### Test rainbow na liniowość modelu modelu 9

*Tabela 26*

```
raintest(m9)
```

```
##  Rainbow test
##
## data:  m9
## Rain = 0.95654, df1 = 132, df2 = 126, p-value = 0.5999
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 26) to nie ma podstaw do odrzucenia $H_0$, mówiącej o liniowości modelu.

## Sprawdzenie założeń modelów linowych 7 i 9

Sprawdzenie, czy zmienne objaśniające nie są współliniowe. Miara współliniowości zmiennych objaśniających VIF - czynnik rozdęcia wariancji.

*Tabela 27*

```
vif(m7)
```

```
##      age      bmi
## 1.008423 1.008423
```

```
vif(m9)
```

```
##            GVIF Df GVIF^(1/(2*Df))
## age    1.017466  1        1.008695
## bmi    1.107292  1        1.052280
## region 1.113680  3        1.018107
```

**Wniosek:** Ponieważ wszystkie wartości czynnika vif w tabeli 27 w obu modelach mają niskie wartości tzn. nie przekraczają wartości = 5.0, nie ma podejrzeń, że zmienne objaśniające są współliniowe.

## Podsumowanie:

W pierwszej kolejności stworzono ***modele (1-5)*** objaśniające koszty leczenia rozliczane przez ubezpieczenie zdrowotne dla beneficjentów palących, w zależności od każdej ze zmiennych z osobna (*age, sex, bmi, childrem i region*). Ponieważ w największym stopniu badane zjawisko wyjaśnia zmienna *bmi* wybrano ją jako główną zmienną objaśniającą w procesie budowania modelu. W związku z tym powstał ***model 6***, gdzie do zmiennej objaśniającej *bmi* dodano zmienną *age*. Ponieważ reszty modelu nie miały rozkładu normalnego oraz odnotowano wiele jednostek odstających w modelu stworzono ***model 7***, który opierał się na tych samych zmiennych objaśniających, ale na danych pozbawionych jednostek odstających. W *modelu 7* reszty nie mają rozkładu normalnego, spełniają resztę założeń statystycznych modelu, a model wyjaśnia 84,07% kształtowania kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów palących. Po dodaniu do zmiennych *bmi* i *age* zmiennej *region* powstał ***model 8*****,** którego reszty również nie miały rozkładu normalnego, ale spełniały resztę założeń statystycznych. Po usunięciu jednostek odstających z *modelu 8* powstał ***model 9***. Dopasowanie tego modelu poprawiło się - reszty modelu mają rozkład normalny, a co za tym idzie model spełnia wszystkie założenia statystyczne, ale lekko zwiększyło się odchylenie standardowe reszt. *Model 9* wyjaśnia w 82,48% kształtowanie kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów palących. Pomimo, że zmienna *region* nie wnosi dużo do modelu to poprawiła wyniki testów statystycznych. Można jednak poszukać lepszych postaci modelu.

### 2.2. Estymacja, a kolejno weryfikacja otrzymanych modeli klasy GLM, objaśniających koszty leczenia rozliczane przez ubezpieczenie zdrowotne dla beneficjentów palących.

## Wybór postaci modelu GLM dla zmiennej Y=charges

Estymujemy modele dla zmiennej Y o rozkładzie normalnym *family = gaussian* z różnymi funkcjami wiążącymi.

## Funkcja wiążąca identycznościowa

Estymacja **modelu 9** GLM z funkcją wiążącą identycznościową (model liniowy) - estymacja metodą MNW.

W poniższej formule wartość domyślna argumentu *link = identity.*

**Estymacja modelu 10 – Klasy GLM -> (analogiczny do modelu liniowego 9)**

*Tabela 28*

```
dane3<-dane1[-c(34,212,263,247,173,116,161,117,231,8),]

#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku, bmi i regionu zamieszkania beneficjenta palącego po wy
eliminowaniu jednostek odstających
m10 <- glm(charges  ~ age + bmi +region , data = dane3, family = gaus
sian)
summary(m10)
```

```
## Call:
## glm(formula = charges ~ age + bmi + region, family = gaussian,
##     data = dane3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9917.6  -3943.1    159.7   3402.5  13559.3
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -23562.68    1665.23 -14.150   <2e-16 ***
## age                261.28      20.68  12.636   <2e-16 ***
## bmi               1505.19      49.47  30.426   <2e-16 ***
## regionnorthwest  -1689.41     856.23  -1.973   0.0496 *
## regionsoutheast  -1965.03     789.04  -2.490   0.0134 *
## regionsouthwest   -505.08     859.74  -0.587   0.5574
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 21727026)
##
##     Null deviance: 3.1999e+10  on 263  degrees of freedom
## Residual deviance: 5.6056e+09  on 258  degrees of freedom
## AIC: 5217.2
##
## Number of Fisher Scoring iterations: 2
```

**Na podstawie tabeli 28 można dokonać poniższych interpretacji.**

Minimalną wartością tego modelu jest -9917.6, a największą 13559.3. Wartość środkowa wynosi 159.7. Kwartyl pierwszy 25% obserwacji położonych jest poniżej -3943.1, a 75% obserwacji jest położonych powyżej tej wartości. Kwartyl trzeci 75% obserwacji położonych jest poniżej 3402.5, a 25% obserwacji położonych jest powyżej tej wartości.

**test t dla wyrazu wolnego :**

- $H_0 = \beta i = 0$
- $H_1 = \beta i \neq 0$

- t= -14.150
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t dla zmiennej age:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 12.636
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wiek statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej bmi:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 30.426
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** BMI statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej regionnorthwest:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -1.973
- p-value: 0.0496
- Wniosek: Brak podstaw do odrzucenia $H_0$

**Wniosek:** Region northwest statystycznie istotnie nie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej regionsoutheast:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -2.490
- p-value: 0.0134
- Wniosek: Brak podstaw do odrzucenia $H_0$

**Wniosek:** Region southeast statystycznie istotnie nie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej regionsouthwest:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -0.587
- p-value: 0.5574
- Wniosek: Brak podstaw do odrzucenia $H_0$

**Wniosek:** Region southwest statystycznie istotnie nie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**AIC - kryterium informacyjne Akaike : pożądane jest, aby wartość kryterium była jak najmniejsza.**

Kryterium Akaike dla tego modelu wynosi 5217.2.

**Postać modelu 10:**

charges = -23562.68+261.28 * *age* + 1505.19 * *bmi*-1689.41* *regionnorthwest*- 1965.03 * *regionsoutheast* - 505.08 * *regionsouthwest*

Ponieważ p-value < alfa = 0.01, to odrzucamy $H_0$ na korzyść $H_1$. Na podstawie testu Walda możemy stwierdzić, że zarówno wiek, bmi oraz region zamieszkania osób palących istotnie wpływają na wysokość składki ubezpieczeniowej.

Dodatkowo wykonamy test istotności wszystkich zmiennych niezależnych w modelu. Można wykonać test ilorazu wiarygodności lub test Walda. Przy czym testami lokalnymi wykonuje się test globalny - najpierw testujemy wszystkie parametry=0, jeśli nie to dopiero robimy testy lokalne.

*Tabela 29*

```
lrtest(m10)
```

```
## Likelihood ratio test
##
## Model 1: charges ~ age + bmi + region
## Model 2: charges ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -2601.6
## 2   2 -2831.5 -5 459.87  < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
waldtest(m10)
```

```
## Wald test
##
## Model 1: charges ~ age + bmi + region
## Model 2: charges ~ 1
##   Res.Df Df      F    Pr(>F)
```

```
## 1    258
## 2    263 -5 242.95 < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
```

**Wnioski na podstawie wyników z tabeli 29:**

W teście ilorazu wiarygodności p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$. W teście Walda p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$.

Odrzucamy hipotezę $H_0$ mówiącą o tym, że parametry przy zmiennych są równe 0 na rzecz $H_1$, mówiącej o tym, że istnieje parametr istotnie różny od 0.Test ten powinien zostać wykonany na początku.

Istnieje parametr beta statystycznie istotny różny od 0, czyli jest zmienna która ma wpływ na wysokość składki ubezpieczeniowej. Ten test powinien poprzedzać testy dla poszczególnych parametrów.

```
dane4<-dane1[-c(34,212,263,247,173,116,161,117,231,8,10,69,113,43,174
,183),]
```

## Estymacja modelu 11 - klasy GLM z funkcją wiążącą logarytmiczną (model log normalny)

*Tabela 30*

```
m11 <- glm(charges  ~ age + bmi +region , data = dane4, family = gaus
sian(link = “log”))
summary(m11)

##
## Call:
## glm(formula = charges ~ age + bmi + region, family = gaussian(link
= “log”),
##     data = dane4)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -9437   -4199    -809    3752   13897
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.6508756  0.0640480 135.069   <2e-16 ***
## age              0.0080495  0.0006911  11.648   <2e-16 ***
## bmi              0.0450499  0.0017403  25.886   <2e-16 ***
## regionnorthwest -0.0213275  0.0306793  -0.695    0.488
## regionsoutheast -0.0631668  0.0267582  -2.361    0.019 *
## regionsouthwest  0.0198589  0.0295741   0.671    0.503
## ---
```

```
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
##
## (Dispersion parameter for gaussian family taken to be 25429128)
##
##     Null deviance: 3.1041e+10  on 257  degrees of freedom
## Residual deviance: 6.4082e+09  on 252  degrees of freedom
## AIC: 5139.4
##
## Number of Fisher Scoring iterations: 5
```

**Na podstawie tabeli 30 można dokonać poniższych interpretacji:**

Minimalną wartością tego modelu jest -9437, a największą 13897. Wartość środkowa wynosi -809. Kwartyl pierwszy 25% obserwacji położonych jest poniżej -4199, a 75% obserwacji jest położonych powyżej tej wartości. Kwartyl trzeci 75% obserwacji położonych jest poniżej 3752, a 25% obserwacji położonych jest powyżej tej wartości.

**test t dla wyrazu wolnego :**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 135.069
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t dla zmiennej age:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 11.648
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wiek statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej bmi:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 25.886
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** BMI statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej regionnorthwest:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -0.695
- p-value:0.488
- Wniosek: Brak podstaw do odrzucenia $H_0$.

**Wniosek:** Region northwest statystycznie istotnie nie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej regionsoutheast:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -2.361
- p-value: 0.019
- Wniosek: Brak podstaw do odrzucenia $H_0$.

**Wniosek:** Region southeast statystycznie istotnie nie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej regionsouthwestt:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 0.671
- p-value: 0.503
- Wniosek: Brak podstaw do odrzucenia $H_0$.

**Wniosek:** Region southwest statystycznie istotnie nie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Kryterium Akaike dla tego modelu wynosi 5139.4.

**Model 11** g(u)=ln u -> model log normalny family=gaussian(link="log")

**Postać modelu 11**:

ln(charges)=8.6508756 + 0.0080495 * *age* + 0.0450499 * *bmi*- 0.0213275 * *regionnorthwest*- 0.0631668 * *regionsoutheast* + 0.0198589 * *regionsouthwest*

Testy istotności:

*Tabela 31*

```
lrtest(m11)
```

```
## Likelihood ratio test
##
## Model 1: charges ~ age + bmi + region
## Model 2: charges ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -2562.7
## 2   2 -2766.2 -5 407.05  < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
waldtest(m11)
```

```
## Wald test
##
## Model 1: charges ~ age + bmi + region
## Model 2: charges ~ 1
##   Res.Df Df      F    Pr(>F)
## 1    252
## 2    257 -5 184.11 < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Wnioski wyciągnięte na podstawie tabeli 31**

W teście ilorazu wiarygodności p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$. W teście Walda p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$.

Odrzucamy hipotezę $H_0$ mówiącą o tym, że parametry przy zmiennych są równe 0 na rzecz $H_1$, mówiącej o tym, że istnieje parametr istotnie różny od 0.Test ten powinien zostać wykonany na początku.

Istnieje parametr beta statystycznie istotny różny od 0, czyli jest zmienna która ma wpływ na wysokość składki ubezpieczeniowej. Ten test powinien poprzedzać testy dla poszczególnych parametrów.

```
dane5<-dane1[-c(34,212,263,247,173,116,161,117,231,8,111,10,113,67,44
,110,43,174,183),]
```

## Estymacja modelu 12 GLM z funkcją wiążącą odwrotną

*Tabela 32*

```
m12 <- glm(charges  ~ age + bmi +region , data = dane5, family = gaus
sian(link = “inverse”))
summary(m12)
## Call:
## glm(formula = charges ~ age + bmi + region, family = gaussian(link
= “inverse”),
##     data = dane5)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -11149   -5342   -1545    5013   13285
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.016e-05  2.425e-06  33.055  < 2e-16 ***
## age             -2.461e-07  2.239e-08 -10.991  < 2e-16 ***
## bmi             -1.235e-06  6.120e-08 -20.183  < 2e-16 ***
## regionnorthwest  6.594e-07  1.017e-06   0.648  0.51748
## regionsoutheast  2.436e-06  8.708e-07   2.798  0.00555 **
## regionsouthwest -1.197e-06  9.755e-07  -1.227  0.22095
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
##
## (Dispersion parameter for gaussian family taken to be 33062748)
##
##     Null deviance: 2.9151e+10  on 254  degrees of freedom
## Residual deviance: 8.2326e+09  on 249  degrees of freedom
## AIC: 5146.6
##
## Number of Fisher Scoring iterations: 8
```

**Interpretacje na podstawie tabeli 32:**

Minimalną wartością tego modelu jest -11149, a największą 13285. Wartość środkowa wynosi -1545. Kwartyl pierwszy 25% obserwacji położonych jest poniżej -5342, a 75% obserwacji jest położonych powyżej tej wartości. Kwartyl trzeci 75% obserwacji położonych jest poniżej 5013, a 25% obserwacji położonych jest powyżej tej wartości.

**test t dla wyrazu wolnego :**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 33.055
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t dla zmiennej age:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -10.991
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wiek statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej bmi:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -20.183
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** BMI statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej regionnorthwest:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 0.648
- p-value: 0.518
- Wniosek: Brak podstaw do odrzucenia $H_0$.

**Wniosek:** Region northwest statystycznie istotnie nie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej regionsoutheast:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 2.798
- p-value: 0.00555
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Region southeast statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej regionsouthwest:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -1.227

- p-value: 0.22095
- Wniosek: Brak podstaw do odrzucenia $H_0$.

**Wniosek:** Region southwest statystycznie istotnie nie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Kryterium Akaike dla tego modelu wynosi 5146.6.

**Postać modelu 12:**

1/charges = 0.00008016 - 2.461e-07 * *age* - 1.235e-06 * *bmi*+ 6.594e-07 * *regionnorthwest*+ 2.436e-06 * *regionsoutheast* -1.197e-06 * *regionsouthwest*

Testy istotności:

*Tabela 33*

```
lrtest(m12)
```

```
## Likelihood ratio test
##
## Model 1: charges ~ age + bmi + region
## Model 2: charges ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   7 -2566.3
## 2   2 -2727.5 -5 322.42  < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
waldtest(m12)
```

```
## Wald test
##
## Model 1: charges ~ age + bmi + region
## Model 2: charges ~ 1
##   Res.Df Df      F    Pr(>F)
## 1    249
## 2    254 -5 110.39 < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Wnioski na podstawie tabeli 33**

W teście ilorazu wiarygodności p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$. W teście Walda p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$.

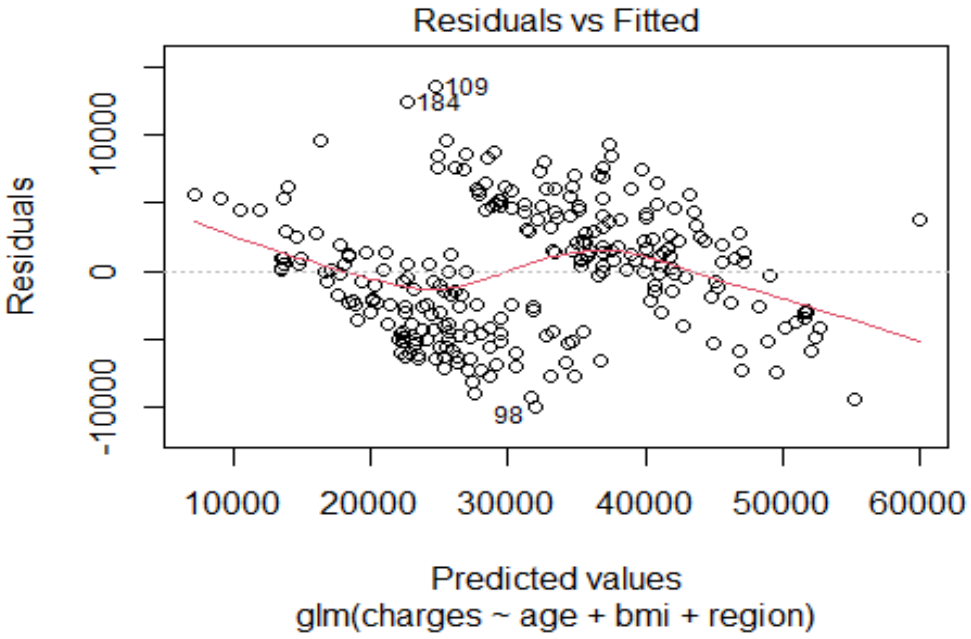
Odrzucamy hipotezę $H_0$ mówiącą o tym, że parametry przy zmiennych są równe 0 na rzecz $H_1$, mówiącej o tym, że istnieje parametr istotnie różny od 0.Test ten powinien zostać wykonany na początku.

Istnieje parametr beta statystycznie istotny różny od 0, czyli jest zmienna która ma wpływ na wysokość składki ubezpieczeniowej. Ten test powinien poprzedzać testy dla poszczególnych parametrów.

**Wykresy diagnostyczne modele 10-12 - wykresy reszt i wartości przewidywane:**

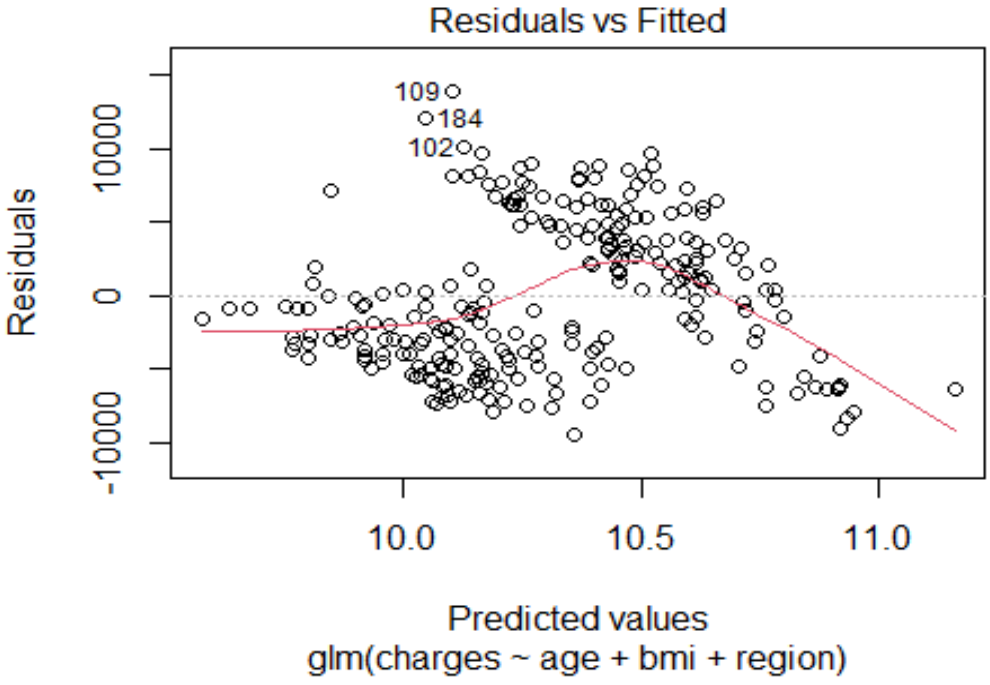
```
plot(m10, which = 1)
```

*Wykres 22*



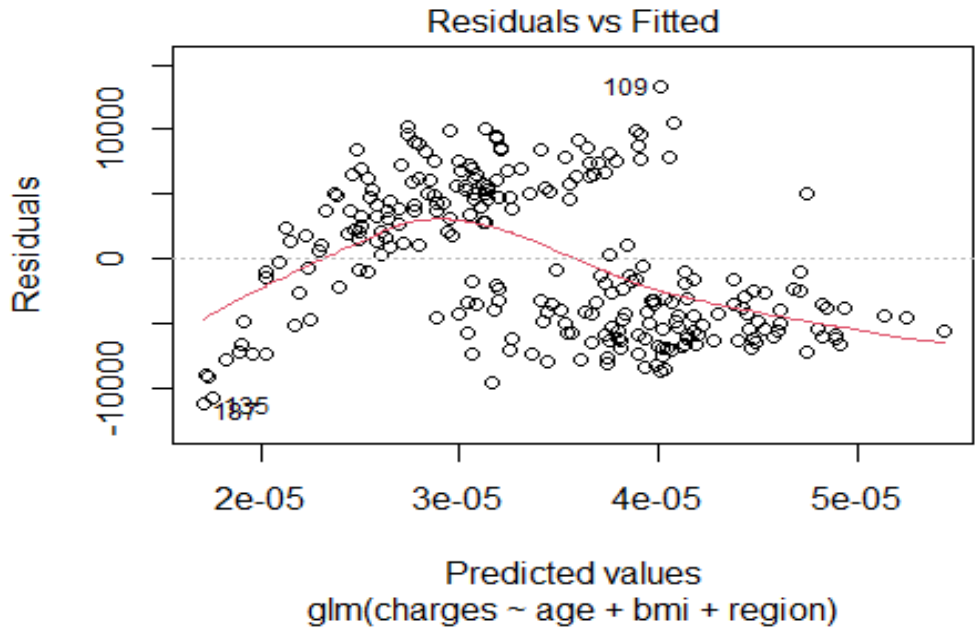

*Wykres 23*

