SIM - Assignment 1

Medical cost

Míriam Méndez, Gabriel Zarate

November 20, 2022

Data Preparation

We load the dataset, we view the data and we take a look to the statistical summary to check that there no were structural errors and all the data had been read correctly.

```
df <- read.csv("insurance.csv")
#View(df)
summary(df)
```

```
##       age            sex                 bmi           children    
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000  
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000  
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095  
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000  
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000  
##     smoker             region             charges     
##  Length:1338        Length:1338        Min.   : 1122  
##  Class :character   Class :character   1st Qu.: 4740  
##  Mode  :character   Mode  :character   Median : 9382  
##                                        Mean   :13270  
##                                        3rd Qu.:16640  
##                                        Max.   :63770  
```

Removing duplicates

```
df[duplicated(df), ]
```

```
##     age  sex   bmi children smoker    region  charges
## 582  19 male 30.59        0     no northwest 1639.563
```

```
df <- df[!duplicated(df),]
```

Checking Data Types

We check the data types of all the features and we casted sex, smoker and region to factors.

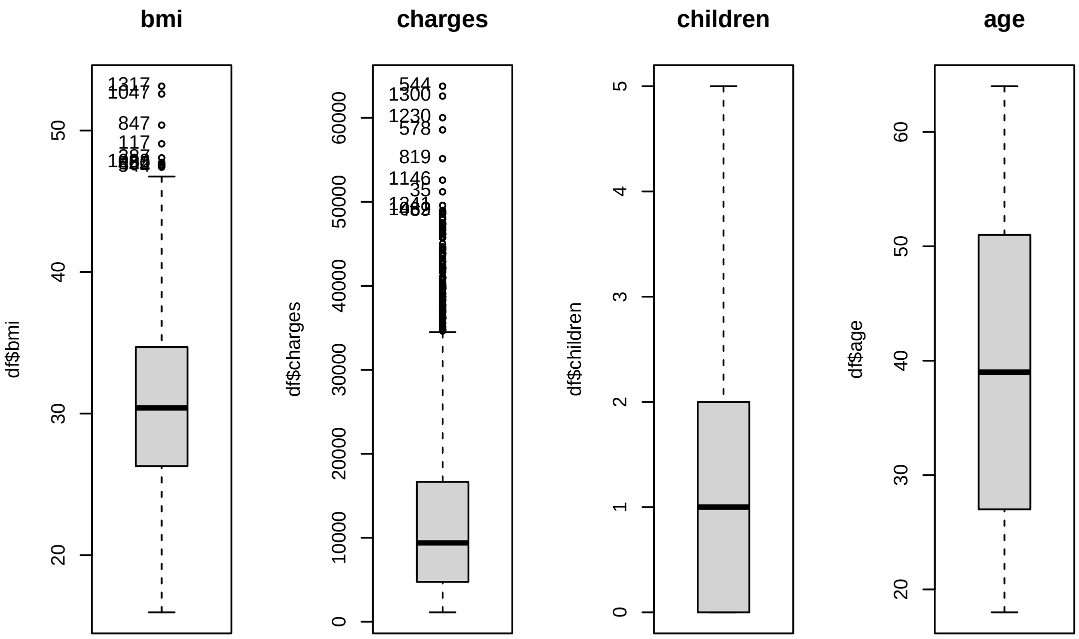
```
categ_cols <- c( 'sex', 'smoker', 'region')
df[categ_cols] = lapply(df[categ_cols], FUN = as.factor)
sapply(df, class)
```

```
##       age       sex       bmi  children    smoker    region   charges
## "integer"  "factor" "numeric" "integer"  "factor"  "factor" "numeric"
```

Univariate Outliers

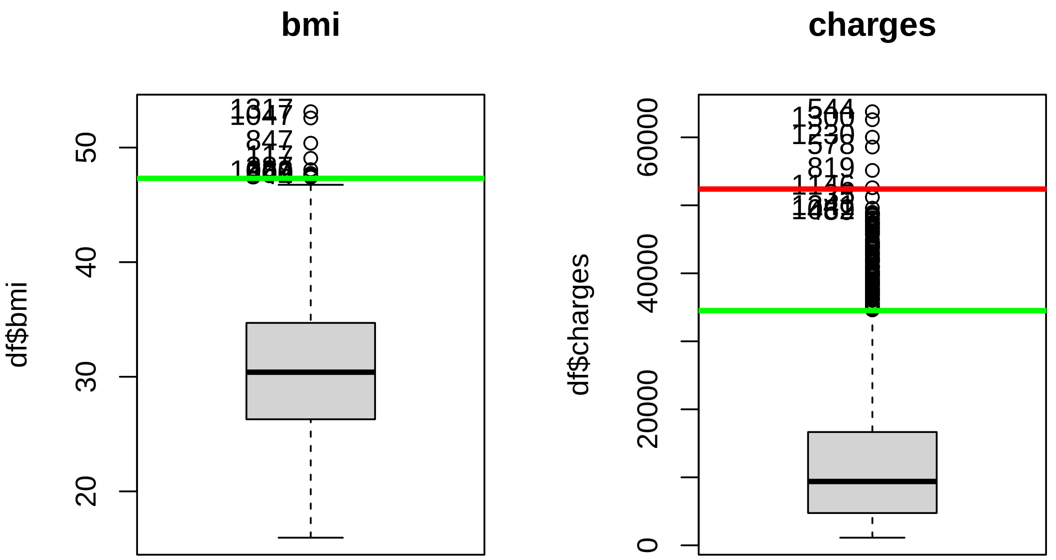
To detect outliers it was decided to check both mild and severe outliers, as it is shown in the boxlpot, the green line show the mild and the red show the severe outliers. Finally it was decided to cast the severe as NA, to be treaten after.

```
par(mfrow=c(1,4))
Boxplot(df$bmi, main="bmi") #bmi  seem to have outliers
Boxplot(df$charges, main="charges") #charges seem to have outliers
Boxplot(df$children, main="children") #Children no outliers
Boxplot(df$age, main ="age") #Age no outliers
```



```
par(mfrow=c(1,2))
#Checking outliers for bmi
ss<-summary(df$bmi);
```

```r
# Upper/lower severe threshold
utso2<-ss[5]+3*(ss[5]-ss[2]);
utsi2<-ss[2]-3*(ss[5]-ss[2]);
# Upper/lower mild threshold
utmo2<-ss[5]+1.5*(ss[5]-ss[2]);
utmi2<-ss[2]-1.5*(ss[5]-ss[2]);
Boxplot(df$bmi, main="bmi")
abline(h=utso2,col="red",lwd=3)
abline(h=utsi2,col="red",lwd=3)
abline(h=utmo2,col="green",lwd=3)
abline(h=utmi2,col="green",lwd=3)
lls.bmi<-which((df$bmi>utso2)|(df$bmi<utsi2));
llm.bmi<-which((df$bmi>utmo2)|(df$bmi<utmi2));


#Checking outliers for charges
ss<-summary(df$charges);
# Upper/lower severe threshold
utso2<-ss[5]+3*(ss[5]-ss[2])
utsi2<-ss[2]-3*(ss[5]-ss[2])
# Upper/lower mild threshold
utmo2<-ss[5]+1.5*(ss[5]-ss[2])
utmi2<-ss[2]-1.5*(ss[5]-ss[2]);
Boxplot(df$charges, main = "charges" )
abline(h=utso2,col="red",lwd=3)
abline(h=utsi2,col="red",lwd=3)
abline(h=utmo2,col="green",lwd=3)
abline(h=utmi2,col="green",lwd=3)
```



```
lls<-which((df$charges>utso2)|(df$charges<utsi2))
llm<-which((df$charges>utmo2)|(df$charges<utmi2))

#Setting severe outliers from charges as NA
df[lls,"charges"]<-NA
```

