Praca zaliczeniowa z przedmiotu:

Analiza współzależności zjawisk

Adam Matuszczyk

Studia Podyplomowe

MSAD 2018/2019

1. Dane i opis źródła

We współczesnym świecie wraz ze wzrostem średniej długości życia koszty leczenia stają się jednym z najważniejszych wydatków społecznych, dlatego wiele krajów z dużą atencją przygląda się wzrastającej średniej długości życia, a także jego komfortowi.

Projekt dotyczy porównania kosztów leczenia z nieokreślonego stanu w USA, a podzielonego na następujące zmienne: **age, sex, bmi, children, smoker, region, charges.**

Dane pochodzą ze strony:

https://www.kaggle.com/mirichoi0218/insurance#insurance.csv

```
library("tidyverse")
library("skimr")
library("mosaic") # favstats
library("car")
library("data.table")
library("psych")

d <- read_csv("path/insurance.csv")
d <- d %>%
  mutate_if(is.character, as.factor)
```

2. Rozkłady zmiennych

```
d %>% count(region) %>% kable()
```

region	n
northeast	324
northwest	325
southeast	364
southwest	325

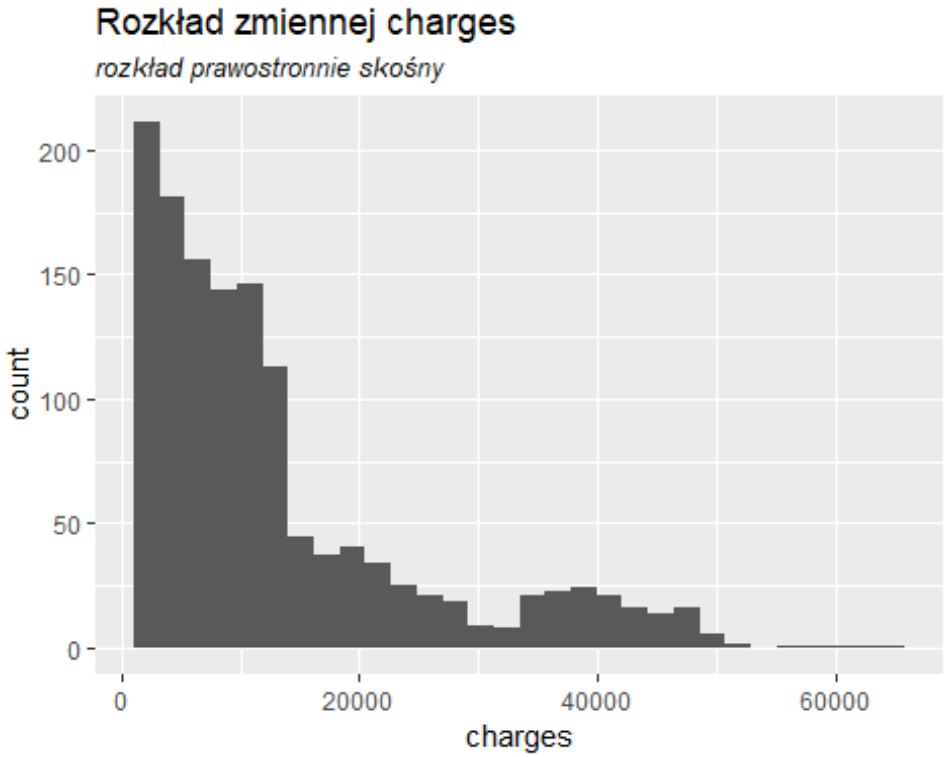
```
d %>% count(sex) %>% kable()
```

sex	n
female	662
male	676

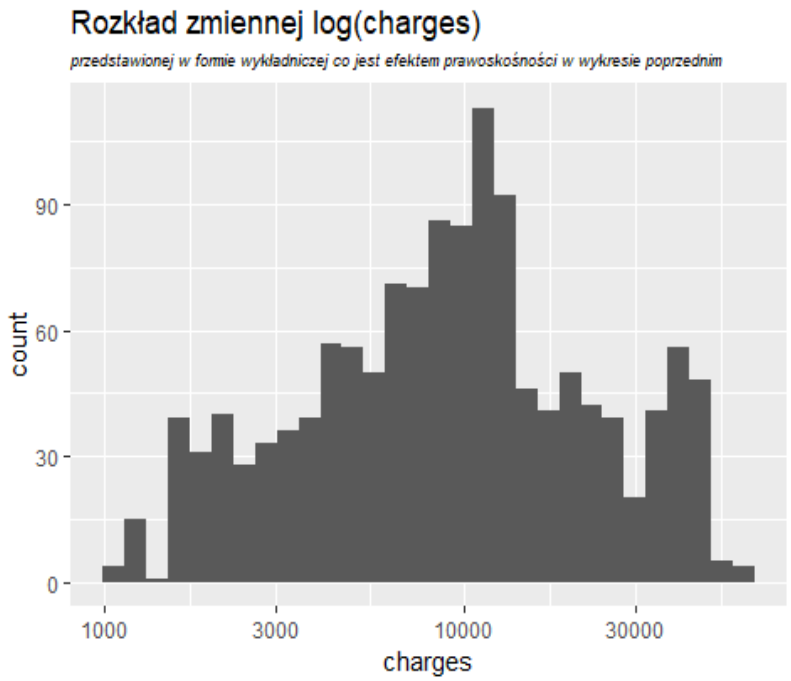
```
d %>% count(smoker) %>% kable()
```

smoker	n
no	1064
yes	274

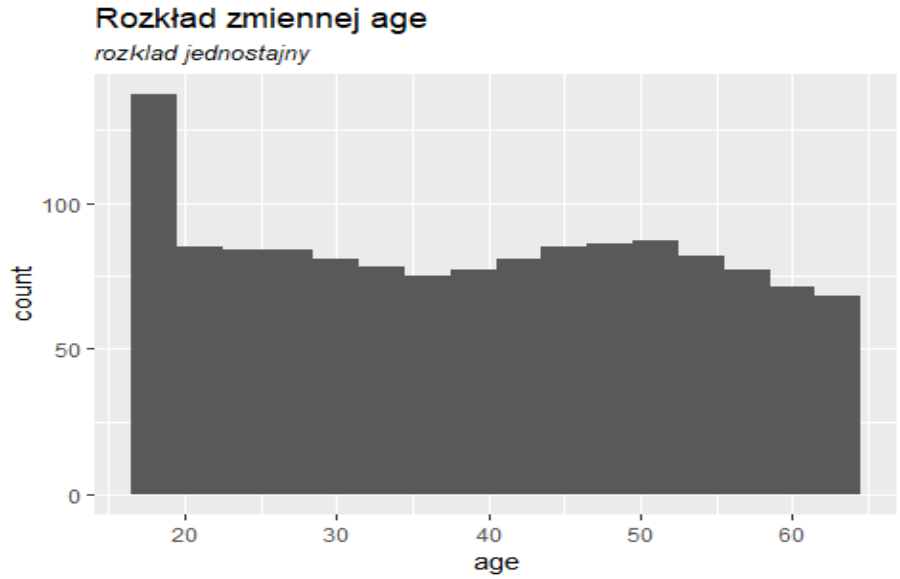
```
ggplot(d, aes(charges)) + geom_histogram() + labs(title = "Rozkład zmiennej c
harges", subtitle = "rozkład prawostronnie skośny")+theme(plot.subtitle = ele
ment_text(size = 10, face = "italic", color = "black"))
```