```
plot(m11, which = 1)
```



*Wykres 24*

```
plot(m12, which = 1)
```




**Wnioski na podstawie wykresów 22-24:**

Na wykresach są zaznaczone 3 największe reszty, ale one niewiele odstają od pozostałych.

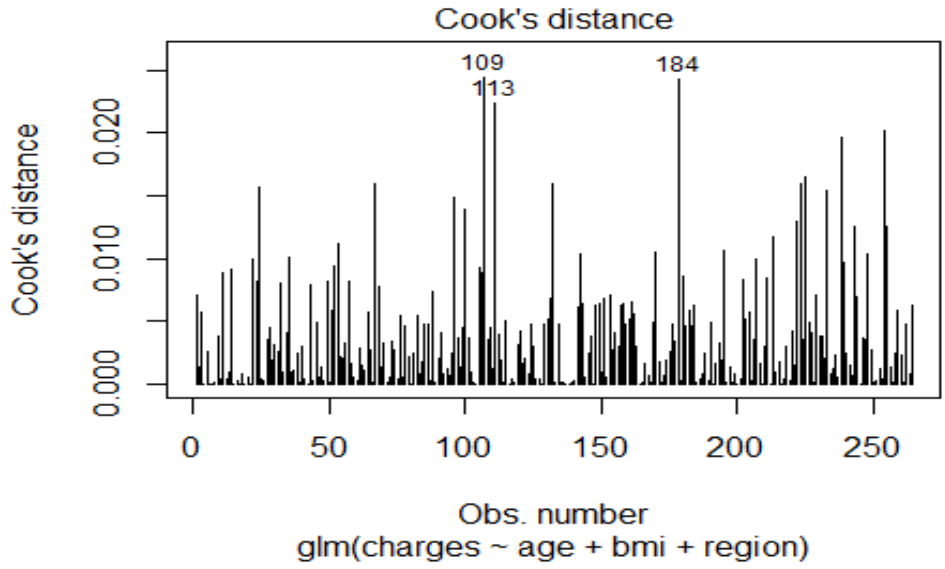
Oś pozioma - Wartości prognozowane

Oś pionowa - Reszty

Identyfikacja obserwacji wpływowych na wykresach (które wpływają w dużym stopniu na oszacowanie parametrów strukturalnych).

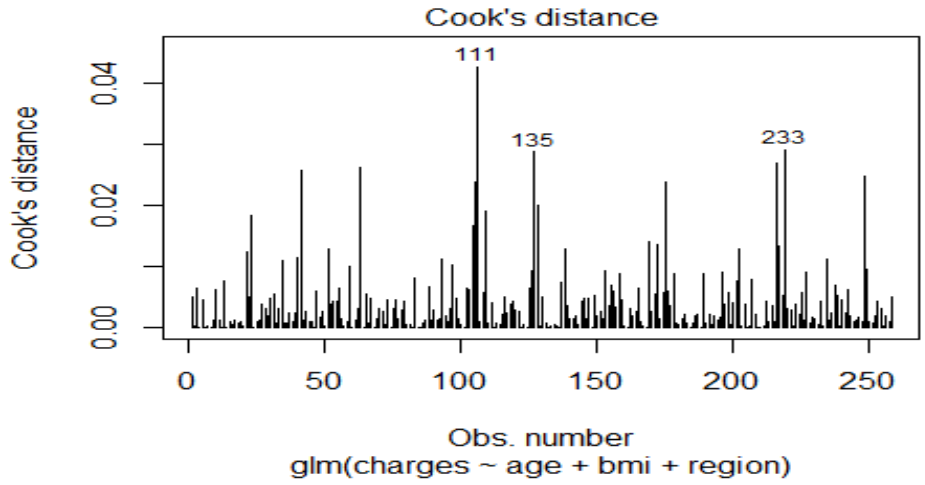
*Wykres 25*

```
plot(m10, which = 4)
```



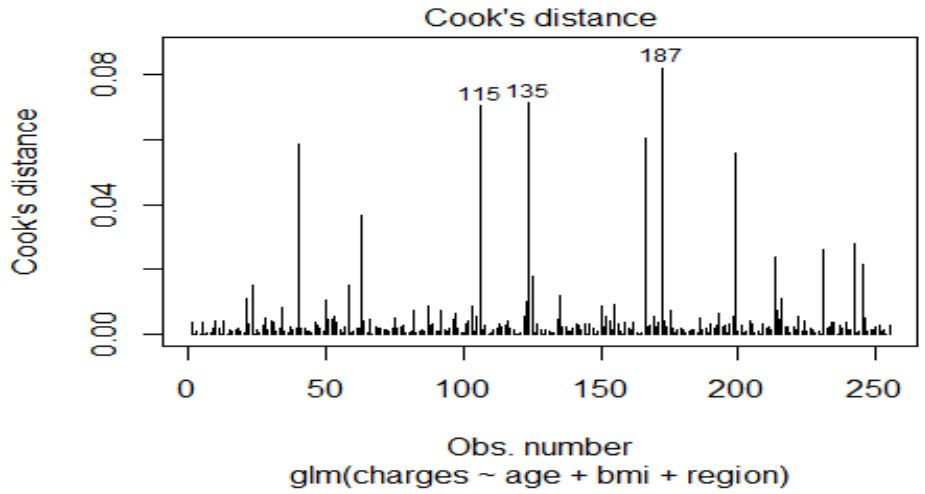
*Wykres 26*

```
plot(m11, which = 4)
```



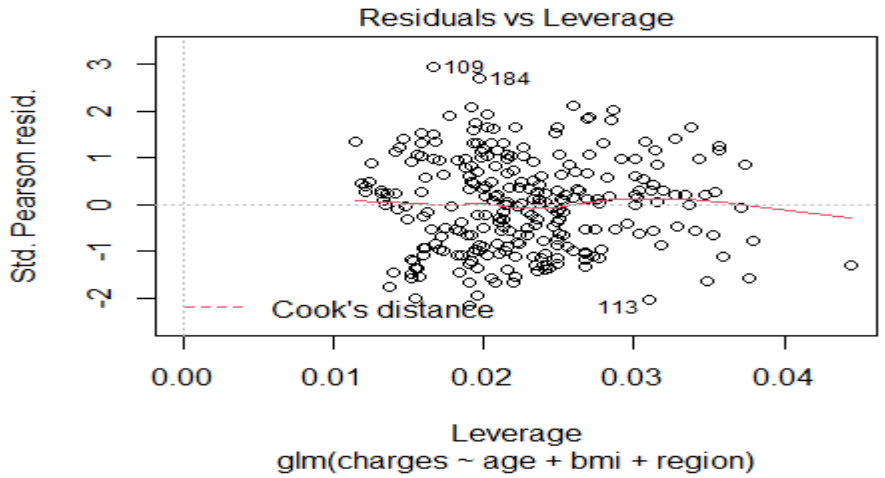

*Wykres 27*

```
plot(m12, which = 4)
```



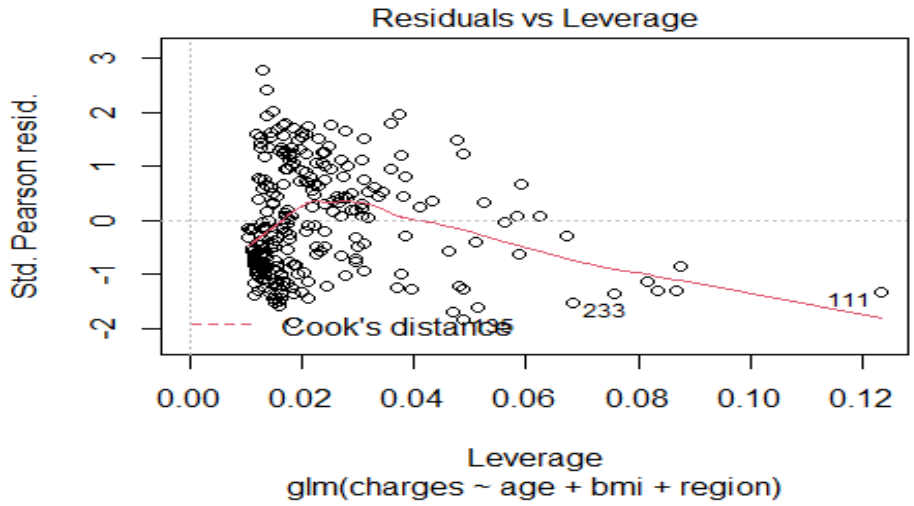

*Wykres 28*

```
plot(m10, which = 5)
```



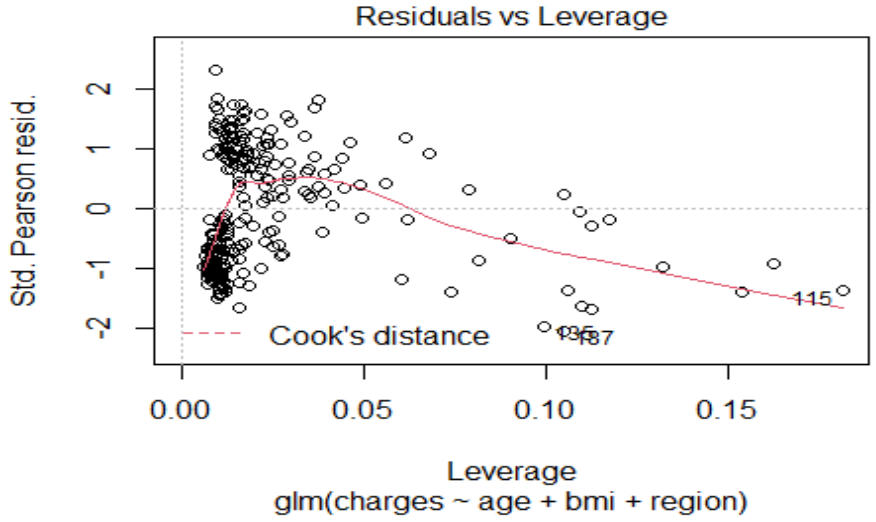
*Wykres 29*

```
plot(m11, which = 5)
```




*Wykres 30*

```
plot(m12, which = 5)
```




**Interpretacje na podstawie wykresów 25-30**

Odległość Cooka jest liczona dla każdej obserwacji (3 największe oznaczone) Oceniając model możemy wyróżnić obserwacje o dużych resztach (te odstające), jak i wpływowe (wpływowe nie zawsze są tymi odstającymi - one mają duży wpływ na oszacowanie parametrów strukturalnych).

Odległość Cooka jest liczona dla każdej obserwacji - liczony jest model na podstawie całego zbioru danych i bez tej obserwacji i budowana miara na podstawie jak zmieniły się te współczynniki beta. Jak jest zbyt duża to jest inny rząd wielkości to nie ma wątpliwości, że jest wpływowa. Stawiamy umowne granice.

Odległość Cooka i wskaźnik wpływu dla każdej zmiennej objaśniającej z osobna, mierzony dla poszczególnych obserwacji odstępstw o zmiennej objaśniającej xi od jej średniego poziomu.

Obserwacje, które są nietypowe mogą być: bo mają dużą resztę, bo x odbiegają, bo wpływ na wskaźniki beta.

Tutaj ta linia Cooka jest poza naszymi zmiennymi.

## Bonferroni Outlier Test

Funkcja OutlierTest - powie nam dla każdej obserwacji czy ona jest wpływowa czy nie. Argument n.max - liczba podawanych obserwacji nietypowych jest nie większa niż n.max

*Tabela 34*

```
outlierTest(m10, n.max = Inf)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 109 2.977943          0.0029019       0.7661

outlierTest(m11, n.max = Inf)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 109 2.811453          0.0049318           NA

outlierTest(m12, n.max = Inf)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 109 2.342093           0.019176           NA
```

**Wnioski na podstawie tabeli 34:**

Ponieważ p-value > 0.01 to odrzucamy $H_0$ na korzyść $H_1$ dla modelu 12.

Ponieważ p-value < 0.01 to nie ma podstaw do odrzucenia $H_0$ dla modeli 10 i 11.

Test Bonferroni Outlier nie wykazał obecności obserwacji nietypowych w analizowanych trzech modelach (test został wykonał dla trzech modeli).

## Porównanie modeli 10-12

Ocena dopasowania modeli GLM: statystyka odchylenia (*deviance*), kryterium informacyjne (AIC), miary pseudo-$R^2$.

W przypadku *family = gaussian* statystyka odchylenia = suma kwadratów reszt, zatem odchylenie standardowe reszt (średni błąd szacunku) = (deviance/df)^0.5.

Zdefiniowana została funkcja *ocena_modelu_GLM* licząca powyższe miary. Argumentem tej funkcji jest obiekt klasy GLM.

*Tabela 35*

```
ocena_modeli
```

```
##          odch_std_reszt kryterium_AIC   McFadden Cragg_Uhler
## model_10       4661.226      5217.164 0.08120556   0.8248182
## model_11       5042.749      5139.366 0.07357605   0.7935562
## model_12       5749.998      5146.634 0.05910556   0.7175938
```

**Interpretacje na podstawie tabeli 35:**

Miary pseudo $R^2$ zbudowane na bazie wiarygodności modelu w porównaniu do wiarygodności modelu 0 (tylko z wyrazem wolnym). $R^2$ jaki procent zmienności Y jest wyjaśniany poprzez zmienność zmiennych objaśniających. Tutaj nie ma takiej interpretacji, dlatego jest nazwa pseudo $R^2$ żeby ni interpretować jako $R^2$ w modelu liniowym. Miary pseudo $R^2$ zwykle w dolnych obszarach się znajdują.

Na bazie statystyki odchylenia mamy kryteria informacyjne. Dla rozkładu normalnego statystyka odchylenia jest równa sumie kwadratów reszt.

Wiarygodność – prawdopodobieństwo łącznego dostania się do próby tych zaobserwowanych przez nas wartości przy założeniu, że model działa. Maksymalizujemy wiarygodność poszukując parametru beta - czyli żeby zaobserwowane prawdopodobieństwo wartości w próbie było jak największe. To jest łączne prawdopodobieństwo dostania się do próby tych obserwacji, które już mamy. To jest liczba z przedziału 0-1, a my logarytmujemy dlatego wychodzi ujemne. Logarytmy wiarygodności reszt ujemne.

**Wnioski:**

Najlepiej dopasowany jest model 11, ponieważ ma najniższe kryterium AIC, niewysokie odchylenie standardowe reszt, a dosyć wysokie kryterium McFaddena i Cragg_Uhlera.

## Interpretacja modelu nr 11

*Tabela 36*

```
m11$coefficients
```

```
##     (Intercept)             age             bmi regionnorthwest regionsoutheast
##     8.650875638     0.008049477     0.045049862    -0.021327548    -0.063166802
## regionsouthwest
##     0.019858875
```

```
## exp(bi)

exp(m11$coefficients)
```

```
##     (Intercept)             age             bmi regionnorthwest regionsoutheast
##    5715.1489451       1.0080820       1.0460800       0.9788983       0.9387869
## regionsouthwest
##       1.0200574
```

**Wnioski na podstawie tabeli 36:**

**Postać modelu 11** można zapisać także jako:

charges = exp(5715.15 - 1.0080820 * *age* + 1.0460800 * *bmi*+ 0.9788983 * *regionnorthwest*+ 0.9387869 * *regionsoutheast* + 1.0200574 * *regionsouthwest*)

Parametry modelu log normalnego posiadają interpretację:

- wyraz wolny $\beta_0$ –> exp($\beta_0$) w tym modelu nie ma interpretacji, ponieważ badanie nie obejmowało osób w wieku 0 lat (niemowląt).
- $\beta_1$ = 0.008049477 –> exp($\beta_1$) = 1.0080820 –> (exp($\beta_1$)-1)*100%

age $\beta_1$ - Jeżeli wiek wzrośnie o jeden rok a pozostałe zmienne nie ulegną zmianie to wysokość składki dla osób palących wzrośnie średnio o 0.81% dla osób z tym samym bmi i mieszających w tym samym regionie cp.

- bmi $\beta_2$ = 0.045049862 –> exp($\beta_2$) = 1.0460800

Jeżeli bmi wzrośnie o jedną jednostkę a pozostałe zmienne nie ulegną zmianie to wysokość składki dla osób palących wzrośnie średnio o 4.6% dla osób w tym samym wieku i mieszających w tym samym regionie cp.

- region northwest $\beta_3$ = -0.021327548 –> exp($\beta_3$) = 0.9788983

Beneficjenci zamieszkujący w regionie northwest mają średnio o 2,11% niższe koszty naliczanie przez ubezpieczalnie zdrowotne, cp.

- region southeast $\beta_4$ = -0.063166802 –> exp($\beta_4$) = -6,31%

Beneficjenci zamieszkujący w regionie southeast mają średnio o 6,31 % niższe koszty naliczanie przez ubezpieczalnie zdrowotne, cp.

- region southwest $\beta_5$ = -0.019858875 –> exp($\beta_5$) = 2%

Beneficjenci zamieszkujący w regionie southwest mają średnio o 2 % wyższe koszty naliczanie przez ubezpieczalnie zdrowotne, cp.

## Podsumowanie:

W pierwszej kolejności stworzono ***modele glm 10, 11 i 12*** oparte na modelu 9. Model 11 ma postać log normalną, natomiast model 12 to model glm z funkcją wiążącą odwrotną Ponieważ model 9 był dobrze dopasowany model glm zbudowany jest w oparciu o zmienne zmienna *bmi, age oraz region*. W każdym z modeli przeprowadzono testy istotności parametrów, testy Walda i Likelihood ratio test, wykresy reszt i wartości przewidywanych, Bonferroni Outlier Test. Na podstawie kryterium informacyjnego i miar pseudo $R^2$ wybrałyśmy model 11, ponieważ miał on najmniejszą wartość kryterium Akaike. Na podstawie modelu log normalnego dokonałyśmy interpretacji parametrów ilorazu szans.

## Estymacja modelu 13 – Klasy GLM -> (analogiczny do modelu liniowego 7)

*Tabela 37*

```
dane6<-dane1[-c(34,212,263,247,173,116,161,117,231,8),]

#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku, bmi i regionu zamieszkania beneficjenta palącego po wy
eliminowaniu jednostek odstających
m13 <- glm(charges  ~ age + bmi , data = dane6, family = gaussian)
summary(m13)

## Call:
## glm(formula = charges ~ age + bmi,family = gaussian,data = dane6)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9767.8  -3686.5     78.9   3567.9  13329.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23658.73    1628.87  -14.53   <2e-16 ***
## age            257.72      20.76   12.41   <2e-16 ***
## bmi           1476.35      47.62   31.00   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 22149492)
##
##     Null deviance: 3.1999e+10  on 263  degrees of freedom
## Residual deviance: 5.7810e+09  on 261  degrees of freedom
## AIC: 5219.3
## Number of Fisher Scoring iterations: 2
```

**Interpretacje na podstawie tabeli 37:**

Minimalną wartością tego modelu jest -9767.8, a największą 13.329. Wartość środkowa wynosi 78.9. Kwartyl pierwszy 25% obserwacji położonych jest poniżej -3686.5, a 75% obserwacji jest położonych powyżej tej wartości. Kwartyl trzeci 75% obserwacji położonych jest poniżej 3567.9, a 25% obserwacji położonych jest powyżej tej wartości.

**test t dla wyrazu wolnego :**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -14.53
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t dla zmiennej age:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 12.41
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wiek statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej bmi:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 31.00
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** BMI statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Kryterium Akaike dla tego modelu wynosi 5219.3.

**Postać modelu 13:** chrages= - 23658.73 + 257.72 * *age* + 1476.35 * *bmi*

**Wnioski:** Ponieważ p-value < alfa = 0.01, to odrzucamy $H_0$ na korzyść $H_1$. Na podstawie testu Walda możemy stwierdzić, że zarówno wiek oraz bmi osób palących istotnie wpływają na wysokość składki ubezpieczeniowej.

Dodatkowo wykonamy test istotności wszystkich zmiennych niezależnych w modelu. Można wykonać test ilorazu wiarygodności lub test Walda. Przy czym testami lokalnymi wykonuje się test globalny - najpierw testujemy wszystkie parametry=0, jeśli nie to dopiero robimy testy lokalne.

*Testy istotności:*

*Tabela 38*

```
lrtest(m13)
```

```
## Likelihood ratio test
##
## Model 1: charges ~ age + bmi
## Model 2: charges ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2605.7
## 2   2 -2831.5 -2 451.73  < 2.2e-16 ***
```

```
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

waldtest(m13)

## Wald test
##
## Model 1: charges ~ age + bmi
## Model 2: charges ~ 1
##   Res.Df Df      F    Pr(>F)
## 1    261
## 2    263 -2 591.83 < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Wnioski na podstawie tabeli 38**

W teście ilorazu wiarygodności p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$. W teście Walda p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$.

Odrzucamy hipotezę $H_0$ mówiącą o tym, że parametry przy zmiennych są równe 0 na rzecz $H_1$, mówiącej o tym, że istnieje parametr istotnie różny od 0.Test ten powinien zostać wykonany na początku.

Istnieje parametr beta statystycznie istotny różny od 0, czyli jest zmienna która ma wpływ na wysokość składki ubezpieczeniowej. Ten test powinien poprzedzać testy dla poszczególnych parametrów.

```
dane7<-dane1[-c(34,212,263,247,173,116,161,117,231,8,111,69,113),]
```

## Estymacja modelu 14 GLM z funkcją wiążącą logarytmiczną (model log normalny)

*Tabela 39*

```
m14 <- glm(charges  ~ age + bmi, data = dane7, family = gaussian(link = “log”))
summary(m14)

## Call:
## glm(formula = charges ~ age + bmi, family =gaussian(link =“log”),
#data = dane7)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -10475.0   -4230.7    -952.7    4203.3   12739.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.6733761  0.0623186  139.18   <2e-16 ***
```

```
## age         0.0078126  0.0006857   11.39   <2e-16 ***
## bmi         0.0439274  0.0016818   26.12   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 25929032)
##
##     Null deviance: 3.0607e+10  on 260  degrees of freedom
## Residual deviance: 6.6897e+09  on 258  degrees of freedom
## AIC: 5201.2
## Number of Fisher Scoring iterations: 5
```

**Interpretacje na podstawie tabeli 39:**

Minimalną wartością tego modelu jest -10475, a największą 12739. Wartość środkowa wynosi -952.7. Kwartyl pierwszy 25% obserwacji położonych jest poniżej -4230.7 , a 75% obserwacji jest położonych powyżej tej wartości. Kwartyl trzeci 75% obserwacji położonych jest poniżej 4203.3 , a 25% obserwacji położonych jest powyżej tej wartości.

**test t dla wyrazu wolnego :**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 139.18
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t dla zmiennej age:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 11.39
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wiek statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej bmi:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 26.12
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** BMI statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Kryterium Akaike dla tego modelu wynosi 5201.2.