Treating missing data

Checking the missing data, it was seen that the only column with missing data was charges (the oultliers casted before). Those values can't be imputed because it is the target variable, so they were deleted.

```
mis_col = colSums(is.na(df)); mis_col
```

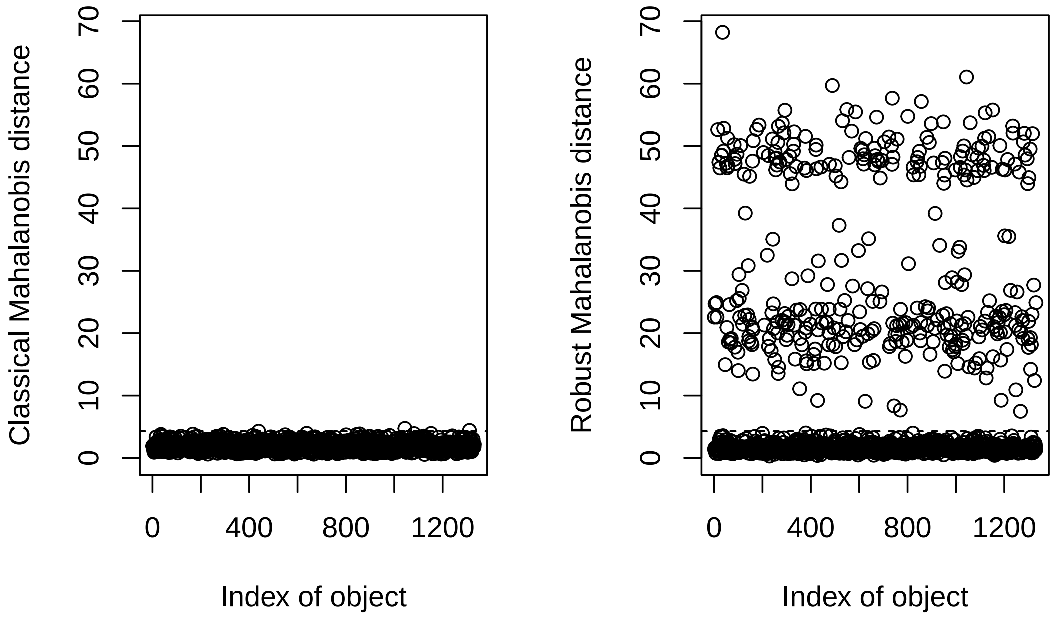
```
##      age      sex      bmi children   smoker   region  charges
##        0        0        0        0        0        0        6
```

```
md<-which(is.na(df$charges))
df <- df[-md,]
```

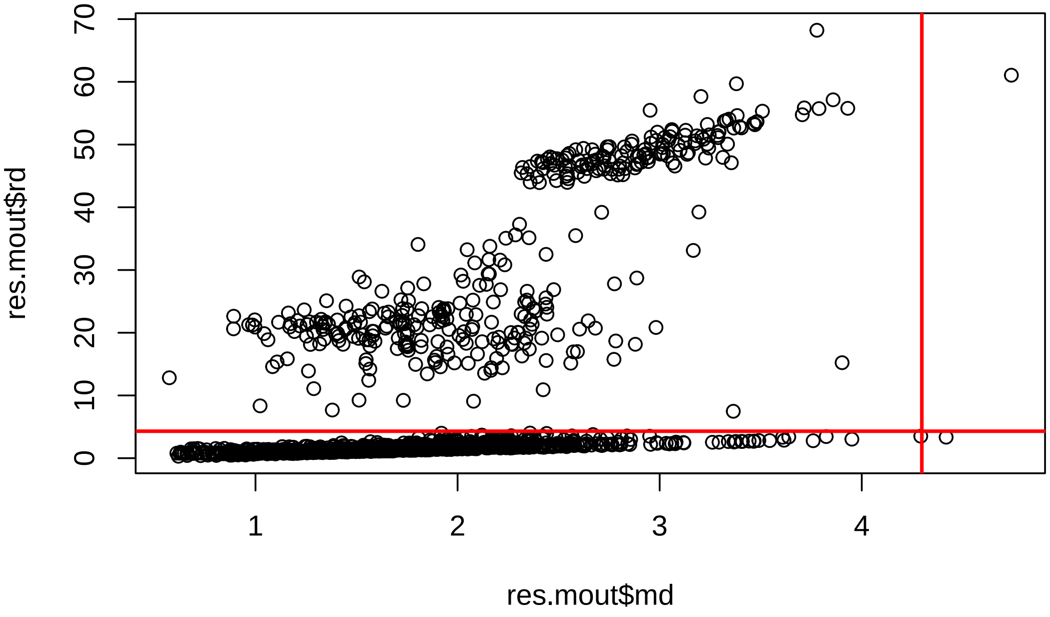
Multivariate Outliers

It was decided to use Mahalanobis distance to detect the multivariate outliers, getting only one, and it was deleted.

```
res.mout <- Moutlier( df[ , c(1,3,4,7)], quantile = 0.999 )
```



```
par(mfrow=c(1,1))
plot( res.mout$md, res.mout$rd )
abline( h=res.mout$cutoff, lwd=2, col="red")
abline( v=res.mout$cutoff, lwd=2, col="red")
```



```
llmout <- which( ( res.mout$md > res.mout$cutoff ) & (res.mout$rd > res.mout$cutoff) );
res.mout$md[llmout]
```

```
##     1048
## 4.740281
```

```
#Since there is only one multivariate outlier, we delete it
df <- df[-llmout,]
```

Data Validation

After checking all the columns of the data that there were no major mistakes in the dataset to be corrected, despite of finding some atypical cases that were decided to keep because they were probable in extreme cases.

Second part

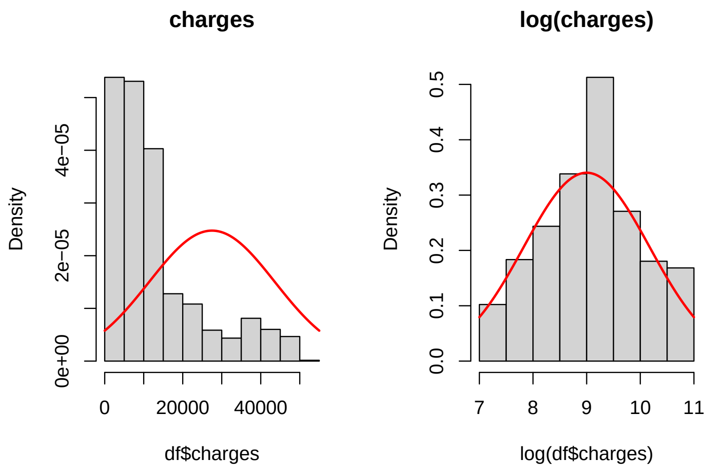
The qualitative variables were casted as factors in pre-processing, therefore we proceed to check the normality of the response variable.

Determine if the response variable (charges) has an acceptably normal distribution

The distribution of charges is right-skewed. We can confirm this visually using a histogram and comparing it with a curve that represents a normal distribution, also the Shapiro Test was applied to check the normality of the distribution, getting a p-value lower that any acceptable alpha, rejecting the H0, so it does not has a normal distribution.