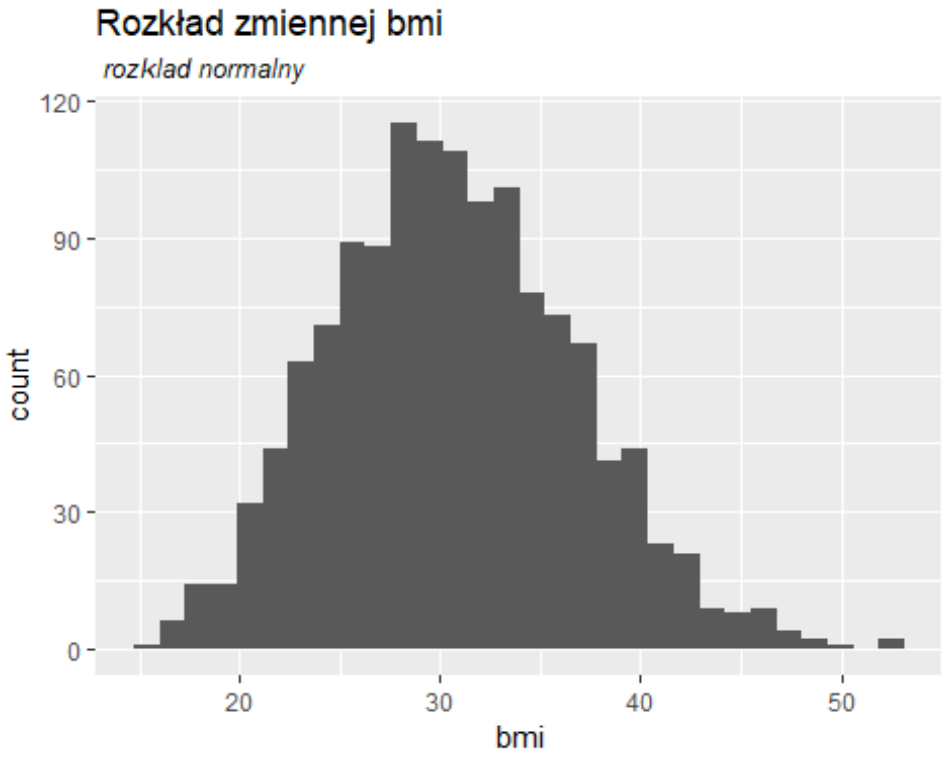
```
ggplot(d, aes(charges)) + geom_histogram()  + scale_x_log10() + labs(title =
"Rozkład zmiennej log(charges)", subtitle = "przedstawionej w formie wykładni
czej co jest efektem prawoskośności w wykresie poprzednim") + theme(plot.subt
itle=element_text(size=7, face="italic", color="black"))
```




```
ggplot(d, aes(age)) + geom_histogram(binwidth = 3)+ labs(title = "Rozkład zmi
ennej age", subtitle = "rozklad jednostajny") + theme(plot.subtitle=element_t
ext(size=10, face="italic", color="black"))
```



```r
# rozklad jednostajny
ggplot(d, aes(bmi)) + geom_histogram() + labs(title = "Rozkład zmiennej bmi"
, subtitle = " rozklad normalny") + theme(plot.subtitle=element_text(size=10,
face="italic", color="black"))  # rozklad normalny
```



2.1 Zależności między zmienną objaśnianą, a wieloma zmiennymi objaśniającymi

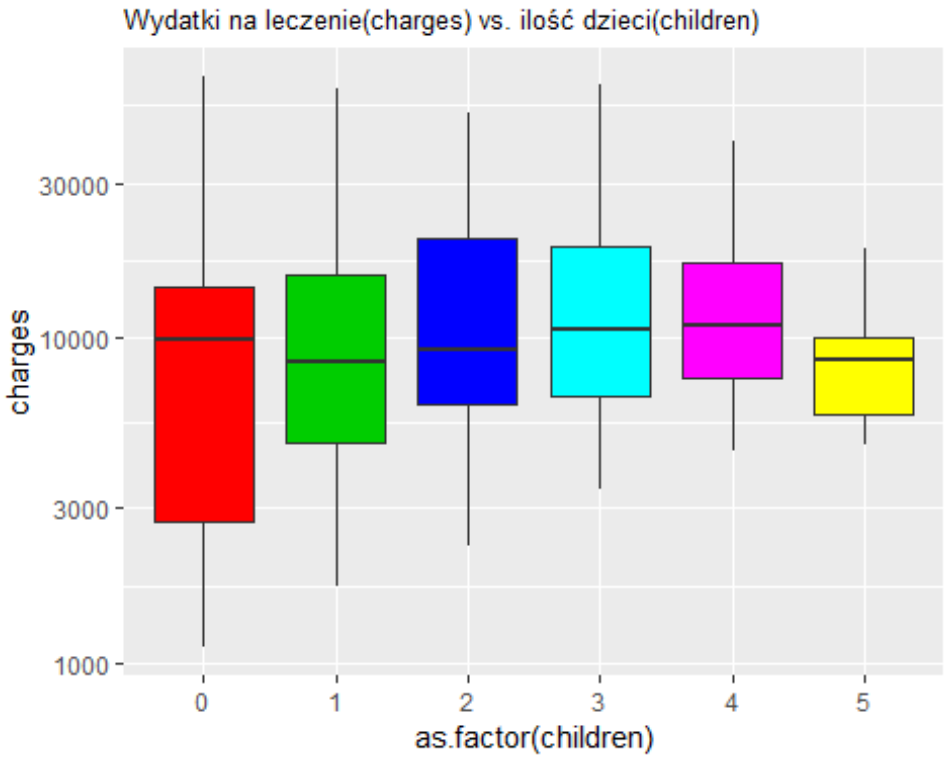
```r
favstats(charges ~ children, data = d) %>% kable()
```

children	min	Q1	median	Q3	max	mean	sd	n	missing
0	1121.874	2734.421	9856.952	14440.12	63770.43	12365.976	12023.294	574	0
1	1711.027	4791.643	8483.870	15632.05	58571.07	12731.172	11823.631	324	0
2	2304.002	6284.939	9264.979	20379.28	49577.66	15073.564	12891.368	240	0
3	3443.064	6652.529	10600.548	19199.94	60021.40	15355.318	12330.869	157	0
4	4504.662	7512.267	11033.662	17128.43	40182.25	13850.656	9139.223	25	0
5	4687.797	5874.974	8589.565	10019.94	19023.26	8786.035	3808.436	18	0

Średnie koszty leczenia wahają się w zależności od ilości dzieci pomiędzy 1127,84, a 4687,797. Mediana natomiast pokazuje, że istnieją inne

czynniki znacznie wpływające na koszty leczenia, które znacznie odbiegają od średnich tej kategorii.

```
ggplot(d, aes(as.factor(children), charges)) + geom_boxplot(fill =c(2 : 7)) + scale_y_log10() + labs(title = "Wydatki na leczenie(charges) vs. ilość dzieci (children)") + theme(plot.title=element_text(size=10))
```

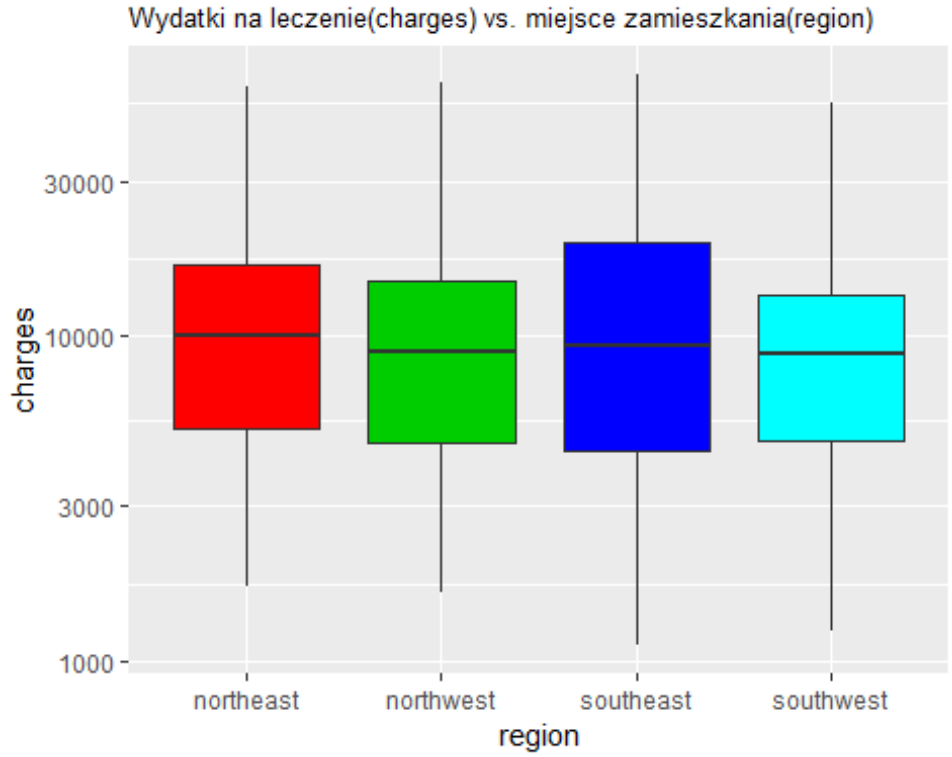



```
favstats(charges ~ region, data = d) %>% kable()
```