**Model 14** g(u)=ln u -> model log normalny family=gaussian(link="log")

**Postać modelu 14:** ln(charges)= 7.943585 + 0.030509 * *age* - 0.005917 * *bmi*

Testy istotności:

*Tabela 40*

```
lrtest(m14)
```

```
## Likelihood ratio test
##
## Model 1: charges ~ age + bmi
## Model 2: charges ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2596.6
## 2   2 -2795.0 -2 396.89  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
waldtest(m14)
```

```
## Wald test
##
## Model 1: charges ~ age + bmi
## Model 2: charges ~ 1
##   Res.Df Df      F    Pr(>F)
## 1    258
## 2    260 -2 427.65 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Wnioski na podstawie tabeli 40**

W teście ilorazu wiarygodności p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$. W teście Walda p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$.

Odrzucamy hipotezę $H_0$ mówiącą o tym, że parametry przy zmiennych są równe 0 na rzecz $H_1$, mówiącej o tym, że istnieje parametr istotnie różny od 0.Test ten powinien zostać wykonany na początku.

Istnieje parametr beta statystycznie istotny różny od 0, czyli jest zmienna która ma wpływ na wysokość składki ubezpieczeniowej. Ten test powinien poprzedzać testy dla poszczególnych parametrów.

```
dane8<-dane1[-c(34,212,263,247,173,116,161,117,231,8,111,110,113),]
```

## Estymacja modelu 15 GLM z funkcją wiążącą odwrotną

*Tabela 41*

```
m15 <- glm(charges  ~ age + bmi , data = dane8, family = gaussian(lin
k = “inverse”))
summary(m15)
```

```
##
## Call:
## glm(formula=charges ~ age + bmi,family =gaussian(link =“inverse”),
data = dane8)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -12051   -5538   -1826    5423   11755
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.653e-05  2.164e-06   35.36   <2e-16 ***
## age         -2.257e-07  2.211e-08  -10.21   <2e-16 ***
## bmi         -1.124e-06  5.367e-08  -20.94   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ‘ 1
##
## (Dispersion parameter for gaussian family taken to be 34595320)
##
##     Null deviance: 3.0420e+10  on 260  degrees of freedom
## Residual deviance: 8.9255e+09  on 258  degrees of freedom
## AIC: 5276.4
##
## Number of Fisher Scoring iterations: 8
```

**Interpretacje na podstawie tabeli 41:**

Minimalną wartością tego modelu jest -12051, a największą 11755. Wartość środkowa wynosi -1826. Kwartyl pierwszy- 25% obserwacji położonych jest poniżej -5538, a 75% obserwacji jest położonych powyżej tej wartości. Kwartyl trzeci - 75% obserwacji położonych jest poniżej 5432, a 25% obserwacji położonych jest powyżej tej wartości.

**test t dla wyrazu wolnego :**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t = 35.36
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t dla zmiennej age:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -10.21
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wiek statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**test t dla zmiennej bmi:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= -20.94
- p-value: 0.276
- Wniosek: Brak podstaw do odrzucenia $H_0$.

**Wniosek:** BMI statystycznie istotnie nie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Kryterium Akaike dla tego modelu wynosi 5276.4 .

**Postać modelu 15:**

1/charges = 7.653e-05 - 2.257e-07 * $age$ - 1.124e-06 * $bmi$

Testy istotności:

*Tabela 42*

```
lrtest(m15)
```

```
## Likelihood ratio test
##
## Model 1: charges ~ age + bmi
## Model 2: charges ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2634.2
## 2   2 -2794.2 -2 320.03  < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
waldtest(m15)
```

```
## Wald test
##
## Model 1: charges ~ age + bmi
## Model 2: charges ~ 1
##   Res.Df Df      F    Pr(>F)
```

```
## 1    258
## 2    260 -2 279.26 < 2.2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Wnioski na podstawie tabeli 42:**

W teście ilorazu wiarygodności p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$. W teście Walda p-value < 0.01, zatem należy odrzucić $H_0$ na korzyść $H_1$.

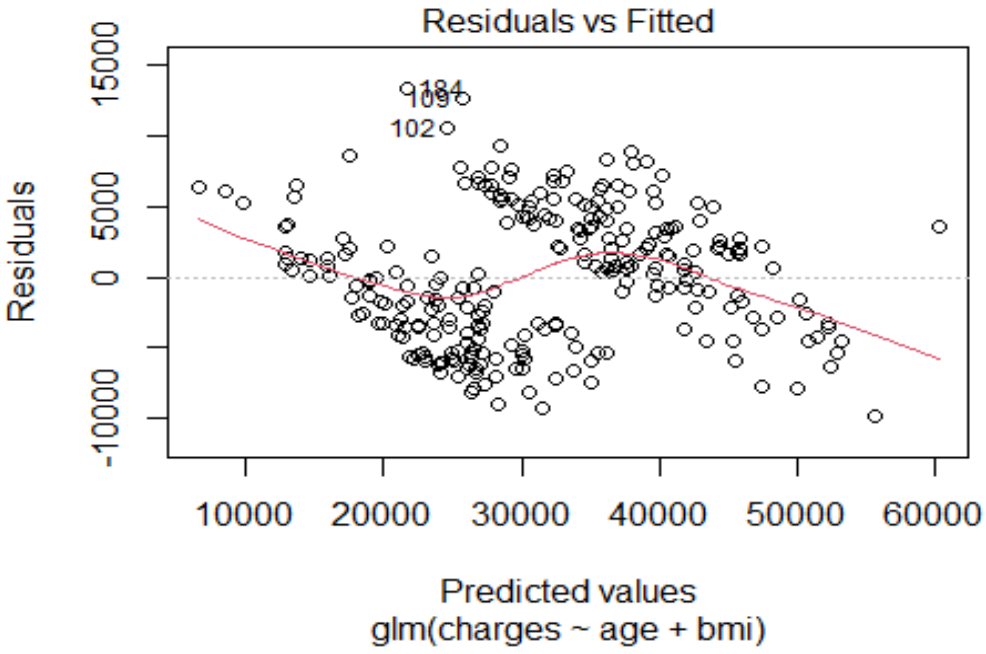
Odrzucamy hipotezę $H_0$ mówiącą o tym, że parametry przy zmiennych są równe 0 na rzecz $H_1$, mówiącej o tym, że istnieje parametr istotnie różny od 0.Test ten powinien zostać wykonany na początku.

**Wykresy diagnostyczne modele 13-15 - wykresy reszt i wartości przewidywane:**

Czy te reszty rozkładają się wzdłuż linii przerywanej (średnia wartość reszt - krzywa wygładzona) czy reszty rosną wraz ze wzrostem wartości prognozowanej? Czy ten rozrzut jest losowy? Sprawdzenie, czy średnia reszt jest bliska 0, czy rozrzut reszt nie zależy od y oraz czy nie ma obserwacji nietypowych o dużych resztach.

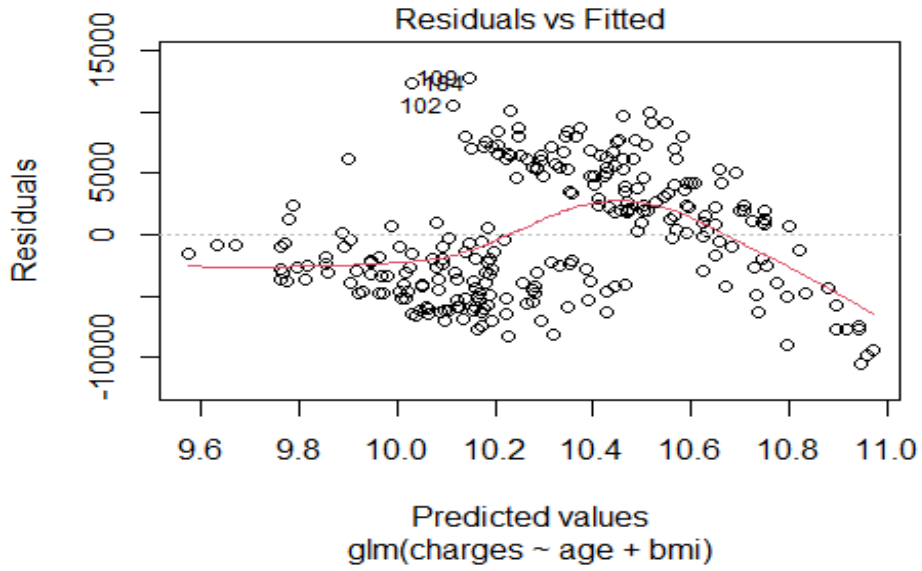
*Wykres 31*

```
plot(m13, which = 1)
```



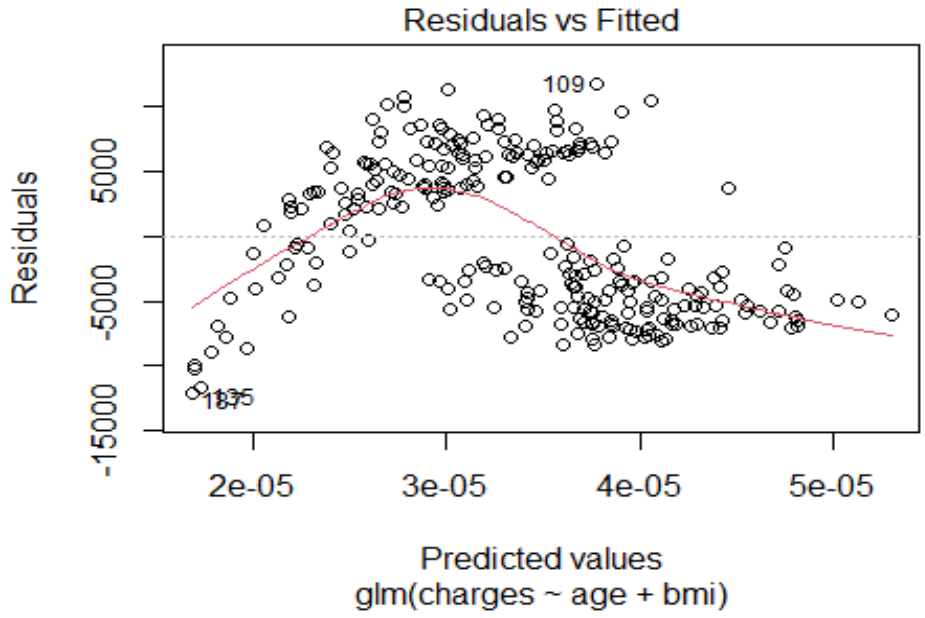
*Wykres 32*

```
plot(m14, which = 1)
```




*Wykres 33*

```
plot(m15, which = 1)
```




**Wnioski na podstawie wykresów 31,32 i 33**

Na wykresach są zaznaczone 3 największe reszty, ale one niewiele odstają od pozostałych.

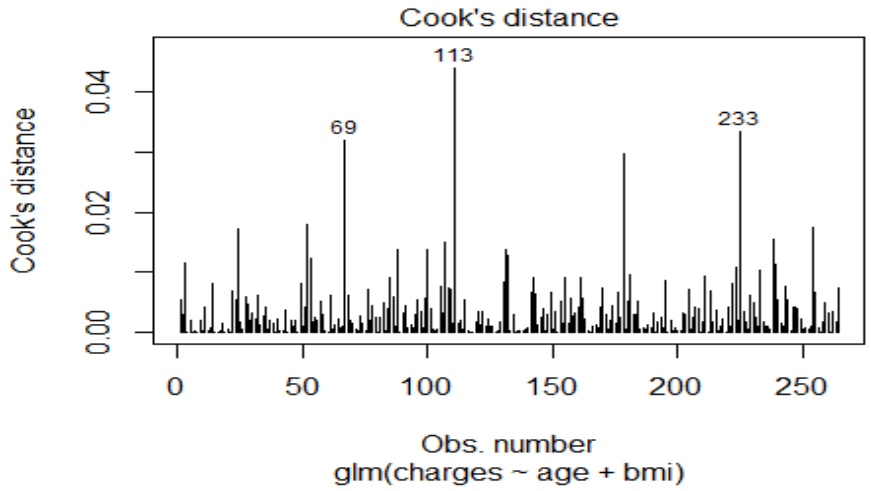
Oś pozioma - Wartości prognozowane

Oś pionowa - Reszty

Identyfikacja obserwacji wpływowych na wykresach (które wpływają w dużym stopniu na oszacowanie parametrów strukturalnych).

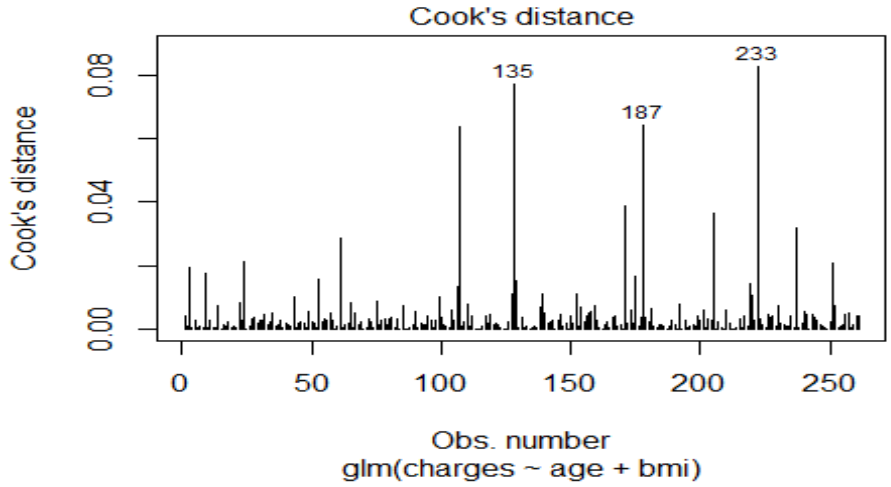
*Wykres 34*

```
plot(m13, which = 4)
```



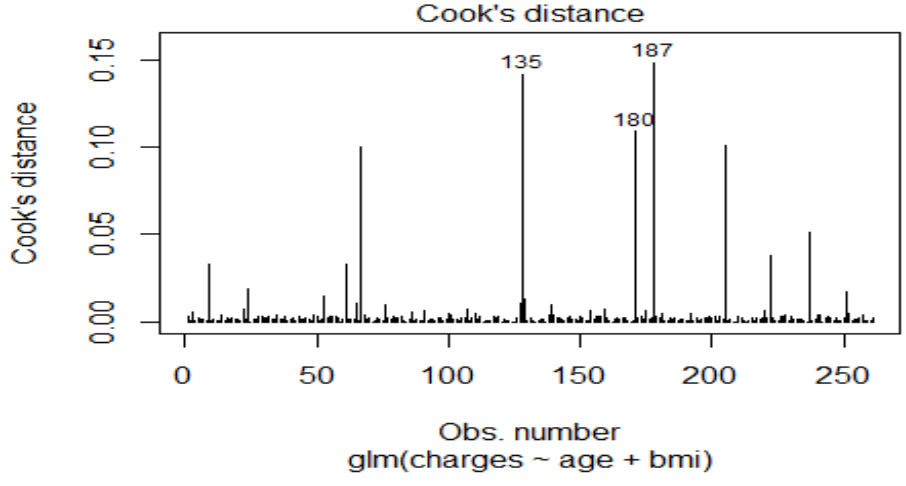

*Wykres 35*

```
plot(m14, which = 4)
```



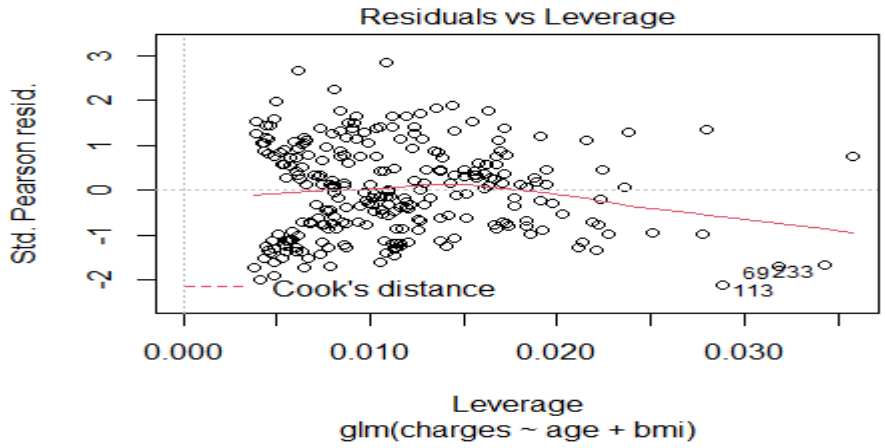

*Wykres 36*

```
plot(m15, which = 4)
```



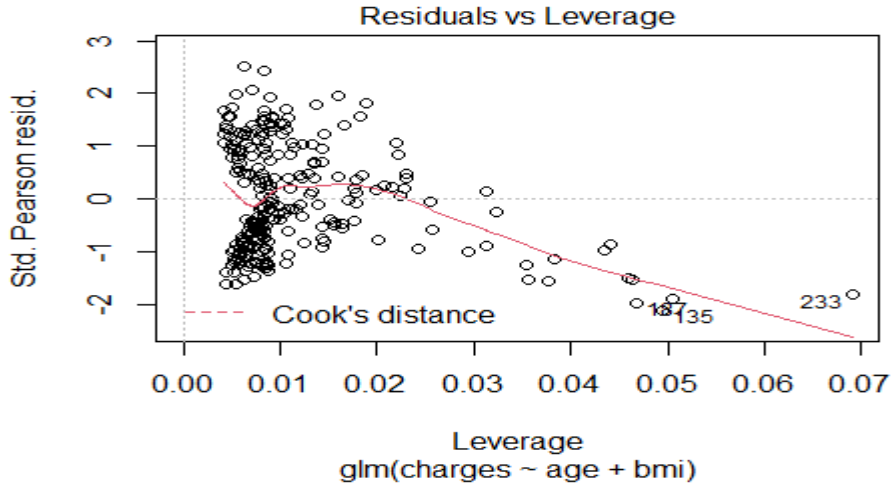
*Wykres 37*

```
plot(m13, which = 5)
```



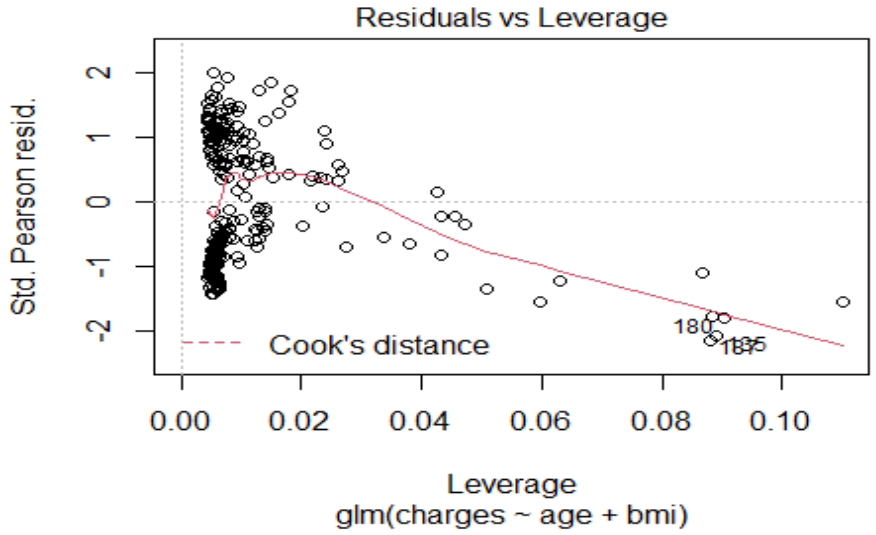

*Wykres 38*

```
plot(m14, which = 5)
```




*Wykres 39*

```
plot(m15, which = 5)
```



**Wnioski na podstawie wykresów 34- 39:**

Odległość Cooka jest liczona dla każdej obserwacji (3 największe oznaczone) Oceniając model możemy wyróżnić obserwacje o dużych resztach (te odstające), jak i wpływowe (wpływowe nie zawsze są tymi odstającymi - one mają duży wpływ na oszacowanie parametrów strukturalnych).

Odległość Cooka jest liczona dla każdej obserwacji - liczony jest model na podstawie całego zbioru danych i bez tej obserwacji i budowana miara na podstawie jak zmieniły się te współczynniki beta. Jak jest zbyt duża to jest inny rząd wielkości to nie ma wątpliwości, że jest wpływowa. Stawiamy umowne granice.

Odległość Cooka i wskaźnik wpływu dla każdej zmiennej objaśniającej z osobna, mierzony dla poszczególnych obserwacji odstępstw o zmiennej objaśniającej xi od jej średniego poziomu.

Obserwacje, które są nietypowe mogą być: bo mają dużą resztę, bo x odbiegają, bo wpływ na wskaźniki beta.

Tutaj ta linia Cooka jest poza naszymi zmiennymi.

## Bonferroni Outlier Test

*Tabela 43*

```
outlierTest(m13, n.max = Inf)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 184 2.887545          0.0038826           NA

outlierTest(m14, n.max = Inf)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 109 2.535831           0.011218           NA

outlierTest(m15, n.max = Inf)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 187 -2.160458           0.030737           NA
```

**Wnioski na podstawie tabeli 43:**

Ponieważ p-value > 0.01 to odrzucamy $H_0$ na korzyść $H_1$ dla modelu 14 i 15. Ponieważ p-value < 0.01 to nie ma podstaw do odrzucenia $H_0$ dla modelu 13. Test Bonferroni Outlier nie wykazał obecności obserwacji nietypowych w analizowanych trzech modelach. (test wykonał dla trzech modeli).

## Porównanie modeli 13-15

Ocena dopasowania modeli GLM: statystyka odchylenia (*deviance*), kryterium informacyjne (AIC), miary pseudo-$R^2$.

W przypadku *family = gaussian* statystyka odchylenia = suma kwadratów reszt, zatem odchylenie standardowe reszt (średni błąd szacunku) = (deviance/df)^0.5.

*Tabela 44*

```
ocena_modeli
```

```
##          odch_std_reszt kryterium_AIC   McFadden Cragg_Uhler
## model_13       4706.325      5219.300 0.07976886   0.8193354
## model_14       5092.075      5201.169 0.07099994   0.7814340
## model_15       5881.741      5276.424 0.05726709   0.7065907
```

**Wnioski na podstawie tabeli 44:**

Najlepiej dopasowany jest model 14, ponieważ ma najniższe kryterium AIC, niewysokie odchylenie standardowe reszt, a dosyć wysokie kryterium McFaddena i Cragg_Uhlera.

