Health Insurance

Contents

Introduction The data we used are related to health insurance that were simulated based on demographic statistics collected by the United States Census Bureau (USCB), available at "https://www.kaggle.com/mirichoi0218/insurance".

The variables in this dataset correspond to parameters that influence the value of medical expenses charged by health insurance. For the implementation of linear regression, we selected the charges as the dependent variable since it depends on all the others. Therefore, with this work, we aim to assess the influence of age, gender, BMI, number of children, smoking status, and region of residence of each insured individual on the individual medical expenses charged by health insurance.

Variables Age: age of the beneficiary/insured;

Gender: gender of the insured (male/female);

BMI: body mass index that provides insight into the body, indicating weights that are relatively high or low compared to height. Ideally, the objective body weight index will be from 18.5 to 24.9 kg/m^2, for which the ratio of height to weight is used;

Children: number of children covered by health insurance (number of dependents);

Smoker: indicator of whether they smoke (yes/no);

Region: region of the US where the beneficiary lives (northeast, southeast, southwest, northwest);

Charges: individual medical expenses charged by health insurance.

Data Importation

```
# Importação do ficheiro csv
insurance = read.csv(file='insurance.csv', header= T, sep=';', dec ='.')
```

After importing the selected file from the aforementioned database, we factorized the age and BMI variables.

For the age variable, the following age groups were defined:

- **Young adult:** Age between 17 and 30
- **Middle age:** Age between 31 and 45
- **Elderly:** Age 45 or older

For the BMI variable, the following divisions were made:

- **Underweight:** BMI is less than 18.5
- **Normal weight:** BMI is from 18.5 to 24.9
- **Overweight:** BMI is from 25 to 29.9
- **Obese:** BMI is 30 or more

```
# Fatorização da variável idade em relação às taxas
jovem_adulto = (insurance$charges[0:444])
idade_media = (insurance$charges[445:838])
idade_avancada = (insurance$charges[839:1338])

values = c(jovem_adulto,idade_media,idade_avancada)
ind = c(rep('1',length(jovem_adulto)),
      rep('2',length(idade_media)),
      rep('3',length(idade_avancada)))

idade = factor(ind)

# Fatorização da variável bmi em relação às taxas

indice_massa = insurance[with(insurance, order(insurance$bmi, insurance$charges)), ]
baixo_peso = (indice_massa$charges[0:21])
peso_normal = (indice_massa$charges[22:245])
sobrepeso = (indice_massa$charges[246:631])
obeso = (indice_massa$charges[632:1338])

values2 = c(baixo_peso,peso_normal,sobrepeso,obeso)
ind2 = c(rep('1',length(baixo_peso)),
      rep('2',length(peso_normal)),
      rep('3',length(sobrepeso)),
      rep('4',length(obeso)))

bmi = factor(ind2)
```

Informations

```
# Informação preliminar do dataset
str(insurance)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age     : int  18 18 18 18 18 18 18 18 18 18 ...
##  $ sex     : chr  "male" "male" "female" "female" ...
##  $ bmi     : num  33.8 34.1 26.3 38.7 35.6 ...
##  $ children: int  1 0 0 2 0 2 0 0 0 0 ...
##  $ smoker  : chr  "no" "no" "no" "no" ...
##  $ region  : chr  "southeast" "southeast" "northeast" "northeast" ...
##  $ charges : num  1726 1137 2198 3393 2211 ...
```

```r
#Verificação da hipótese de existir dados em falta
any(is.na(insurance))
```

```
## [1] FALSE
```

The dataset does not contain missing data (NA).

This dataset contains data from 1338 individuals, distributed across 7 variables as explained earlier, with 3 of them being continuous, one discrete (number of children), and 3 categorical. Among the categorical variables, two are binary (gender and smoker), and the other has 4 levels (region). The predictive variables include age, gender, BMI, number of children, smoker, and region.

Exploratory Analysis

```r
#install.packages("summarytools")
library(summarytools)

dfSummary(insurance,na.col=F,valid.col=F)
```

```
## Data Frame Summary
## insurance
## Dimensions: 1338 x 7
## Duplicates: 1
##
## ----------------------------------------------------------------------------------------------
## No   Variable      Stats / Values              Freqs (% of Valid)     Graph
## ---- ------------- --------------------------- ---------------------- ---------------------
## 1    age           Mean (sd) : 39.2 (14)       47 distinct values     :
##      [integer]     min < med < max:                                   :
##                    18 < 39 < 64                                       : : . : : . : . : .
##                    IQR (CV) : 24 (0.4)                                : : : : : : : : : :
##                                                                       : : : : : : : : : :
##
## 2    sex           1. female                   662 (49.5%)            IIIIIIIII
##      [character]   2. male                     676 (50.5%)            IIIIIIIIII
##
## 3    bmi           Mean (sd) : 30.7 (6.1)      548 distinct values        : :
##      [numeric]     min < med < max:                                       : :
##                    16 < 30.4 < 53.1                                     . : : :
##                    IQR (CV) : 8.4 (0.2)                                 : : : :
##                                                                       . : : : : :
##
## 4    children      Mean (sd) : 1.1 (1.2)       0 : 574 (42.9%)        IIIIIIII
##      [integer]     min < med < max:            1 : 324 (24.2%)        IIII
##                    0 < 1 < 5                   2 : 240 (17.9%)        III
##                    IQR (CV) : 2 (1.1)          3 : 157 (11.7%)        II
##                                                4 :  25 ( 1.9%)
##                                                5 :  18 ( 1.3%)
##
## 5    smoker        1. no                       1064 (79.5%)           IIIIIIIIIIIIIII
##      [character]   2. yes                       274 (20.5%)           IIII
```

```
##
## 6    region        1. northeast               324 (24.2%)            IIII
##      [character]   2. northwest               325 (24.3%)            IIII
##                    3. southeast               364 (27.2%)            IIIII
##                    4. southwest               325 (24.3%)            IIII
##
## 7    charges       Mean (sd) : 13270 (12110)  1337 distinct values   :
##      [numeric]     min < med < max:                                  : .
##                    1122 < 9382 < 63770                               : :
##                    IQR (CV) : 11900 (0.9)                            : :
##                                                                      : : : : . . . .
## ----------------------------------------------------------------------------------------
```

The descriptive statistics are presented above.

Age Variable

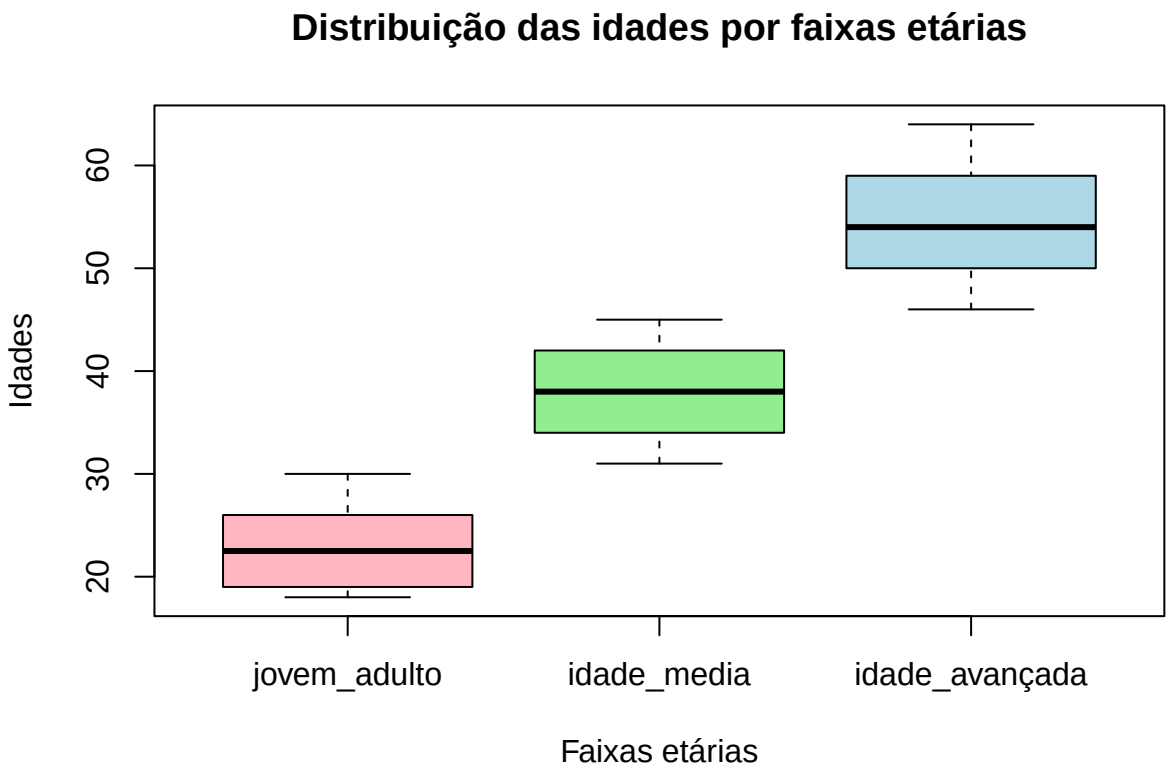
```
var(insurance$age)
```

```
## [1] 197.4
```

```
sd(insurance$age)
```

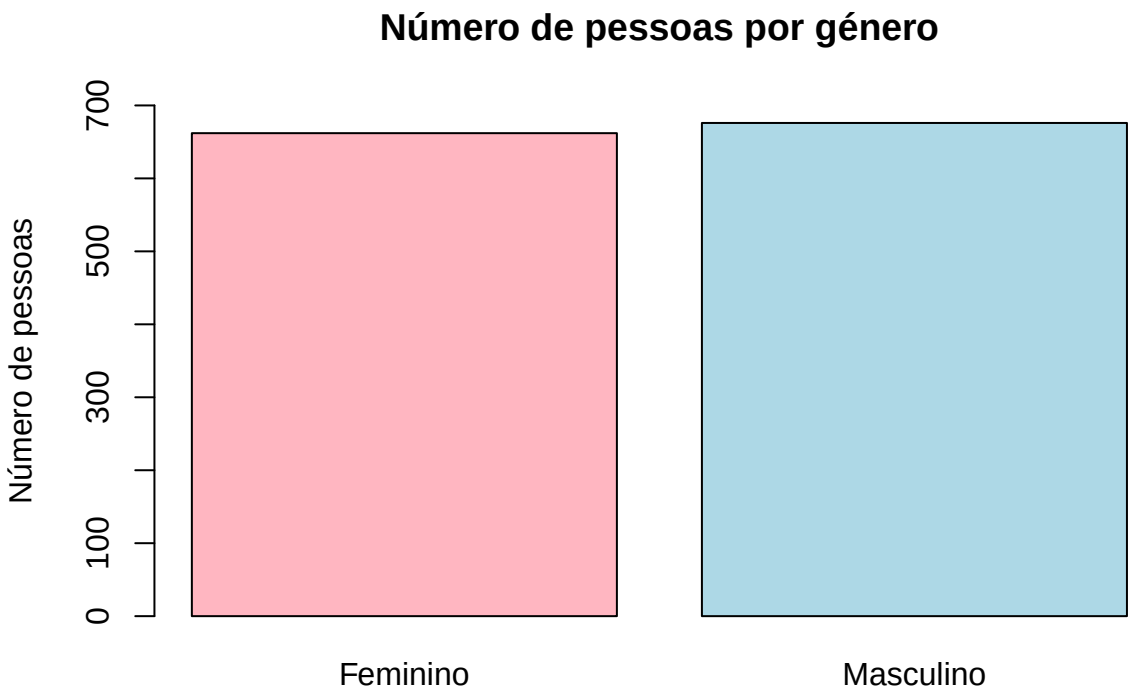
```
## [1] 14.05
```

```
boxplot (values3~idade2, names = c('jovem_adulto','idade_media','idade_avançada'), xlab = 'Faixas etári
```



Gender variable

```
genero = table(insurance$sex)
barplot (genero, ylim = c (0,700), names.arg = c('Feminino','Masculino'), col = c('lightpink','lightblu
```