Additionally, the normality of the logarithmic transformation of charge was tested, getting the same results by checking the histogram, comparing it with the curve, and by applying the Shapiro Test, rejecting the H0, so it does not has a log-normal distribution.

```
par(mfrow=c(1,2))
#Normal Check
hist(df$charges, freq=F, main="charges")
curve(dnorm(x,mean(x),sd(x)),lwd=2,add=T,col="red")
#Log normal check
hist(log(df$charges), freq=F, main="log(charges)")
curve(dnorm(x,mean(x),sd(x)),lwd=2,add=T,col="red")
```



```
shapiro.test(df$charges)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df$charges
## W = 0.81764, p-value < 2.2e-16
```

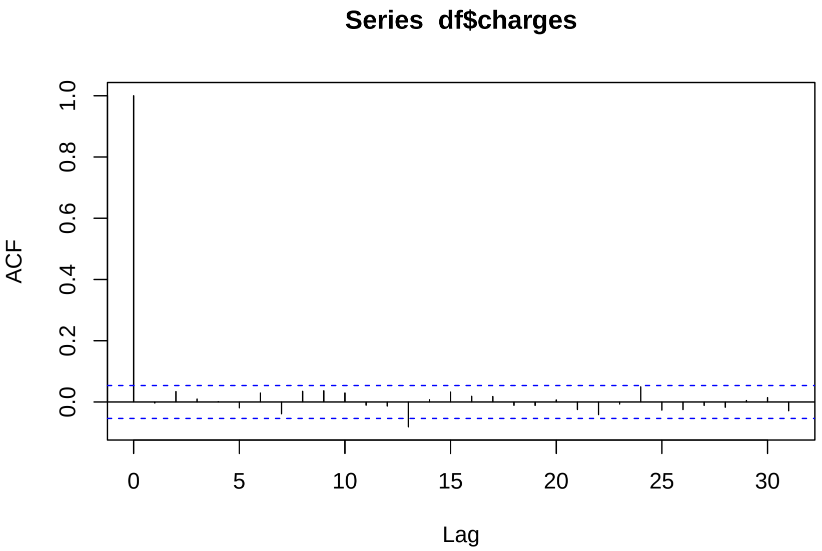
```
shapiro.test(log(df$charges))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(df$charges)
## W = 0.98185, p-value = 7.028e-12
```

Address test to discard serial correlation

By using the acf() method to plot the autocorrelation in charges, it can be seen that there is no serial correlation in it. Also the Durbin-Watson Test was realized, and it gets a p-value = 0.53, so the H0 fails to be rejected, so it can be said that there is no autocorrelation present in charges.

```
# autocorrelation
acf(df$charges)
```



```
# Durbin-Watson Test for serial correlation
dwtest(df$charge ~ 1)
```

```
##
##  Durbin-Watson test
##
## data:  df$charge ~ 1
## DW = 2.005, p-value = 0.5364
## alternative hypothesis: true autocorrelation is greater than 0
```

Preliminary exploratory analysis

First of all, we studied the correlations and non of them were considered strong, nevertheless all had a positive correlation, meaning that as *feature 1* increase, *feature 2* also increase.

```
cor(df[c(7,1,3,4)])
```

```
##            charges        age        bmi   children
## charges  1.0000000 0.30408298 0.18275794 0.07624490
## age      0.3040830 1.00000000 0.11170476 0.04224779
## bmi      0.1827579 0.11170476 1.00000000 0.01509009
## children 0.0762449 0.04224779 0.01509009 1.00000000
```

Also checking the condes() function, we can see the following:

- With the numeric variables: There is small positive correlation with age, an smaller with bmi, and there is almost no correlation with children
- With the categorical: there is a moderately high coefficient of determination with smoker, so we can say that it is an influential variable

```
res.con <- condes( df, num.var=7, proba = 0.01 )
res.con$quanti
```

```
##          correlation      p.value
## age        0.3040830 7.497946e-30
## bmi        0.1827579 1.875552e-11
## children   0.0762449 5.402109e-03
```

```
res.con$quali
```

```
##                R2       p.value
## smoker 0.6169441 5.382469e-279
```

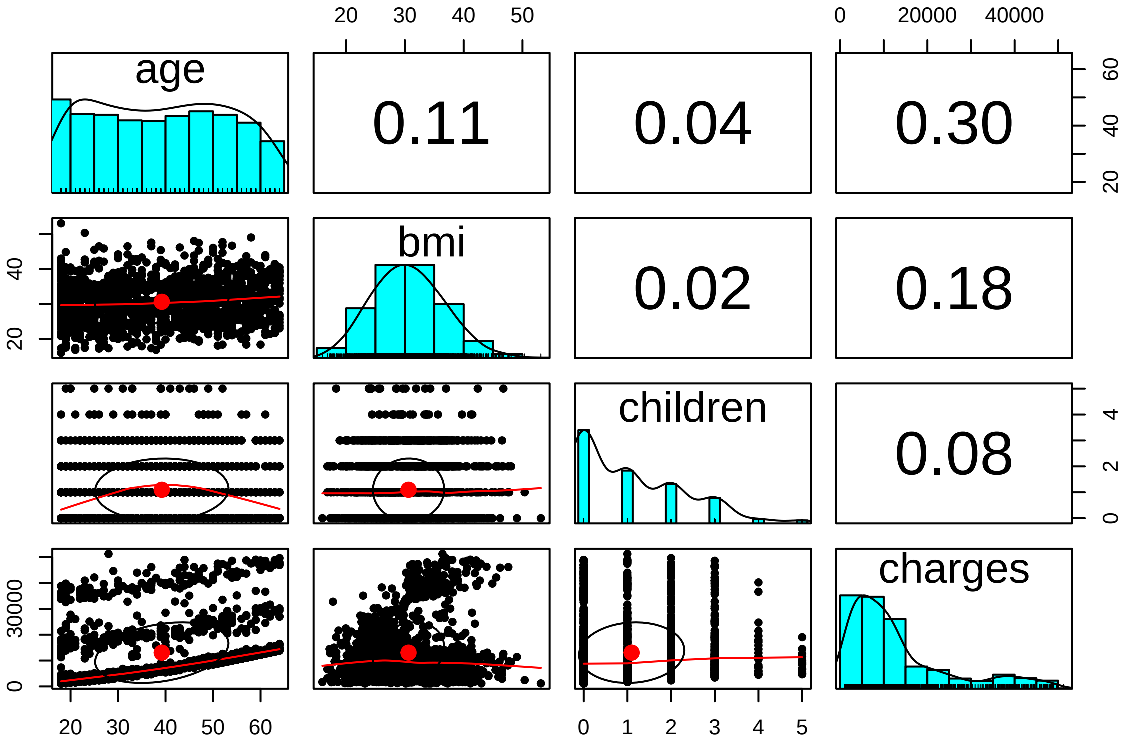
```
res.con$category
```

```
##              Estimate       p.value
## smoker=yes  11481.13 5.382469e-279
## smoker=no  -11481.13 5.382469e-279
```

Checking the relationships and the distributions of the numerical variables, we found some interesting relationships with the target:

- age it can be seen that there seems to be three patterns, but all of them follow a exponential increase, with only the difference that each pattern has a different domain in charges.
- bmi it can bee seen that there is no clear pattern, but it can be seen that there are higher charges (>30000) more frequently in cases where the bmi is higher than 30
- children there is no clear pattern, only that there are higher charges for people with less than 3 children than the ones with 4 or 5