## Interpretacja modelu nr 14

*Tabela 45*

```
m14$coefficients
```

```
## (Intercept)         age         bmi
##  8.67337612  0.00781260  0.04392745
```

```
## exp(bi)
```

```
exp(m14$coefficients)
```

```
## (Intercept)         age         bmi
## 5845.200159    1.007843    1.044907
```

**Wnioski na podstawie tabeli 45:**

**Postać modelu 14** można zapisać także jako:

charges = exp(7.943585420 + 0.030508538 * *age* -0.005917269 * *bmi*)

Parametry modelu log normalnego posiadają interpretację:

- wyraz wolny $\beta_0$ –> exp($\beta_0$) w tym modelu nie ma interpretacji, ponieważ badanie nie obejmowało osób w wieku 0 lat (niemowląt).
- age $\beta_1$ = 0.00781260 –> exp($\beta_1$)=1.007843 –> (exp($\beta_1$)-1)*100%

Jeżeli wiek wzrośnie o jeden rok a pozostałe zmienne nie ulegną zmianie to wysokość składki dla osób palących wzrośnie średnio o 0,78% dla osób z tym samym bmi cp.

- bmi $\beta_2$ = 0.04392745 –> exp(($\beta_2$)-1)=4,4%

Jeżeli bmi wzrośnie o jedną jednostkę pozostałe zmienne nie ulegną zmianie to wysokość składki dla osób palących wzrośnie średnio o 4,4% dla osób w tym samym wieku cp.

## Podsumowanie:

W pierwszej kolejności stworzono ***modele glm 13, 14 i 15*** oparte na modelu 7. Model 14 ma postać log normalną, natomiast model 15 to model glm z funkcją wiążącą odwrotną Ponieważ model 7 był dobrze dopasowany model glm zbudowany jest w oparciu o zmienne zmienna *bmi* oraz *age*. W każdym z modeli przeprowadzono testy istotności parametrów, testy Walda i Likelihood ratio test, wykresy reszt i wartości przewidywanych, Bonferroni Outlier Test. Na podstawie kryterium informacyjnego i miar pseudo $R^2$ wybrałyśmy model 14, ponieważ miał on najmniejszą wartość kryterium Akaike. Na podstawie modelu log normalnego dokonałyśmy interpretacji parametrów ilorazu szans.

### 2.3. Budowa i weryfikacja modeli logitowych i probitowego objaśniających koszty leczenia rozliczane przez ubezpieczenie zdrowotne dla beneficjentów palących. Zmienna *charges* została podzielona na dwie części: 1-powyżej średniej, 0-poniżej średniej.

## Wczytanie danych i statystyki opisowe dla poszczególnych zmiennych:

*Tabela 46*

```
dane9 <- read.table("palacyl.csv", header = TRUE, sep = ";",dec=",")
dane9$sex<-as.factor(dane9$sex)
dane9$region<-as.factor(dane9$region)
dane9$charges<-as.factor(dane9$charges)
summary(dane9)
```

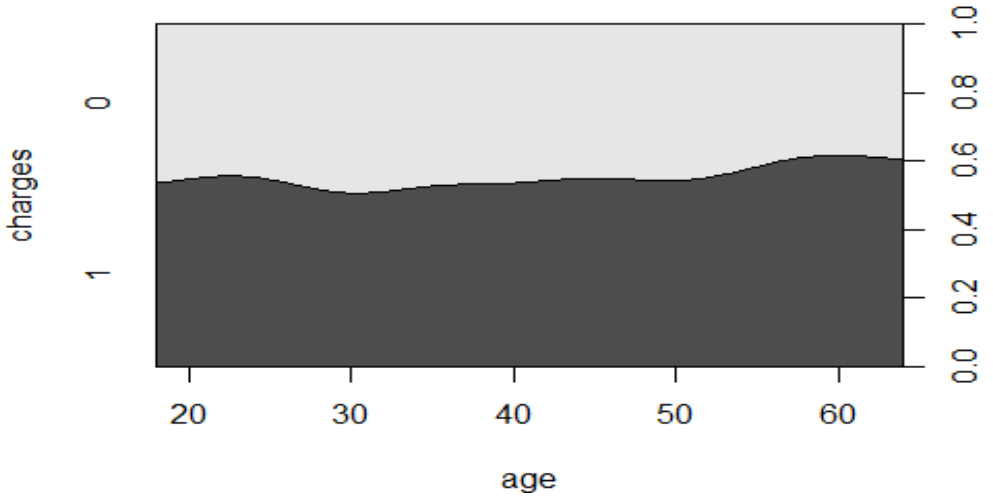
```
##       age            sex           bmi           children         region
##  Min.   :18.00   female:115   Min.   :17.20   Min.   :0.000   northeast:67
##  1st Qu.:27.00    male  :159   1st Qu.:26.08   1st Qu.:0.000    northwest:58
##  Median :38.00                 Median :30.45   Median :1.000   southeast:91
##  Mean   :38.51                 Mean   :30.71   Mean   :1.113   southwest:58
##  3rd Qu.:49.00                 3rd Qu.:35.20   3rd Qu.:2.000
##  Max.   :64.00                 Max.   :52.58   Max.   :5.000

    charges
##  0:124
##  1:150
```

Jaki jest związek między kosztami leczenia rozliczanymi przez ubezpieczenie zdrowotne powyżej średniej, a potencjalnymi predyktorami?

**Wykresy warunkowych prawdopodobieństw wystąpienia wariantów cechy jakościowej** (kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej) **pod warunkiem, że zmienna ilościowa przyjmuje określony poziom:**

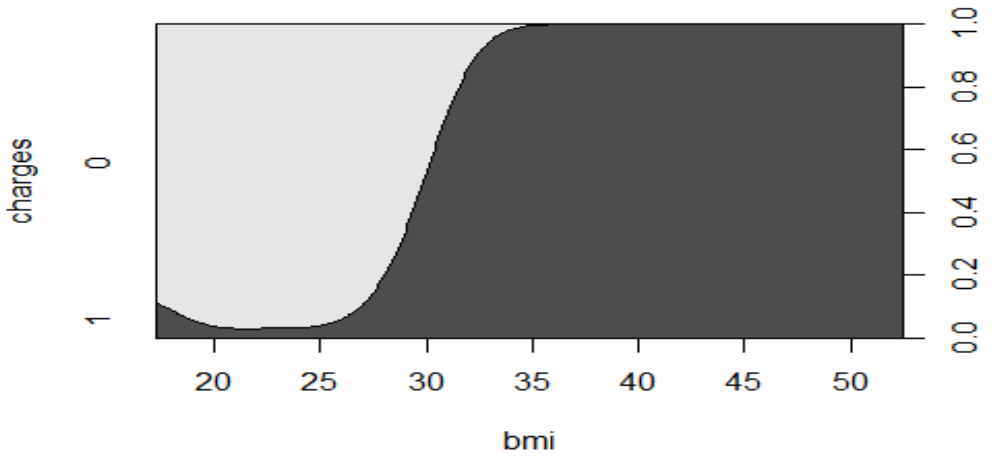
*Wykres 40*

```
cdplot(dane9$age,dane9$charges,xlab="age",ylab="charges")
```



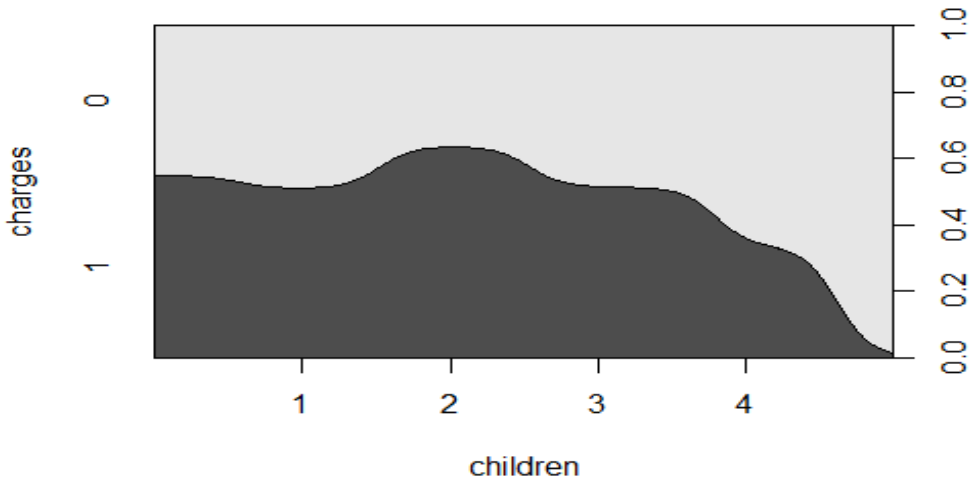

*Wykres 41*

```
cdplot(dane9$bmi,dane9$charges,xlab="bmi",ylab="charges")
```




*Wykres 42*

```
cdplot(dane9$children,dane9$charges,xlab="children",ylab ="charges")
```



**Wnioski na podstawie wykresów 40-42:** Wiek beneficjenta palącego nie wpływa na prawdopodobieństwo kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej. Wraz ze wzrostem bmi beneficjenta palącego, wzrasta prawdopodobieństwo na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej. Wraz ze wzrostem ilości osób na utrzymaniu beneficjenta palącego, wzrasta prawdopodobieństwo na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej.

**Podział zbioru na uczący i testowy** Zbiór uczący posłuży do budowy modelu, a zbiór testowy posłuży do oceny modelu. Dokonano losowego podziału w proporcji odpowiednio: 70% i 30%. W celu powtarzalności eksperymentu wykorzystano funkcję `set.seed()`, która inicjuje „ziarno" dla generatora liczb losowych - za każdym razem otrzymuje się ten sam zestaw liczb losowych.

**Proporcje beneficjentów o kosztach leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej (1) i poniżej średniej(0) w podzbiorach danych.**

*Tabela 47 – dzieli zbiór danych na uczący i testowy*

```
table(dane9$charges)/nrow(dane9)

##         0         1
## 0.4525547 0.5474453

table(dane9_uczacy$charges)/nrow(dane9_uczacy)

##         0         1
## 0.4427083 0.5572917

table(dane9_testowy$charges)/nrow(dane9_testowy)

##         0         1
## 0.4756098 0.5243902
```

**Macierz korelacji dla objaśniających zmiennych ilościowych**

*Tabela 48*

```
cor(dane9_uczacy[,c(1,3,4)])

##                  age        bmi   children
## age      1.00000000 0.07014313 0.11934947
## bmi      0.07014313 1.00000000 0.02392168
## children 0.11934947 0.02392168 1.00000000
```

**Wniosek na podstawie tabeli 48:** Żadna ze zmiennych w modelu nie jest nadmiernie skorelowana, tzn. nie przekracza `|r|>=0.7`. Wszystkie zmienne mogą się znaleźć w jednym modelu.

## Estymacja modeli dwumianowych logitowych jednoczynnikowych

Estymujemy model dla zmiennej dychotomicznej/binarnej Y `family = binomial` z domyślną funkcją wiążącą probit `link = logit`.

*Tabela 49*

```
logit1 <- glm(charges ~ age, data = dane9_uczacy, family = binomial)
summary(logit1)$coefficients

##                Estimate Std. Error   z value  Pr(>|z|)
## (Intercept) 0.081091171 0.43277661 0.1873742 0.8513673
## age         0.003901959 0.01068073 0.3653269 0.7148674

logit2 <- glm(charges ~ sex, data = dane9_uczacy, family = binomial)
summary(logit2)$coefficients

##               Estimate Std. Error    z value   Pr(>|z|)
## (Intercept) -0.1100009  0.2099742 -0.5238783 0.60036322
## sexmale      0.6579661  0.2945185  2.2340395 0.02548047

logit3 <- glm(charges ~ bmi, data = dane9_uczacy, family = binomial)
summary(logit3)$coefficients

##               Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -24.147463  3.8427639 -6.283879 3.302279e-10
## bmi           0.816106  0.1299532  6.279999 3.385745e-10

logit4 <- glm(charges ~ children, data=dane9_uczacy,family=binomial)
summary(logit4)$coefficients

##               Estimate Std. Error   z value  Pr(>|z|)
## (Intercept) 0.18054434  0.2027665 0.8904053 0.3732483
## children    0.04325545  0.1236222 0.3499005 0.7264134

logit5 <- glm(charges ~ region, data = dane9_uczacy,family=binomial)
summary(logit5)$coefficients

##                    Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)     -0.18232156  0.3027650 -0.6021883 0.54704881
## regionnorthwest -0.05129329  0.4312365 -0.1189447 0.90531917
## regionsoutheast  0.92676203  0.4114586  2.2523820 0.02429814
## regionsouthwest  0.62415431  0.4277108  1.4592906 0.14448514
```

**Poniżej zinterpretowano modele logitowe na podstawie tabeli 49.**

- **logit1, postać modelu:**

charges = 0.081091171 + 0.003901959 * *age*

Wniosek: Ponieważ p-value > 0,01 nie ma podstaw do odrzucenia $H_0$. Wiek beneficjenta palącego nie ma istotnego wpływu na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu nie będziemy włączać zmiennej *age*.

- **logit2, postać modelu:**

charges = -0.1100009 + 0.6579661 * *sex*

Wniosek: Ponieważ p-value > 0,01 nie ma podstaw do odrzucenia $H_0$. Płeć beneficjenta palącego nie ma istotnego wpływu na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu nie będziemy włączać zmiennej *sex*.

- **logit3, postać modelu:**

charges = -24.147463 + 0.816106 * *bmi*

Wniosek: Ponieważ p-value < 0,01 odrzucamy $H_0$ na korzyść $H_1$. Bmi beneficjenta palącego ma istotny wpływ na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu będziemy włączać zmienną *bmi*.

- **logit4, postać modelu:**

charges = 0.18054434 + 0.04325545 * *children*

Wniosek: Ponieważ p-value > 0,01 nie ma podstaw do odrzucenia $H_0$. Ilość osób na utrzymaniu beneficjenta palącego nie ma istotnego wpływu na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu nie będziemy włączać zmiennej *children*.

- **logit5, postać modelu:**

charges = -0.18232156 - 0.05129329 * *regionnorthwest* + 0.92676203 *regionsoutheast* *+ 0.62415431 * regionsouthwest*

Wniosek: Ponieważ p-value > 0,01 nie ma podstaw do odrzucenia $H_0$. Region zamieszkania beneficjenta palącego nie ma istotnego wpływu na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu nie będziemy włączać zmiennej *region*.

## Porównanie dobroci dopasowania modeli logitowych 1-5

*Tabela 50*

```
wyniki_oceny_logit
```

```
##         kryterium_AIC     McFadden  Cragg_Uhler
## model_1     267.50855 0.0005067153 0.0009315088
## model_2     262.58731 0.0191730780 0.0347985788
## model_3      89.59333 0.6753427706 0.8094329145
## model_4     267.51946 0.0004653048 0.0008554070
## model_5     263.38931 0.0313031421 0.0563458114
```

**Wnioski na podstawie tabeli 50:**

Najlepszym modelem jest model 3, ponieważ dla kryterium AIC przyjmuje najmniejsze wartości natomiast dla kryterium McFadden i Cragg Uhlera przyjmuje największe wartości.

Do modelu nie można dodać żadnej dodatkowej zmiennej objaśniającej, ponieważ żadna ze zmiennych poza *bmi* nie jest statystycznie istotna.

**Wybór i interpretacja modelu** Wybieramy model charges ~ *bmi* -> logit3.

logit(p)= -24.147463 + 0.816106 * *bmi*

logit(p)=ln(p/1-p) p/(1-p)=exp(-24.147463+0.816106 * *bmi)*

*Tabela 51*

```
logit3$coefficients

## (Intercept)         bmi
##  -24.147463    0.816106

##  exp(bi)

exp(logit3$coefficients)

##  (Intercept)          bmi
## 3.257543e-11 2.261676e+00

##  exp(0.5*bi)

exp(0.5*logit3$coefficients[2])

##      bmi
## 1.503887

##  exp(2*bi)

exp(2*logit3$coefficients[2])

##      bmi
## 5.115177
```

## Interpretacje zmiennych na podstawie tabeli 51:

- *exp($\beta_0$)=3.257543e-11*, gdzie $\beta_0$ to wyraz wolny => interpretuje się jako szansę zdarzenia w grupie referencyjnej (xi=0). Nie posiada interpretacji.
- *exp($\beta_1$) = 2.261676 => (exp($\beta_1$)-1)*100%= 126.16%*

Jeżeli bmi beneficjenta palącego wzrośnie o 1 jednostkę w skali bmi, to szansa na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej, wzrośnie średnio o 126.16%.

- *exp(0.5* $\beta_2$) = 1.503887 => (exp(0.5* $\beta_2$)-1)*100%= 50.39%*

Jeżeli bmi beneficjenta palącego wzrośnie o 0.5 jednostki w skali bmi, to szansa na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej, wzrośnie średnio o 50.39%.

- *exp(2* $\beta_3$) = 5.115177 => (exp(2* $\beta_3$)-1)*100%= 411.52%*

Jeżeli bmi beneficjenta palącego wzrośnie o 2 jednostki w skali bmi, to szansa na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej, wzrośnie średnio o 411.52%.

*Tabela 52*

```
predict(logit3,data.frame(bmi=c(17,22,27,32,37,42)),type="response")
```

```
##            1            2            3            4            5            6
## 3.452953e-05 2.039251e-03 1.078780e-01 8.773884e-01 9.976441e-01 9.999601e-01
```

**Wnioski na podstawie tabeli 52:**
Spodziewamy się, że u beneficjentów palących z bmi wynoszącym:

- *17 jednostek (niedowaga), prawdopodobieństwo kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.00003452953,*
- 22 jednostki (waga prawidłowa), prawdopodobieństwo kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.002039251,
- *27 jednostek (nadwaga), prawdopodobieństwo kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.107878,*
- 32 jednostki (otyłość I stopnia), prawdopodobieństwo kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.8773884,
- *37 jednostek (otyłość II stopnia), prawdopodobieństwo kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.9976441,*
- 42 jednostki (otyłość III stopnia), prawdopodobieństwo kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.9999601.

## Estymacja modelu dwumianowego probitowego

Estymujemy model dla zmiennej dychotomicznej/binarnej Y *family = binomial* z funkcją wiążącą probit *link = probit*

*Tabela 53*

```
probit1 <- glm(charges ~ bmi, data = dane9_uczacy, family = binomial(
link=probit))
summary(probit1)$coefficients
```

```
##                Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -10.1935998 1.36664092 -7.458872 8.726625e-14
## bmi           0.3468573 0.04623625  7.501847 6.292477e-14
```

Model probitowy – interpretacja parametrów sprowadza się do stwierdzenia, czy dana zmienna jest stymulantą (gdy $\beta_i > 0$), czy destymulantą modelu (gdy $\beta_i < 0$).

**Wniosek:** Jak widać w tabeli 53 zmienna BMI jest istotnie statystyczną. Zmienna BMI jest stymulantą.

## Porównanie dobroci dopasowania modeli logit3 i probit1

*Tabela 54*

```
wyniki_oceny_logit_probit
```

```
##                kryterium_AIC  McFadden Cragg_Uhler
## model_logit_3       89.59333 0.6753428   0.8094329
## model_probit_1     102.00675 0.6282584   0.7740472
```

**Wnioski na podstawie tabeli 54:**

Lepszym modelem okazał się model logitowy, ponieważ posiada zdecydowanie mniejszą wartość kryterium Akaike oraz wyraźnie wyższe wartości dla kryterium McFadden i Cragg Uhler.

## Porównanie jakości predykcji modeli logit3 i probit1

Tablice trafności dla wybranego punktu odcięcia p*
Niech p* = proporcja z próby uczącej

*Tabela 55*

```
p <- table(dane9_uczacy$charges)[2]/nrow(dane9_uczacy)
```

```
## Tablica trafności dla modelu logitowego - próba ucząca
```

```
##            przewidywane
## obserwowane   0   1
##           0  82   3
##           1   5 102
```

```
## Tablica trafności dla modelu probitowego - próba ucząca
```

```
##            przewidywane
## obserwowane   0   1
##           0  80   5
##           1   5 102
```

```
## Tablica trafności dla modelu logitowego - próba testowa
```

```
##            przewidywane
## obserwowane  0  1
##           0 39  0
##           1  1 42
```

```
## Tablica trafności dla modelu probitowego - próba testowa
```

```
##            przewidywane
## obserwowane  0  1
##           0 37  2
##           1  1 42
```

## Miary jakości predykcji

Miary oparte na tablicy trafności dla wybranego punktu odcięcia p*

Poniższa funkcja *miary_pred* została została określona dla argumentów: *model* (model dwumianowy), *dane* (np. zbiór uczący, testowy), *Y* (obserwowane Y 0-1 w analizowanym zbiorze danych).

**Ocena zdolności predykcyjnej na zbiorze uczącym**

*Tabela 56*

```
##                   ACC          ER      SENS      SPEC       PPV       NPV
## model_logit  0.9583333 0.04166667 0.953271 0.9647059 0.9714286 0.9425287
## model_probit 0.9479167 0.05208333 0.953271 0.9411765 0.9532710 0.9411765
```

**Ocena zdolności predykcyjnej na zbiorze testowym**

*Tabela 57*

```
##                   ACC          ER        SENS      SPEC       PPV       NPV
## model_logit    0.9878049  0.01219512  0.9767442  1.0000000  1.0000000 0.9750000
## model_probit   0.9634146  0.03658537  0.9767442  0.9487179  0.9545455 0.9736842
```

**Wnioski na podstawie tabel 56 i 57**

Dla zbioru uczącego (tabela 56) i testowego (tabela 57) lepszy okazał się model logitowy pod względem jakości predykcji.

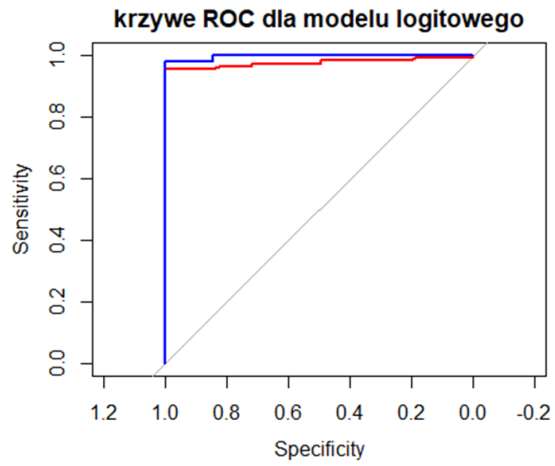
Poprzez porównanie wyników można stwierdzić, że model nie był przeuczony (przystosowany tylko dla zbioru uczącego).