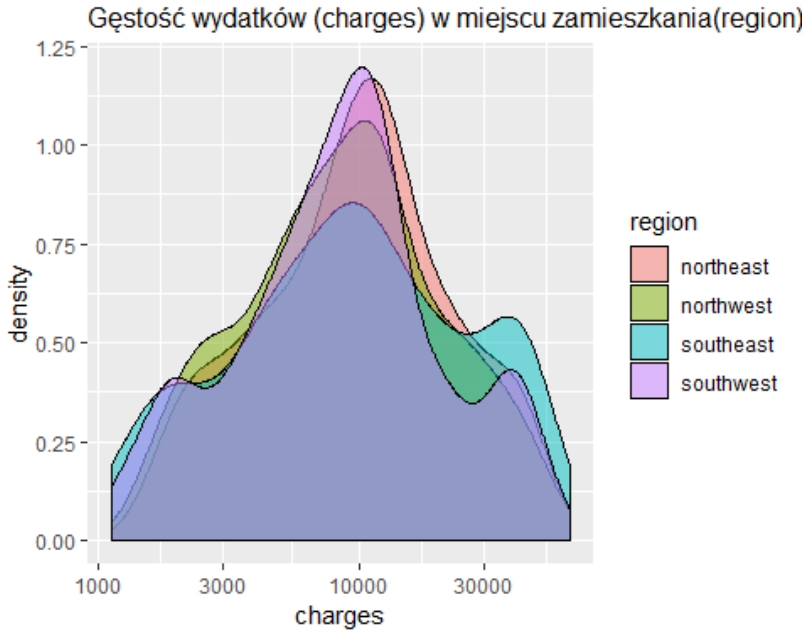
region	min	Q1	median	Q3	max	mean	sd	n	missing
northeast	1694.796	5194.322	10057.652	16687.36	58571.07	13406.38	11255.80	324	0
northwest	1621.340	4719.737	8965.796	14711.74	60021.40	12417.58	11072.28	325	0
southeast	1121.874	4440.886	9294.132	19526.29	63770.43	14735.41	13971.10	364	0
southwest	1241.565	4751.070	8798.593	13462.52	52590.83	12346.94	11557.18	325	0

Chociaż wydawać się powinno, że miejsce zamieszkania nie powinno specjalnie wpływać na koszty to zauważyć można żyjącym w południowych regionach jest średnio taniej niż w północnych. Mediana wyrównuje tą dysproporcję.

```
ggplot(d, aes(region, charges)) + geom_boxplot(fill =c(2:5)) + scale_y_log10(
) + labs(title = "Wydatki na leczenie(charges) vs. miejsce zamieszkania(regio
n)") + theme(plot.title=element_text(size=10))
```



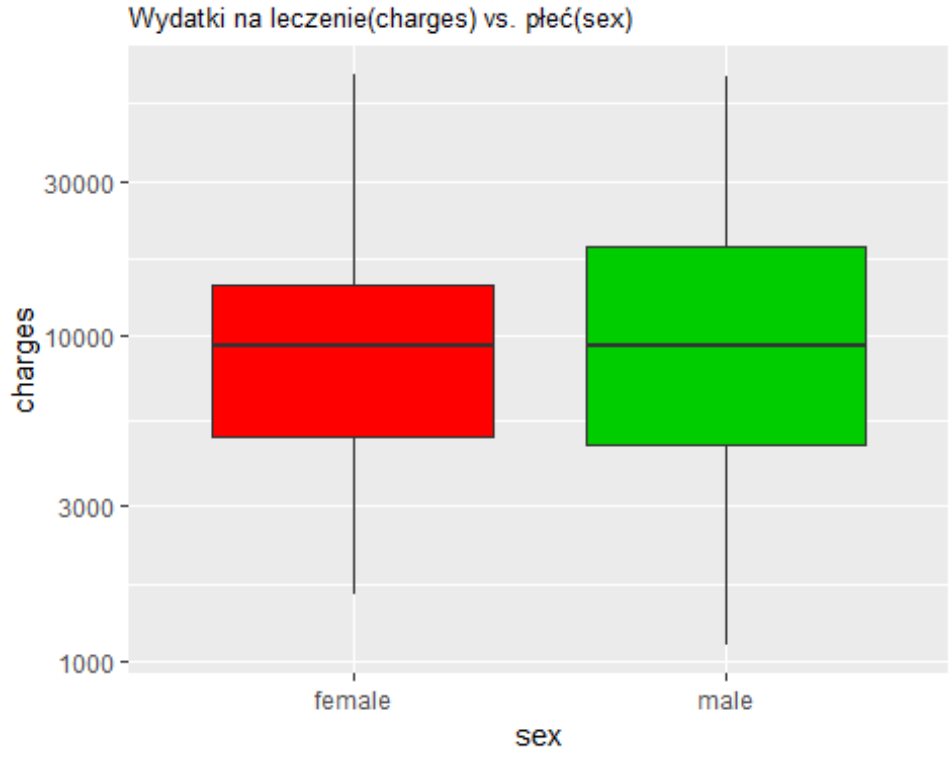

```
ggplot(d, aes(charges, fill = region)) + geom_density(alpha = 0.5) + scale_x_
log10() + labs(title = "Gęstość wydatków (charges) w miejscu zamieszkania(reg
ion)") + theme(plot.title=element_text(size=12))
```



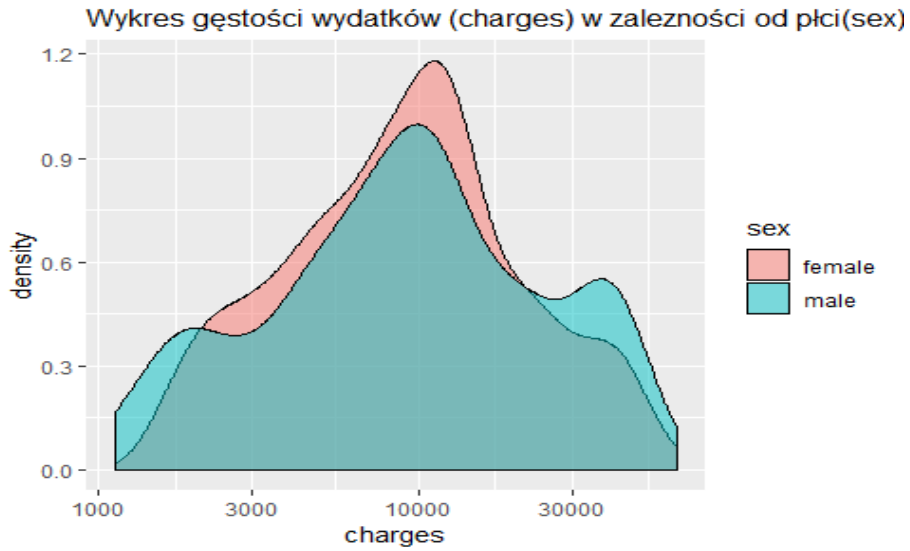
```
favstats(charges ~ sex, data = d) %>% kable()
```

sex	min	Q1	median	Q3	max	mean	sd	n	missing
female	1607.510	4885.159	9412.962	14454.69	63770.43	12569.58	11128.70	662	0
male	1121.874	4619.134	9369.616	18989.59	62592.87	13956.75	12971.03	676	0

```
ggplot(d, aes(sex, charges)) + geom_boxplot(fill = c(2:3)) + scale_y_log10()
+ labs(title = "Wydatki na leczenie(charges) vs. płeć(sex)") + theme(plot.tit
le=element_text(size=10))
```




```
ggplot(d, aes(charges, fill = sex)) + geom_density(alpha = 0.5) + scale_x_log
10() + labs(title = "Wykres gęstości wydatków (charges) w zależności od płci(
sex)") + theme(plot.title=element_text(size=12))
```



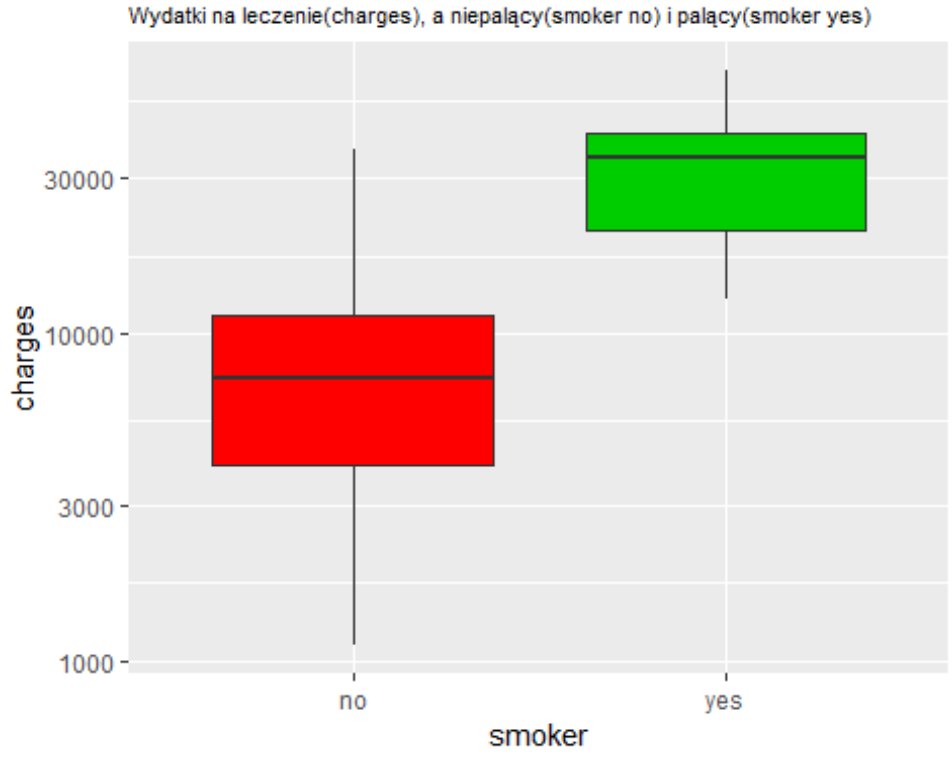
Typowy wykres gęstości charakterystyczny dla mężczyzn i kobiet. Mężczyźni mniej dbają o swoje zdrowie, natomiast w ogonach pojawiają się zdecydowanie kosztowniejsi pacjenci.

```
favstats(charges ~ smoker, data = d) %>% kable()
```