BMI variable

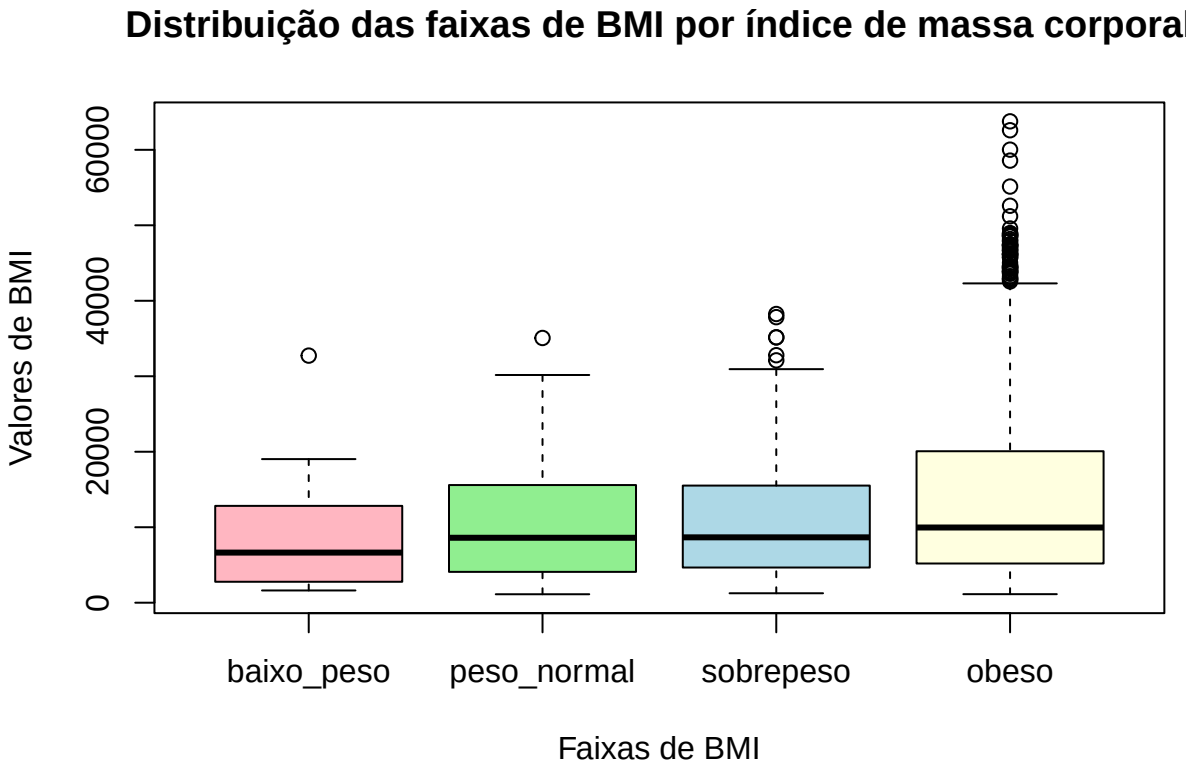
```
var(insurance$bmi)
```

```
## [1] 37.19
```

```
sd(insurance$bmi)
```

```
## [1] 6.098
```

```
boxplot (values4~bmi2, names = c('baixo_peso','peso_normal','sobrepeso','obeso'), xlab = 'Faixas de BMI
```



Children variable

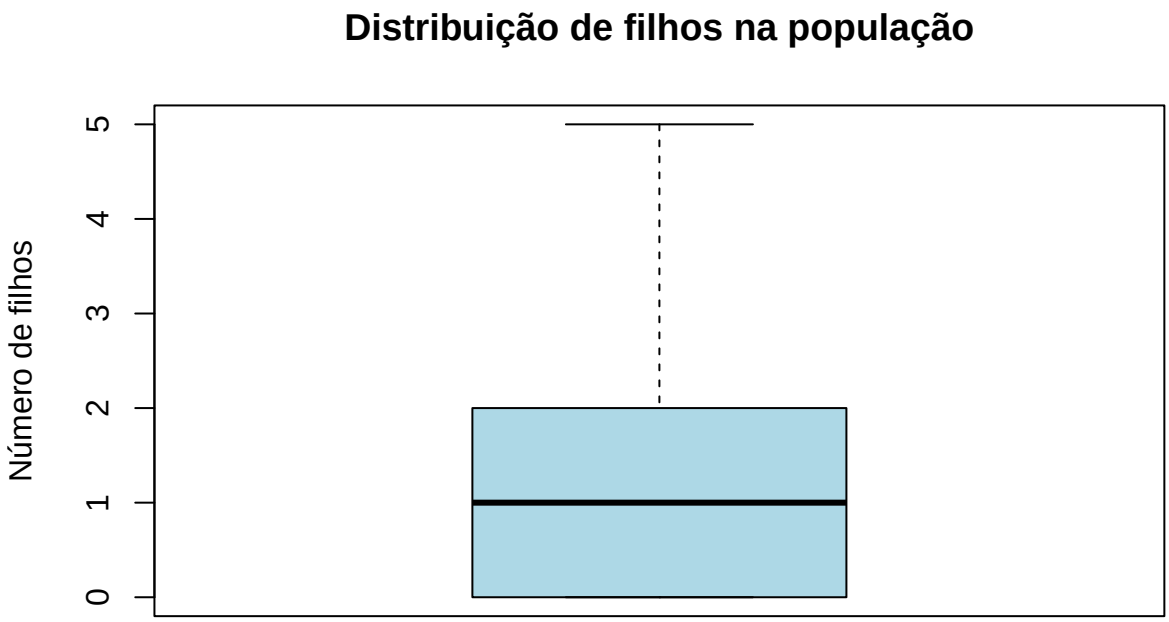
```
var(insurance$children)
```

```
## [1] 1.453
```

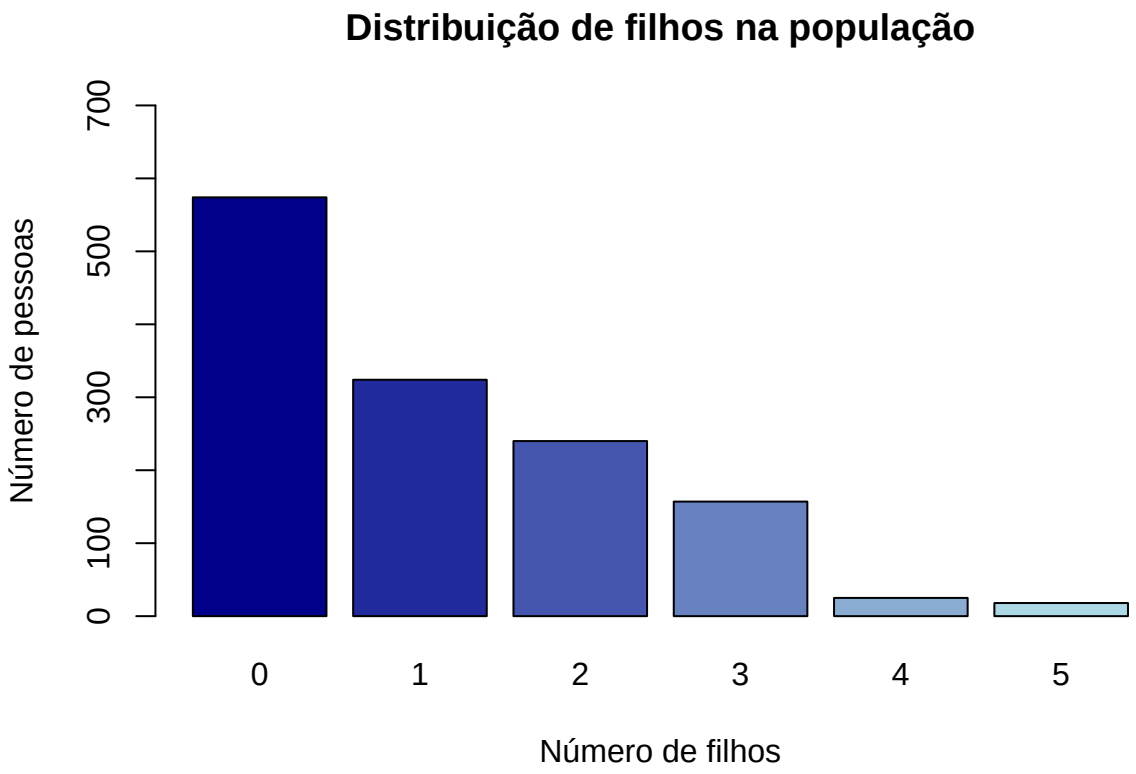
```
sd(insurance$children)
```

```
## [1] 1.205
```

```
boxplot(insurance$children, ylab = 'Número de filhos', main = 'Distribuição de filhos na população', col
```

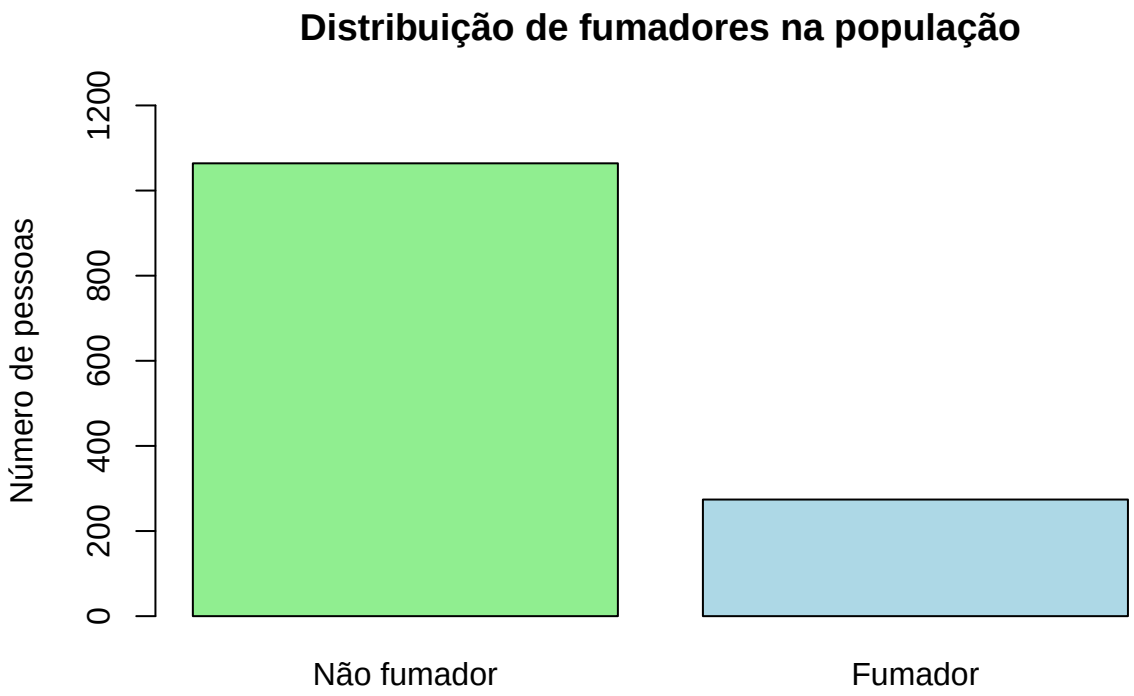



```
color <- colorRampPalette(c("darkblue","lightblue"))
barplot(table (insurance$children), ylim = c(0,700), xlab = 'Número de filhos', ylab = 'Número de pesso
```



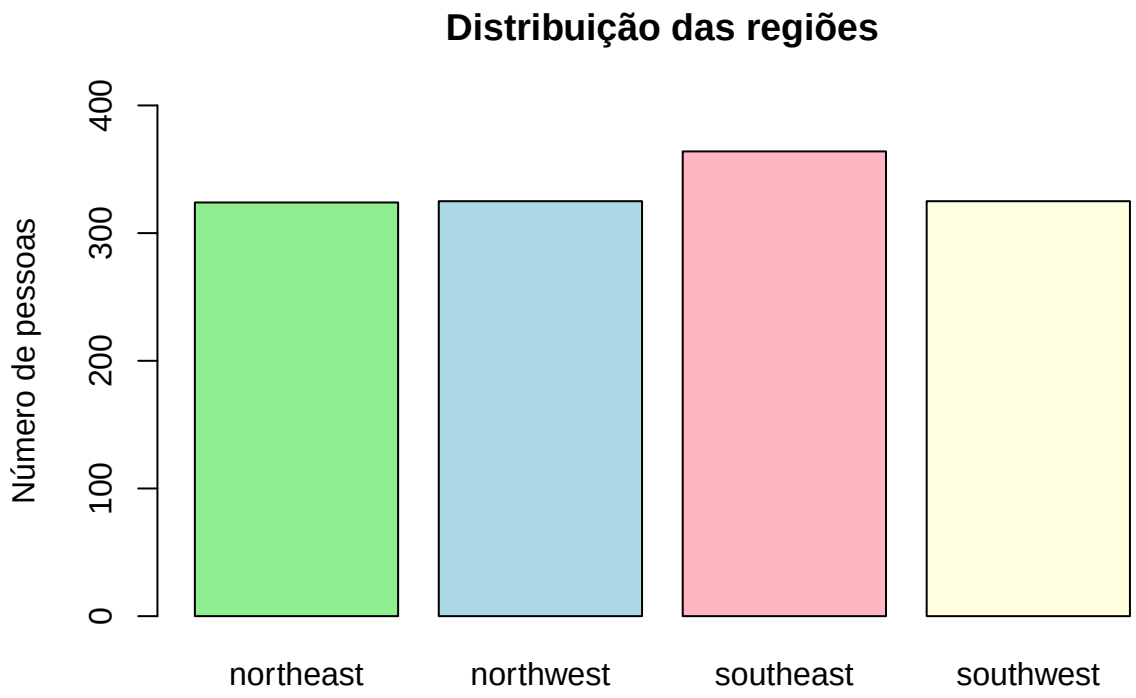
Smokers variable

```
barplot (table(insurance$smoker), ylim = c(0,1200), names = c('Não fumador','Fumador'), main = 'Distribu
```



Region variable

```
regiao = table (insurance$region)
barplot (regiao, ylim = c(0,400), main = 'Distribuição das regiões',col =c('lightgreen','lightblue','lig
```