## Krzywa ROC

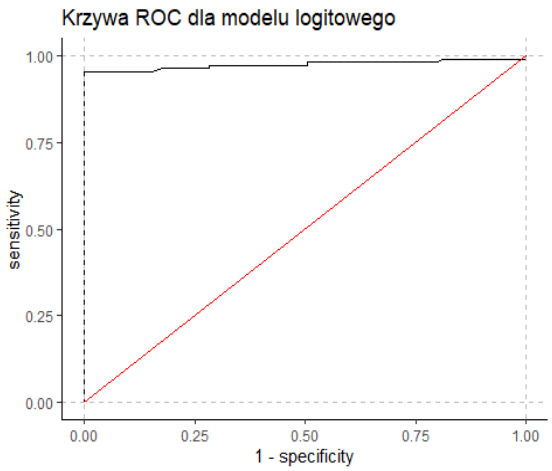
Krzywa ROC prezentuje jakość predykcji modelu dla wszystkich możliwych punktów odcięcia p* (jest niezależna od wyboru p*). Dla modeli oszacowanych na zbiorze uczącym porównana została poniżej jakość predykcji na zbiorze uczącym i testowym. Proszę sprawdzić, czy jakość predykcji dla zbioru testowego nie pogorszyła się znacząco w stosunku do jakości predykcji dla zbioru uczącego.

krzywa czerwona - ROC wyznaczona na zbiorze uczącym
krzywa niebieska - ROC wyznaczona na zbiorze testowym

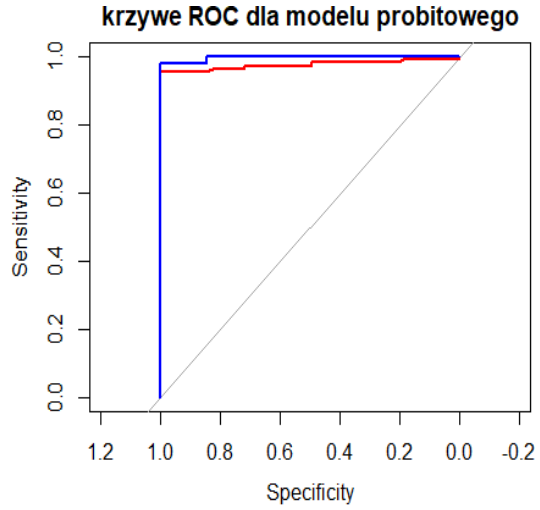
*Wykres 43*



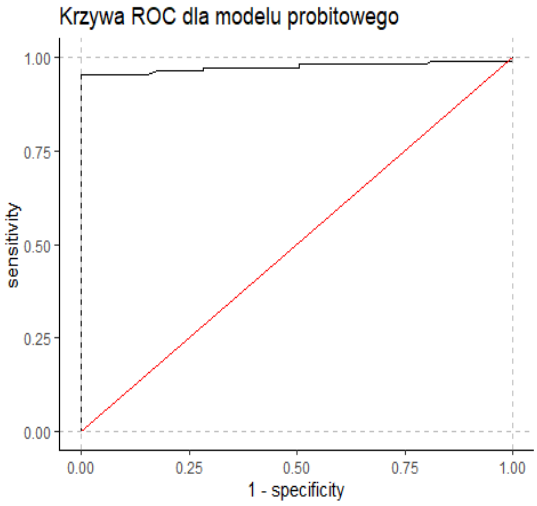

*Wykres 44*




*Wykres 45*




*Wykres 46*




**Pole powierzchni pod krzywą ROC dla zbioru uczącego (model logitowy i probitowy)**
*Tabela 58*

```
## pole_AUC_logit  0.9742166
## pole_AUC_probit 0.9742166
```

**Pole powierzchni pod krzywą ROC dla zbioru testowego (model logitowy i probitowy)**

*Tabela 59*

```
## pole_AUC_logit  0.9964222
## pole_AUC_probit 0.9964222
```

**Wnioski na podstawie wykresów 40-43 oraz tabel 58 i 59**

Dla zbioru uczącego (tabela 58) i testowego (tabela 59):

Na podstawie pola pod krzywą ROC można stwierdzić, że oba modele: `logit` oraz `probit` mają wystarczająco dobrą zdolność predykcyjną. Wartości spełniają równanie 0,5<=AUC<=1.

## Podsumowanie:

W pierwszej kolejności stworzono *zbiór uczący i testowy na podstawie zmiennej objaśnianej charges.* Zbiór uczący posłużył do budowy modelu, a zbiór testowy posłużył do oceny modelu. Następnie stworzono modele logitowe na podstawie każdej ze zmiennych. Najlepszy model logitowy powstał w oparciu o zmienną bmi, bo tylko przy niej odrzucamy $H_0$ oraz zmienna miała najniższe kryterium Akaike. Na podstawie zmiennej bmi stworzono również model probitowy. Po porównaniu dobroci dopasowania modeli logit3 i probit1, lepszy okazał się model logit, ponieważ miał niższe kryterium AIC oraz wyższe miary pseudo $R^2$. Po przeprowadzeniu miar jakości predykcji lepszym modelem również okazał się model logitowy. Podczas wyliczania pola powierzchni pod krzywą ROC dla zbioru uczącego i testowego obydwa modele spełniają równanie 0,5<=AUC<=1. Pole powierzchni okazało się lepsze na zbiorze testowym.

# 3. ETAPY BUDOWY MODELI DLA NIEPALĄCYCH BENEFICJENTÓW

- Każdy wykres oraz tabela zostały wygenerowane w programie R Studio (źródło).
- W testach statystycznych przyjmujemy poziom istotności alfa = 0.01.
- W projekcie zastosowano kryterium dopasowania modelu wg. kryterium informacyjnego Akaike.

## 3.1. W pierwszym etapie dokonano budowy, estymacji, a kolejno weryfikacji otrzymanych modeli liniowych objaśniających koszty leczenia rozliczane przez ubezpieczenie zdrowotne dla beneficjentów niepalących. Estymacji dokonano klasyczną metodą MNK - metoda najmniejszych kwadratów.

**Niezbędne pakiety:**

```
library("car") # funkcja vif()
library("ggplot2") # wykresy - funkcja ggplot()
library("lmtest") # testy diagnostyczne modeli lm
library("pscl")
library("pROC") #funkcje roc, auc
```

### Wczytanie danych i podstawowe statystyki opisowe dla poszczególnych zmiennych zostały zaprezentowane w tabeli 60:

*Tabela 60*

```
dane1 <-  read.table("niepalacy.csv", header=TRUE, sep =";",dec=",")
dane1$sex<-as.factor(dane1$sex)
dane1$region<-as.factor(dane1$region)
summary(dane1)
```

```
##       age            sex           bmi           children          region
##  Min.   :18.00   female:134   Min.   :15.96   Min.   :0.000   northeast:68
##  1st Qu.:25.25   male  :140   1st Qu.:26.03   1st Qu.:0.000   northwest:70
##  Median :38.00                Median :31.04   Median :1.000   southeast:73
##  Mean   :38.28                Mean   :30.86   Mean   :1.047   southwest:63
##  3rd Qu.:50.75                3rd Qu.:34.85   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##     charges
##  Min.   : 1136
##  1st Qu.: 3560
##  Median : 6734
##  Mean   : 8156
##  3rd Qu.:11246
##  Max.   :35160
```

Zmienną objaśnianą są koszty leczenia rozliczane przez ubezpieczenie zdrowotne (zmienna charges). Na początku stworzono modele 1-5 objaśniające koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależności od każdej ze zmiennych objaśniających ze z osobna.

## Model liniowy 1 ze zmienną *age*

*Tabela 61*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku beneficjenta niepalącego
m1 <- lm(charges  ~ age, data = dane1)
summary(m1)
```

```
## Call:
## lm(formula = charges ~ age, data = dane1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3184.5 -1966.5 -1427.0  -582.1 22532.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2077.75     833.83  -2.492   0.0133 *
## age           267.37      20.42  13.094   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4809 on 272 degrees of freedom
## Multiple R-squared:  0.3866, Adjusted R-squared:  0.3844
## F-statistic: 171.5 on 1 and 272 DF,  p-value: < 2.2e-16
```

**Postać modelu:** charges = -2077.75 + 267.37 * *age*

Miary dopasowania z tabeli 61:

- Odchylenie standardowe reszt: Se = 4809
- Współczynnik determinacji $R^2$ = 0.3866
- $R^2$ skorygowany = 0.3844

**Wniosek:** Model 1 wyjaśnia 38,66% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

## Model liniowy 2 ze zmienną *sex*

*Tabela 62*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od płci beneficjenta niepalącego
m2 <- lm(charges  ~ sex, data = dane1)
summary(m2)
```

```
## Call:
## lm(formula = charges ~ sex, data = dane1)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
##  -7032  -4586  -1434   3080  27017
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8143.47     530.42  15.353   <2e-16 ***
## sexmale        25.24     742.04   0.034    0.973
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6140 on 272 degrees of freedom
## Multiple R-squared:  4.253e-06,  Adjusted R-squared:  -0.003672
## F-statistic: 0.001157 on 1 and 272 DF,  p-value: 0.9729
```

**Postać modelu:** charges = 8143.47 + 25.25 * *sex*

Miary dopasowania z tabeli 62:

- Odchylenie standardowe reszt: Se = 6140
- Współczynnik determinacji $R^2$ = 4.253e-06
- $R^2$ skorygowany = -0.003672

**Wniosek:** Zmienna *sex* ma nieistotny wpływ na koszty leczenia rozliczane przez ubezpieczenie zdrowotne.

## Model liniowy 3 ze zmienną *bmi*

*Tabela 63*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od bmi beneficjenta niepalącego
m3 <- lm(charges  ~ bmi, data = dane1)
summary(m3)
```

```
## Call:
## lm(formula = charges ~ bmi, data = dane1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7574  -4659  -1284   3111  27110
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7351.65    1850.47   3.973  9.1e-05 ***
## bmi            26.08      58.76   0.444    0.657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6138 on 272 degrees of freedom
## Multiple R-squared:  0.0007238,  Adjusted R-squared:  -0.00295
## F-statistic: 0.197 on 1 and 272 DF,  p-value: 0.6575
```

**Postać modelu:** charges=7351.65 + 26.08 * *bmi*

Miary dopasowania z tabeli 63:

- Odchylenie standardowe reszt: Se = 6138
- Współczynnik determinacji $R^2$ = 0.0007238
- $R^2$ skorygowany = -0.00295

**Wniosek:** Zmienna *bmi* ma nieistotny wpływ na koszty leczenia rozliczane przez ubezpieczenie zdrowotne.

## Model 4 liniowy ze zmienną *children*

*Tabela 64*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od liczby osób na utrzymaniu beneficjenta niepalącego
m4 <- lm(charges  ~ children, data = dane1)
summary(m4)
```

```
## Call:
## lm(formula = charges ~ children, data = dane1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6524  -4797  -1656   3068  27026
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7660.2      488.7  15.676   <2e-16 ***
## children       473.7      305.5   1.551    0.122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6113 on 272 degrees of freedom
## Multiple R-squared:  0.008763,   Adjusted R-squared:  0.005118
## F-statistic: 2.405 on 1 and 272 DF,  p-value: 0.1221
```

**Postać modelu:** charges = 7660.2 + 473.7 * *children*

Miary dopasowania z tabeli 64:

- Odchylenie standardowe reszt: Se = 6113
- Współczynnik determinacji $R^2$ = 0.008763
- $R^2$ skorygowany = 0.005118

**Wniosek:** Model wyjaśnia 0,887% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

## Model liniowy 5 ze zmienną *region*

*Tabela 65*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od obszaru zamieszkania w USA beneficjenta niepalącego
m5 <- lm(charges  ~ region, data = dane1)
summary(m5)
```

```
## Call:
## lm(formula = charges ~ region, data = dane1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7341  -4269  -1454   2890  27447
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       8495.8      743.5  11.426   <2e-16 ***
## regionnorthwest    477.4     1044.0   0.457    0.648
## regionsoutheast  -1056.2     1033.3  -1.022    0.308
## regionsouthwest   -782.8     1072.2  -0.730    0.466
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6131 on 270 degrees of freedom
## Multiple R-squared:  0.01018,    Adjusted R-squared:  -0.0008136
## F-statistic: 0.926 on 3 and 270 DF,  p-value: 0.4286
```

**Postać modelu:**

charges= 8495.8 + 477.4 * *regionnorthwest* - 1056.2 * *regionsoutheast* -782.8 * *regionsouthwest*

Miary dopasowania z tabeli 65:

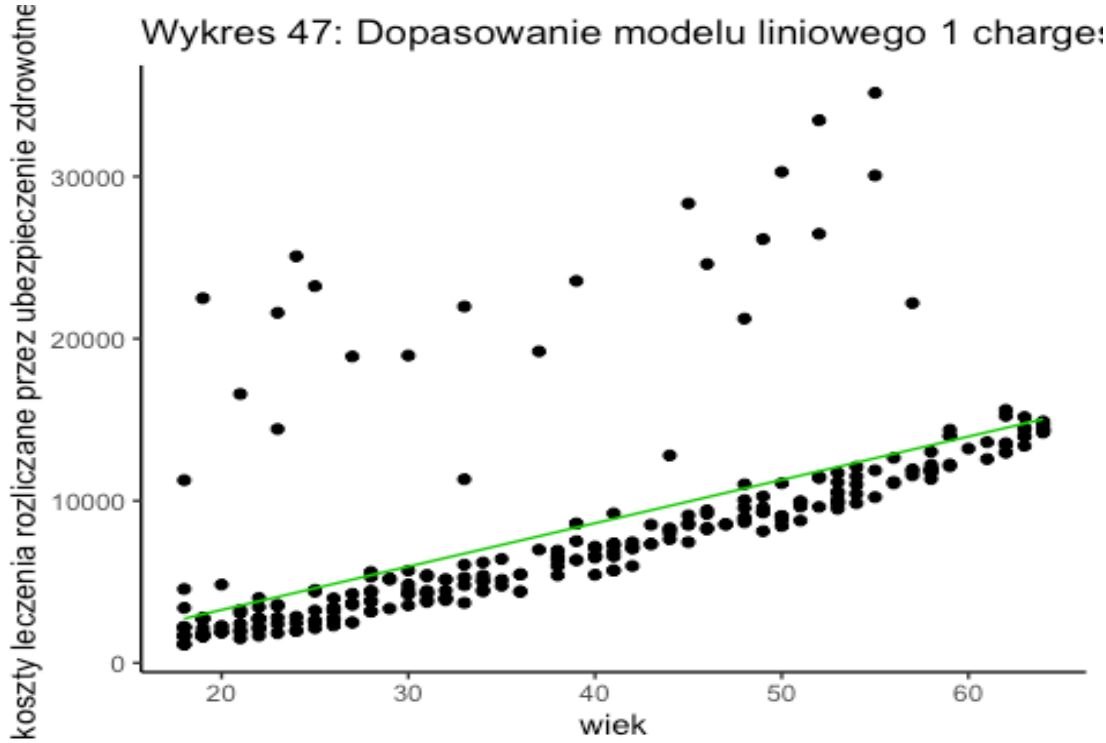
- Odchylenie standardowe reszt: Se = 6131
- Współczynnik determinacji $R^2$ = 0.01018
- $R^2$ skorygowany = -0.0008136

**Wniosek:** Model 5 wyjaśnia 1,02% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

Model liniowy m2, m3, m4 oraz m5 wyjaśniają bardzo mały procent kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne, natomiast model m1 wyjaśniają w większym stopniu w porównaniu do reszty badane zjawisko.

Do dalszej części budowy modelu wybrano tylko zmienną *age*.

**Ponieważ model 1 ma tylko jedną zmienną objaśniającą można pokazać go na wykresie:**

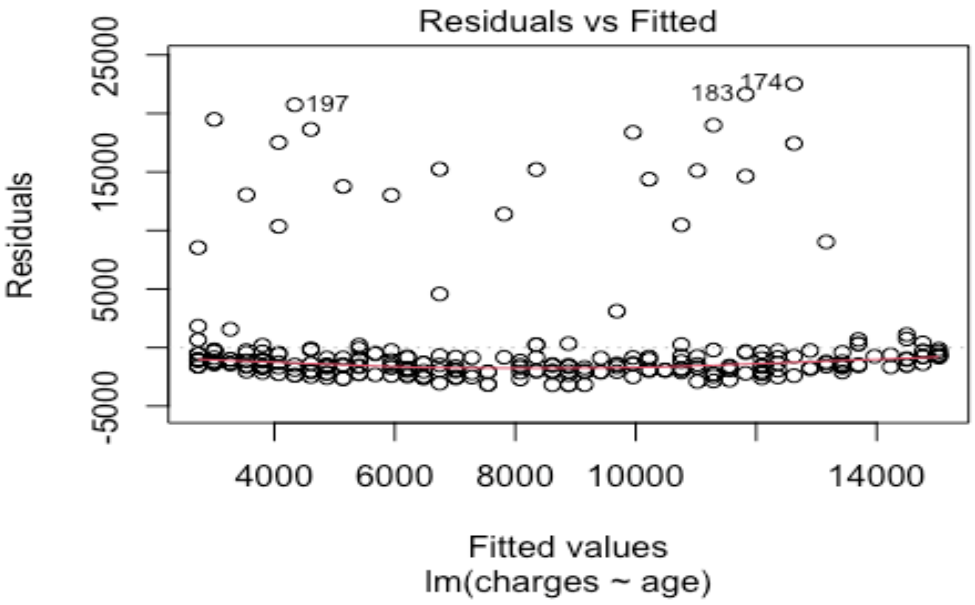



**Wniosek:** Analizując powyższy wykres 1 można zauważyć, że dopasowanie modelu jest dosyć dobre, jednak jest kilka jednostek odstających, które wpływają na obniżone wyjaśnienie badanego zjawiska.

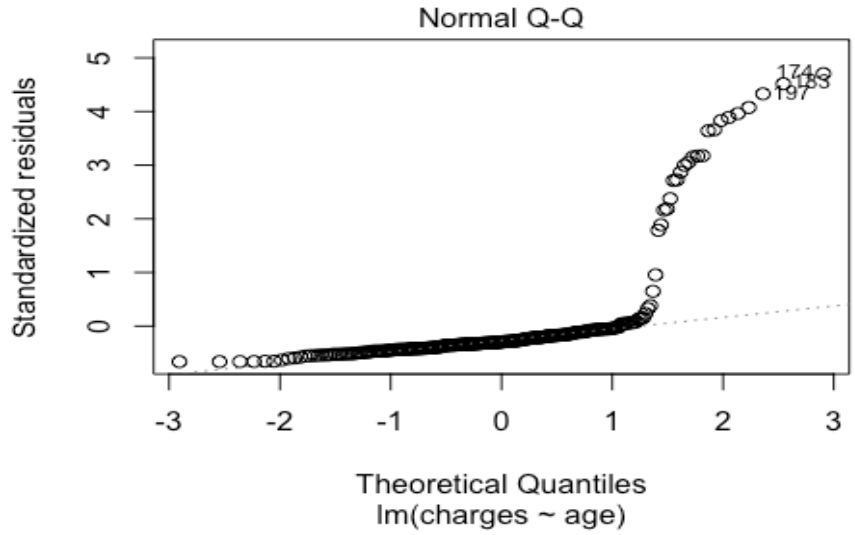
**W celu lepszego zbadania modelu 1 (w tym jednostek wpływowych) wykonano dla niego wykresy diagnostyczne:**

```
plot(m1, which = 1:5)
```

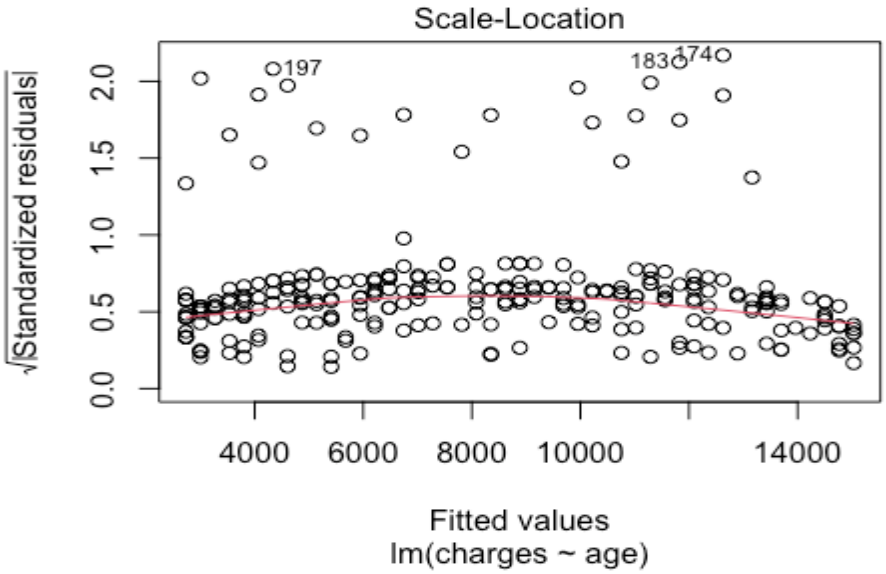
*Wykres 48*



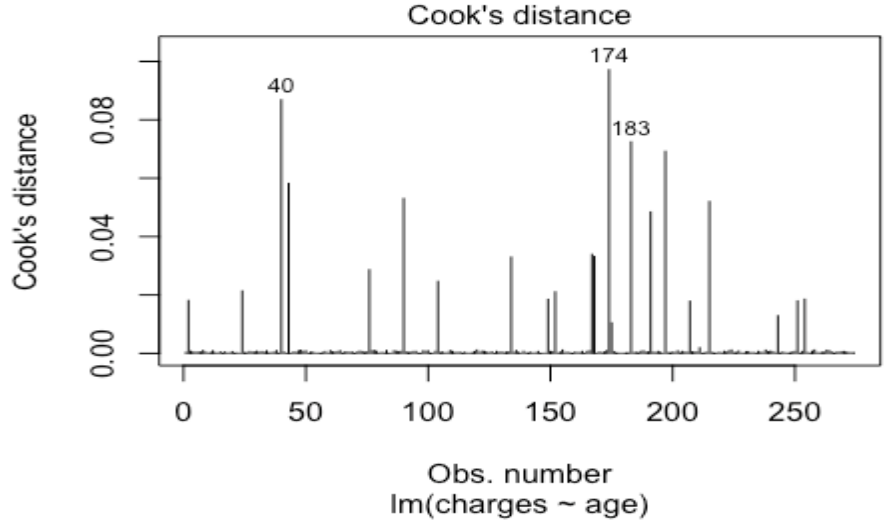
*Wykres 49*



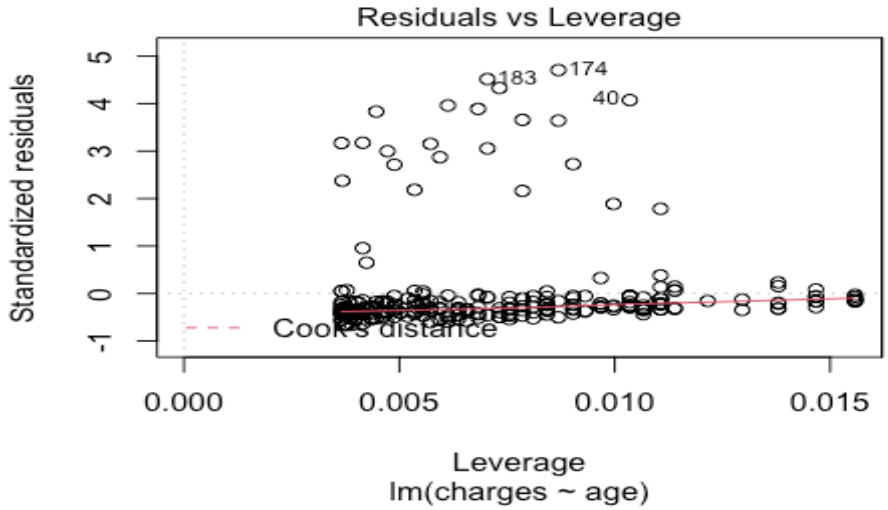

*Wykres 50*




*Wykres 51*



*Wykres 52*




**Wniosek:** Wariancja (rozrzut) reszt jest równomierne rozmieszczony wzdłuż linii poziomej, na poziomie =0, jednak można też zauważyć jednostki odstające. Rozkład reszt nie jest normalny, można zauważyć obserwacje nietypowe o dużych resztach. Wartość średnia reszt jest bliska 0, co wskazuje właściwą postać funkcyjną modelu, jednak występuje duża ilość jednostek odstających. Za pomocą miar: odległość Cooka oraz wskaźnik wpływu leverage (dźwignia) dokonano oceny wpływu poszczególnych obserwacji na parametry strukturalne modelu 1, które zostały kolejno usunięte i zapisane jako *dane2*.

**Usunięcie jednostek odstających z modelu 1**

```
dane2<-dane1[-c(40,174,183,43,90,197,167,191,215,76,134,168,24,104,15
2,207,255,254,149,2,251,175,211,243,50,136,230,28,171,213,177,117,14,
102,127,212,54,179,252),]
```

**Kolejno wykonano testy istotności dla parametrów modelu 1**

Interpretacja testów istotności parametrów **modelu 1** (test F i test t)

**test F:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- F= 171.5
- p-value: <2.2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t= 13.094
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Wiek beneficjenta niepalącego statystycznie istotnie wpływa na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

## Model liniowy 6, który jest modelem 1 po usunięciu jednostek odstających

*Tabela 66*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku beneficjenta niepalącego po wyeliminowaniu jednostek od
sających
m6 <- lm(charges  ~ age , data = dane2)
summary(m6)

## Call:
## lm(formula = charges ~ age, data = dane2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1673.95  -516.71   -27.66   526.82  1848.62
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3449.393    142.406  -24.22   <2e-16 ***
## age           263.980      3.552   74.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 748.8 on 233 degrees of freedom
## Multiple R-squared:  0.9595, Adjusted R-squared:  0.9593
## F-statistic:  5522 on 1 and 233 DF,  p-value: < 2.2e-16
```

**Postać modelu:**

charges = -3449.393  +263.980 * *age*

Miary dopasowania z tabeli 66:

- Odchylenie standardowe reszt: Se = 748.8
- Współczynnik determinacji $R^2$ = 0.9595
- $R^2$ skorygowany = 0.9593

Model wyjaśnia 95.95% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne.