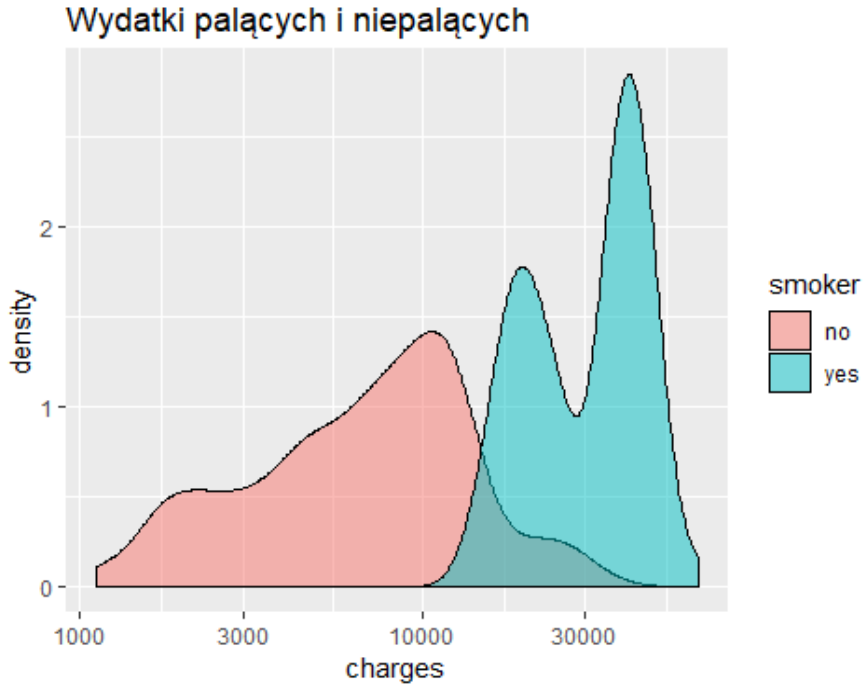
smoker	min	Q1	median	Q3	max	mean	sd	n	missing
no	1121.874	3986.439	7345.405	11362.89	36910.61	8434.268	5993.782	1064	0
yes	12829.455	20826.244	34456.348	41019.21	63770.43	32050.232	11541.547	274	0

Jak się należało spodziewać średnia kosztów leczenia palących jest znacznie przewyższająca niepalących. Mediana jest ponad cztery razy wyższa.

```
ggplot(d, aes(smoker, charges)) + geom_boxplot(fill = c(2:3)) + scale_y_log10
() + labs(title = "Wydatki na leczenie(charges), a niepalący(smoker no) i pal
ący(smoker yes)")+ theme(plot.title=element_text(size=8))
```

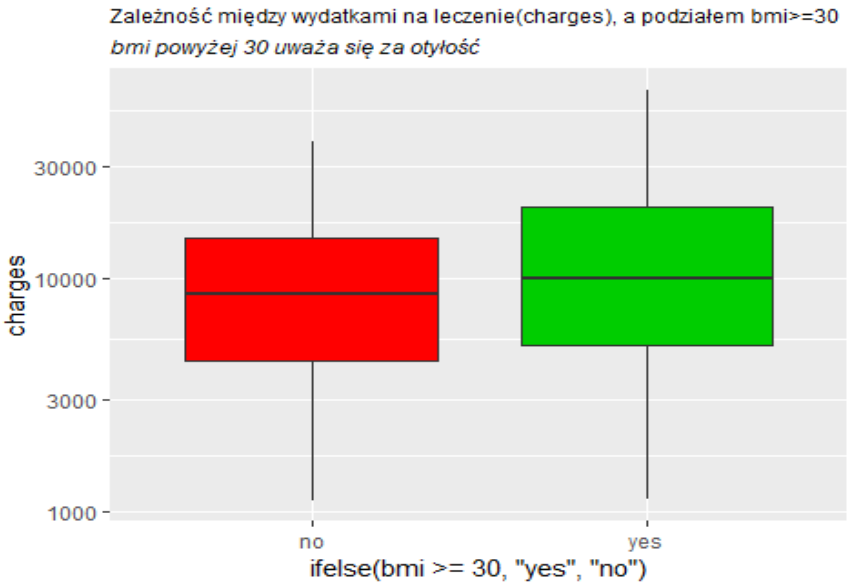



```
ggplot(d, aes(charges, fill = smoker)) + geom_density(alpha = 0.5) + scale_x_
log10() + labs(title = "Wydatki palących i niepalących")
```



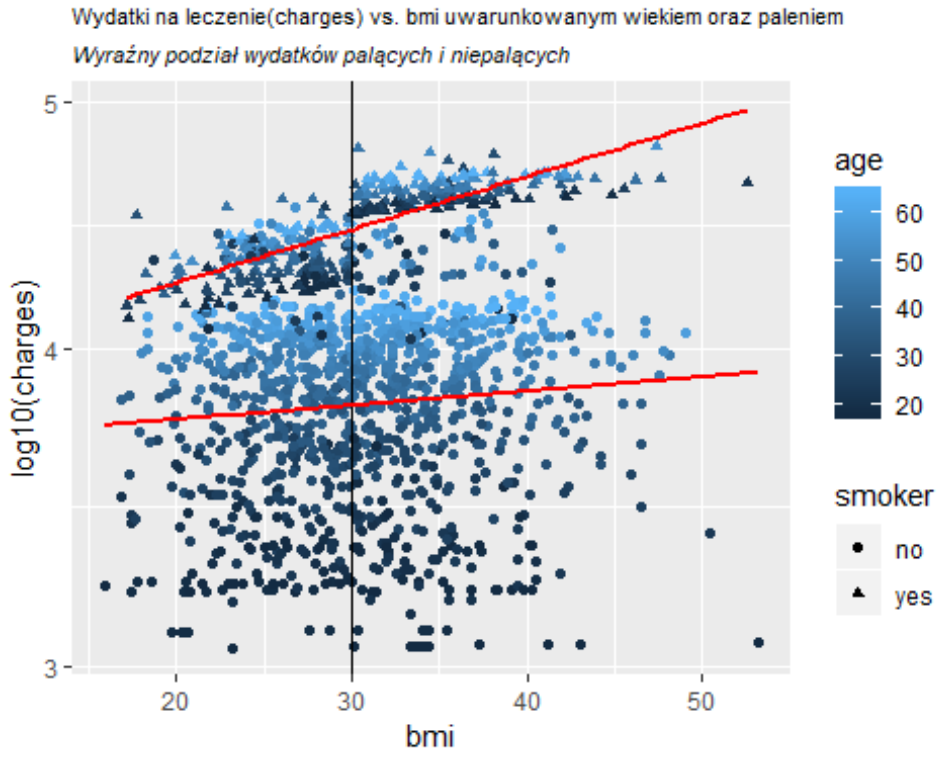
Ciekawostką jest podwójny pik wśród palących. Prawdopodobnie jest to związane z chorobami, które są efektem palenia papierosów, a pojawiają się z biegiem czasu.

```
ggplot(d, aes(ifelse(bmi >= 30, "yes", "no"), charges)) + geom_boxplot(fill=c(2:3), outlier.colour = "red", outlier.shape = 1) + scale_y_log10() + labs(title = "Zależność między wydatkami na leczenie(charges), a podziałem bmi>=30", subtitle = "bmi powyżej 30 uważa się za otyłość") + theme(plot.title = element_text(size=9, color="black") ) + theme(plot.subtitle = element_text(size=9, face="italic", color="black") )
```