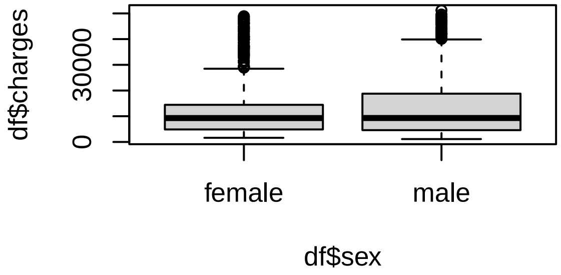
```
pairs.panels(df[c(1,3,4,7)])
```

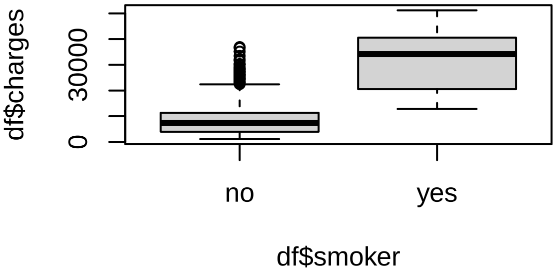


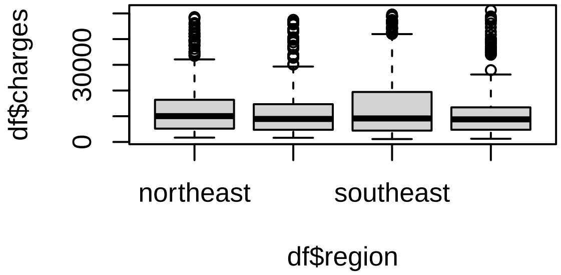
Checking the relationship between charges and the other categorical variables we found by using graphics:

- sex seems to not be any significative differences between sex, only that the charges for the 3rd Qu. of male is higher than the female one.
- smoker it can be seen that there is a significant difference in charges depending on this variable, if the person smokes it tends to get higher charges. So this variable is going to be impactful to charges.
- region seems to not be any significative differences between them

```
par( mfrow = c(2,2))
plot(df$charges ~ df$sex)
tapply(df$charges, df$sex, summary)
plot(df$charges ~ df$smoker)
tapply(df$charges, df$smoker, summary)
plot(df$charges ~ df$region)
tapply(df$charges, df$region, summary)
```



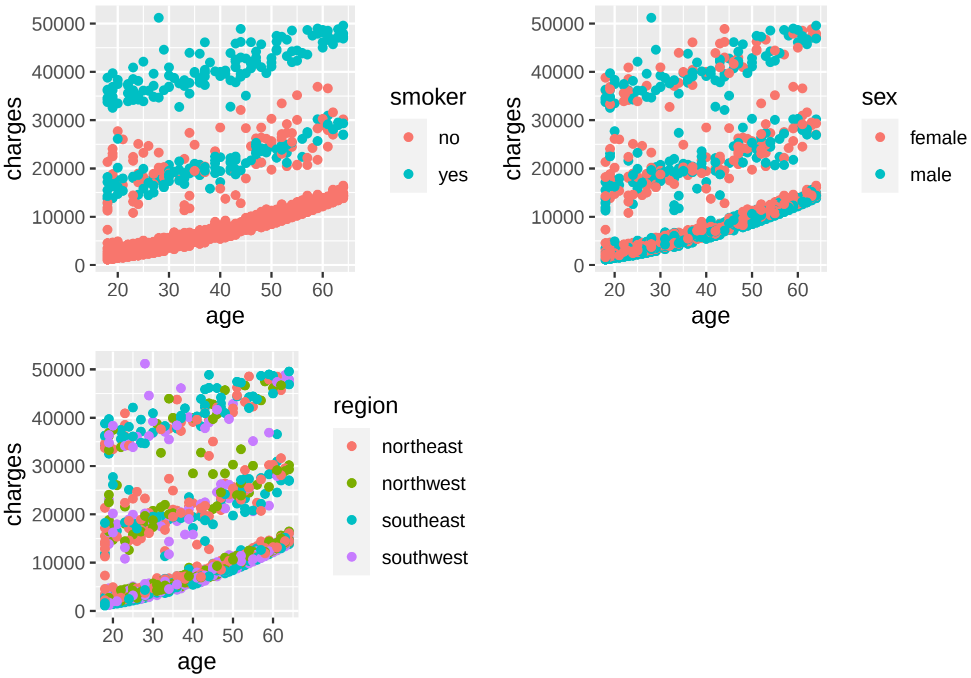




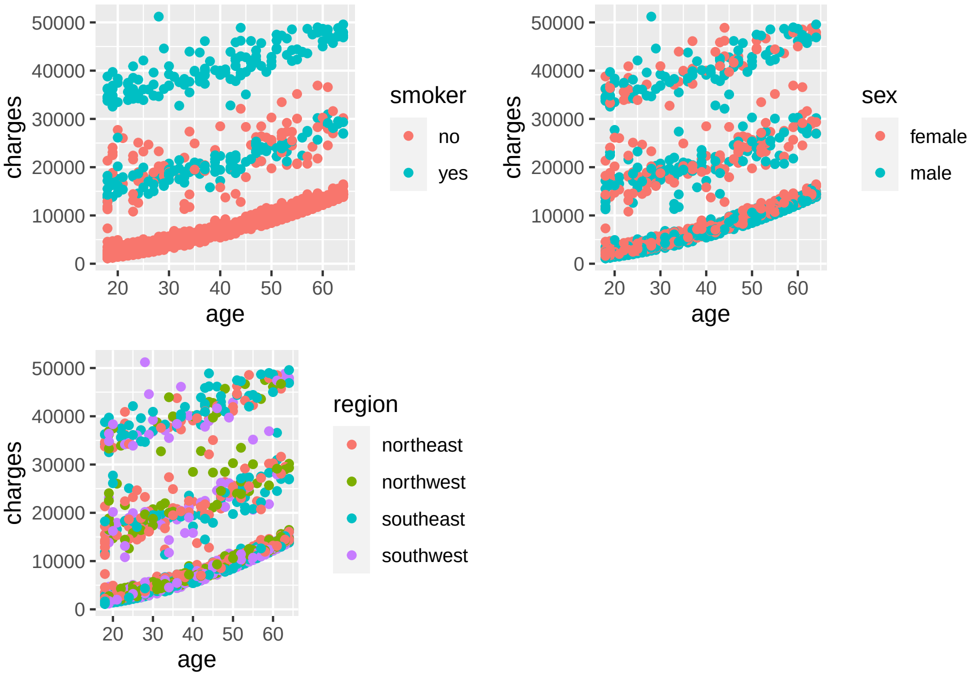
Checking the relation of charges with the interaction of other variables:

- With this plots it can be confirmed that smoker affects every variable, because its combination with any numerical variable affects the charges, because all the higher charges are associated with the people being smokers
- Sex and region does not have impact in the charges despite combining it with the other numerical variables
- In the graphic of charges ~ age coloring by smoker it can bee seen that the three patterns previously detected are affected, one pattern (less charges) is only for none smokers, the mid one is a mix of smokers and none smokers, and the higher one is only smokers

```
plot1 <- ggplot(df, aes(x=age , y = charges, color = smoker)) + geom_point()
plot2 <- ggplot(df, aes(x=age , y = charges, color = sex)) + geom_point()
plot3 <- ggplot(df, aes(x=age , y = charges, color = region)) + geom_point()
grid.arrange( plot1, plot2,plot3 , ncol=2)
```