Charges variable

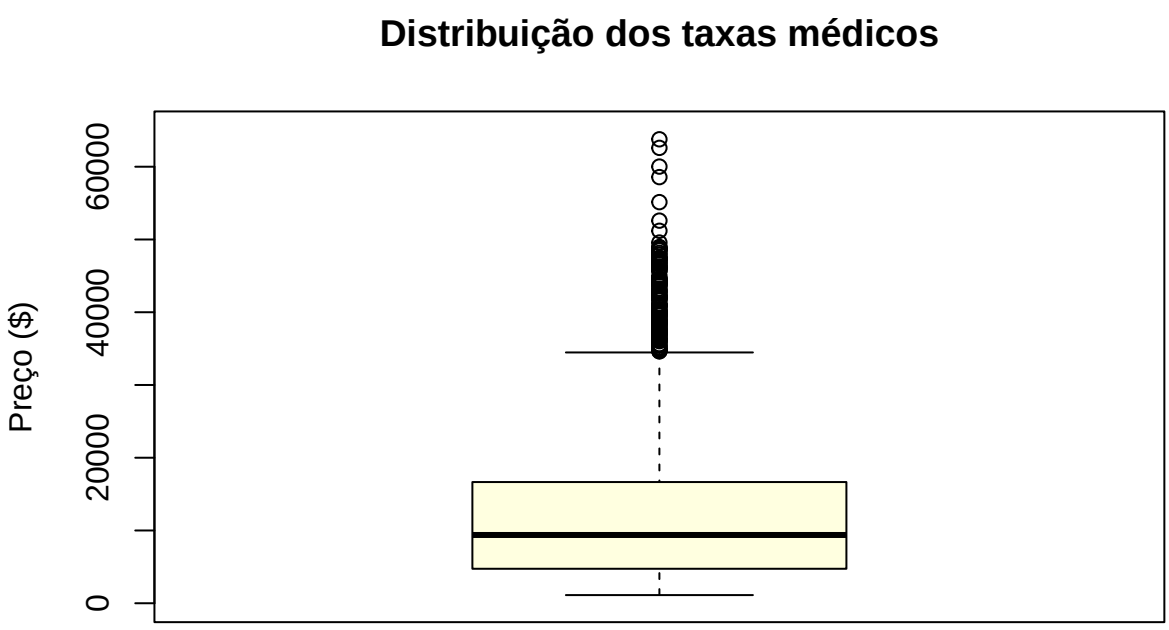
```
var(insurance$charges)
```

```
## [1] 146652372
```

```
sd(insurance$charges)
```

```
## [1] 12110
```

```
boxplot (insurance$charges, ylim= c(0,65000), main = 'Distribuição dos taxas médicos',ylab = 'Preço ($)
```

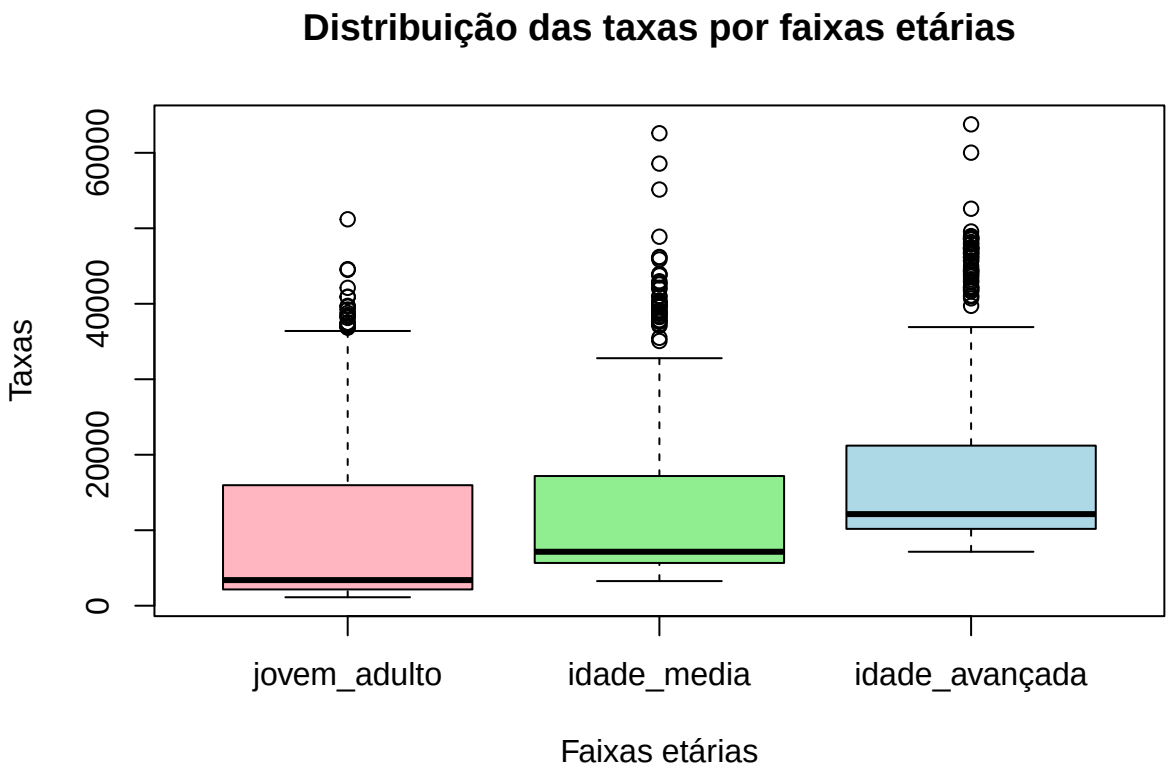


ANOVA Analysis

Since the dependent variable we aim to study is "charges", an analysis of this variable was conducted compared to the independent variables (gender, age, region, and BMI).

Age Variable

```
# Boxplot da distribuição das taxas por faixas etárias
boxplot(values~idade,names = c('jovem_adulto','idade_media','idade_avançada'), xlab = 'Faixas etárias',
```




```
#Procede-se à execução do fligner.test para comprovar a homogeneidade das variâncias.
fligner.test(insurance$charges~idade)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  insurance$charges by idade
## Fligner-Killeen:med chi-squared = 7.4, df = 2, p-value = 0.02
```

```
oneway.test(insurance$charges~idade,var.equal=F)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  insurance$charges and idade
## F = 55, num df = 2, denom df = 867, p-value <2e-16
```

```
model= aov(insurance$charges~idade)
summary(model)
```

```
##               Df   Sum Sq  Mean Sq F value Pr(>F)
## idade          2 1.45e+10 7.27e+09    53.4 <2e-16 ***
## Residuals   1335 1.82e+11 1.36e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = insurance$charges ~ idade)
##
## $idade
##     diff  lwr  upr p adj
## 2-1 3250 1356 5144 2e-04
## 3-1 7803 6019 9587 0e+00
## 3-2 4553 2710 6396 0e+00
```

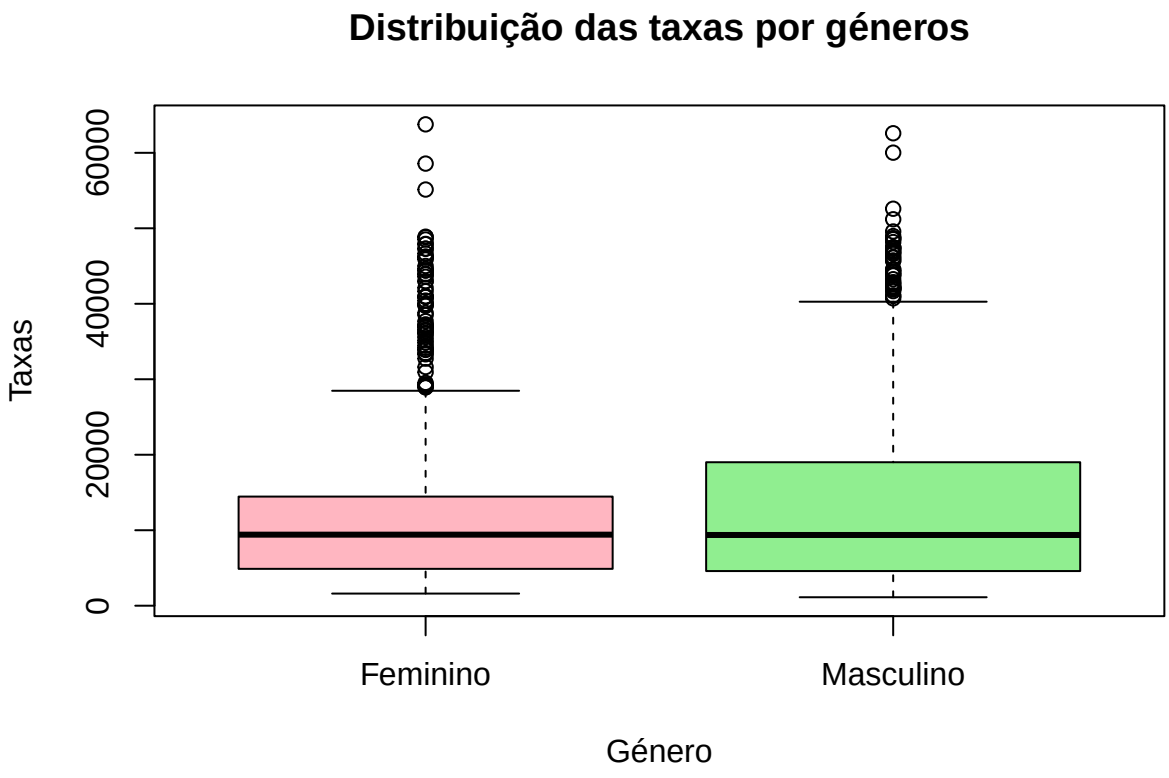
For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis that declares the homogeneity of variances.

For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis, thus assuming that there are significant differences. Therefore, we proceed with Tukey's Honestly Significant Difference (TukeyHSD) test, which is a multiple comparison test.

Outliers were observed for all factors in the boxplot of charges by age groups. The factor with a relatively higher mean concerning charges is elderly age group. Regarding the ANOVA analysis, significant differences were observed between middle age and young adult, elderly and young adult, and between elderly and middle age.

Gender variable

```r
boxplot(charges~sex, data=insurance, col = c('lightpink','lightgreen'), main = 'Distribuição das taxas
```




```
#Procede-se à execução do fligner.test para comprovar a homogeneidade das variâncias.
fligner.test(insurance$charges~insurance$sex)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  insurance$charges by insurance$sex
## Fligner-Killeen:med chi-squared = 9.4, df = 1, p-value = 0.002
```

```
oneway.test(insurance$charges~insurance$sex,var.equal=F)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  insurance$charges and insurance$sex
## F = 4.4, num df = 1, denom df = 1313, p-value = 0.04
```

```
model= aov(insurance$charges~insurance$sex)
summary(model)
```

```
##                 Df   Sum Sq  Mean Sq F value Pr(>F)
## insurance$sex    1 6.44e+08 6.44e+08     4.4  0.036 *
## Residuals     1336 1.95e+11 1.46e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = insurance$charges ~ insurance$sex)
##
## $`insurance$sex`
##             diff   lwr  upr  p adj
## male-female 1387 89.81 2685 0.0361
```

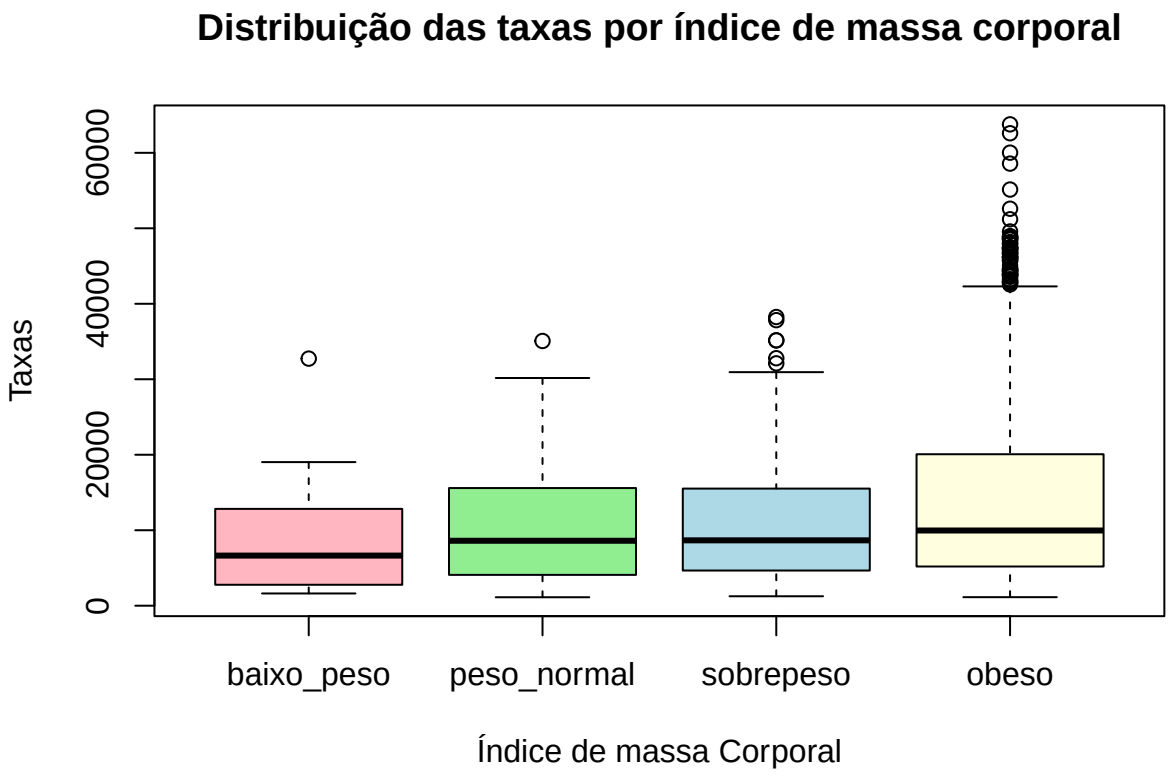
For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis that declares the homogeneity of variances.