**Interpretacja modelu 6**

- Wyraz wolny $\beta_0$ = -3449.393 w tym modelu nie ma interpretacji, ponieważ badanie nie obejmowało osób w wieku 0 lat (niemowląt).

- Współczynnik przy zmiennej “age” $\beta_1 = 263.980$:
  - Jeżeli wiek beneficjenta niepalącego wzrośnie o 1 rok, to koszty leczenia rozliczane przez ubezpieczenie zdrowotne wzrosną średnio o 263.98 $ (ceteris paribus).
  - Jeżeli wiek beneficjenta niepalącego wzrośnie o 10 lat, to koszty leczenia rozliczane przez ubezpieczenie zdrowotne wzrosną średnio o 2639.8 $ (ceteris paribus).

Interpretacja testów istotności parametrów **modelu 6** (test F i test t)

**test F:**

- $H_0 = \beta i = 0$
- $H_1 = \beta i \neq 0$
- F= 5522
- p-value: < 2.2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t:**

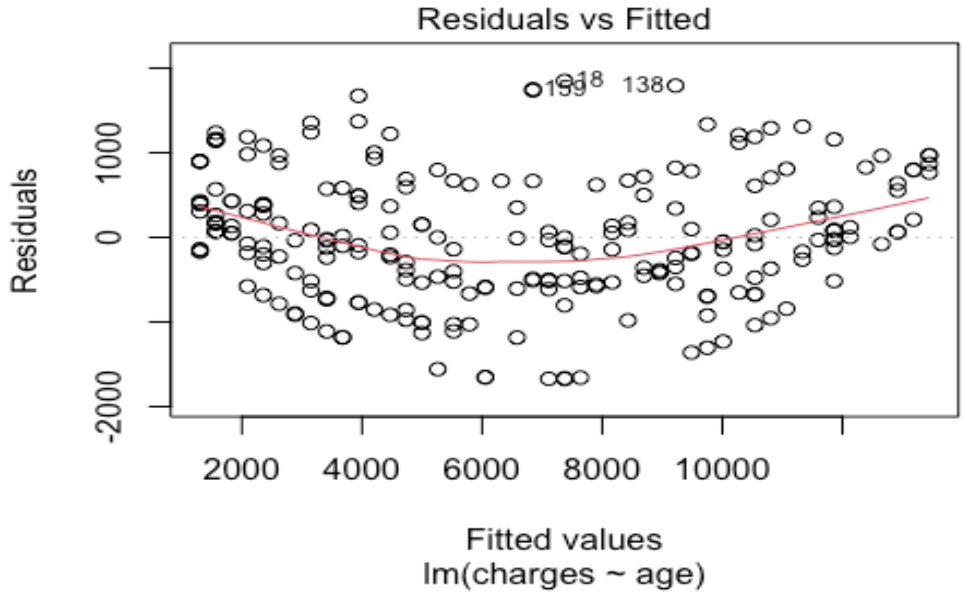
- $H_0 = \beta i = 0$
- $H_1 = \beta i \neq 0$
- t= -24.22
- p-value: <2e-16
- Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**Wniosek:** Parametr *age* w modelu istotnie różnią się od 0 na wszystkich poziomach istotności.

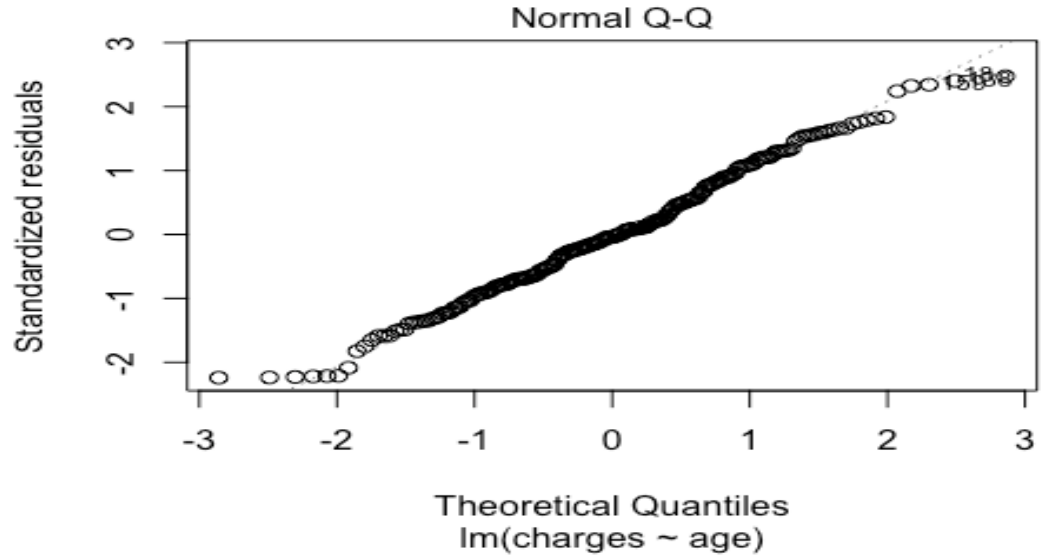
**Wykresy diagnostyczne modelu liniowego 6:**

```
plot(m6, which = 1:3)
```

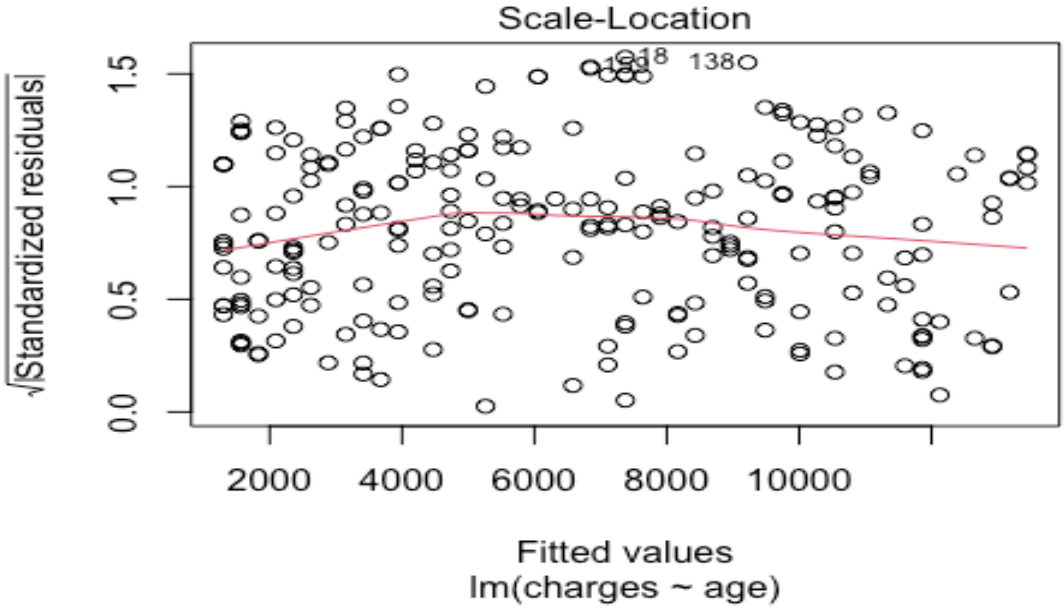
*Wykres 53*



*Wykres 54*



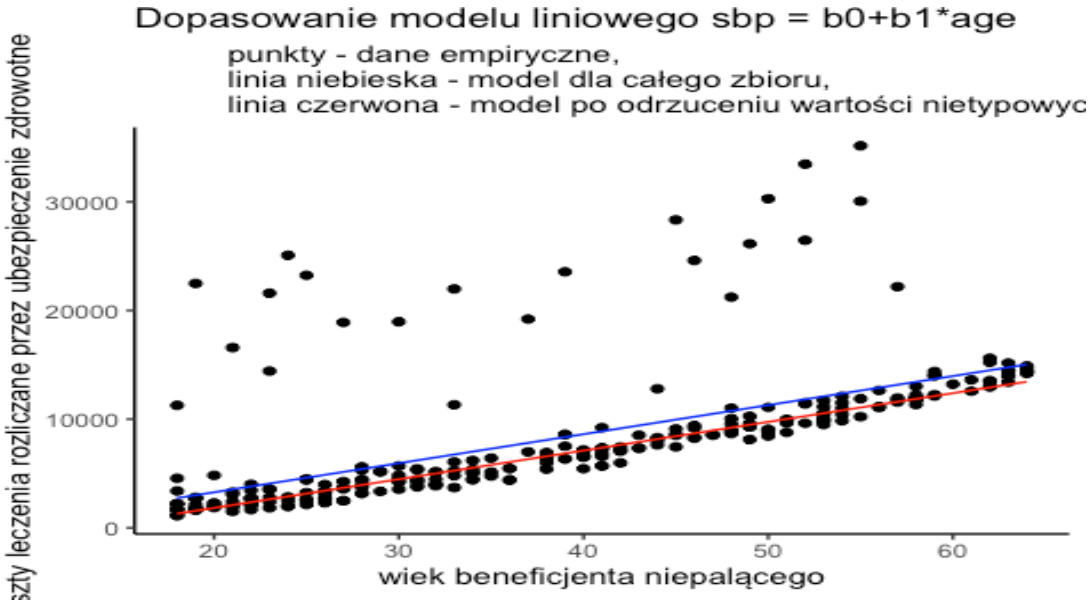

*Wykres 55*




**Wniosek:** Dzięki wykresom 53-55 można stwierdzić, że wariancja (rozrzut) reszt jest równomierne rozmieszczony wzdłuż linii poziomej, na poziomie=0. Rozkład reszt jest normalny. Wartość średnia reszt jest bliska 0, co wskazuje właściwą postać funkcyjną modelu.

**Porównanie na wykresie modelu 1 (przed odrzuceniem jednostek nietypowych) oraz modelu 6 (model 1 po odrzuceniu jednostek nietypowych):**

*Wykres 56*




**Testy statystyczne dla modelu 6**

## Test normalności Shapiro-Wilka dla reszt modelu 6

*Tabela 67*

```
shapiro.test(m6$residuals)
```

```
##  Shapiro-Wilk normality test
##
## data:  m6$residuals
## W = 0.99153, p-value = 0.192
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 67) to nie ma podstaw do odrzucenia $H_0$. Reszty modelu mają rozkład normalny.

## Test Breuscha-Pagana jednorodności wariancji reszt modelu 6

*Tabela 68*

```
bptest(m6)
```

```
##  studentized Breusch-Pagan test
##
## data:  m6
## BP = 0.052493, df = 1, p-value = 0.8188
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 68) to nie ma podstaw do odrzucenia $H_0$ o jednorodności wariancji reszt.

### Test Durbina-Watsona niezależności reszt modelu 6

*Tabela 69*

```
dwtest(m6, order.by = ~age, data = dane2)
```

```
##  Durbin-Watson test
##
## data:  m6
## DW = 1.7408, p-value = 0.01964
## alternative hypothesis: true autocorrelation is greater than 0
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 69) to nie ma podstaw do odrzucenia $H_0$, mówiącej o niezależności reszt.

### Test rainbow na liniowość modelu modelu 6

*Tabela 70*

```
raintest(m6)
```

```
##  Rainbow test
##
## data:  m6
## Rain = 0.83803, df1 = 118, df2 = 115, p-value = 0.8295
```

**Wniosek:**

Ponieważ p-value > 0.01 (tabela 70) to nie ma podstaw do odrzucenia $H_0$, mówiącej o liniowości modelu.

## Podsumowanie:

W pierwszej kolejności stworzono ***modele (1-5)*** objaśniające koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależności od każdej ze zmiennych z osobna (*age, sex, bmi, childrem i region*). ***Model 1*** z jedną zmienną objaśniającą *age* wyjaśnia 38.66% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów niepalących. Ponieważ w największym stopniu badane zjawisko wyjaśnia jako jedyna zmienna *age* wybrano ją jako główną i jedyną zmienną objaśniającą w procesie budowania modelu. W związku z tym powstał ***model 6***, który jest rozwinięciem ***modelu 1*** po usunięciu jednostek odstających. ***Model 6*** wyjaśnia 95,95% kształtowania się kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów niepalących. W *modelu 6* znacznie zmniejszyło się odchylenie standardowe reszt oraz model spełnia wszystkie założenia statystyczne. Model wyjaśnia w bardzo wysokim stopniu badane zjawisko, jednak można poszukać lepszych postaci modelu.

### 3.2. Estymacja, a kolejno weryfikacja otrzymanych modeli klasy GLM, objaśniających koszty leczenia rozliczane przez ubezpieczenie zdrowotne dla beneficjentów niepalących.

**Wybór postaci GLM dla zmiennej Y=charges.**

## Estymacja modelu 7 klasy GLM z funkcją wiążącą identycznościową (model liniowy) - estymacja metodą MNW.

*Tabela 71*

```
#h: koszty leczenia rozliczane przez ubezpieczenie zdrowotne w zależn
ości od wieku beneficjenta niepalącego po wyeliminowaniu jednostek od
stających
m7 <- glm(charges  ~ age , data = dane2, family = gaussian)
summary(m7)

##
## Call:
## glm(formula = charges ~ age, family = gaussian, data = dane2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1673.95   -516.71    -27.66    526.82   1848.62
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3449.393    142.406  -24.22   <2e-16 ***
## age           263.980      3.552   74.31   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 560631.3)
##
##     Null deviance: 3226407289  on 234  degrees of freedom
## Residual deviance:  130627101  on 233  degrees of freedom
## AIC: 3781.5
##
## Number of Fisher Scoring iterations: 2
```

**Wnioski na podstawie tabeli 71:**

Minimalną wartością tego modelu jest -1673.95, a najwyższą: 1848.62. Wartość środkowa wynosi -27.66. Kwartyl pierwszy- 25% obserwacji wynosi -516.71, a 75% obserwacji jest położonych powyżej tej wartości. Kwartyl trzeci - 75% obserwacji położonych jest poniżej 526.82, a 25% obserwacji położonych jest powyżej tej wartości.

**test t dla wyrazu wolnego:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t=-24.22
- p-value <2e-16

Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t dla zmiennej age:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t=74.31
- p-value: <2e-16

**Wniosek:** Odrzucamy $H_0$ na korzyść $H_1$. Wiek statystycznie istotnie wpływa na poziom kosztów leczenia osób niepalących rozliczanych przez ubezpieczenie zdrowotne.

Kryterium AIC wynosi 3781.5.

**Postać modelu:** charges = -3449.39 + 263.98 * age

Ponieważ p-value < 0.01 to odrzucamy $H_0$ na korzyść $H_1$, wiek statystycznie istotnie wpływa na poziom kosztów leczenia osób niepalących rozliczanych przez ubezpieczenie zdrowotne.

## Estymacja modelu 8 - klasy GLM z funkcją wiążącą logarytmiczną (model log normalny)

*Tabela 72*

```
dane3<-dane1[-c(40,174,183,43,90,197,167,191,215,76,134,168,24,104,152,207,255,254,149,2,251,175,211,243,50,136,230,28,171,213,177,117,14,102,127,212,54,179,252,163,226,72,122,144,140,246,13,203,223,33,105,48,18,235,20),]

m8 <- glm(charges  ~ age , data = dane3, family = gaussian(link = "log"))
summary(m8)

##
## Call:
## glm(formula = charges ~ age, family = gaussian(link = "log"),
##     data = dane3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1447.64   -802.90    -89.84    456.54   2586.25
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.133537   0.039369  181.20   <2e-16 ***
## age         0.040198   0.000808   49.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 715580.2)
##
##     Null deviance: 2453855375  on 218  degrees of freedom
## Residual deviance:  155282707  on 217  degrees of freedom
## AIC: 3577.8
##
## Number of Fisher Scoring iterations: 5
```

**Na podstawie tabeli 72 można stwierdzić:**

Minimalną wartością tego modelu jest -1447.64, a najwyższą: 2586.25. Wartość środkowa wynosi -89.84. Kwartyl pierwszy- 25% obserwacji wynosi -802.90, a 75% obserwacji jest położonych powyżej tej wartości. Kwartyl trzeci - 75% obserwacji położonych jest poniżej 456.54, a 25% obserwacji położonych jest powyżej tej wartości.

**test t dla wyrazu wolnego:**

- $H_0 = \beta i = 0$
- $H_1 = \beta i \neq 0$
- t=181.20
- p-value <2e-16

Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t dla zmiennej age:**

- $H_0 = \beta i = 0$
- $H_1 = \beta i \neq 0$
- t=49.75.
- p-value: <2e-16

**Wniosek:** Odrzucamy $H_0$ na korzyść $H_1$. Wiek statystycznie istotnie wpływa na poziom kosztów leczenia osób niepalących rozliczanych przez ubezpieczenie zdrowotne.

Kryterium AIC wynosi 3577.8.

**Postać modelu**: charges = 7.133537 + 0.040198 * age

Ponieważ p-value < 0.01 to odrzucamy $H_0$ na korzyść $H_1$, wiek statystycznie istotnie wpływa na poziom kosztów leczenia osób niepalących rozliczanych przez ubezpieczenie zdrowotne.

## Estymacja modelu 9 GLM z funkcją wiążącą odwrotną

*Tabela 73*