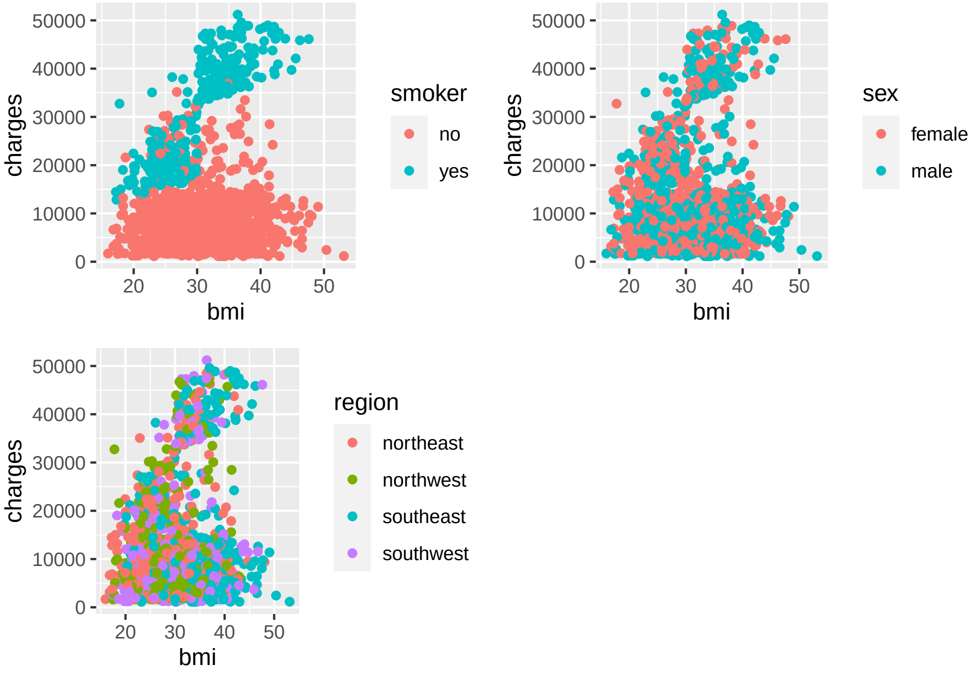


```
plot1 <-ggplot(df, aes(x=age , y = charges, color = smoker)) + geom_point()
plot2 <-ggplot(df, aes(x=age , y = charges, color = sex)) + geom_point()
plot3 <-ggplot(df, aes(x=age , y = charges, color = region)) + geom_point()
grid.arrange( plot1, plot2,plot3 , ncol=2)
```

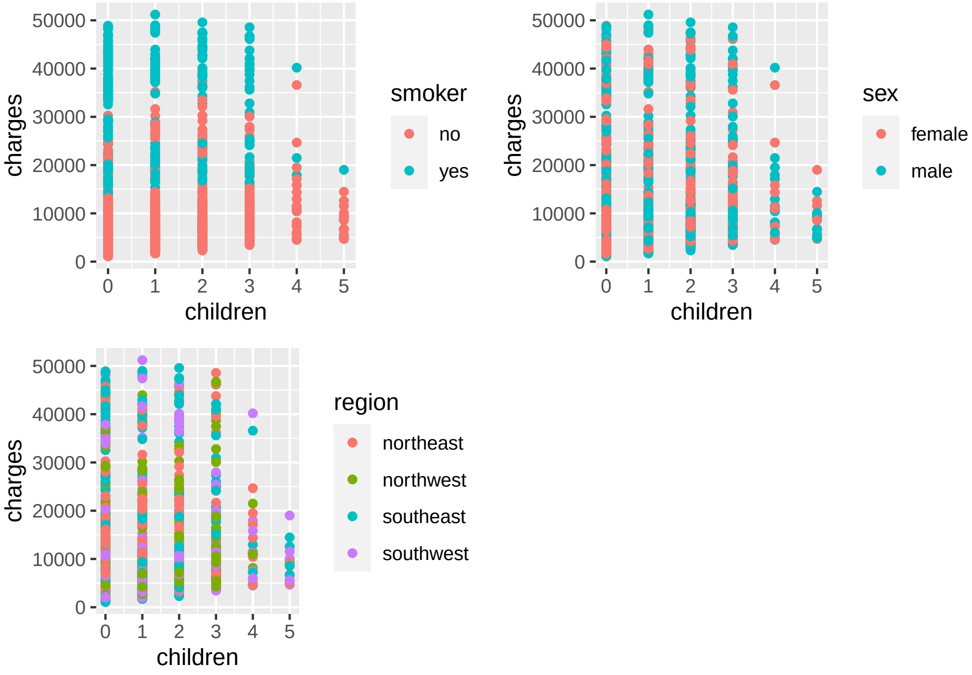


```
plot1 <-ggplot(df, aes(x=bmi , y = charges, color = smoker)) + geom_point()
plot2 <-ggplot(df, aes(x=bmi , y = charges, color = sex)) + geom_point()
```

```
plot3 <-ggplot(df, aes(x=bmi , y = charges, color = region)) + geom_point()
grid.arrange( plot1, plot2,plot3 , ncol=2)
```



```
plot1 <-ggplot(df, aes(x=children , y = charges, color = smoker)) + geom_point()
plot2 <-ggplot(df, aes(x=children , y = charges, color = sex)) + geom_point()
plot3 <-ggplot(df, aes(x=children , y = charges, color = region)) + geom_point()
grid.arrange( plot1, plot2,plot3 , ncol=2)
```



Conclusions

- The variable age has three cuadratic patterns in terms of charges
- The variable smoke is of high impact to charges, if the people smokes it tends to have higher charges

Model

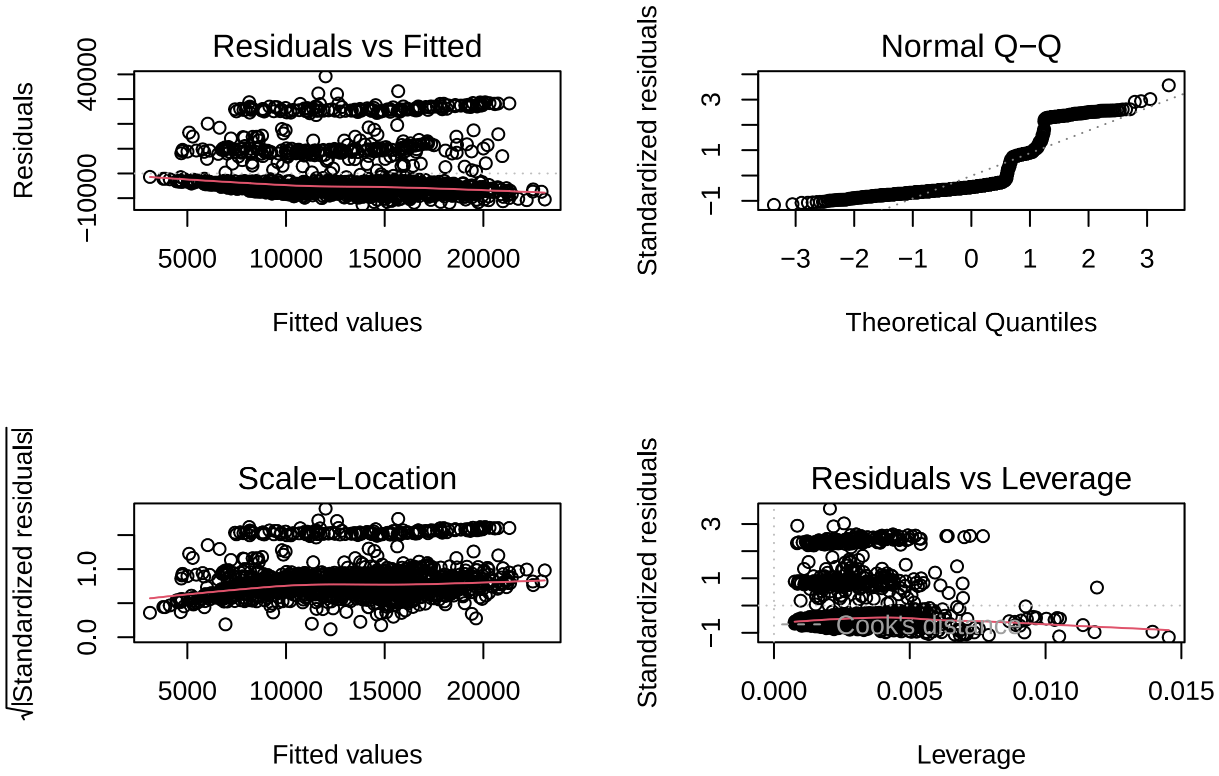
Model with only numerical variables

To start a model with only numerical variables was created. As it was seen in the EDA, age has an strange pattern in terms of charge because it has 3 parabolic lines that has to be converted, so that variable will be the main focus. So first with the result of the boxcox graphic it was decided to transform charges to log(charges) because lambda was near 0. Then boxTidwell was applied, getting that lambda for age was 0.5, so an squared root transformation was suggested. To confirm that three models were developed, one was with a polynomic approach for age, the second similar to the fist one but only keeping the quadratic form of age and a third one with the squared root of age. Finally comparing them by BIC it was decided to keep m3, that has R2 = 0.30847.