For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis, thus assuming that there are significant differences. Therefore, we proceed with Tukey's Honestly Significant Difference (TukeyHSD) test, which is a multiple comparison test.

Through the boxplot of charges by gender, it is evident that both factors have outliers. The ANOVA analysis reveals significant differences between the levels (Female-Male) and the dependent variable.

BMI variable

```r
# Boxplot da distribuição das taxas por Índice de massa corporal
boxplot(values2~bmi, names = c('baixo_peso', 'peso_normal', 'sobrepeso', 'obeso'), xlab = 'Índice de mas
```



```r
#Procede-se à execução do fligner.test para comprovar a homogeneidade das variâncias.
fligner.test(indice_massa$charges~bmi)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  indice_massa$charges by bmi
## Fligner-Killeen:med chi-squared = 34, df = 3, p-value = 2e-07
```

```r
oneway.test(indice_massa$charges~bmi,var.equal=F)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  indice_massa$charges and bmi
## F = 20, num df = 3, denom df = 98, p-value = 3e-10
```

```r
model= aov(indice_massa$charges~bmi)
summary(model)
```

```
##               Df   Sum Sq  Mean Sq F value  Pr(>F)
## bmi            3 7.94e+09 2.65e+09    18.8 6.3e-12 ***
## Residuals   1334 1.88e+11 1.41e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = indice_massa$charges ~ bmi)
##
## $bmi
##     diff     lwr   upr  p adj
## 2-1 1777 -5194.5  8748 0.9136
## 3-1 2330 -4515.0  9175 0.8175
## 4-1 6895   130.5 13659 0.0438
## 3-2  553 -2012.8  3119 0.9454
## 4-2 5118  2775.7  7460 0.0000
## 4-3 4565  2631.6  6498 0.0000
```

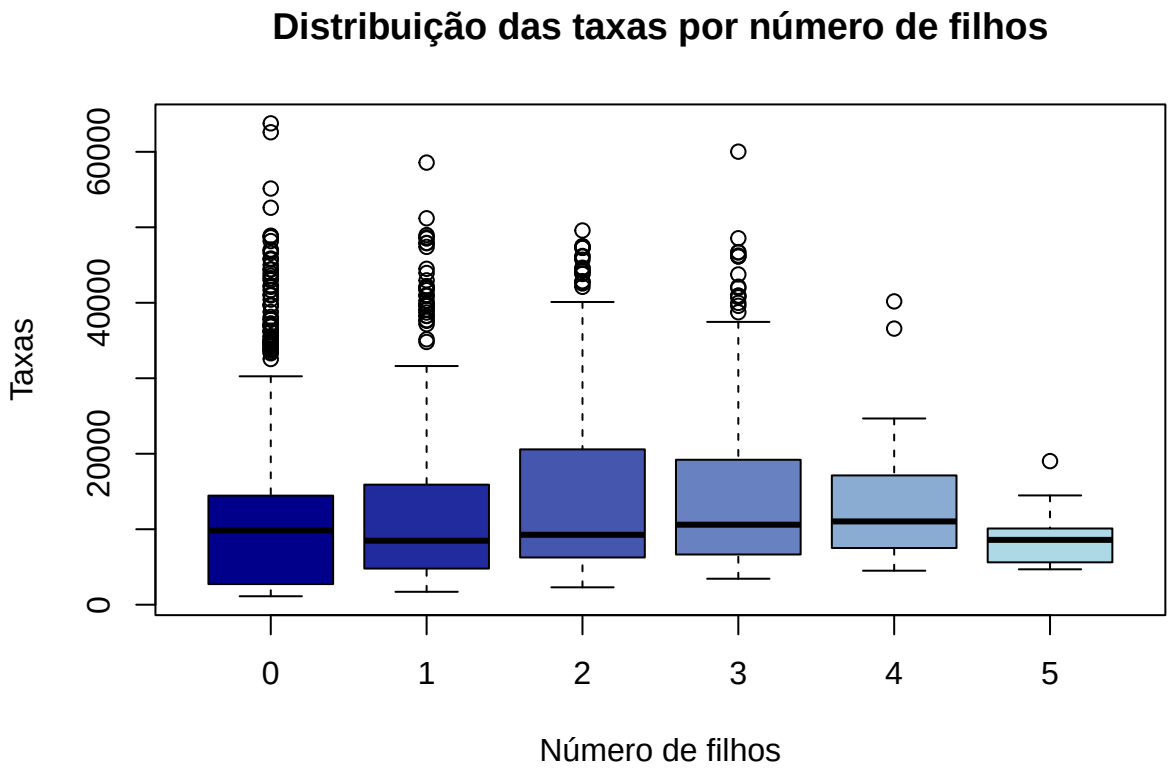
For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis that declares the homogeneity of variances.

For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis, thus assuming that there are significant differences. Therefore, we proceed with Tukey's Honestly Significant Difference (TukeyHSD) test, which is a multiple comparison test.

According to the analysis of the boxplot of charges by BMI, outliers are only present for "obesity", which also has the highest mean regarding charges. Through the ANOVA results analysis, significant differences are observed only between obesity and the other factors (underweight, normal weight, and overweight). Thus, the other factors do not show differences between them.

Children variable

```
boxplot(charges~children, data=insurance, col = color (6), main = 'Distribuição das taxas por número de
```




```
#Procede-se à execução do fligner.test para comprovar a homogeneidade das variâncias.
children = factor (insurance$children)
fligner.test(insurance$charges~children)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  insurance$charges by children
## Fligner-Killeen:med chi-squared = 22, df = 5, p-value = 5e-04
```

```
oneway.test(insurance$charges~children,var.equal=F)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  insurance$charges and children
## F = 6.9, num df = 5, denom df = 122, p-value = 1e-05
```

```r
model= aov(insurance$charges~children)
summary(model)
```

```
##               Df   Sum Sq  Mean Sq F value Pr(>F)
## children       5 2.40e+09 4.79e+08     3.3 0.0058 **
## Residuals   1332 1.94e+11 1.45e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = insurance$charges ~ children)
##
## $children
##        diff       lwr  upr  p adj
## 1-0   365.2  -2026.08 2756 0.9980
## 2-0  2707.6     62.31 5353 0.0413
## 3-0  2989.3   -110.03 6089 0.0660
## 4-0  1484.7  -5546.18 8516 0.9909
## 5-0 -3579.9 -11817.33 4657 0.8169
## 2-1  2342.4   -588.38 5273 0.2025
## 3-1  2624.1   -722.20 5970 0.2211
## 4-1  1119.5  -6023.69 8263 0.9978
## 5-1 -3945.1 -12278.59 4388 0.7562
## 3-2   281.8  -3250.57 3814 0.9999
## 4-2 -1222.9  -8455.07 6009 0.9968
## 5-2 -6287.5 -14697.39 2122 0.2705
## 4-3 -1504.7  -8914.98 5906 0.9924
## 5-3 -6569.3 -15132.83 1994 0.2434
## 5-4 -5064.6 -15702.35 5573 0.7517
```

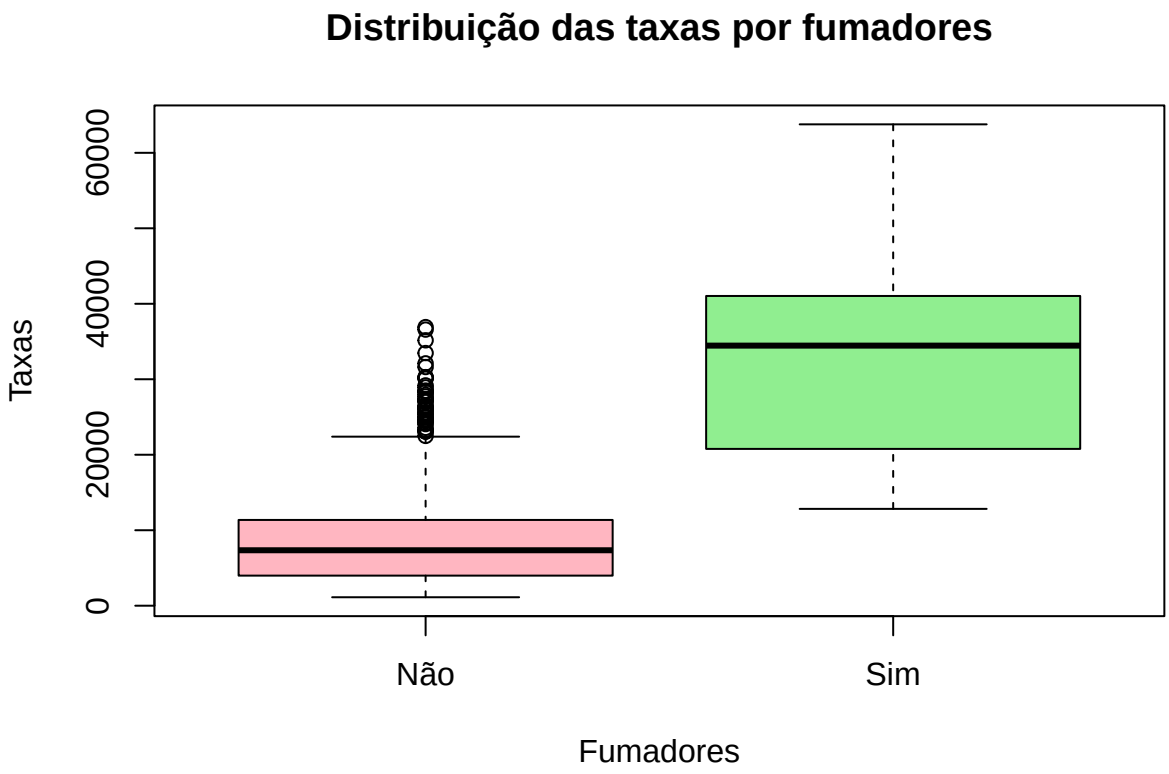
For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis that declares the homogeneity of variances.

For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis, thus assuming that there are significant differences. Therefore, we proceed with Tukey's Honestly Significant Difference (TukeyHSD) test, which is a multiple comparison test.

According to the analysis of the boxplot of charges by the number of children variable, outliers are present. The factor 4 (number of children) has the highest mean regarding charges. Through the ANOVA results analysis, significant differences are observed between 0 and 2 children.

Smoker variable

```r
boxplot(charges~smoker, data=insurance, col = c('lightpink','lightgreen'), main = 'Distribuição das taxa
```




```
#Procede-se à execução do fligner.test para comprovar a homogeneidade das variâncias.
fligner.test(insurance$charges~insurance$smoker)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  insurance$charges by insurance$smoker
## Fligner-Killeen:med chi-squared = 238, df = 1, p-value <2e-16
```

```
oneway.test(insurance$charges~insurance$smoker,var.equal = F)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  insurance$charges and insurance$smoker
## F = 1073, num df = 1, denom df = 312, p-value <2e-16
```

```
oneway.test(insurance$charges~insurance$smoker,var.equal = T)
```

```
##
##  One-way analysis of means
##
## data:  insurance$charges and insurance$smoker
## F = 2178, num df = 1, denom df = 1336, p-value <2e-16
```

```r
model= aov(insurance$charges~insurance$smoker)
summary(model)
```

```
##                     Df   Sum Sq  Mean Sq F value Pr(>F)
## insurance$smoker     1 1.22e+11 1.22e+11    2178 <2e-16 ***
## Residuals         1336 7.46e+10 5.58e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = insurance$charges ~ insurance$smoker)
##
## $`insurance$smoker`
##         diff   lwr   upr p adj
## yes-no 23616 22623 24609     0
```

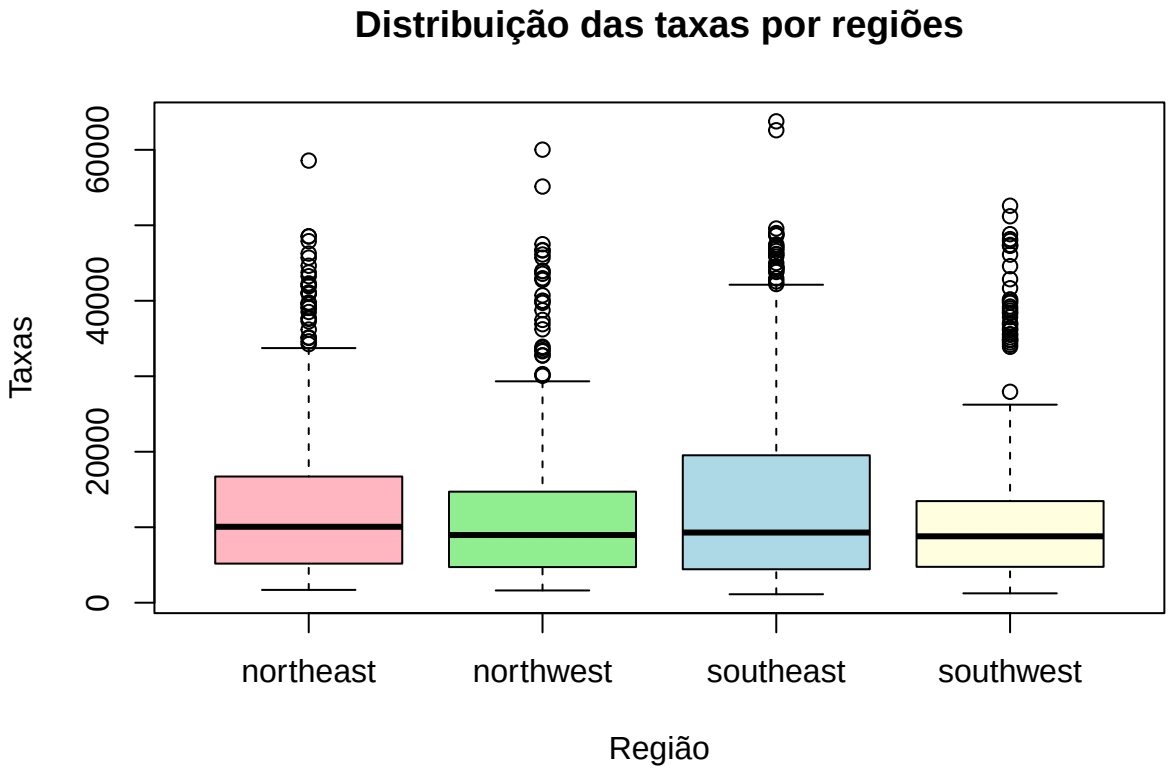
For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis that declares the homogeneity of variances.