```
dane4<-dane1[-c(40,174,183,43,90,197,167,191,215,76,134,168,24,104,152,207,255,254,149,2,251,175,211,243,50,136,230,28,171,213,177,117,14,102,127,212,54,179,252,144,140,226,246,163,13,203,72,122,33,105,223,260,199,235,20,48,3,150,194,241,5,231,269,59,63,66,219,165,240),]

m9 <- glm(charges  ~ age , data = dane4, family = gaussian(link = "inverse"))
summary(m9)
## Call:
## glm(formula = charges ~ age, family = gaussian(link = "inverse"),
##     data = dane4)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2582.3  -1062.1   -195.7    720.7   3139.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.172e-04  1.146e-05   36.40   <2e-16 ***
## age         -6.166e-06  2.270e-07  -27.16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1315315)
##
##     Null deviance: 1870810698  on 204  degrees of freedom
## Residual deviance:  266997056  on 203  degrees of freedom
## AIC: 3474.1
##
## Number of Fisher Scoring iterations: 6
```

**Na podstawie tabeli 73 można stwierdzić:**

Minimalną wartością tego modelu jest -2582.3, a najwyższą: 3139.7. Wartość środkowa wynosi -195.7. Kwartyl pierwszy- 25% obserwacji wynosi -1062.1, a 75% obserwacji jest położonych powyżej tej wartości. Kwartyl trzeci - 75% obserwacji położonych jest poniżej 720.7, a 25% obserwacji położonych jest powyżej tej wartości.

**test t dla wyrazu wolnego:**

- $H_0 = \beta i = 0$
- $H_1 = \beta i \neq 0$
- t=36.40
- p-value <2e-16

Wniosek: Odrzucamy $H_0$ na korzyść $H_1$.

**test t dla zmiennej age:**

- $H_0$= βi = 0
- $H_1$= βi ≠ 0
- t=-27.16.
- p-value: <2e-16

**Wniosek:** Odrzucamy $H_0$ na korzyść $H_1$. Wiek statystycznie istotnie wpływa na poziom kosztów leczenia osób niepalących rozliczanych przez ubezpieczenie zdrowotne.

Kryterium AIC dla tego modelu wynosi 3474.1.

**Postać modelu:** charges= 0.00004172 - 0.0000006166 *age

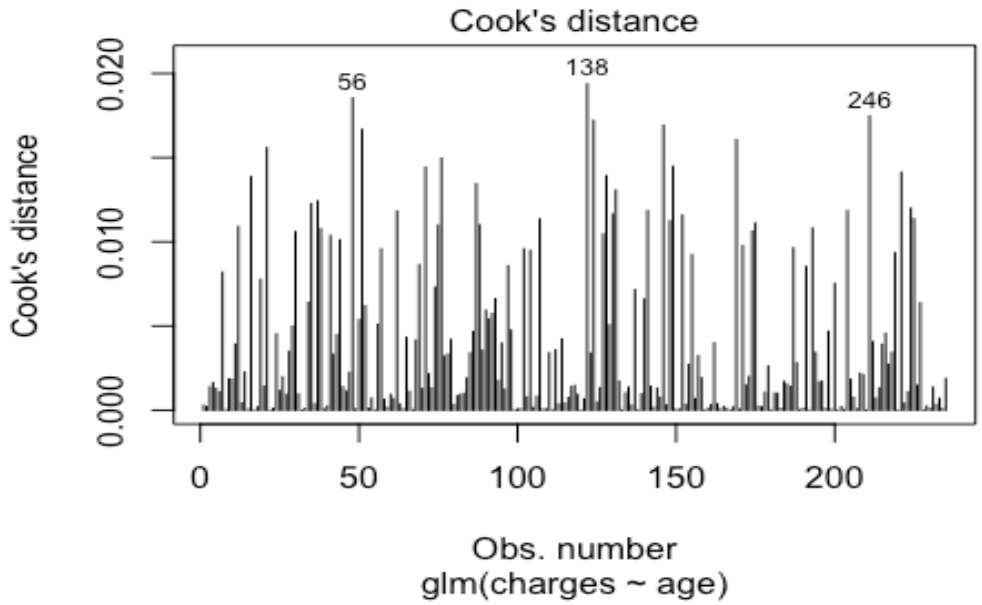
**Wnioski:**

Ponieważ p-value < 0.01 to odrzucamy $H_0$ na korzyść $H_1$. wiek statystycznie istotnie wpływa na poziom kosztów leczenia osób niepalących rozliczanych przez ubezpieczenie zdrowotne.

**Wykresy diagnostyczne dla modeli 7-9:**

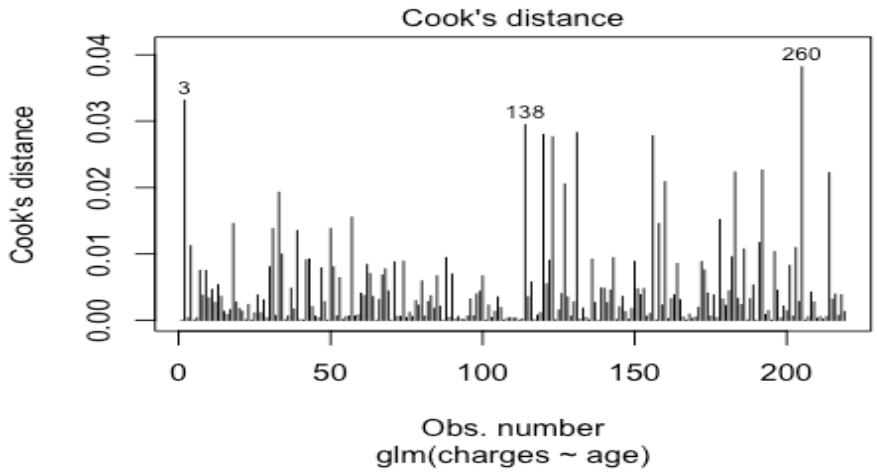
*Wykres 57*

```
plot(m7, which = 4)
```



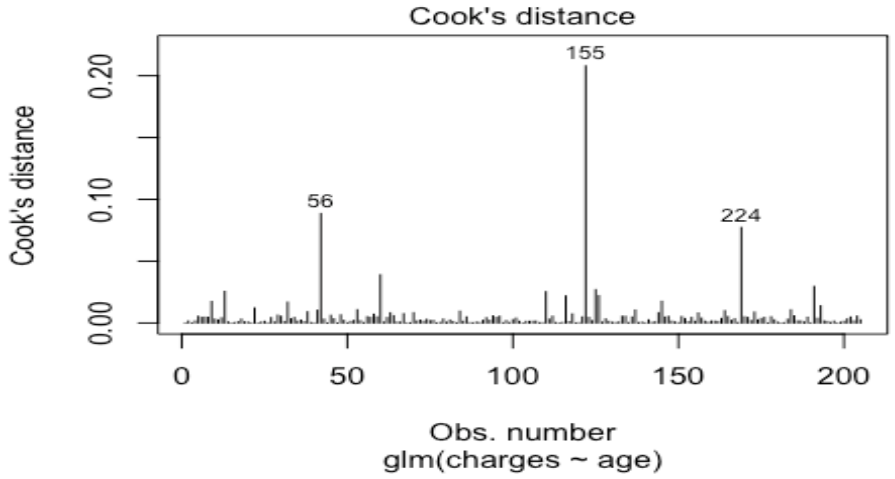
*Wykres 58*

```
plot(m8, which = 4)
```



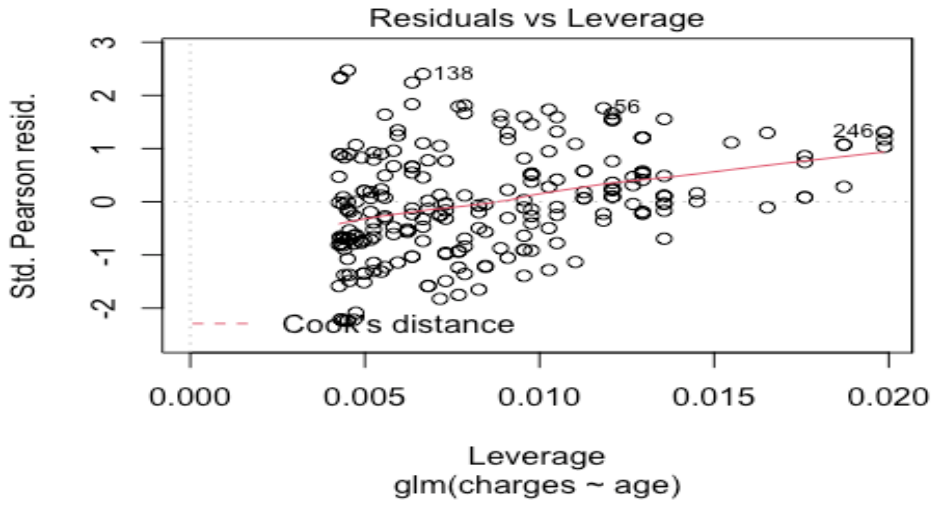

*Wykres 59*

```
plot(m9, which = 4)
```



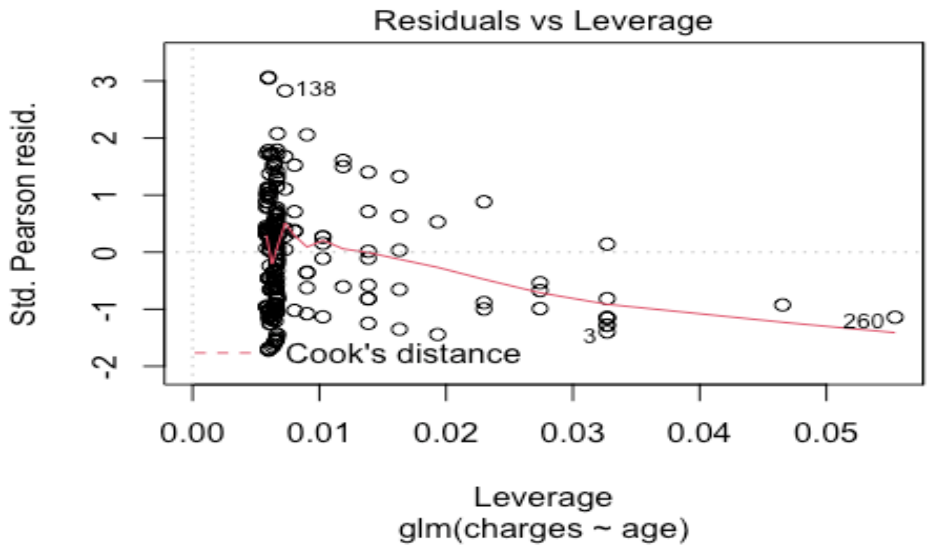

*Wykres 60*

```
plot(m7, which = 5)
```



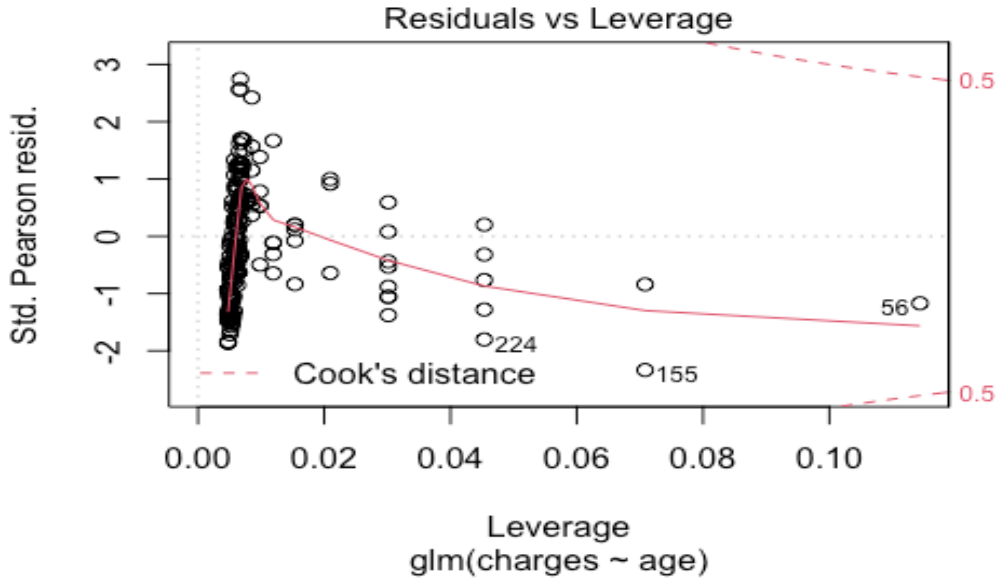
*Wykres 61*

```
plot(m8, which = 5)
```




*Wykres 62*

```
plot(m9, which = 5)
```




**Interpretacje do wykresów 57-62:**

Odległość Cooka jest liczona dla każdej obserwacji (3 największe oznaczone) Oceniając model możemy wyróżnić obserwacje o dużych resztach (te odstające), jak i wpływowe (wpływowe nie zawsze są tymi odstającymi - one mają duży wpływ na oszacowanie parametrów strukturalnych).

Odległość Cooka jest liczona dla każdej obserwacji - liczony jest model na podstawie całego zbioru danych i bez tej obserwacji i budowana miara na podstawie jak zmieniły się te współczynniki beta. Jak jest zbyt duża to jest inny rząd wielkości to nie ma wątpliwości, że jest wpływowa. Stawiamy umowne granice.

Odległość Cooka i wskaźnik wpływu dla każdej zmiennej objaśniającej z osobna, mierzony dla poszczególnych obserwacji odstępstw o zmiennej objaśniającej xi od jej średniego poziomu.

Obserwacje, które są nietypowe mogą być: bo mają dużą resztę, bo x odbiegają, bo wpływ na wskaźniki beta. Tutaj Linia Cooka jest poza naszymi zmiennymi.

## Bonferroni Outlier Test

*Tabela 74*

```
outlierTest(m7, n.max = Inf)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 18 2.502293           0.012339           NA

outlierTest(m8, n.max = Inf)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 159  3.12794          0.0017604      0.38552

outlierTest(m9, n.max = Inf)

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 18 2.792498          0.0052303           NA
```

**Wnioski na podstawie tabeli 74:**

Ponieważ p-value > 0.01 to odrzucamy $H_0$ na korzyść $H_1$ dla modelu 7.

Ponieważ p-value < 0.01 to nie ma podstaw do odrzucenia $H_0$ dla modeli 8 i 9.

Test Bonferroni Outlier nie wykazał obecności obserwacji nietypowych w analizowanych trzech modelach. (test został wykonany dla trzech modeli).

**Porównanie modeli 7-9**

Ocena dopasowania modeli GLM: statystyka odchylenia (deviance), kryterium informacyjne (AIC), miary pseudo-$R^2$. W przypadku family = gaussian statystyka odchylenia = suma kwadratów reszt, zatem odchylenie standardowe reszt (średni błąd szacunku) = (deviance/df)^0.5 Zdefiniowana została funkcja ocena_modelu_GLM licząca powyższe miary. Argumentem tej funkcji jest obiekt klasy GLM.

*Tabela 75*

```
ocena_modeli
```

```
##         odch_std_reszt kryterium_AIC  McFadden Cragg_Uhler
## model_7       748.7532      3781.545 0.1663877   0.9595131
## model_8       845.9246      3577.794 0.1447408   0.9367189
## model_9      1146.8463      3474.111 0.1032038   0.8572827
```

**Interpretacje na podstawie tabeli 75**

Miary pseudo $R^2$ zbudowane na bazie wiarygodności modelu w porównaniu do wiarygodności modelu 0 (tylko z wyrazem wolnym). $R^2$ jaki procent zmienności Y jest wyjaśniany poprzez zmienność zmiennych objaśniających. Tutaj nie ma takiej interpretacji, dlatego jest nazwa pseudo $R^2$ żeby ni interpretować jako $R^2$ w modelu liniowym. Miary pseudo $R^2$ zwykle w dolnych obszarach się znajdują. Na bazie statystyki odchylenia mamy kryteria informacyjne.

Dla rozkładu normalnego statystyka odchylenia jest równa sumie kwadratów reszt.

**Wnioski:**

Najlepiej dopasowany jest model 9, ponieważ ma najniższe kryterium AIC.

*Tabela 76*

```
m9$coefficients
```

```
##   (Intercept)           age
##  4.172154e-04 -6.166228e-06
```

```
## exp(bi)
```

```
exp(m9$coefficients)
```

```
## (Intercept)         age
##   1.0004173   0.9999938
```

**Interpretacje na podstawie tabeli 76**

**Postać modelu 9** można zapisać także jako:

charges = exp( 1.0004173 + 0.9999938 * *age*)

Parametry modelu log normalnego posiadają interpretację:

- Wyraz wolny $\beta_0$ –> exp($\beta_0$) w tym modelu nie ma interpretacji, ponieważ badanie nie obejmowało osób w wieku 0 lat (niemowląt).
- age $\beta_1$ = -6.166228e-06 –> exp($\beta_1$)=0.9999938 –> (exp($\beta_1$)-1)*100%

Jeżeli wiek wzrośnie o jeden rok a pozostałe zmienne nie ulegną zmianie to wysokość składki dla osób niepalących spadnie średnio o 0.01% dla osób tej samej płci.

## Podsumowanie:

W pierwszej kolejności stworzono ***modele glm 7,8 i 9*** oparte na ***modelu 6***. ***Model 8*** ma postać log normalną, natomiast ***model 9*** to model glm z funkcją wiążącą odwrotną. Ponieważ model 6 był dobrze dopasowany – model glm zbudowany jest w oparciu o zmienną age. W każdym z modeli przeprowadzono testy istotności parametrów, wykresy diagnostyczne, Bonferroni Outlier Test. Na podstawie kryterium informacyjnego i miar pseudo $R^2$ został wybrany model 9, ponieważ miał on najmniejszą wartość kryterium Akaike. Na podstawie modelu log odwrotnego została dokonana interpretacja parametrów ilorazu szans.

### 3.3. Budowa i weryfikacja modeli logitowych i probitowego objaśniających koszty leczenia rozliczane przez ubezpieczenie zdrowotne dla beneficjentów niepalących. Zmienna *charges* została podzielona na dwie części: 1-powyżej średniej, 0-poniżej średniej.

**Wczytanie danych i statystyki opisowe dla poszczególnych zmiennych:**

*Tabela 77*

```
dane5 <-  read.table("niepalacyl.csv", header=TRUE,sep =";",dec=",")
dane5$sex<-as.factor(dane5$sex)
dane5$region<-as.factor(dane5$region)
dane5$charges<-as.factor(dane5$charges)
summary(dane5)
```

```
##       age            sex           bmi           children          region
##  Min.   :18.00   female:134   Min.   :15.96   Min.   :0.000   northeast:68
##  1st Qu.:25.25   male  :140   1st Qu.:26.03   1st Qu.:0.000   northwest:70
##  Median :38.00                Median :31.04   Median :1.000   southeast:73
##  Mean   :38.28                Mean   :30.86   Mean   :1.047   southwest:63
##  3rd Qu.:50.75                3rd Qu.:34.85   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##
```

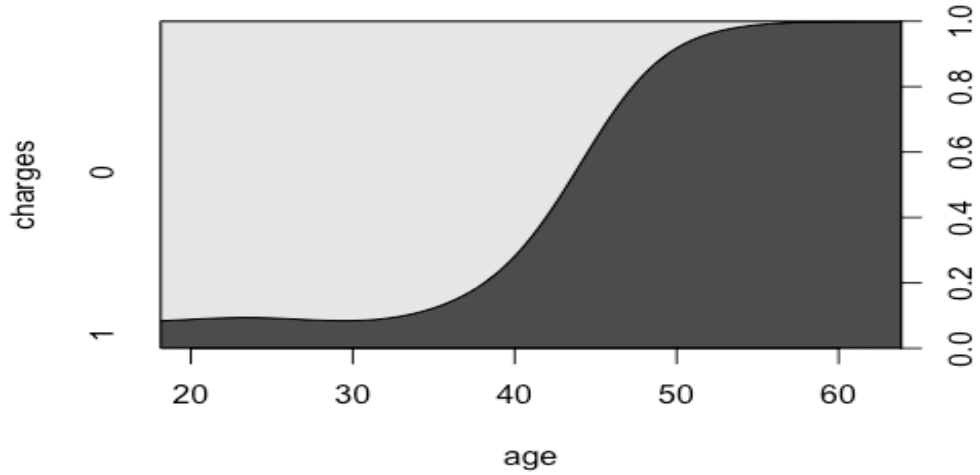
```
charges
##  0:155
##  1:119
```

Jaki jest związek między kosztami leczenia rozliczanymi przez ubezpieczenie zdrowotne powyżej średniej, a potencjalnymi predyktorami?

**Wykresy warunkowych prawdopodobieństw wystąpienia wariantów cechy jakościowej** (kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej) **pod warunkiem, że zmienna ilościowa przyjmuje określony poziom:**

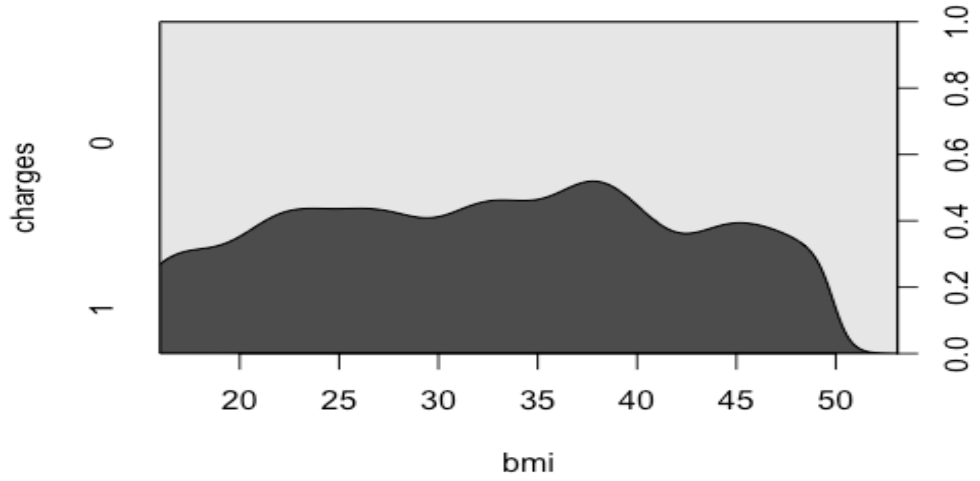
*Wykres 63*

```
cdplot(dane5$age, dane5$charges, xlab = "age", ylab = "charges")
```



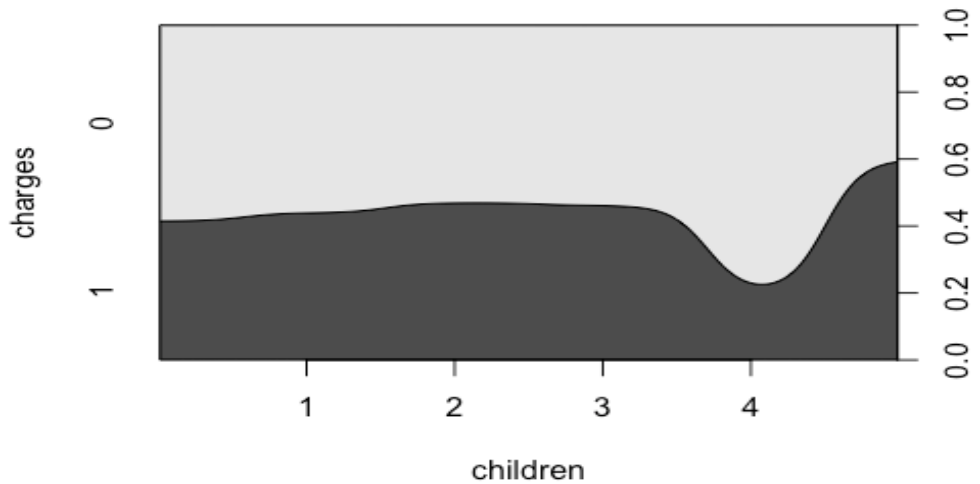

*Wykres 64*

```
cdplot(dane5$bmi, dane5$charges, xlab = "bmi", ylab = "charges")
```




*Wykres 65*

```
cdplot(dane5$children, dane5$charges,xlab="children",ylab="charges")
```



**Wnioski na podstawie wykresów 63-65:** Wraz ze wzrostem wieku beneficjenta niepalącego, wzrasta prawdopodobieństwo na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej. Wraz ze wzrostem bmi beneficjenta niepalącego, wzrasta prawdopodobieństwo na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej. Wraz ze wzrostem ilości osób na utrzymaniu beneficjenta niepalącego, wzrasta prawdopodobieństwo na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej.

**Podział zbioru na uczący i testowy** Zbiór uczący posłuży do budowy modelu, a zbiór testowy posłuży do oceny modelu. Dokonano losowego podziału w proporcji odpowiednio: 70% i 30%. W celu powtarzalności eksperymentu wykorzystano funkcję `set.seed()`, która inicjuje „ziarno" dla generatora liczb losowych - za każdym razem otrzymuje się ten sam zestaw liczb losowych.

**Proporcje beneficjentów o kosztach leczenia rozliczanych przez ubezpieczenie zdrowotne powyżej średniej (1) i poniżej średniej(0) w podzbiorach danych.**

*Tabela 78*