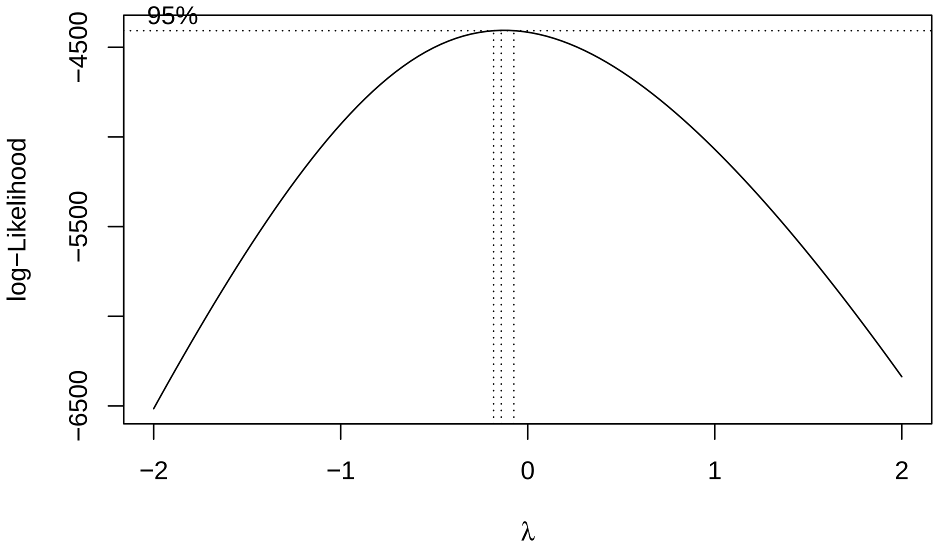
```
m0 <- lm(charges ~ age + bmi + children , data = df)
summary(m0)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12716  -6748  -5085   6506  39188
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5793.54    1710.44  -3.387 0.000727 ***
## age           237.42      21.65  10.965  < 2e-16 ***
## bmi           289.84      50.12   5.782 9.17e-09 ***
## children      601.73     250.63   2.401 0.016495 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11010 on 1326 degrees of freedom
## Multiple R-squared:  0.1187, Adjusted R-squared:  0.1167
## F-statistic: 59.54 on 3 and 1326 DF,  p-value: < 2.2e-16
```

```
par( mfrow = c(2,2))
plot(m0, id.n = 0)
```



```
#The intercept is difficult to interpret because it is impossible to have 0's for all values
par( mfrow = c(1,1))
boxcox( charges ~ age + bmi + children , data=df) #lambda = 0 so we transform charges with log
```



```r
boxTidwell( log(charges) ~ age + bmi + I(children+0.5) , data=df) #lambda age = 0.5
```

```
##                    MLE of lambda Score Statistic (z) Pr(>|z|)
## age                      0.51596             -1.4720  0.14103
## bmi                     -1.22298             -1.5094  0.13119
## I(children + 0.5)        0.25754             -1.9146  0.05555 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  14
```

```r
# First we will try a polinomic convertion with age
m1 <- lm(log(charges) ~ age + I(age^2) + bmi + I(children+0.5) , data = df)
#summary(m1) # R2 = 0.3081

m2 <- lm(log(charges) ~  I(age^2) + bmi + I(children+0.5) , data = df)
#summary(m2) #R2 = 0.2946

m3 <- lm(log(charges) ~  sqrt(age) + bmi + I(children+0.5) , data = df)
summary(m3) #R2 = 0.3085
```

```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5),
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3506 -0.4527 -0.3095  0.5224  2.2594
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.113260   0.147927  41.326  < 2e-16 ***
## sqrt(age)         0.411057   0.018174  22.618  < 2e-16 ***
## bmi               0.009474   0.003453   2.744  0.00616 **
## I(children + 0.5) 0.098376   0.017294   5.689 1.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7587 on 1326 degrees of freedom
## Multiple R-squared:  0.3085, Adjusted R-squared:  0.3069
## F-statistic: 197.2 on 3 and 1326 DF,  p-value: < 2.2e-16
```

```r
anova(m2,m1) # models are not equivalent
```

```
## Analysis of Variance Table
##
## Model 1: log(charges) ~ I(age^2) + bmi + I(children + 0.5)
## Model 2: log(charges) ~ age + I(age^2) + bmi + I(children + 0.5)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1326 778.53
## 2   1325 763.61  1    14.923 25.894 4.126e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(m3,m2,m1, k=log(nrow(df))) # m3 has the best BIC (3071.687)
```

```
##    df      AIC
## m3  5 3071.687
## m2  5 3098.094
## m1  6 3079.546
```

Influential data and residual outliers

Apriori influent data

The hat values were used to check the model leverage, and the threshold used was 2*p/n because this is an small dataset, so a new model was generated without those values just to check the impact, and it was decided to don't do nothing with them.

```
llev <- which( hatvalues(m1) > 2*(length(coef(m1))/nrow(df)))
length(llev)
```

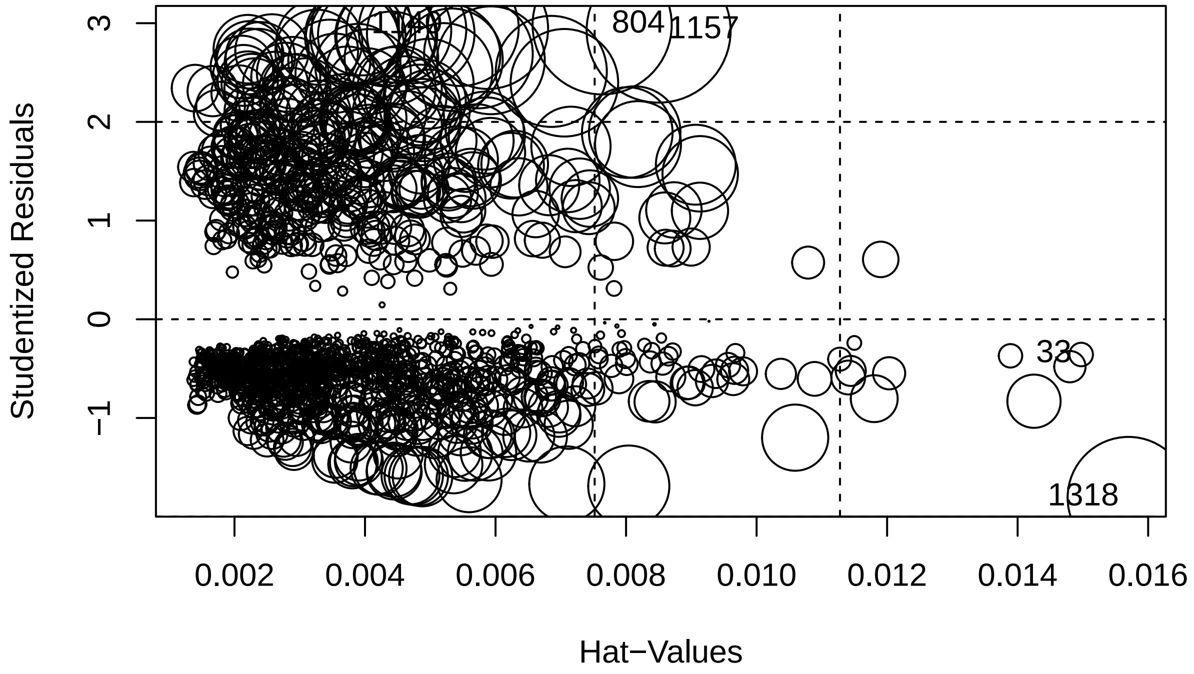
```
## [1] 70
```

```
# We try a model without those values only to check, but it won't be kept
m3 <- lm(log(charges) ~  sqrt(age) + bmi + I(children+0.5) , data = df[-llev,])
# summary(m3) # R2 = 0.3051
```