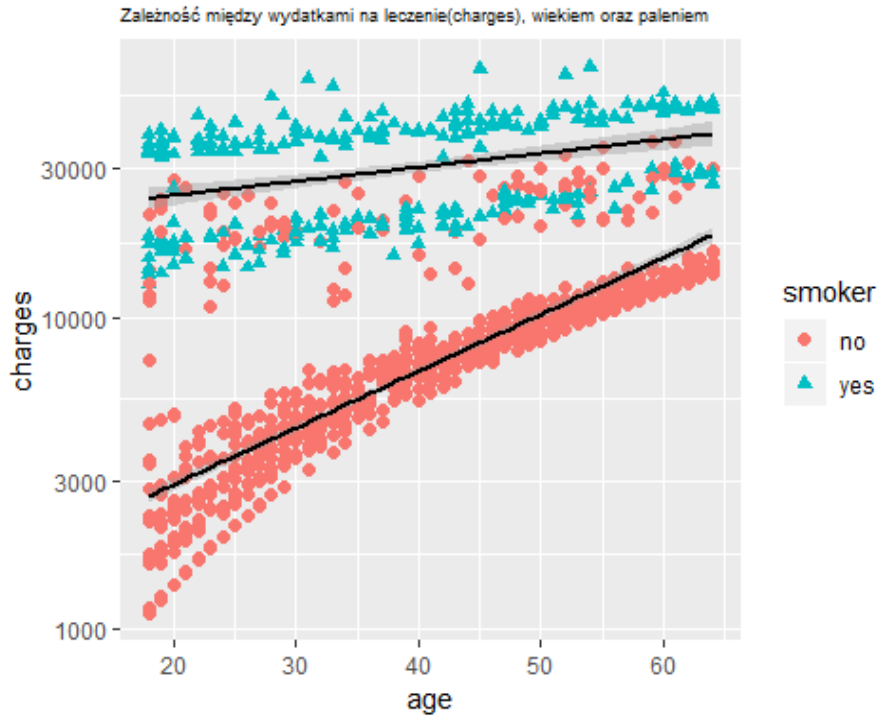
Dla mnie największe zaskoczenie. Specjalnie bmi powyżej 30 nie wpływa na koszty leczenia, chociaż jest wyższe.

```
ggplot(d, aes(bmi, log10(charges), colour = age, shape = smoker)) + geom_poin
t() + scale_y_log10()  + geom_vline(xintercept = 30) + geom_smooth(method = "
lm", se = FALSE, col = "red")+ labs(title = "Wydatki na leczenie(charges) vs.
bmi uwarunkowanym wiekiem oraz paleniem", subtitle = "Wyraźny podział wydatkó
w palących i niepalących") + theme(plot.title = element_text(size=8, color="b
lack")) + theme(plot.subtitle = element_text(size=8, face="italic", color="bl
ack"))
```



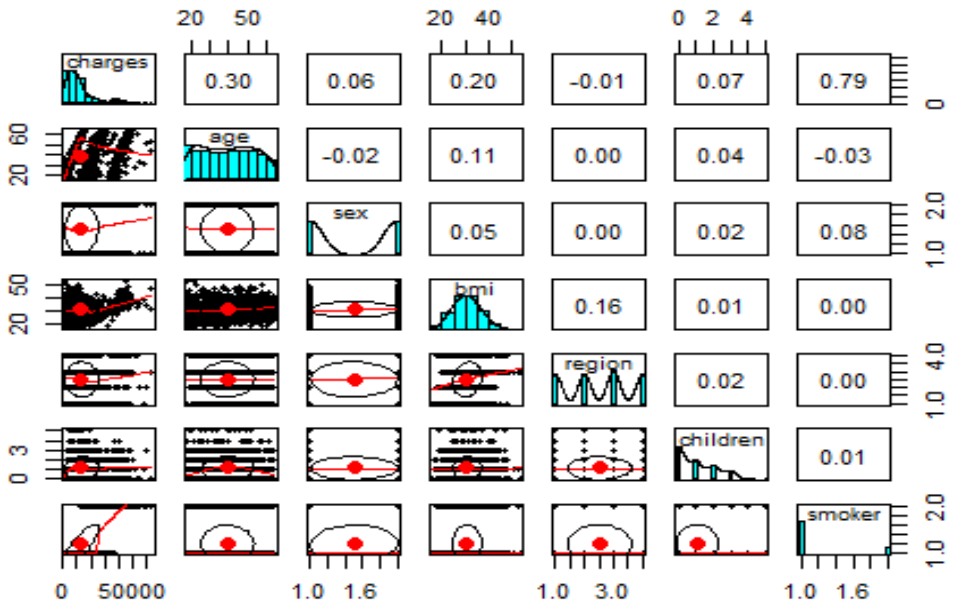
Jeden z ciekawszych wykresów gdzie jak widzimy „ggplot” wyznaczył sobie tak naprawdę dwie grupy ludzi i narysował linie, które dzielą nasz zbiór „bmi” na palących(czerwona górna linia) i niepalących. Widać też grupę z „bmi” ponad 30 równocześnie palącą, której wydatki na zdrowie są zdecydowanie wyższe i odbiegają dość znacznie od pozostałych grup.

```
ggplot(d, aes(age, charges, col = smoker, shape = smoker)) + geom_point(size
= 2) + scale_y_log10() + geom_smooth(method = "lm",col = "black") + labs(titl
e = "Zależność między wydatkami na leczenie(charges), wiekiem oraz paleniem")
+ theme(plot.title = element_text(size=7, color="black"))
```



Wyraźna różnica między palącymi, a niepalącymi maleje z wiekiem, ale jak widać palenie papierosów jest dość kosztowne, nie tylko ze względu na ich kupno.

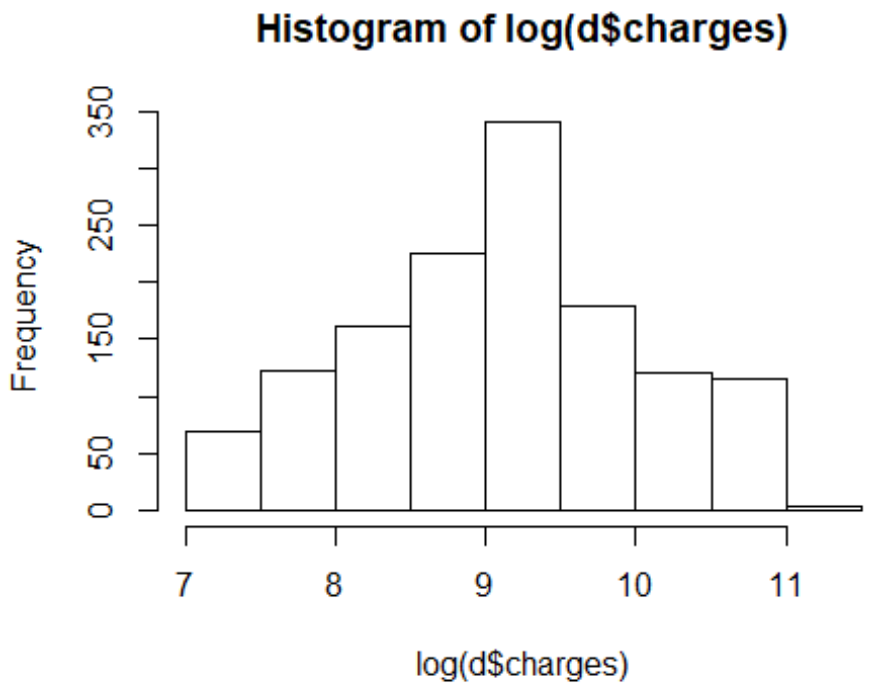
```
charges.subset <- subset(d, select = c(charges,age, sex, bmi,region, children
, smoker))
pairs.panels(charges.subset)
```



Ostatni wykres „korelacji pearsona” między zmiennymi. Widać, że największa korelacja ze zmienną „charges” jest dla zmiennych „smoker’, ”age” oraz „bmi”.

3. Wstępna analiza modelu

```
hist(log(d$charges))
```




```
m2 <- lm(log(charges) ~ ., d)
summary(m2) #summary(lm(log(charges) ~ ., d))

##
## Call:
## lm(formula = log(charges) ~ ., data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07186 -0.19835 -0.04917  0.06598  2.16636
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.0305581  0.0723960  97.112  < 2e-16 ***
## age              0.0345816  0.0008721  39.655  < 2e-16 ***
## sexmale         -0.0754164  0.0244012  -3.091 0.002038 **
## bmi              0.0133748  0.0020960   6.381 2.42e-10 ***
## children         0.1018568  0.0100995  10.085  < 2e-16 ***
## smokeryes        1.5543228  0.0302795  51.333  < 2e-16 ***
## regionnorthwest -0.0637876  0.0349057  -1.827 0.067860 .
## regionsoutheast -0.1571967  0.0350828  -4.481 8.08e-06 ***
## regionsouthwest -0.1289522  0.0350271  -3.681 0.000241 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4443 on 1329 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7666
## F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Pierwszy rozpatrywany model m2 uwzględnia wszystkie zmienne z danych. Funkcja. Funkcja *lm* uprościła już model do płci „mężczyźni" oraz „palący".

Widać po wartościach *p-wartość,* że poza graniczną „regionnorthwest", który usuwam, wszystkie pozostałe zmienne należy brać pod uwagę przy dalszej analizie modelu.

Współczynnik determinancji *R-squared* jest na poziomie 0,768 czyli otrzymujemy informację, że 77% zmienności zostaje wyjaśniona dzięki temu modelowi.

P-wartość dla *testu-F* również jest bardzo małe więc należy rozpatrywać **H1: przynajmniej jedna zmienna objaśniająca jest istotna.** Równocześnie można założyć, że występują interakcje między zmiennymi objaśniającymi.

Wstępny model możemy opisać równaniem:

log(charges) = 7,0305581 + 0,0345876age – 0,0754164sexmale + 0.0133748bmi + 0,1018568children + 1,5543228smokeryes – 0,1571967regionsoutheast – 0,1289522regionsouthwest + ε