For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis, thus assuming that there are significant differences. Therefore, we proceed with Tukey's Honestly Significant Difference (TukeyHSD) test, which is a multiple comparison test.

According to the analysis of the boxplot of charges by the smoker variable, outliers are only present for "no", which also has the lowest mean regarding charges. Through the ANOVA results analysis, significant differences are observed between the levels (No/Yes).

Region variable

```r
boxplot(charges~region, data=insurance, col = c('lightpink','lightgreen', 'lightblue', 'lightyellow'), m
```




```
#Procede-se à execução do fligner.test e shapiro.test para comprovar a homogeneidade das variâncias
fligner.test(insurance$charges~insurance$region)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  insurance$charges by insurance$region
## Fligner-Killeen:med chi-squared = 19, df = 3, p-value = 2e-04
```

```
oneway.test(insurance$charges~insurance$region,var.equal=F)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  insurance$charges and insurance$region
## F = 2.6, num df = 3, denom df = 741, p-value = 0.05
```

```
model= aov(insurance$charges~insurance$region)
summary(model)
```

```
##                    Df   Sum Sq  Mean Sq F value Pr(>F)
## insurance$region    3 1.30e+09 4.34e+08    2.97  0.031 *
## Residuals        1334 1.95e+11 1.46e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(model)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = insurance$charges ~ insurance$region)
##
## $`insurance$region`
##                           diff     lwr     upr  p adj
## northwest-northeast  -988.81 -3428.9 1451.32 0.7245
## southeast-northeast  1329.03 -1044.9 3703.00 0.4745
## southwest-northeast -1059.45 -3499.6 1380.68 0.6792
## southeast-northwest  2317.84   -54.2 4689.87 0.0583
## southwest-northwest   -70.64 -2508.9 2367.61 0.9999
## southwest-southeast -2388.47 -4760.5  -16.44 0.0477
```

For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis that declares the homogeneity of variances.

For an alpha of 0.05, since the p-value is less than alpha, we reject the null hypothesis, thus assuming that there are significant differences. Therefore, we proceed with Tukey's Honestly Significant Difference (TukeyHSD) test, which is a multiple comparison test.

Regarding the region variable, outliers are present in all factors. It is noteworthy that the factor with the highest mean regarding charges is "northeast". Considering the results from the ANOVA analysis, significant differences are observed only between southwest and southeast. Therefore, the other factors do not show differences between them.

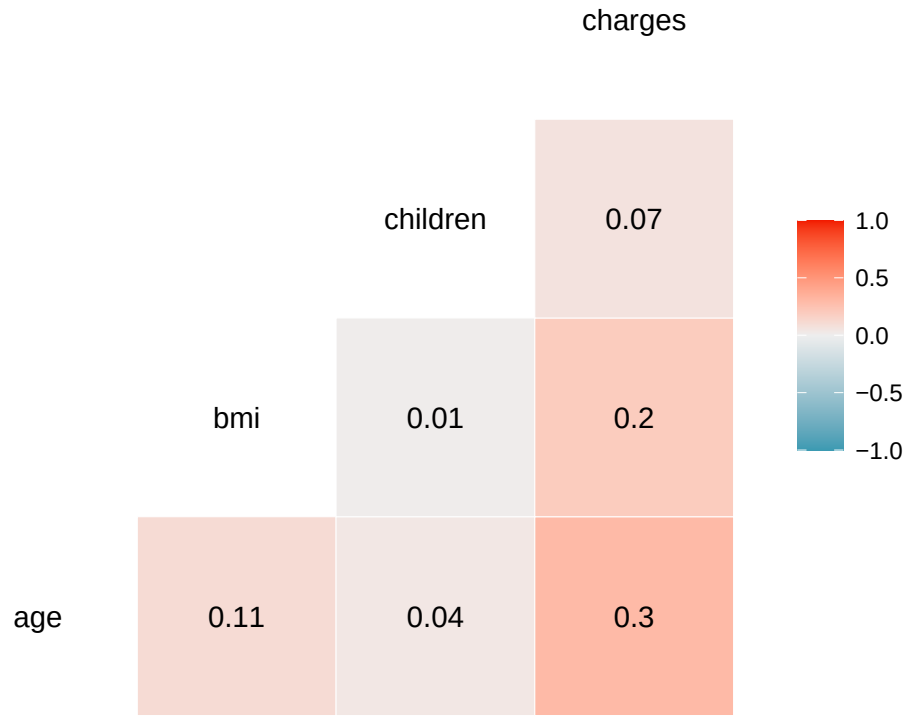
Correlations

Based on the Pearson correlation coefficient, the correlation between the quantitative variables was performed, which in this case are also continuous.

```r
# Criação de dataset com apenas variáveis numéricas
insurance_num <- insurance
insurance_num$sex<- NULL
insurance_num$region<- NULL
insurance_num$smoker<- NULL

# Correlações
#install.packages('ggplot2')
library (ggplot2)
library (GGally)

ggcorr(insurance_num, geom="tile", label= T, label_alpha=F, label_round=2)
```



Therefore, only correlations with values greater than 0.5 were considered relevant for the case study, assuming them as strong and positive correlations (1 is considered a perfect linear relationship).

Since none of the correlations obtained a value greater than 0.5, we can verify that none of the variables has a strong correlation with the insurance charges. However, "charges" has a weak positive correlation with "age", "bmi", and "children", with the highest correlation being between "charges" and "age" (0.3), which is logical and expected.

Linear Regression (Full Model)

```
full.model<- lm(charges~., data=insurance)
summary(full.model)
```

```
##
## Call:
## lm(formula = charges ~ ., data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11305  -2848   -982   1394  29993
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8  -12.09  < 2e-16 ***
```

```
## age               256.9       11.9   21.59  < 2e-16 ***
## sexmale          -131.3      332.9   -0.39  0.69335
## bmi               339.2       28.6   11.86  < 2e-16 ***
## children          475.5      137.8    3.45  0.00058 ***
## smokeryes       23848.5      413.2   57.72  < 2e-16 ***
## regionnorthwest   -353.0      476.3   -0.74  0.45877
## regionsoutheast  -1035.0      478.7   -2.16  0.03078 *
## regionsouthwest   -960.1      477.9   -2.01  0.04476 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1329 degrees of freedom
## Multiple R-squared:  0.751,  Adjusted R-squared:  0.749
## F-statistic:  501 on 8 and 1329 DF,  p-value: <2e-16
```

```
#Teste a multicolinearidade
car::vif(full.model)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## age      1.017  1           1.008
## sex      1.009  1           1.004
## bmi      1.107  1           1.052
## children 1.004  1           1.002
## smoker   1.012  1           1.006
## region   1.099  3           1.016
```

Diagnostic measures: The p-value and F-statistics have very low values, indicating that at least one predictor variable is related to the charges.

R^2 -> 1, the value tends to 1, indicating that the predictor variables are well fitted in the model.

RSE -> 0, values below 0 produce lower model errors.

The value of RSE is 6062 and R^2 is 75%. The intercept is -11938.5 and almost all predictor variables, except gender and the northwest region, are significant according to their p-values. The interpretation of categorical variables, e.g., "smoker", can be interpreted as "average charges increase by 23848.5 if the individual is a smoker - with all other variables held constant". The coefficient value, when significant, is the average change in charges with a one-unit increase in the predictor variable - with the others held constant. While correlation measures the strength of the relationship, the coefficient quantifies the relationship and allows predictions from an equation. For example, for each unit added to age, the expected average charge is 256.9 higher - after controlling for other variables.

```
library (MASS)

step.model <- stepAIC(full.model, direction="both", trace=FALSE)
summary(step.model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -11367  -2835   -980   1362  29936
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11990.3      978.8  -12.25  < 2e-16 ***
## age                257.0       11.9   21.61  < 2e-16 ***
## bmi                338.7       28.6   11.86  < 2e-16 ***
## children           474.6      137.7    3.45  0.00059 ***
## smokeryes        23836.3      411.9   57.88  < 2e-16 ***
## regionnorthwest   -352.2      476.1   -0.74  0.45962
## regionsoutheast  -1034.4      478.5   -2.16  0.03083 *
## regionsouthwest   -959.4      477.8   -2.01  0.04485 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.751,  Adjusted R-squared:  0.75
## F-statistic:  573 on 7 and 1330 DF,  p-value: <2e-16
```

```
AIC(step.model)
```

```
## [1] 27114
```

```
step.modelf <- stepAIC(full.model, direction='forward', trace=FALSE)
summary(step.modelf)
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11305  -2848   -982   1394  29993
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8  -12.09  < 2e-16 ***
## age                256.9       11.9   21.59  < 2e-16 ***
## sexmale           -131.3      332.9   -0.39  0.69335
## bmi                339.2       28.6   11.86  < 2e-16 ***
## children           475.5      137.8    3.45  0.00058 ***
## smokeryes        23848.5      413.2   57.72  < 2e-16 ***
## regionnorthwest   -353.0      476.3   -0.74  0.45877
## regionsoutheast  -1035.0      478.7   -2.16  0.03078 *
## regionsouthwest   -960.1      477.9   -2.01  0.04476 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1329 degrees of freedom
## Multiple R-squared:  0.751,  Adjusted R-squared:  0.749
## F-statistic:  501 on 8 and 1329 DF,  p-value: <2e-16
```

```
AIC(step.modelf)
```

```
## [1] 27116
```

Plots

The Residuals vs Fitted plot linearly relates the residuals and the predictor variable.

The Normal Q-Q plot visually assesses whether the residuals follow a normal distribution. Later, a Shapiro test was performed to verify normality.

The Scale-Location plot checks the homogeneity of residuals and the constant variance regression criterion.

The Residuals vs Leverage plot identifies the values that most influence the regression.

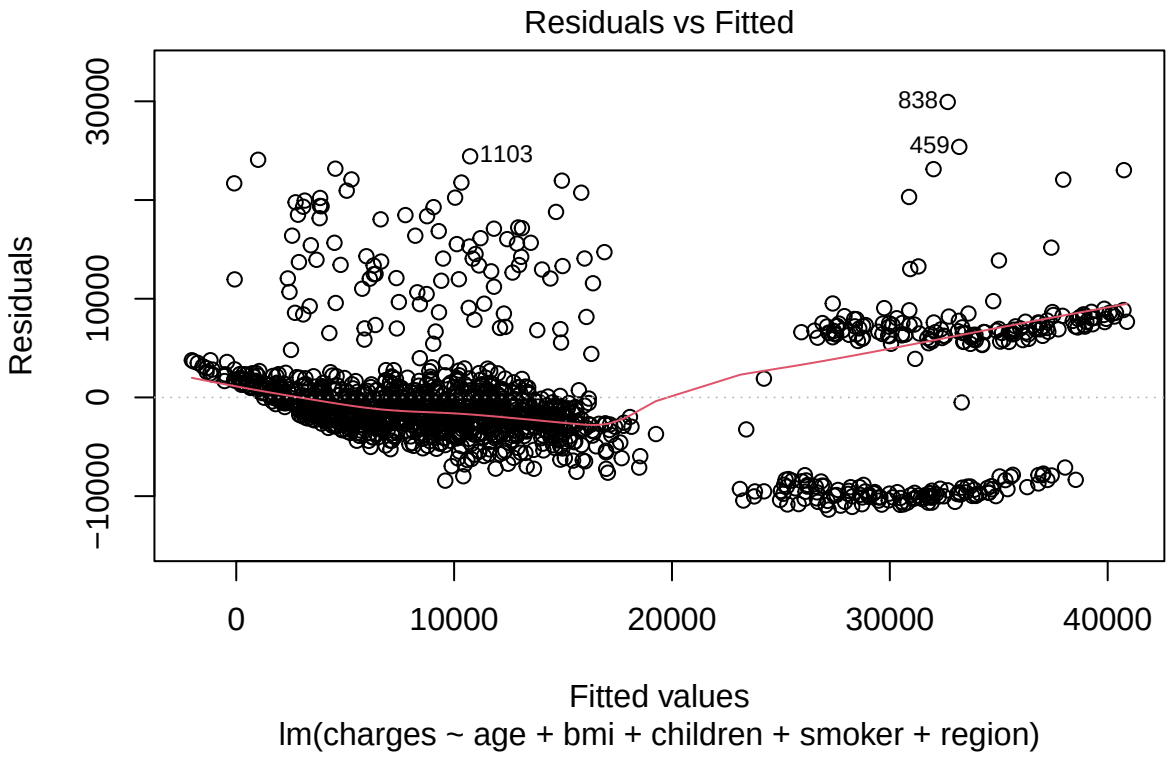
```
step.modelb <- stepAIC(full.model, direction='backward', trace=FALSE)
summary(step.modelb)
```