Aposteriori influent data

To check the actual influent data it was decided to use the Chatterjee-Hadi's threshold, to trim the outliers in the cook distances of the model. And it was decided that from now on all the models would not use those values.

```
par(mfrow=c(1,1))
influencePlot(m1)
```

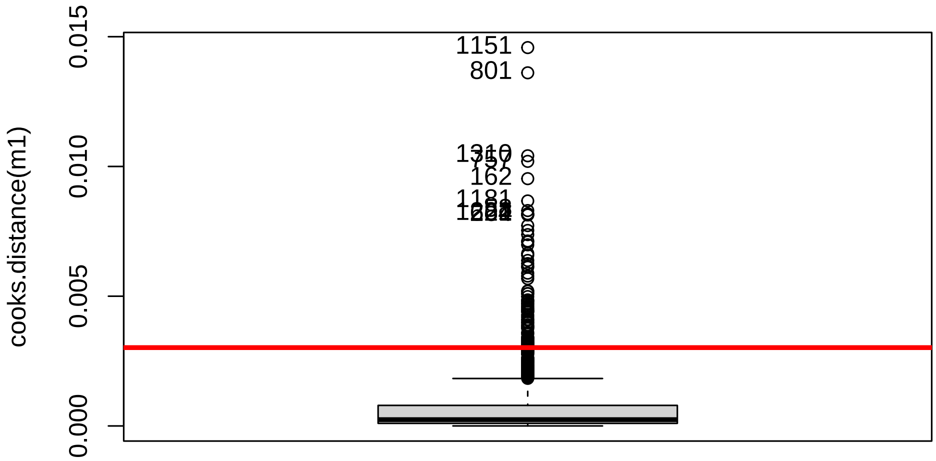


```
##          StudRes         Hat        CookD
## 33   -0.3571086 0.014973667 0.0003879691
## 804   2.9830916 0.007633409 0.0136090642
## 1140  2.9760299 0.003947828 0.0069793206
## 1157  2.9246446 0.008498814 0.0145805031
## 1318 -1.8082205 0.015698144 0.0104114093
```

```
# Threshold Chatterjee-Hadi
thChH <- 4/ (nrow(df) - length(coef(m1)));thChH
```

```
## [1] 0.003018868
```

```
# Actual influent data Cook´s distance: outliers in cook´s distance
Boxplot(cooks.distance(m1))
abline(h=thChH,col="red",lwd=3)
```

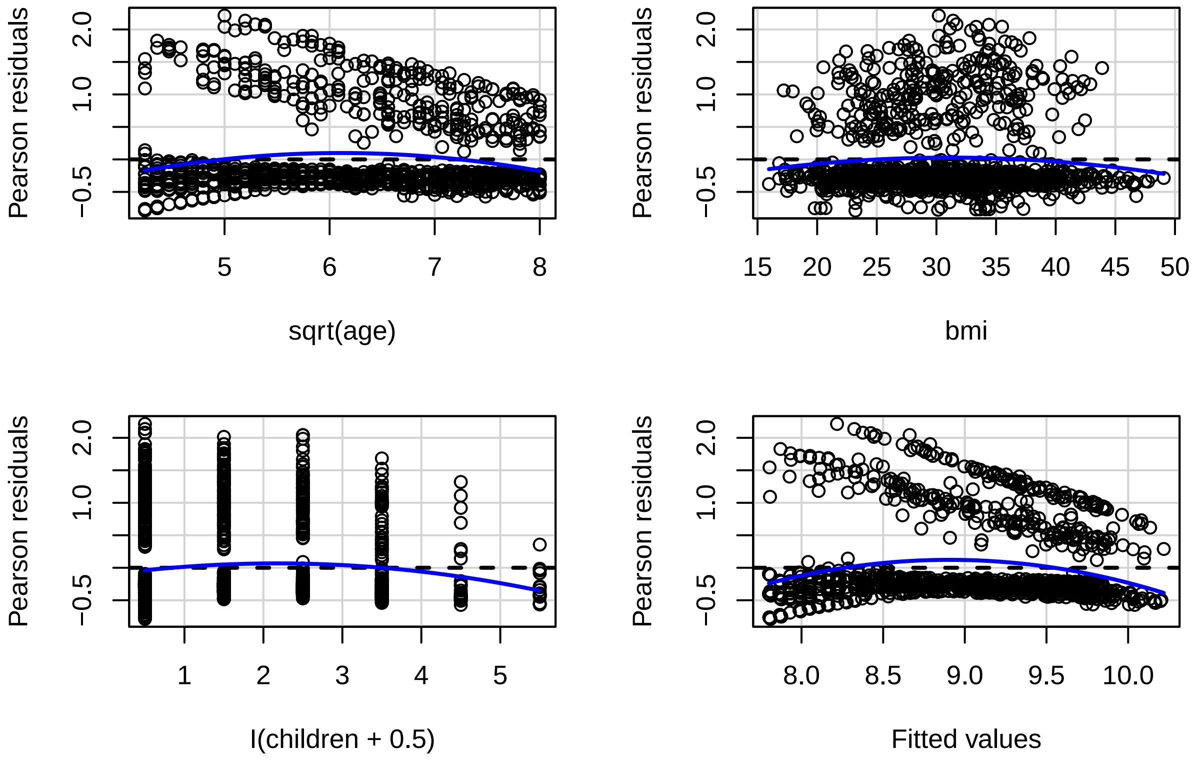


```
resout <- which( cooks.distance(m1) > thChH)
# length(resout) # 78
```

```
# We try a model with no Cook's distance outliers
m4 <- lm(log(charges) ~  sqrt(age) + bmi + I(children+0.5) , data = df[-resout,])
summary(m4) # R2 = 0.513
```