```
res_m2 <- m2 %>% summary() %>% coef() %>% as.data.frame()
res_m2[, 1] <- exp(res_m2[, 1])
res_m2[, 2] <- exp(res_m2[, 2])
kable(res_m2 %>% round(3))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1130.661	1.075	97.112	0.000
age	1.035	1.001	39.655	0.000
sexmale	0.927	1.025	-3.091	0.002
bmi	1.013	1.002	6.381	0.000
children	1.107	1.010	10.085	0.000
smokeryes	4.732	1.031	51.333	0.000
regionnorthwest	0.938	1.036	-1.827	0.068
regionsoutheast	0.855	1.036	-4.481	0.000
regionsouthwest	0.879	1.036	-3.681	0.000

Do ciekawych wniosków można dojść zmieniając chwilowo naszą zmienną log(charges) na „charges" i wtedy jeśli palisz, koszty leczenie wzrastają średnio 4.73 razy (przy ustalonych pozostałych zmiennych) czyli 373%. Wraz ze wzrostem „children" o 1, koszty leczenia wzrastają średnio 1,11razy czyli o

11%. Jeśli mieszkasz w region „southwest”, koszty leczenia "wzrastają 0.88razy" czyli maleją o 12%.

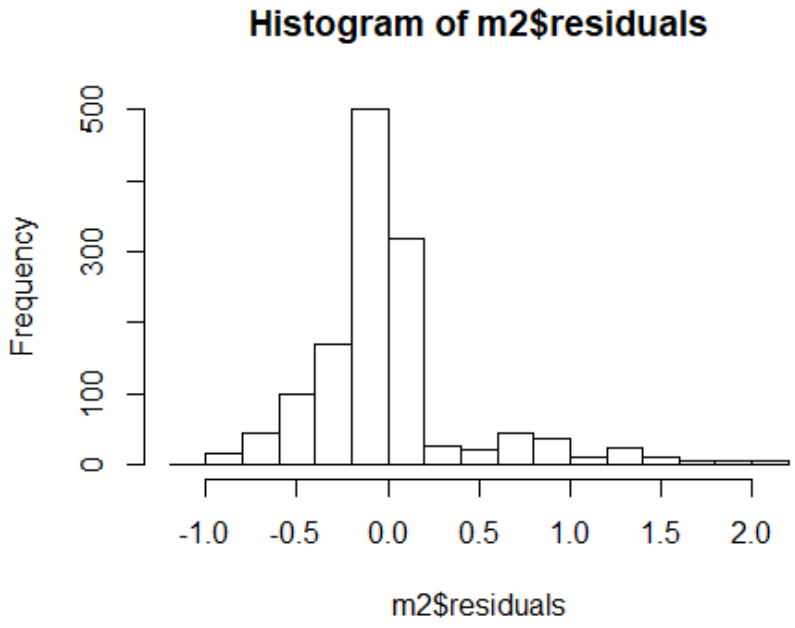
```
vif(m2)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.016822  1        1.008376
## sex      1.008900  1        1.004440
## bmi      1.106630  1        1.051965
## children 1.004011  1        1.002003
## smoker   1.012074  1        1.006019
## region   1.098893  3        1.015841
```

jeśli GVIF^(1/(2*Df)) > sqrt(10), to jest problem ze współliniowością

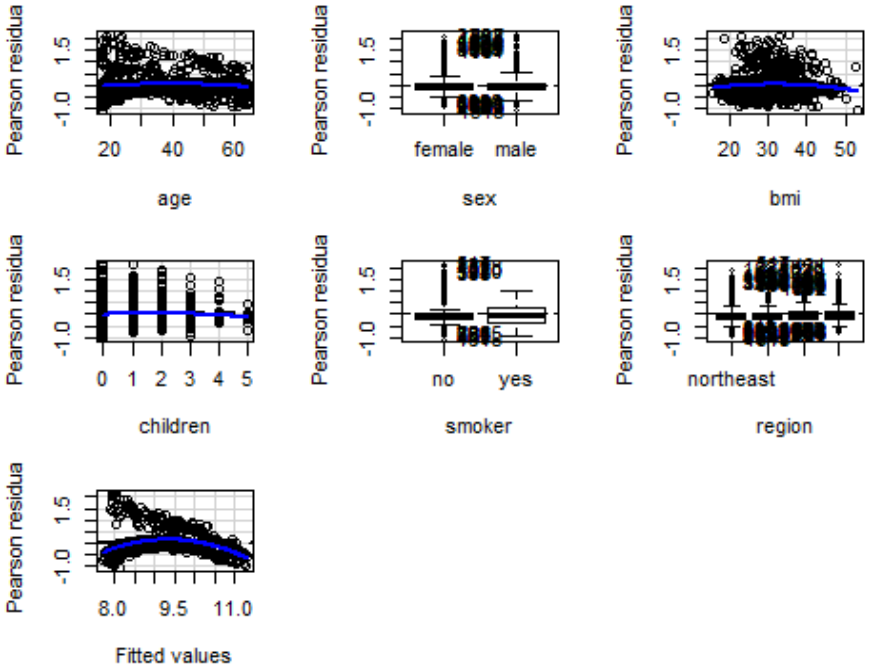
Ze względu na zmienną ”region”, która ma więcej kategorii niż dwie, liczymy tzw. „uogólniony VIF. Otrzymane wyniki są dużo niższe niż √10 więc nie ma problemu ze współliniowością.

```
hist(m2$residuals)
```



Pierwszy wykres ogólnie sprawdza czy reszty mają rozkład normalny

```
residualPlots(m2)
```

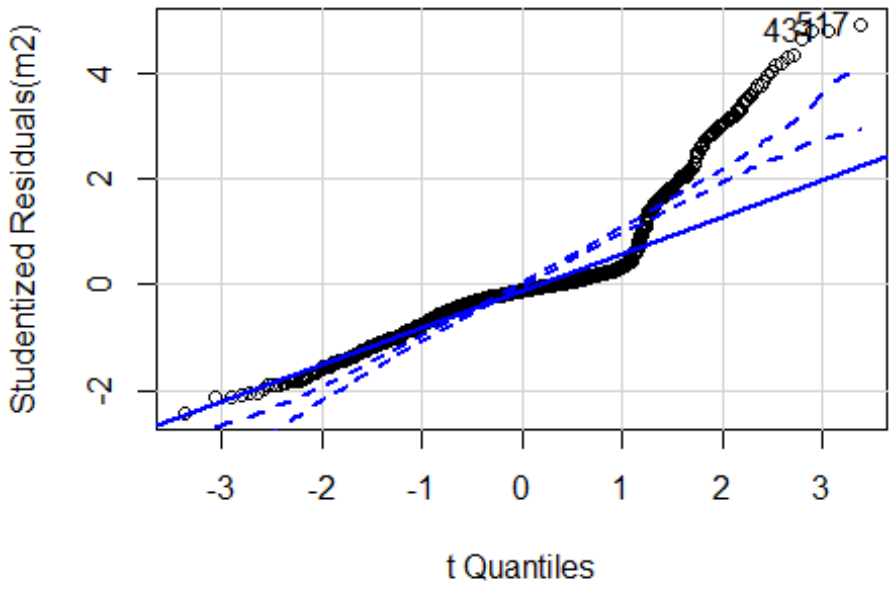


Niestety po wykresach widać, że nie są specjalnie zachowane założenia regresji, które są bardzo restrykcyjne jeżeli chodzi o liniowość. Szczególnie ostatni wykres pokazuje, że rozjeżdża się wszystko w jakimś parabolicznym kształcie.

Możliwe, że wiek należałoby podnieść do kwadratu aby uwzględnić nieliniową zależność z „charges”, chociaż zależność może być bardziej skomplikowana i należałoby podnieść do trzeciej. Ewidentnie zależności są bardziej skomplikowane niż zakłada ten model regresji.

```
##             Test stat Pr(>|Test stat|)
## age           -3.1063         0.001934 **
## sex
## bmi           -2.7069         0.006879 **
## children      -1.9867         0.047166 *
## smoker
## region
## Tukey test   -16.9453        < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
qqPlot(m2)
```



Kolejny wykres rozkład reszt w naszym modelu, który potwierdza, że nie ma rozkładu normalnego. W pewnym momencie pojawiają się dane, które wyraźnie odstają od krzywej

```
## [1] 431 517
```

```
outlierTest(m2)#zmienna odstająca(obserwacja wpływowa)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 517  4.931762         9.1785e-07    0.0012281
## 431  4.829232         1.5296e-06    0.0020466
## 220  4.819643         1.6037e-06    0.0021457
## 1028 4.652732         3.6029e-06    0.0048207
## 103  4.347975         1.4789e-05    0.0197880
## 1040 4.327581         1.6206e-05    0.0216830
## 527  4.206679         2.7653e-05    0.0370000
## 1020 4.163362         3.3380e-05    0.0446620
```

Bonferroni p pokazuje konkretne pozycje, które mogą mieć wpływ na mode, ale wyniki nie są specjalnie odstające na tyle aby można założyć, że ich usunięcie wyraźnie zmieni model.