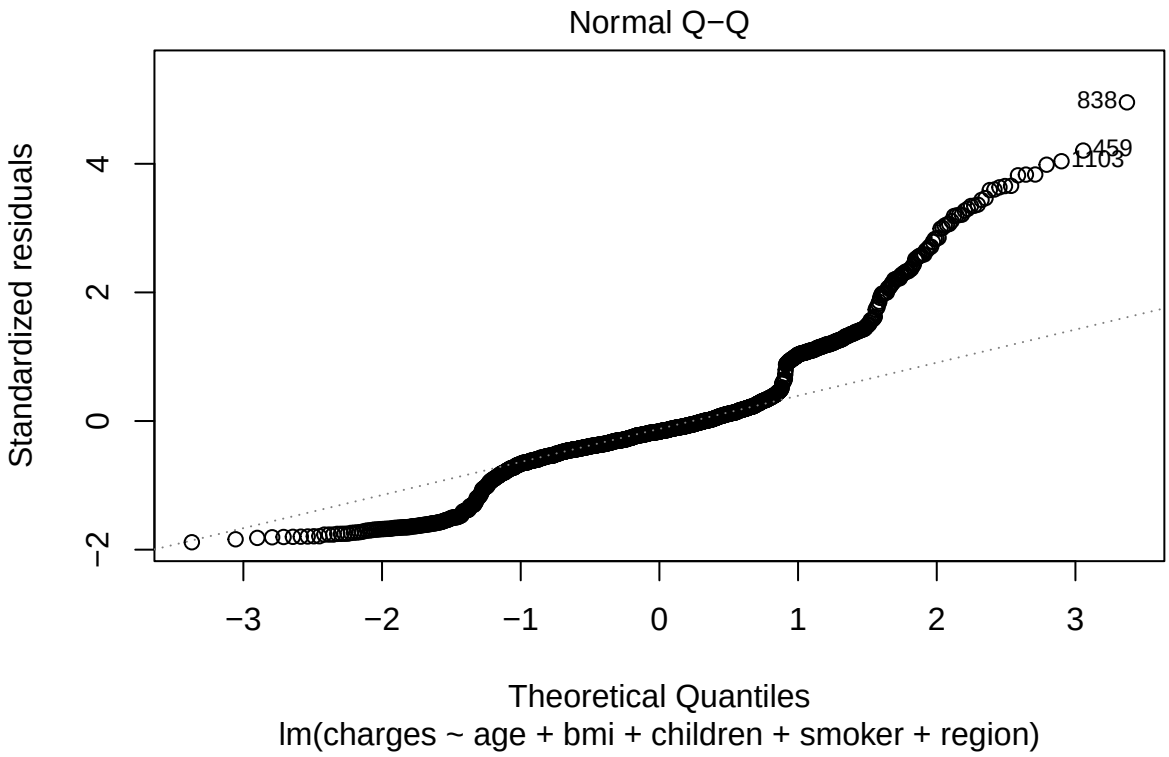
```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11367  -2835   -980   1362  29936
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11990.3      978.8  -12.25  < 2e-16 ***
## age                257.0       11.9   21.61  < 2e-16 ***
## bmi                338.7       28.6   11.86  < 2e-16 ***
## children           474.6      137.7    3.45  0.00059 ***
## smokeryes        23836.3      411.9   57.88  < 2e-16 ***
## regionnorthwest   -352.2      476.1   -0.74  0.45962
## regionsoutheast  -1034.4      478.5   -2.16  0.03083 *
## regionsouthwest   -959.4      477.8   -2.01  0.04485 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.751,  Adjusted R-squared:  0.75
## F-statistic:  573 on 7 and 1330 DF,  p-value: <2e-16
```

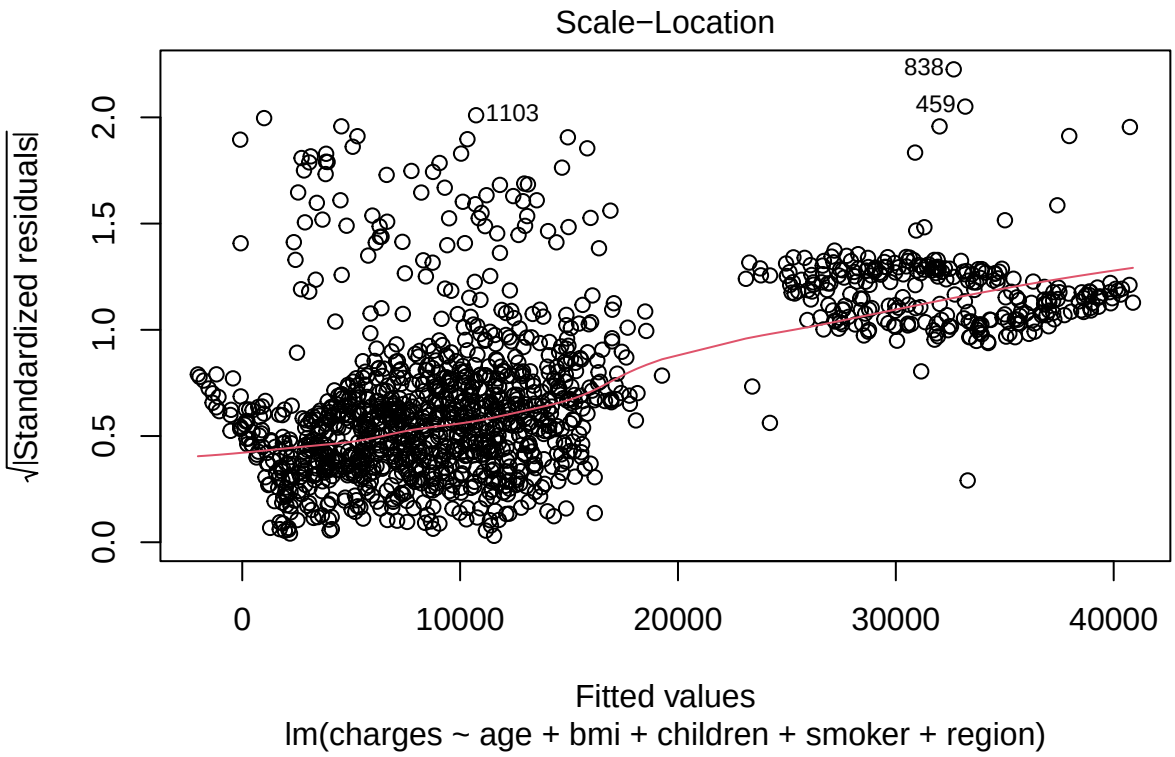
```
AIC(step.modelb)
```

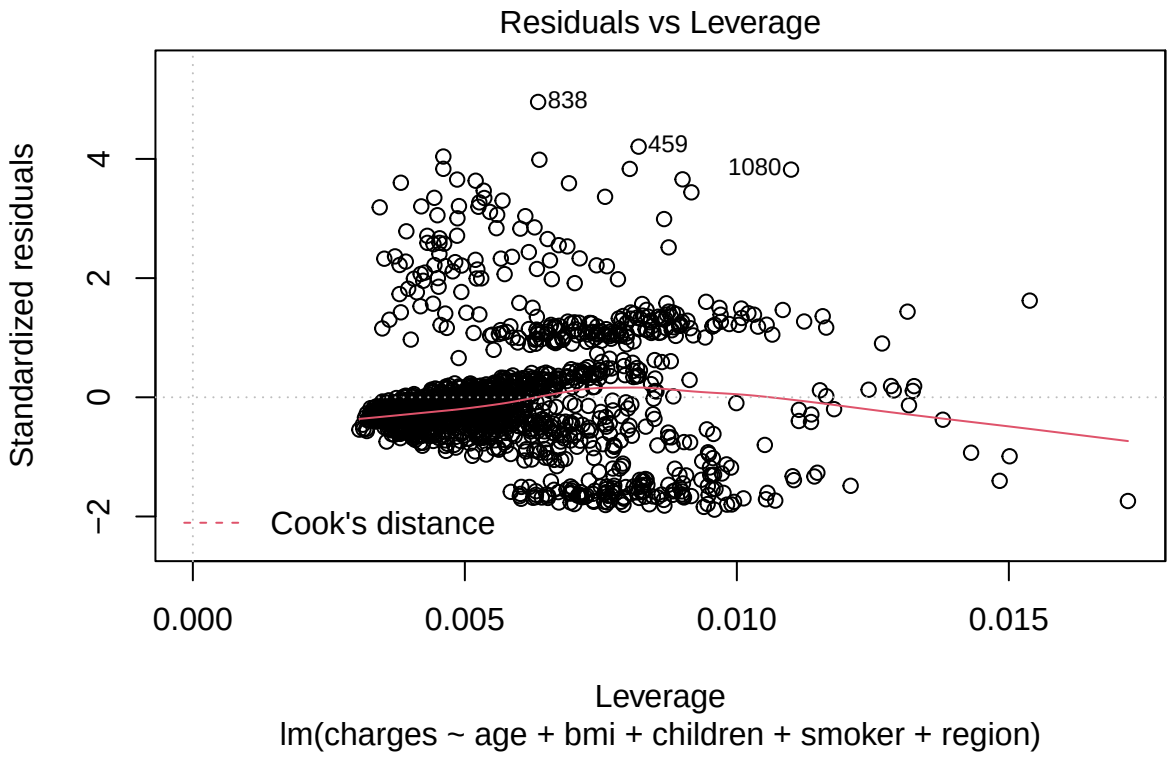
```
## [1] 27114
```

```
plot(step.modelb)
```


Residuals vs Fitted
Residuals
30000
10000
0
−10000
1103
838
459
0
10000
20000
30000
40000
Fitted values
lm(charges ~ age + bmi + children + smoker + region)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ age + bmi + children + smoker + region)
838
459
1103
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1103
838
459
0
10000
20000
30000
40000
Fitted values
lm(charges ~ age + bmi + children + smoker + region)



No gráfico Residuals vs Fitted, the non-horizontal red line may indicate a non-linear relationship.

In the Normal Q-Q plot, we see that the residuals are not exactly on the straight line, indicating that they are not normally distributed.

In the Scale-Location plot, the non-straight line indicates heteroscedasticity (different variances for all observations).

The contradictory assumptions obtained indicate that this model does not allow for reliable conclusions.

Finally, the stepwise selection method was performed from the full model fit. For this, the 'backward', 'forward', and 'both' direction methods were used. Then, a comparison of the AIC values for the three aforementioned methods was conducted, and the 'backward' method was chosen because it has the lowest AIC value.

```
anova(full.model)
```

```
## Analysis of Variance Table
##
## Response: charges
##             Df   Sum Sq  Mean Sq F value  Pr(>F)
## age          1 1.75e+10 1.75e+10  477.02 < 2e-16 ***
## sex          1 7.92e+08 7.92e+08   21.54 3.8e-06 ***
## bmi          1 5.26e+09 5.26e+09  143.07 < 2e-16 ***
## children     1 5.51e+08 5.51e+08   15.00 0.00011 ***
## smoker       1 1.23e+11 1.23e+11 3343.50 < 2e-16 ***
## region       3 2.33e+08 7.78e+07    2.12 0.09622 .
## Residuals 1329 4.88e+10 3.67e+07
```

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

```
anova(step.modelf)
```

```
## Analysis of Variance Table
##
## Response: charges
##             Df   Sum Sq  Mean Sq F value  Pr(>F)
## age          1 1.75e+10 1.75e+10  477.02 < 2e-16 ***
## sex          1 7.92e+08 7.92e+08   21.54 3.8e-06 ***
## bmi          1 5.26e+09 5.26e+09  143.07 < 2e-16 ***
## children     1 5.51e+08 5.51e+08   15.00 0.00011 ***
## smoker       1 1.23e+11 1.23e+11 3343.50 < 2e-16 ***
## region       3 2.33e+08 7.78e+07    2.12 0.09622 .
## Residuals 1329 4.88e+10 3.67e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(step.modelb)
```

```
## Analysis of Variance Table
##
## Response: charges
##             Df   Sum Sq  Mean Sq F value  Pr(>F)
## age          1 1.75e+10 1.75e+10  477.33 < 2e-16 ***
## bmi          1 5.45e+09 5.45e+09  148.30 < 2e-16 ***
## children     1 5.72e+08 5.72e+08   15.56 8.4e-05 ***
## smoker       1 1.23e+11 1.23e+11 3361.34 < 2e-16 ***
## region       3 2.33e+08 7.77e+07    2.12   0.096 .
## Residuals 1330 4.88e+10 3.67e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Distância de cook para visualizar se há pontos influentes na regressão do step.modelb
n=1338
cook = cooks.distance(step.modelb)
pontInf = which(cook>4/n) # Para ver os pontos com distância de cook superior a uma unidade
pontInf
```

```
##    7   15   51   53   67   98  105  127  131  148  173  176  189  216  240  249  254  256  266  270
##    7   15   51   53   67   98  105  127  131  148  173  176  189  216  240  249  254  256  266  270
##  313  316  343  344  358  364  367  373  374  385  407  410  417  435  459  483  497  498  513  517
##  313  316  343  344  358  364  367  373  374  385  407  410  417  435  459  483  497  498  513  517
##  536  561  562  615  626  669  681  683  695  768  773  793  794  803  811  830  838  849  880  914
##  536  561  562  615  626  669  681  683  695  768  773  793  794  803  811  830  838  849  880  914
##  947  962  963 1015 1017 1026 1032 1036 1049 1051 1056 1063 1071 1080 1087 1103 1121 1126 1160 1223
##  947  962  963 1015 1017 1026 1032 1036 1049 1051 1056 1063 1071 1080 1087 1103 1121 1126 1160 1223
## 1229 1247 1257 1261 1266 1275 1280 1317
## 1229 1247 1257 1261 1266 1275 1280 1317
```

```r
summary(cook)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000021 0.000112 0.000851 0.001024 0.020287
```

A Cook's distance is one method used to identify influential points in regression analysis. If the values of the distances are greater than 4/n, where n is the sample size, it indicates that these values exist and should be considered.