```
table(dane5$charges)/nrow(dane5)

##         0         1
## 0.5656934 0.4343066

table(dane5_uczacy$charges)/nrow(dane5_uczacy)

##      0      1
## 0.5625 0.4375

table(dane5_testowy$charges)/nrow(dane5_testowy)

##         0         1
## 0.5731707 0.4268293
```

**Macierz korelacji dla objaśniających zmiennych ilościowych**

*Tabela 79*

```
cor(dane5_uczacy[,c(1,3,4)])

##                   age         bmi     children
## age       1.000000000  0.16674569 -0.007688545
## bmi       0.166745695  1.00000000 -0.026334032
## children -0.007688545 -0.02633403  1.000000000
```

**Wniosek na podstawie tabeli 79:**

Żadna ze zmiennych w modelu nie jest nadmiernie skorelowana, tzn. nie przekracza `|r|>=0.7`. Wszystkie zmienne mogą się znaleźć w jednym modelu.

**Estymacja modeli dwumianowych logitowych jednoczynnikowych** Estymujemy model dla zmiennej dychotomicznej/binarnej Y *family = binomial* z domyślną funkcją wiążącą probit *link = logit*

*Tabela 80*

```
logit1 <- glm(charges ~ age, data = dane5_uczacy, family = binomial)
summary(logit1)$coefficients
```

```
##              Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -8.467379 1.16937124 -7.240967 4.454961e-13
## age          0.202675 0.02744977  7.383487 1.541964e-13
```

```
logit2 <- glm(charges ~ sex, data = dane5_uczacy, family = binomial)
summary(logit2)$coefficients
```

```
##                Estimate Std. Error    z value  Pr(>|z|)
## (Intercept) -0.26826398  0.2127144 -1.2611461 0.2072562
## sexmale      0.03187521  0.2915694  0.1093229 0.9129464
```

```
logit3 <- glm(charges ~ bmi, data = dane5_uczacy, family = binomial)
summary(logit3)$coefficients
```

```
##                 Estimate Std. Error      z value  Pr(>|z|)
## (Intercept) -0.001510450  0.7619659 -0.001982306 0.9984183
## bmi         -0.008028243  0.0240519 -0.333788343 0.7385393
```

```
logit4 <- glm(charges ~ children, data=dane5_uczacy,family=binomial)
summary(logit4)$coefficients
```

```
##                Estimate Std. Error   z value  Pr(>|z|)
## (Intercept) -0.28509379  0.1938888 -1.470398 0.1414539
## children     0.03186636  0.1205967  0.264239 0.7915958
```

```
logit5 <- glm(charges ~ region, data = dane5_uczacy,family=binomial)
summary(logit5)$coefficients
```

```
##                   Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)      0.1541507  0.2781743  0.5541514 0.57947526
## regionnorthwest -0.2341934  0.3968742 -0.5900947 0.55512715
## regionsoutheast -0.7866732  0.4092120 -1.9224100 0.05455419
## regionsouthwest -0.7041970  0.4272109 -1.6483592 0.09927898
```

- **logit1**, **postać modelu:**

charges = -8.467379 + 0.202675 * *age*

Wniosek: Ponieważ p-value < 0,01 odrzucamy $H_0$ na korzyść $H_1$. Wiek beneficjenta niepalącego ma istotny wpływ na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu będziemy włączać zmienną *age*.

- **logit2, postać modelu:**

charges= -0.26826398 + 0.03187521 * *sex*

Wniosek: Ponieważ p-value > 0,01 nie ma podstaw do odrzucenia $H_0$. Płeć beneficjenta niepalącego nie ma istotnego wpływu na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu nie będziemy włączać zmiennej *sex*.

- **logit3, postać modelu:**

charges= -0.001510450 - 0.008028243 * *bmi*

Wniosek: Ponieważ p-value > 0,01 nie ma podstaw do odrzucenia $H_0$. Bmi beneficjenta niepalącego nie ma istotnego wpływu na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu nie będziemy włączać zmiennej *bmi*.

- **logit4**, p**ostać modelu:**

charges =-0.28509379 + 0.03186636 * *children*

Wniosek: Ponieważ p-value > 0,01 nie ma podstaw do odrzucenia $H_0$. Ilość osób na utrzymaniu beneficjenta niepalącego nie ma istotnego wpływu na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu nie będziemy włączać zmiennej *children*.

- **logit5, postać modelu:**

charges= 0.1541507 - 0.2341934 * regionnorthwest - 0*.7866732 * regionsoutheast* - 0.7041970 * *regionsouthwest*

Wniosek: Ponieważ p-value > 0,01 nie ma podstaw do odrzucenia $H_0$. Region zamieszkania beneficjenta niepalącego nie ma istotnego wpływu na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu nie będziemy włączać zmiennej *region*.

**Porównanie dobroci dopasowania modeli logitowych 1-5:**

*Tabela 81*

```
wyniki_oceny_logit
```

```
##         kryterium_AIC     McFadden  Cragg_Uhler
## model_1      119.0569 0.5627882026 7.206226e-01
## model_2      267.1487 0.0000454193 8.344055e-05
## model_3      267.0491 0.0004237993 7.783668e-04
## model_4      267.0909 0.0002650489 4.868524e-04
## model_5      266.1264 0.0191298634 3.468814e-02
```

**Wnioski na podstawie tabeli 81:**

Najlepszym modelem jest model 1, ponieważ dla kryterium AIC przyjmuje najmniejsze wartości natomiast dla kryterium McFadden i Cragg Uhlera przyjmuje największe wartości.

Do modelu nie można dodać żadnej dodatkowej zmiennej objaśniającej, ponieważ żadna ze zmiennych poza *age* nie jest statystycznie istotna.

**Wybór i interpretacja modelu** Wybieramy model charges ~ *age* -> logit1.

logit(p)=-8.467379+0.202675age

*logit(p)=ln(p/1-p) p/(1-p)=exp(-8.467379+0.202675age*)

*Tabela 82*

```
logit1$coefficients
```

```
## (Intercept)         age
##   -8.467379    0.202675
```

```
##  exp(bi)
```

```
exp(logit1$coefficients)
```

```
##  (Intercept)          age
## 0.0002102152 1.2246743885
```

```
##  exp(5*bi)
```

```
exp(5*logit1$coefficients[2])
```

```
##      age
## 2.754883
```

```
##  exp(10*bi)
```

```
exp(10*logit1$coefficients[2])
```

```
##      age
## 7.589381
```

**Interpretacje na podstawie tabeli 82:**

- *exp($\beta_0$)=0.0002102152*, gdzie $\beta_0$ to wyraz wolny => interpretuje się jako szansę zdarzenia w grupie referencyjnej (xi=0). Nie posiada interpretacji.
- *exp($\beta_1$) = 1.2246743885 => (exp($\beta_1$)-1)*100%= 22.47%*

Jeżeli wiek beneficjenta niepalącego wzrośnie o 1 rok, to szansa na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej, wzrośnie średnio o 22.47%.

- *exp(5* $\beta_2$) = 2.754883 => (exp(5* $\beta_2$)-1)*100%= 175.49%*

Jeżeli wiek beneficjenta niepalącego wzrośnie o 5 lat, to szansa na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej, wzrośnie średnio o 175.49%.

- $exp(10*\beta_3) = 7.589381 \Rightarrow (exp(10*\beta_3)-1)*100\% = 658.94\%$

Jeżeli wiek beneficjenta niepalącego wzrośnie o 10 lat, to szansa na koszty leczenia rozliczane przez ubezpieczenie zdrowotne powyżej średniej, wzrośnie średnio o 658.94%.

*Tabela 83*

```
predict(logit1, data.frame(age=c(20,30,40,50,60)), type="response")
```

```
##          1          2          3          4          5
## 0.01196327 0.08415948 0.41086790 0.84109143 0.97571049
```

**Wnioski do tabeli 83:**

Spodziewamy się, że u beneficjentów niepalących w wieku:

- 20 lat, prawdopodobieństwo rozliczanych kosztów leczenia przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.01196327,
- 30 lat, prawdopodobieństwo rozliczanych kosztów leczenia przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.08415948,
- 40 lat, prawdopodobieństwo rozliczanych kosztów leczenia przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.41086790,
- 50 lat, prawdopodobieństwo rozliczanych kosztów leczenia przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.84109143,
- 60 lat, prawdopodobieństwo rozliczanych kosztów leczenia przez ubezpieczenie zdrowotne powyżej średniej będzie wynosić 0.97571049.

## Estymacja modelu dwumianowego probitowego

Estymujemy model dla zmiennej dychotomicznej/binarnej Y `family = binomial` z funkcją wiążącą probit `link = probit.`

*Tabela 84*

```
probit1 <- glm(charges ~ age, data = dane5_uczacy, family = binomial(
link=probit))
summary(probit1)$coefficients
```

```
##              Estimate Std. Error   z value     Pr(>|z|)
## (Intercept) -4.282626 0.50148873 -8.539824 1.344265e-17
## age          0.104618 0.01198336  8.730265 2.540757e-18
```

Model probitowy – interpretacja parametrów sprowadza się do stwierdzenia, czy dana zmienna jest stymulantą (gdy $\beta_i > 0$), czy destymulantą modelu (gdy $\beta_i < 0$).

Według powyższej tabeli wiek jest istotnie statystyczny. Zmienna age jest stymulantą.

## Porównanie dobroci dopasowania modeli logit1 i probit1

*Tabela 85*

```
wyniki_oceny_logit_probit
```

```
##                kryterium_AIC  McFadden Cragg_Uhler
## model_logit_1         119.0569 0.5627882   0.7206226
## model_probit_1        123.5825 0.5455913   0.7058409
```

**Wnioski z tabeli 85:**

Lepszym modelem jest logitowy, ponieważ ma lepsze możliwości interpretacyjne, posiada mniejszą wartość kryterium Akaike oraz wyższe wartości dla kryterium McFadden i Cragg Uhler.

**Porównanie jakości predykcji modeli logit1 i probit1**

Tablice trafności dla wybranego punktu odcięcia p* (p* = proporcja z próby uczącej).

*Tabela 86*

```
p <- table(dane5_uczacy$charges)[2]/nrow(dane5_uczacy)
```

```
## Tablica trafności dla modelu logitowego - próba ucząca
```

```
##             przewidywane
## obserwowane  0  1
##           0 94 14
##           1  7 77
```

```
## Tablica trafności dla modelu probitowego - próba ucząca
```

```
##             przewidywane
## obserwowane  0  1
##           0 89 19
##           1  7 77
```

```
## Tablica trafności dla modelu logitowego - próba testowa
```

```
##             przewidywane
## obserwowane  0  1
##           0 44  3
##           1  8 27
```

```
## Tablica trafności dla modelu probitowego - próba testowa
```

```
##             przewidywane
## obserwowane  0  1
##           0 43  4
##           1  8 27
```

**Miary jakości predykcji - miary oparte na tablicy trafności dla wybranego punktu odcięcia p***

Poniższa funkcja *miary_pred* została została określona dla argumentów: *model* (model dwumianowy), *dane* (np. zbiór uczący, testowy), *Y* (obserwowane Y 0-1 w analizowanym zbiorze danych).

**Ocena zdolności predykcyjnej na zbiorze uczącym**

*Tabela 87*

```
##                     ACC        ER      SENS      SPEC       PPV       NPV
## model_logit  0.8906250 0.1093750 0.9166667 0.8703704 0.8461538 0.9306931
## model_probit 0.8645833 0.1354167 0.9166667 0.8240741 0.8020833 0.9270833
```

**Ocena zdolności predykcyjnej na zbiorze testowym**

*Tabela 88*

```
##                     ACC        ER      SENS      SPEC       PPV       NPV
## model_logit  0.8658537 0.1341463 0.7714286 0.9361702 0.9000000 0.8461538
## model_probit 0.8536585 0.1463415 0.7714286 0.9148936 0.8709677 0.8431373
```

**Wnioski do tabel 87-88:**

Dla danego progu odcięcia:

- Na podstawie powyższych miar jakości predykcji można stwierdzić, że dla zbioru uczącego model logitowy i probitowy w podstawowych miarach mają taki sam wynik.
- Dla zbioru testowego odrobinę lepszy okazał się model logitowy pod względem jakości predykcji.
- Poprzez porównanie wyników można stwierdzić, że model nie był przeuczony (przystosowany tylko dla zbioru uczącego).
- Należy sprawdzić, czy na zbiorze testowym nie pogorszyły się znacząco miary jakości predykcji w stosunku do zbioru uczącego.

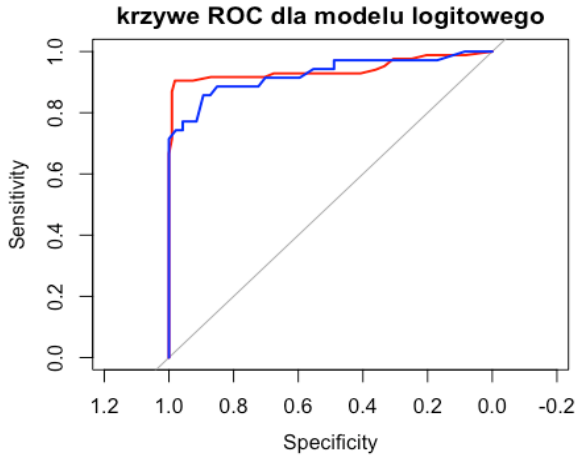
**Krzywa ROC**

Krzywa ROC prezentuje jakość predykcji modelu dla wszystkich możliwych punktów odcięcia p* (jest niezależna od wyboru p*). Dla modeli oszacowanych na zbiorze uczącym porównana została poniżej jakość predykcji na zbiorze uczącym i testowym. Proszę sprawdzić, czy jakość predykcji dla zbioru testowego nie pogorszyła się znacząco w stosunku do jakości predykcji dla zbioru uczącego.

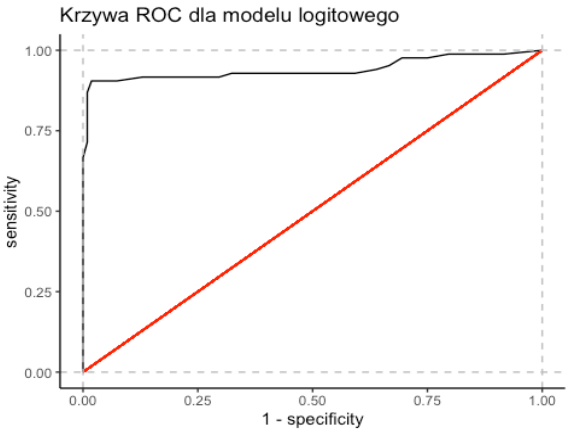
krzywa czerwona - ROC wyznaczona na zbiorze uczącym

krzywa niebieska - ROC wyznaczona na zbiorze testowym

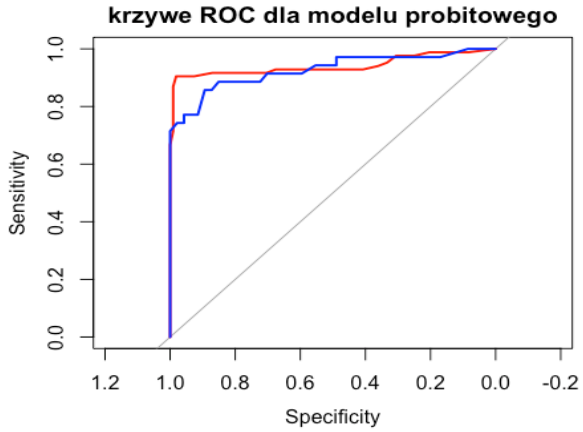
*Wykres 66*



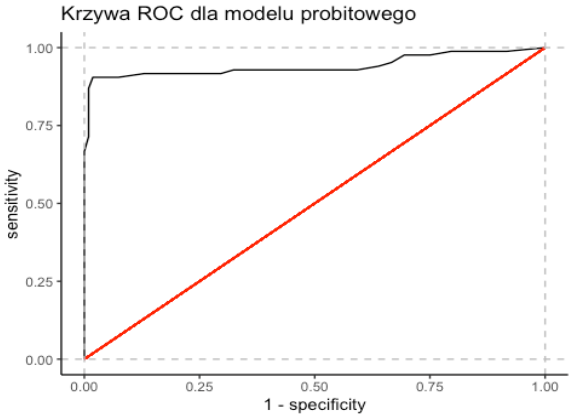

*Wykres 67*




*Wykres 68*




*Wykres 69*




**Pole powierzchni pod krzywą ROC dla zbioru uczącego (model logitowy i probitowy)**

*Tabela 89*

```
## pole_AUC_logit  0.9410273
## pole_AUC_probit 0.9410273
```

**Pole powierzchni pod krzywą ROC dla zbioru testowego (model logitowy i probitowy)**

*Tabela 90*

```
## pole_AUC_logit  0.9264438
## pole_AUC_probit 0.9264438
```

**Wnioski na podstawie wykresów 66-69 oraz tabel 89 i 90:**

Na podstawie pola pod krzywą ROC można stwierdzić, że oba modele: `logit` oraz `probit` mają wystarczająco dobrą zdolność predykcyjną. Wartości spełniają równanie 0.5<=AUC<=1.

## Podsumowanie:

W pierwszej kolejności stworzono zbiór uczący i testowy na podstawie zmiennej objaśnianej charges. Zbiór uczący posłużył do budowy modelu, a zbiór testowy posłużył do oceny modelu. Następnie stworzono modele logitowe na podstawie zmiennej age. Został wybrany jako najlepszy model – ***model logit 1***. Stwierdzono, że wiek beneficjenta niepalącego ma istotny wpływ na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Do budowania modelu będziemy włączać zmienną age. Na podstawie zmiennej age. stworzono również model probitowy. W porównaniu dobroci dopasowania ***model logit 1*** i ***model probit 1***, lepszym modelem okazał się model logitowy, ponieważ ma lepsze możliwości interpretacyjne, posiada mniejszą wartość kryterium Akaike oraz wyższe wartości dla kryterium McFadden i Cragg Uhler. Podczas wyliczania pola powierzchni pod krzywą ROC dla zbioru uczącego i testowego obydwa modele spełniają równanie 0,5<=AUC<=1. Pole powierzchni okazało się lepsze na zbiorze uczącym.

# 4. ZAKOŃCZENIE – PODSUMOWANIE I WNIOSKI

## Wybrane modele dla grupy beneficjentów palących:

| Nazwa modelu | Postać modelu | Uzasadnienie |
|---|---|---|
| **Model 7** | charges= -28320.58+242.90*age+ *1646.64*bmi* | Model 7 wyjaśnia 84,07% kształtowania kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów palących. Reszty nie mają rozkładu normalnego. |
| **Model 9** | charges=-23562.68+261.28*age+ *1505.19*bmi*- 1689.41*regionnorthwest- *1965.034*regionsoutheast*- 505.08**regionsouthwest* | Model 9 wyjaśnia w 82,48% kształtowanie kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów palących. Model spełnia wszystkie założenia statystyczne. |
| **Model 11** | ln(charges)=8.6508756 + 0.0080495 * *age* + 0.0450499 * *bmi*- 0.0213275 * *regionnorthwest*- 0.0631668 * *regionsoutheast* + 0.0198589 * *regionsouthwest* | Zmienne (poza nortwest i southwest) statystycznie istotnie wpływają na kształtowanie kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów palących, niskie kryterium Akaike. |
| **Model 14** | ln(charges)= 7.943585 + 0.030509 * *age* - 0.005917 * *bmi* | Zmienne statystycznie istotnie wpływają na kształtowanie kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów palących, niskie kryterium Akaike. |
| **Logit_3** | charges=-24.147463+0.816106**bmi* | Lepsze miary dobroci dopasowania, statystycznie wpływa na zmienną charges, dobra jakość i zdolność predykcyjna. |
| **Probit_1** | charges= -10.193599 + 0.3468573 * *bmi* | Gorszy od modelu logitowego, statystycznie wpływa na zmienną charges, dobra jakość i zdolność predykcyjna. |

**Wniosek:**

Najlepszym modelem okazał się model logitowy, ponieważ spełnia wszystkie swoje założenia.

## Wybrane modele dla grupy beneficjentów niepalących:

| Nazwa modelu | Postać modelu | Uzasadnienie |
|---|---|---|
| **Model 6** | charges= -3449.393 + 263.980 * age | Model 6 wyjaśnia 95,95% kształtowania kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów niepalących. Model spełnia wszystkie założenia statystyczne. Zmienna age została wybrana do dalszej analizy ze względu na wysoką istotność statystyczną. |
| **Logit_1** | charges = -8.467379 +0.202675*$age$ | Wiek beneficjenta niepalącego ma istotny wpływ na poziom kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne. Ten model dla kryterium Akaike przyjmuje najmniejsze wartości natomiast dla kryterium McFadden i Cragg Uhlera przyjmuje największe wartości. |
| **Probit_1** | charges = -4.282626+0.104618*$age$ | Według analizy modelu probitowego wiek jest istotnie statystyczny. Minimalnie lepszy okazał się model logitowy. |

**Wniosek:**

Najlepszym modelem okazał się model 6, ponieważ wyjaśnia on aż 95,95% kształtowania kosztów leczenia rozliczanych przez ubezpieczenie zdrowotne dla beneficjentów niepalących i spełnia wszystkie założenia statystyczne.

**Podsumowanie grup beneficjentów:**

- Modele liniowe i klasy GLM:

Modele dla grupy beneficjentów palących okazały się ciekawsze, ponieważ charakteryzują się większą różnorodnością doboru zmiennych objaśniających wpływających na zmienną objaśnianą (charges) podczas budowy modeli, natomiast dla beneficjentów niepalących możliwe było użycie jedynie jednej zmiennej – age.

- Modele logitowe i probitowe:

Modele logitowe i probitowe dla obu grup beneficjentów, były tak samo budowane – z jedną zmienną objaśniającą, zatem ich postać oraz stopień wyjaśnienia zjawiska jest podobny.

# 5. BIBLIOGRAFIA


- Medical Cost Personal Datasets
  https://www.kaggle.com/datasets/mirichoi0218/insurance?select=insurance.csv
- Harrell F.E., „*Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis”*, Second Edition, Springer 2015
- Skala BMI https://www.zikodlazdrowia.org/blog/sprawdz-swoje-bmi/
- Systemowe reformy zdrowotne w Stanach Zjednoczonych.
  https://www.ejournals.eu/pliki/art/10704/