4. Interakcje

Dodawanie interakcji aby polepszyć model Dodanie interakcji dzięki funkcji *"step()"*, model sam sobie poszuka, które są istotne.

Odejmujemy region w interakcji drugiego stopnia, żeby nie komplikować modelu jeszcze bardziej, **ale region zostaje,** .

```
m4 <- lm(log(charges) ~ (. - region)^2 + region, d) %>% step()
```

```
## Start:  AIC=-2600.84
## log(charges) ~ ((age + sex + bmi + children + smoker + region) -
##     region)^2 + region
##
##                   Df Sum of Sq    RSS     AIC
## - sex:bmi          1     0.007 186.19 -2602.8
## - bmi:children     1     0.013 186.19 -2602.8
## - sex:children     1     0.020 186.20 -2602.7
## - age:bmi          1     0.057 186.24 -2602.4
## <none>                         186.18 -2600.8
## - sex:smoker       1     0.633 186.81 -2598.3
## - age:sex          1     1.625 187.80 -2591.2
## - region           3     4.871 191.05 -2572.3
## - children:smoker  1     4.636 190.81 -2569.9
## - age:children     1     5.090 191.27 -2566.8
## - bmi:smoker       1    20.654 206.83 -2462.1
## - age:smoker       1    44.223 230.40 -2317.7
##
## Step:  AIC=-2602.79
## log(charges) ~ age + sex + bmi + children + smoker + region +
##     age:sex + age:bmi + age:children + age:smoker + sex:children +
##     sex:smoker + bmi:children + bmi:smoker + children:smoker
##
##                   Df Sum of Sq    RSS     AIC
## - bmi:children     1     0.013 186.20 -2604.7
## - sex:children     1     0.020 186.21 -2604.7
## - age:bmi          1     0.059 186.24 -2604.4
## <none>                         186.19 -2602.8
## - sex:smoker       1     0.632 186.82 -2600.3
## - age:sex          1     1.670 187.85 -2592.8
## - region           3     4.866 191.05 -2574.3
## - children:smoker  1     4.629 190.81 -2571.9
## - age:children     1     5.099 191.28 -2568.6
## - bmi:smoker       1    20.676 206.86 -2463.9
## - age:smoker       1    44.362 230.55 -2318.8
##
```

```
## Step:  AIC=-2604.7
## log(charges) ~ age + sex + bmi + children + smoker + region +
##     age:sex + age:bmi + age:children + age:smoker + sex:children +
##     sex:smoker + bmi:smoker + children:smoker
##
##                   Df Sum of Sq    RSS     AIC
## - sex:children     1     0.020 186.22 -2606.6
## - age:bmi          1     0.061 186.26 -2606.3
## <none>                         186.20 -2604.7
## - sex:smoker       1     0.626 186.82 -2602.2
## - age:sex          1     1.677 187.87 -2594.7
## - region           3     4.864 191.06 -2576.2
## - children:smoker  1     4.616 190.81 -2573.9
## - age:children     1     5.273 191.47 -2569.3
## - bmi:smoker       1    20.691 206.89 -2465.7
## - age:smoker       1    44.372 230.57 -2320.7
##
## Step:  AIC=-2606.56
## log(charges) ~ age + sex + bmi + children + smoker + region +
##     age:sex + age:bmi + age:children + age:smoker + sex:smoker +
##     bmi:smoker + children:smoker
##
##                   Df Sum of Sq    RSS     AIC
## - age:bmi          1     0.063 186.28 -2608.1
## <none>                         186.22 -2606.6
## - sex:smoker       1     0.625 186.84 -2604.1
## - age:sex          1     1.696 187.91 -2596.4
## - region           3     4.872 191.09 -2578.0
## - children:smoker  1     4.597 190.81 -2575.9
## - age:children     1     5.338 191.56 -2570.7
## - bmi:smoker       1    20.785 207.00 -2467.0
## - age:smoker       1    44.352 230.57 -2322.7
##
## Step:  AIC=-2608.1
## log(charges) ~ age + sex + bmi + children + smoker + region +
##     age:sex + age:children + age:smoker + sex:smoker + bmi:smoker +
##     children:smoker
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         186.28 -2608.1
## - sex:smoker       1     0.641 186.92 -2605.5
## - age:sex          1     1.674 187.95 -2598.1
## - region           3     4.871 191.15 -2579.6
## - children:smoker  1     4.602 190.88 -2577.4
## - age:children     1     5.375 191.66 -2572.1
## - bmi:smoker       1    20.891 207.17 -2467.9
## - age:smoker       1    44.423 230.70 -2323.9
```

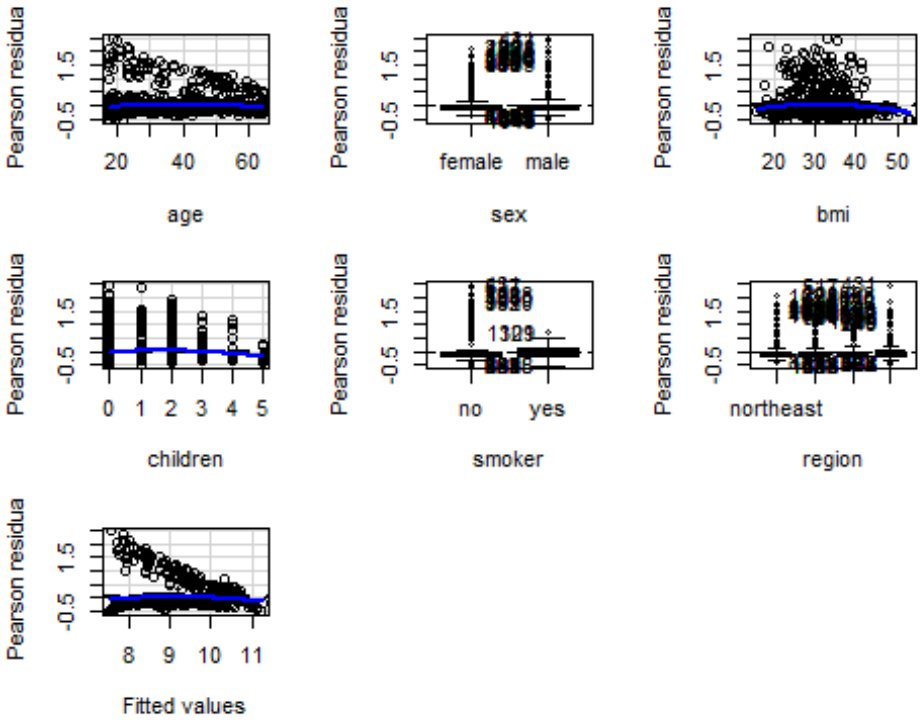
```
summary(m4)
```

```
##
## Call:
## lm(formula = log(charges) ~ age + sex + bmi + children + smoker +
##     region + age:sex + age:children + age:smoker + sex:smoker +
##     bmi:smoker + children:smoker, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51136 -0.15405 -0.08433 -0.01707  2.47390
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.0888514  0.0776978  91.236  < 2e-16 ***
## age                 0.0421160  0.0012071  34.890  < 2e-16 ***
## sexmale            -0.3080542  0.0621635  -4.956 8.15e-07 ***
## bmi                 0.0014020  0.0019885   0.705 0.480881
## children            0.2883382  0.0273731  10.534  < 2e-16 ***
## smokeryes           1.3570030  0.1438525   9.433  < 2e-16 ***
## regionnorthwest    -0.0560587  0.0295126  -1.899 0.057718 .
## regionsoutheast    -0.1404290  0.0296686  -4.733 2.45e-06 ***
## regionsouthwest    -0.1505651  0.0296077  -5.085 4.20e-07 ***
## age:sexmale         0.0050583  0.0014670   3.448 0.000583 ***
## age:children       -0.0040252  0.0006515  -6.178 8.60e-10 ***
## age:smokeryes      -0.0327328  0.0018428 -17.762  < 2e-16 ***
## sexmale:smokeryes   0.1110171  0.0520458   2.133 0.033103 *
## bmi:smokeryes       0.0502169  0.0041226  12.181  < 2e-16 ***
## children:smokeryes -0.1253703  0.0219296  -5.717 1.34e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3752 on 1323 degrees of freedom
## Multiple R-squared:  0.8352, Adjusted R-squared:  0.8335
## F-statistic:   479 on 14 and 1323 DF,  p-value: < 2.2e-16
```

```
##            Test stat Pr(>|Test stat|)
## age          -3.4054        0.0006806 ***
## sex
## bmi          -3.3228        0.0009156 ***
## children     -3.3737        0.0007630 ***
## smoker
## region
## Tukey test   -7.7792        7.299e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

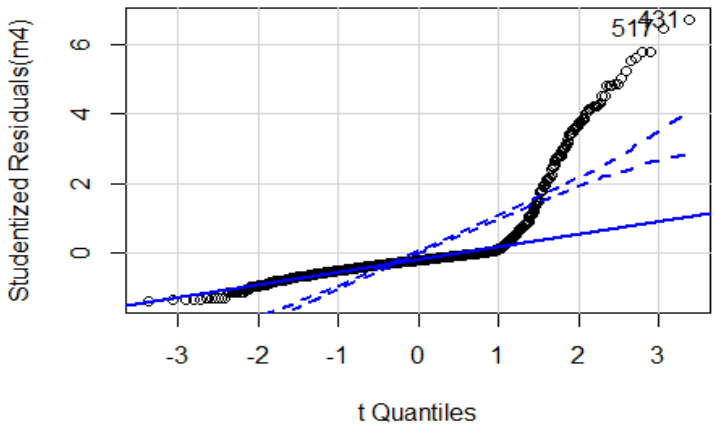
Dołączenie interakcji podniosło rozwiązanie modelu do 83,5%, niestety zamiast się upraszczać coraz bardziej się komplikuje.