The Cook's distance was used to check for the existence of points that may influence the regression. To determine if the distance is large enough to consider influential values, it must have a value greater than 4/n, which in this case is 0.003. Although on average the values of the distances are less than 0.003, we can see, for example, that the maximum value is much higher than the reference value, proving that there is at least one value influencing the results.

```r
n= 1338
cook2 = cooks.distance(step.modelf)
pontInf = which(cook2>4/n) # Para ver os pontos com distância de cook superior a 1 unidade

summary(cook2)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000021 0.000114 0.000842 0.001014 0.019857
```

```r
#leverage : hj>2(p+1)/n para que na observação possam ser considerados outliers
n = 1338
p = 7

limhj = (2*(p+1))/n
limhj
```

```
## [1] 0.01196
```

As we obtained a value greater than 0, it means that outliers are being considered for the regressions. Given that we observe the existence of outliers, these may be influencing other results such as correlation.

```r
# Consideração de outliers no comportamento do step.modelb
hj = hatvalues(step.modelb)
out = which(hj>limhj)
out
```

```
##   67   72  142  156  176  216  285  387  507  615  669  670  781  816 1023
##   67   72  142  156  176  216  285  387  507  615  669  670  781  816 1023
```

```r
summary(hj)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00306 0.00452 0.00554 0.00598 0.00706 0.01719
```

From these results, we can see which outliers are influencing the regression in the step.modelb.

```r
# Consideração de outliers no comportamento do step.modelf
hj2=hatvalues(step.modelf)
out=which(hj2>limhj)
out
```

```
##   67   72  142  156  176  216  244  285  387  448  507  514  577  615  664  665  669  670  721  781
##   67   72  142  156  176  216  244  285  387  448  507  514  577  615  664  665  669  670  721  781
##  846  861 1023 1080 1330
##  846  861 1023 1080 1330
```

```r
summary(hj2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00382 0.00526 0.00629 0.00673 0.00785 0.01812
```

From these results, we can see which outliers are influencing the regression in the step.modelf.

Conclusions

Residuals Conditions Verification

```r
shapiro.test(residuals(step.modelb))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(step.modelb)
## W = 0.9, p-value <2e-16
```

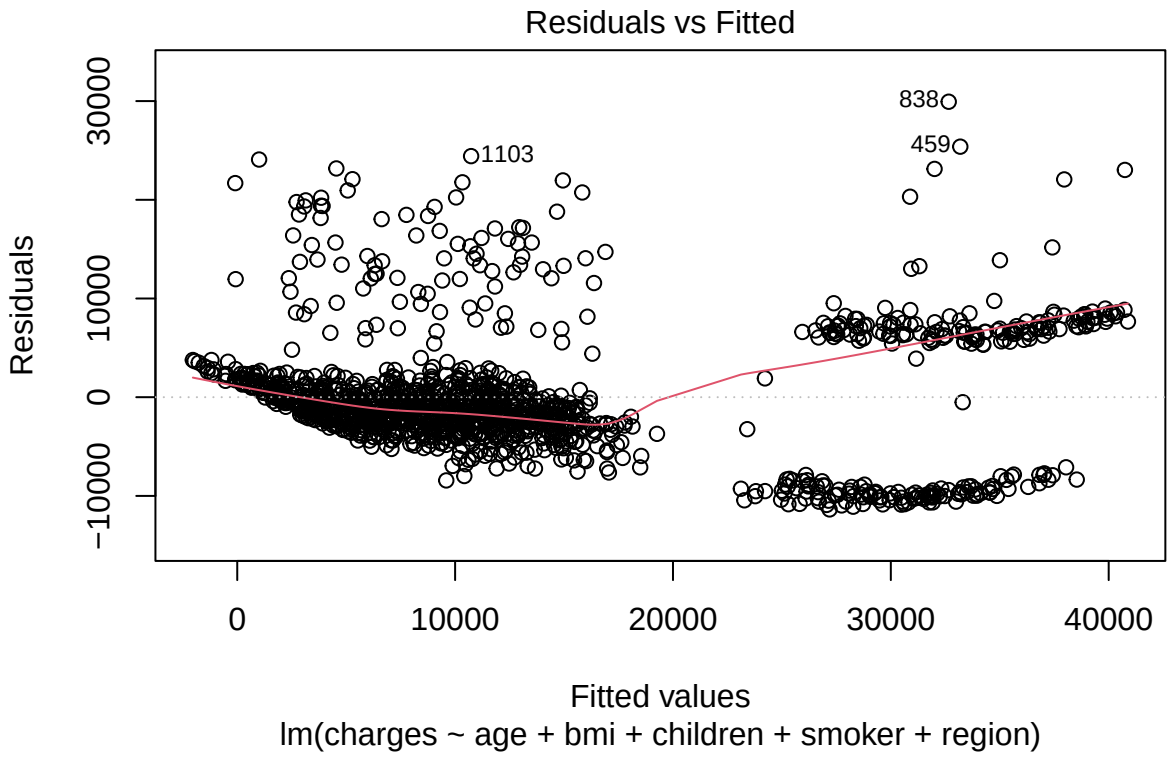
```r
#shapiro.test(residuals(cook))
t.test(residuals(step.modelb))
```

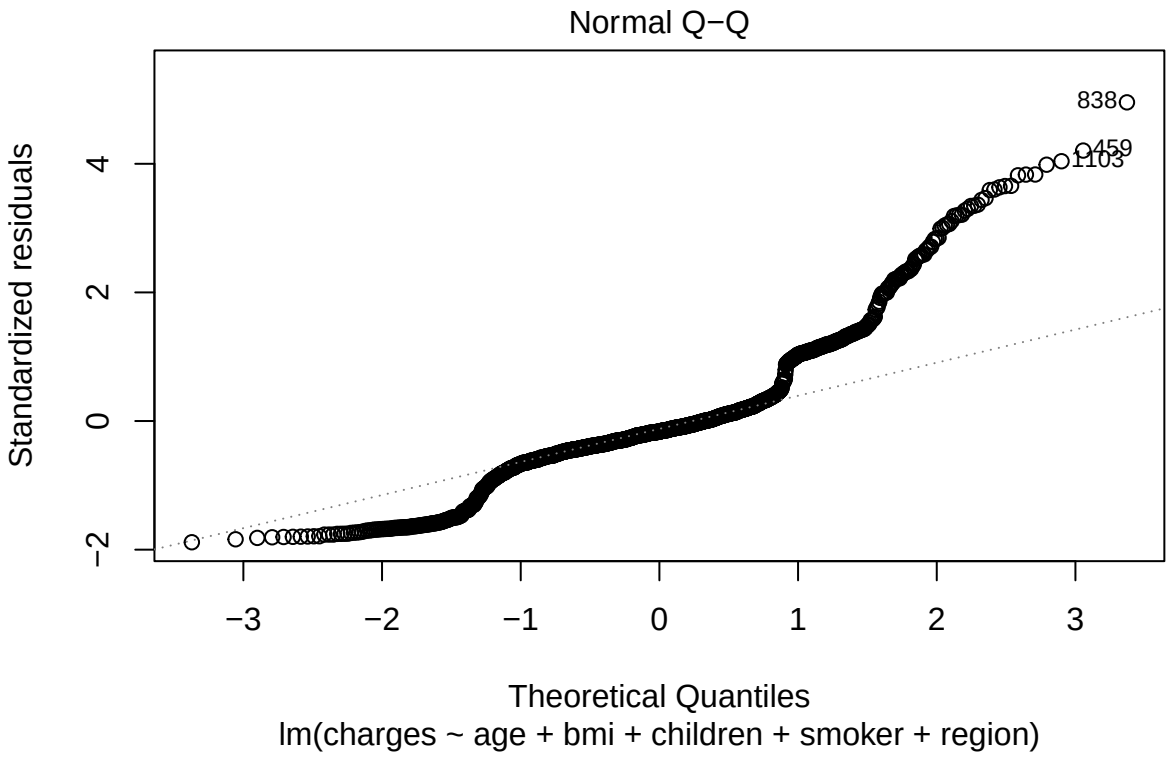
```
##
##  One Sample t-test
##
## data:  residuals(step.modelb)
## t = 1.1e-15, df = 1337, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -324.2  324.2
## sample estimates:
## mean of x
## 1.798e-13
```

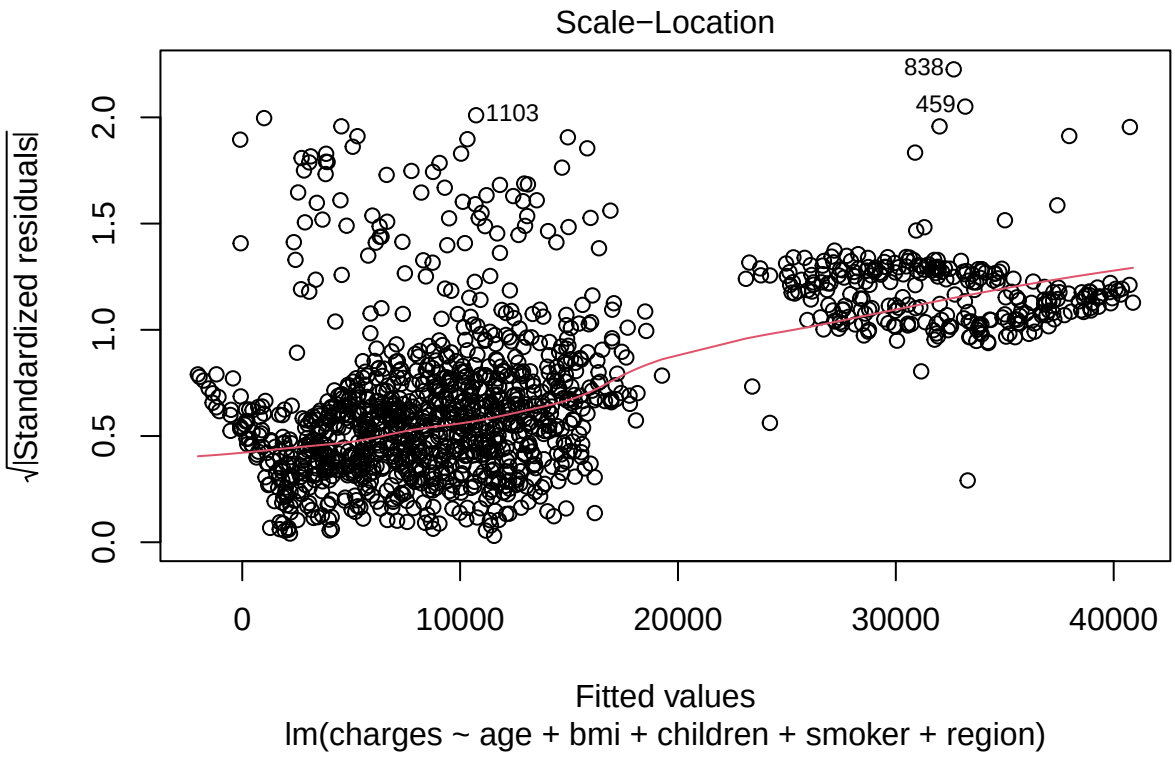
```r
sigma(step.modelb)/mean(insurance$charges)
```

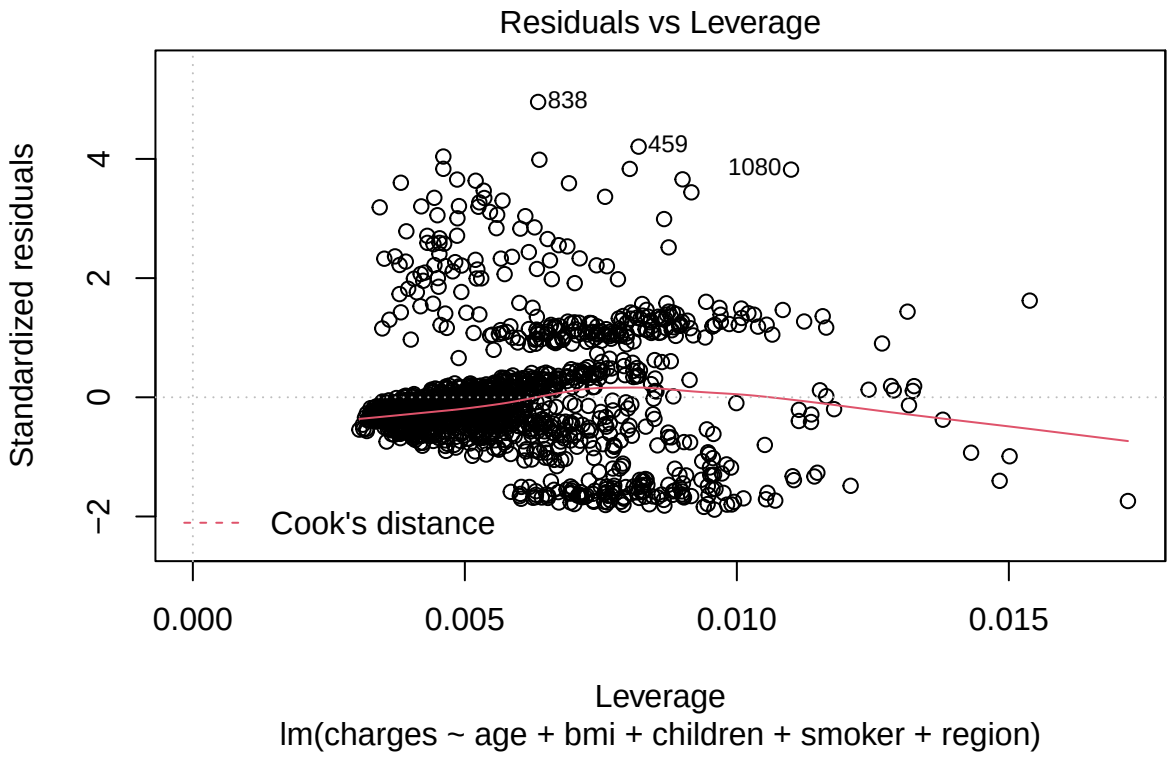
```
## [1] 0.4567
```

```r
plot(step.modelb)
```

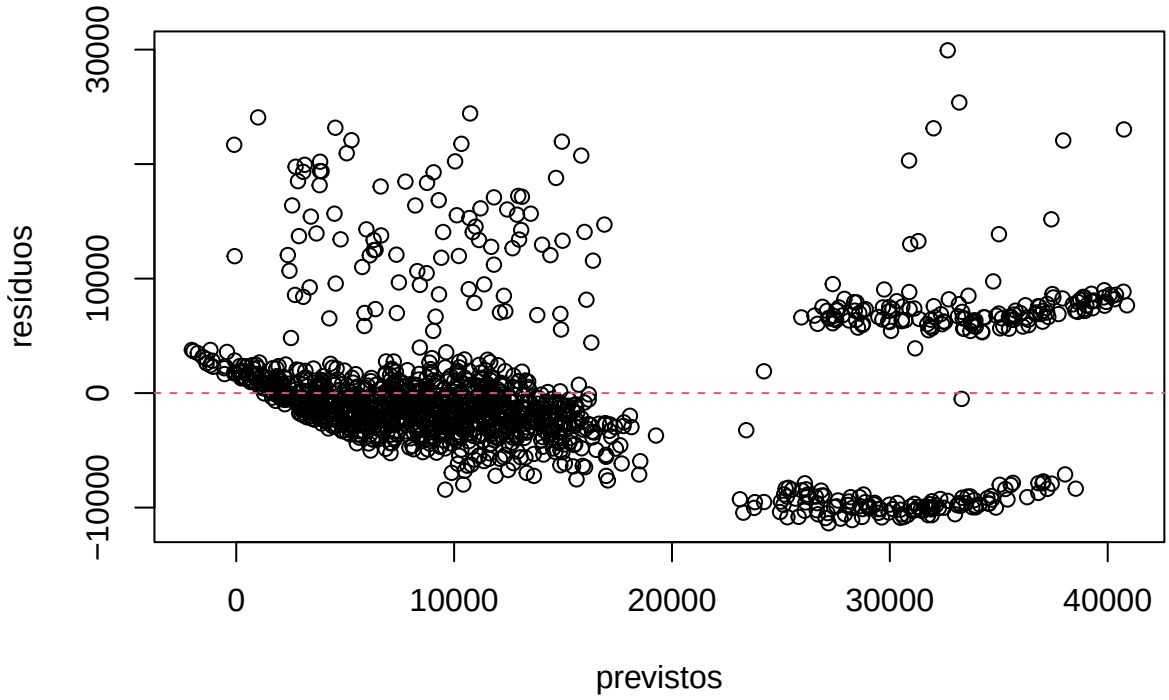



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ age + bmi + children + smoker + region)
838
459
1103
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
1103
838
459
0
10000
20000
30000
40000
Fitted values
lm(charges ~ age + bmi + children + smoker + region)

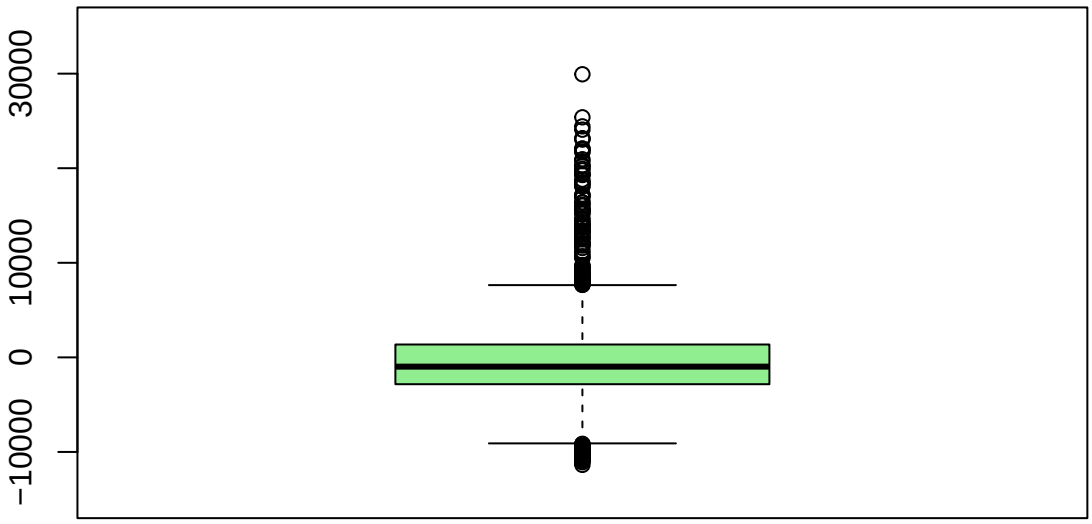



```
plot(step.modelb$fitted.values, step.modelb$residuals,ylab="resíduos",xlab="previstos")
abline(h=0,lty=2,col=2)
```



Based on the alpha value being 0.05, since the p-value < alpha, we reject the null hypothesis that declares the normality of residuals. With a p-value = 1, we can conclude that the mean is 0, which aligns with one of the assumptions of ANOVA.

```
boxplot(step.modelb$residuals,col= 'lightgreen', ylim = c(-15000,35000))
```




```
boxplot.stats(step.modelb$residuals)
```

```
## $stats
##     626     694     735     495     838
## -9093.0 -2837.8  -979.7  1362.0  7641.1
##
## $n
## [1] 1338
##
## $conf
## [1] -1161.1  -798.3
##
## $out
##      7      9     10     15     18     28     39     41     51     53     63     66     93     98
##  18511  -9479   8232 -10447   9568   8408   7944   9050 -10027  11958  -9524   8575  -9274  19940  20
##    126    127    129    130    131    136    138    148    165    173    179    189    216    240
##  -9215  19771   9519   8829  15416   7928  -9514  23177   8026  13709  -9758  20948   9755  21689   8
##    245    248    249    254    256    266    269    270    276    283    289    292    293    297
##  12059  10683  19322  24079  16390   9664   9238 -10391  -9467   8504   9450 -10272  13429  19397  18
##    316    334    336    343    344    352    354    357    358    360    364    367    373    374
## -10844  -9614  -9554 -10808  12490  -9376  -9773  11014 -10224  12510  20311  -9889  15670  19373  13
##    394    398    399    406    407    410    417    421    433    435    439    442    444    459
##  12029  14303  12091 -10483  13275 -10913 -10583  -9597   7763 -10024 -10576  10654  -9872  25383  -9
##    466    483    485    486    497    498    510    511    513    517    527    528    536    542
##  -9523  13944  -9587  -9945  -9666  18157  -9978 -10819  23129 -10966  22087 -10094  13007 -10123 -10
```

```
##    561    562    568    581    586    609    613    615    620    623    626    652    654    656
##  14061 -10138  -9744 -10132 -10289  10475  -9721   8649  13770 -10186 -11367 -10061  -9885 -10663 -10
##    672    673    681    683    684    695    729    730    731    736    752    754    762    764
##  14068 -10015 -10121 -11091  -9791  16038 -10589  -9947 -10434  -9695 -10662   7671   8355   7874 -10
##    769    773    774    778    779    793    794    797    803    811    830    832    838    844
##  -9874 -10860  -9977  -9449 -10157 -10886  13877   8269  21773 -10662  19282   8624  29936   7753  10
##    863    872    876    880    884    886    889    892    895    897    900    911    914    916
## -10123  -9991   9499  18472  -9249  -9845  -9636  -9171  -9347  -9610 -10193  13412  15583  11823  -9
##    936    938    946    947    954    962    963    972    989   1004   1005   1015   1017   1020
##  -9979  -9791  -9830  16846   9077  20239  15532  12651 -10034   8068  -9978  -9999  15290  -9729  18
##   1028   1030   1032   1034   1036   1037   1041   1049   1051   1056   1063   1071   1074   1076
##  12047  11217  18361  -9615  22062   7990  -9410  -9673  16134  15669 -10314  14244  12776   7665  -9
##   1080   1087   1103   1107   1121   1126   1132   1157   1160   1180   1208   1209   1222   1223
##  23026  14534  24425   8497  14076 -10580  -9290   8474  11974   8154   7964  13308   8590  21964  17
##   1227   1229   1243   1244   1247   1252   1255   1257   1261   1266   1268   1275   1280   1296
##   8356  17148   7789   8585  15178   8199   7968  13369  11566  20744   8074  12964  14720   8171   8
##   1301   1317   1318   1320   1326   1337
##   8375  17223   8409   8674   8157   8854
```

In conclusion, we can observe that:

Regarding exploratory analysis, the mean and median in the age variable present very similar values, and ages are distributed similarly without outliers detected. The youngest patient is 18 years old, and the oldest is 64.

There are slightly more men than women, with this difference being around 1%.

In BMI, there is an asymmetric distribution of results as patients fit into different categories. The mean and median have very close values, almost identical. Some outliers can be detected in all groups, with a greater presence in the obese group, as expected since this group has the most samples. The minimum BMI in the sample is 16, and the maximum is 53.1.

In the case of children, the mean and median have very close values, but the distribution of values is highly disproportional, with about 43% having no children. Among patients who have children, most (24%) have 1 child, with the maximum in this sample being 5 children (1%).

There is a significant difference between smoking and non-smoking patients regarding their distribution (20% and 80%, respectively).

Regarding the regions variable, the values are distributed similarly except for the South East group, which has a slightly higher value (27%).

For the charges variable, we can see that the mean and median are very different from each other, showing a high skewness in the sample. From the boxplot, several outliers can be detected. The minimum value is 1121.9, and the maximum is 63770.4.

Through the ANOVA test, it was possible to see if there were significant differences in prices within each of the variables. It was demonstrated that, for example, being a smoker or being older causes an increase in the insurance value since they are presumably people who need greater medical attention due to a weaker health condition. The results obtained in this test were in line with those of the t-test performed at the end.

We also observed that when correlating quantitative variables with the charges variable, there were not very significant results except for one, which, although weak, was relatively close to 0.5 (threshold used to consider a strong correlation), which was age. This was an expected result because, as mentioned before, older people need much more medical attention than younger patients.

Finally, by examining the graphs obtained from the linear regression analysis and the Shapiro test, we saw that we did not obtain normality in our results. This was confirmed by checking the Cook's distance and observing the information from the residuals boxplot, which showed the presence of several values that are

significantly influencing the results. These outliers may have a significant impact, affecting the correlations previously analyzed considerably.

In addition to the outliers that are causing changes in the final results, some variables such as smoker and children present highly skewed distributions of values, which also influence the impact of these during the analysis.

We were able to verify that smoking is a significant indicator for the increase in insurance prices, caused by the need for greater medical attention, and we also observed that insurance prices do not increase gradually with the number of children. From the boxplot made of prices based on the number of children, we see that having 3 children covered by insurance seemed to be cheaper than having only 2 children, and having 5 children seems to lead to a smaller price increase than having 4. This can be explained by the unequal number of values being analyzed for each group, as we have, for example, 18 patients with 5 children and 324 with only 1.