```
residualPlots(m4)
```



Ciekawie wyglądają wykresy po uwzględnieniu interakcji. Powoli można wysnuć wnioski, że jest jakaś podgrupa ludzi, których koszty leczenia są po prostu niedoszacowane. Reszty są mniej więcej z przedziału **-3<reszty<3**. Z wykresu wynika, że jest bardzo duża grupa osób z resztą ”dodatnią”. To może wynikać z braku informacji na temat konkretnych schorzeń, które powodują odstawanie. Widać, że grupa osób nie jest losowa.

```
qqPlot(m4)
```



```
## [1] 431 517
```

```
outlierTest(m4)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 431  6.736167         2.4197e-11   3.2375e-08
## 517  6.489237         1.2168e-10   1.6281e-07
## 1028 5.815184         7.5858e-09   1.0150e-05
## 220  5.794173         8.5724e-09   1.1470e-05
## 103  5.632659         2.1654e-08   2.8973e-05
## 398  5.524293         3.9793e-08   5.3243e-05
## 1040 5.242712         1.8404e-07   2.4624e-04
## 341  5.026402         5.6831e-07   7.6039e-04
## 355  4.851804         1.3688e-06   1.8314e-03
## 1020 4.850980         1.3744e-06   1.8389e-03
```

4. Model otrzymany metodą Hellwiga

Na koniec chciałem stworzyć model w oparciu o metodę Hellwiga (wykorzystam funkcję Hellwig z MSAD) wybiera nie tylko najwyższe korelacje, ale również eliminuje współliniowość dlatego może uznać, że jeżeli zmienne korelują z wydatkami to dodatkowo może jakaś zmienna opisywać również inną, która jest wysoko skorelowana z „charges”.

```
Hellwig <- function(LiczbaZm, ZbiorZm, RodzajKor) {
   require(gtools)
  Zmienne_opt <- array();
  chwilowe_H <- 0;
  maxH <- 0;


 Macierz_kor <- cor(ZbiorZm, method=RodzajKor);

  Lista_pelna <- list();

  for( i in 1: (LiczbaZm-1) ) {
     Lista_pelna[[i]] <- combinations( (LiczbaZm-1), i, 2:LiczbaZm,
repeats=FALSE);
    }
  for(Liczba_el_w_komb in 1:length(Lista_pelna)) {
    H<-array();
    for( Numer_komb in 1: length(Lista_pelna[[ Liczba_el_w_komb ]][, 1]) ) {
     h<-array();
      for( Index_zm_w_komb in 1:length( Lista_pelna[[ Liczba_el_w_komb ]][ Nu
mer_komb, ])) {
 Zmienna <- Lista_pelna[[ Liczba_el_w_komb ]][ Numer_komb, Index_zm_w_komb ];
  RXY <- ( Macierz_kor[ Zmienna, 1] )^2;
  Zmienne <- Lista_pelna[[ Liczba_el_w_komb ]][ Numer_komb, ];
```

```
   Suma_kor_mianownik <- 0;
        for(k in Zmienne) {
    Suma_kor_mianownik <- Suma_kor_mianownik + abs( Macierz_kor[k, Zmienna]);
    }
    h[ Index_zm_w_komb ] <- RXY / Suma_kor_mianownik;
      }

chwilowe_H <- sum(h);
      if(chwilowe_H > maxH) {

        maxH = chwilowe_H;

        Zmienne_opt <- Zmienne-1;

      }

    }

  }
  cat("Koszta optymalne wynoszą",maxH,"\n")

  for(d in 1:length(Zmienne_opt)) {
    cat("Należy wybrać zmienną X o indeksie",Zmienne_opt[d],"\n")
  }
  }



# W celu wyeliminowania problemu ze zmienną „region”, która opisana jest
# przez cztery zmienne jakościowe utworzę macierz, która wprowadzi dodatkowe
#zmienne objaśniające


matrix <- model.matrix(charges ~ ., d)
matrix <- as.data.frame(matrix[, -1])
Dane<-data.frame(charges = log(d$charges), matrix)


Hellwig(ncol(Dane), Dane, "pearson")

## Koszta optymalne wynoszą 0.7146584
## Należy wybrać zmienną X o indeksie 1
## Należy wybrać zmienną X o indeksie 4
## Należy wybrać zmienną X o indeksie 5

# Metoda Hellwiga sugeruje wybór zmiennych o numerach 1,4,5
# czyli age, children, smoker
```

```r
xnam0<-colnames(d)[c(1,4,5)]
formula0 <- as.formula(paste("log(charges) ~ ", paste(xnam0, collapse= "+")))
formula0
```

```
## log(charges) ~ age + children + smoker
```

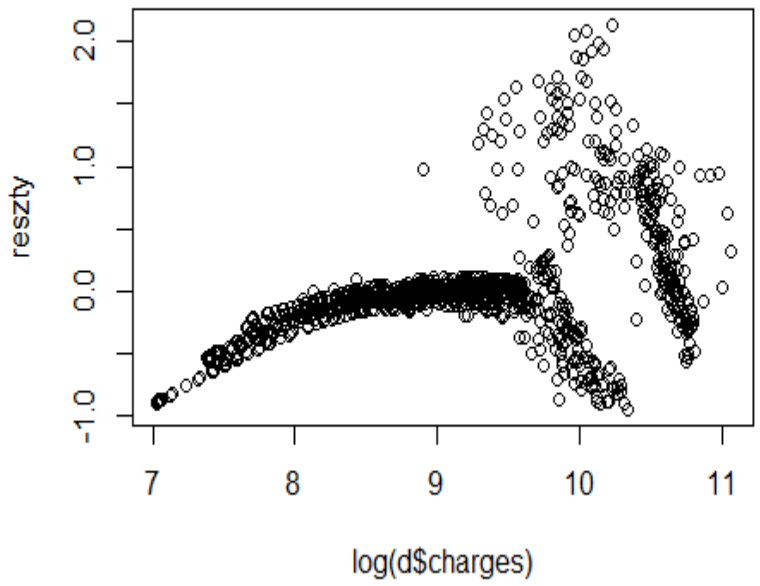
```r
#Tworzymy model dla części zmiennych

restrykcja<-lm(formula0, d)

#Ocena dopasowania

summary(restrykcja)
```

```
##
## Call:
## lm(formula = formula0, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94939 -0.17632 -0.04368  0.04252  2.13501
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.2877234  0.0387040 188.294   <2e-16 ***
## age         0.0352849  0.0008839  39.919   <2e-16 ***
## children    0.1016311  0.0102990   9.868   <2e-16 ***
## smokeryes   1.5442724  0.0307364  50.242   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4535 on 1334 degrees of freedom
## Multiple R-squared:  0.7573, Adjusted R-squared:  0.7567
## F-statistic:  1387 on 3 and 1334 DF,  p-value: < 2.2e-16
```

```r
# Wykres rozrzutu
```

Model uzyskany metodą *Hellwiga* rozwiązuje nam problem również na poziomie 75,7% czyli jest zbliżony do modelu **m2**.

```
reszty<-restrykcja$residuals
plot(log(d$charges),reszty)
```



5. Podsumowanie

Zgodnie z metodą *Hellwiga* nasze równanie przyjmuje najprostsze wzór

Log(charges)= 7,2877234 + 0.0352849age + 0.1016311children + 1.5442724smokeryes

Natomiast model **m4** rozwiązuje równanie w niecałych 84%, ale jest bardzo skomplikowany.

Niestety wykresy wskazują na bardziej skomplikowane zależności. Możliwe, że mamy do czynienia z pewnym brakiem informacji ze względu na dane dyskretne, które nie są ujawnione.