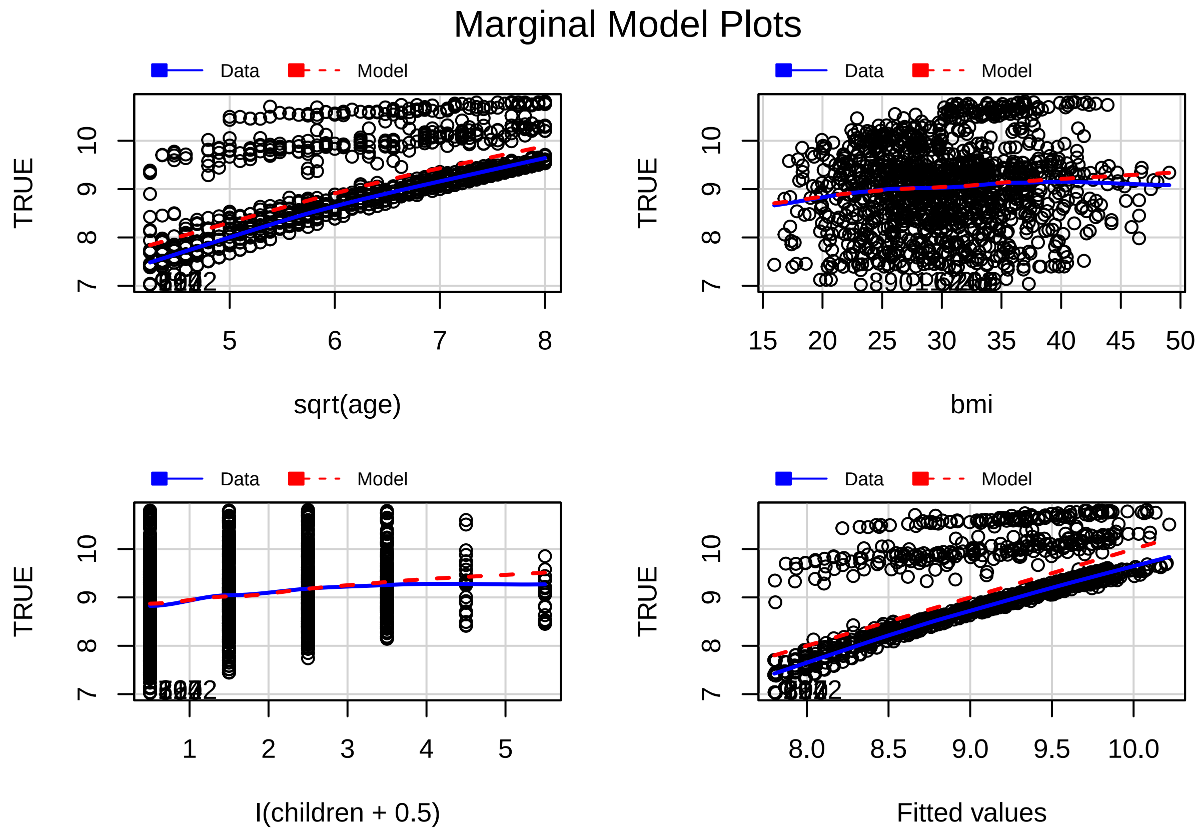
```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5),
##     data = df[-resout, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7879 -0.3448 -0.2632 -0.1008  2.2137
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.4626087  0.1252842  43.602  < 2e-16 ***
## sqrt(age)          0.5425257  0.0157220  34.507  < 2e-16 ***
## bmi               -0.0005796  0.0029662  -0.195    0.845
## I(children + 0.5)  0.1194819  0.0144412   8.274  3.3e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6152 on 1248 degrees of freedom
## Multiple R-squared:  0.513,  Adjusted R-squared:  0.5118
## F-statistic: 438.2 on 3 and 1248 DF,  p-value: < 2.2e-16
```

```
residualPlots( m4)
```



```
##                    Test stat Pr(>|Test stat|)
## sqrt(age)            -5.3092        1.303e-07 ***
## bmi                  -2.0439        0.0411761 *
## I(children + 0.5)    -3.7259        0.0002033 ***
## Tukey test           -6.4875        8.724e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
marginalModelPlots(m4, id=list(n=5, labels=rownames(df)))
```



Adding factors

After adding the factors to the models, it was seen that all the variables were significant to charges, so all of them were kept. Region was tested because it was multiclass but with the anova it was seen that the models with and without this variable were not equivalent so it can not be deleted.

```
names(df)
```

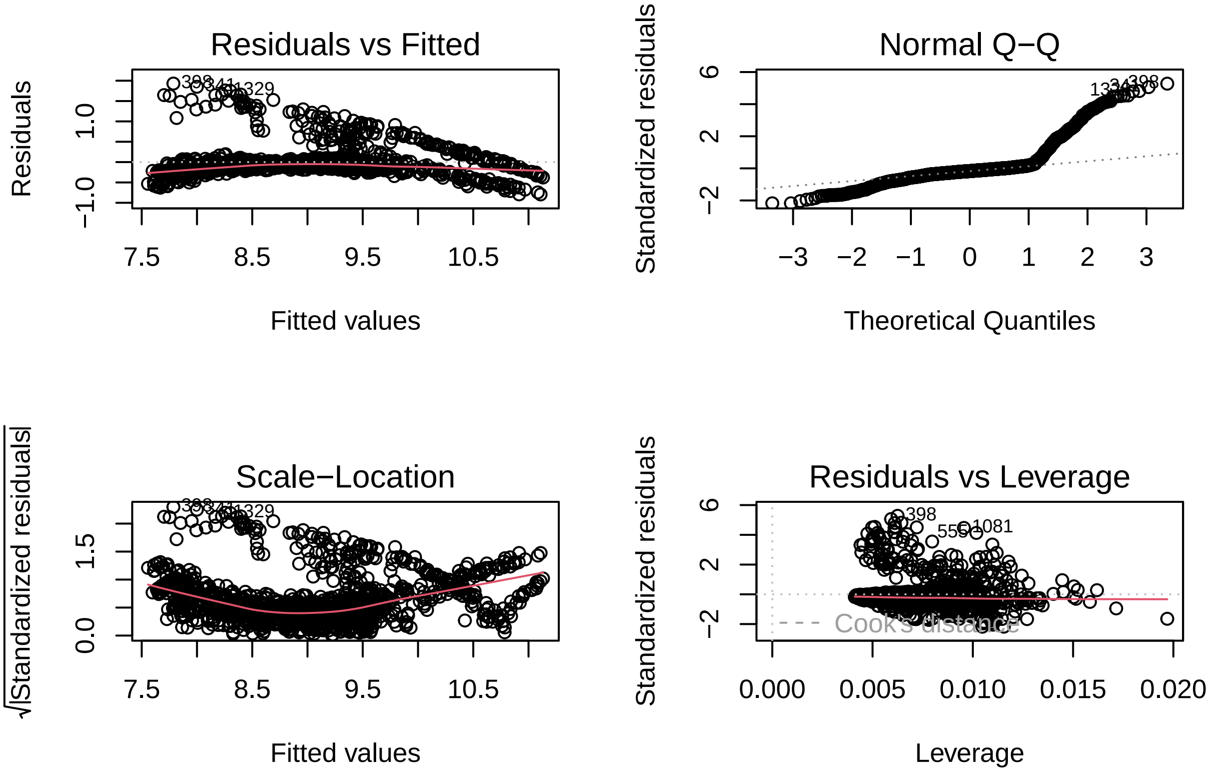
```
## [1] "age"      "sex"      "bmi"      "children" "smoker"   "region"   "charges"
```

```
m6 <- lm(log(charges) ~  sqrt(age) + bmi + I(children+0.5) + sex + smoker + region,
         data = df[-resout,])
summary(m6) # R2 = 0.8275
```

```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5) +
##     sex + smoker + region, data = df[-resout, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79428 -0.13989 -0.06244  0.01223  1.92975
##
## Coefficients:
```

```
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.441203   0.076936  70.724  < 2e-16 ***
## sqrt(age)           0.492680   0.009443  52.173  < 2e-16 ***
## bmi                 0.007951   0.001845   4.309 1.77e-05 ***
## I(children + 0.5)   0.103565   0.008630  12.001  < 2e-16 ***
## sexmale            -0.060708   0.020837  -2.913  0.00364 **
## smokeryes           1.344887   0.028460  47.255  < 2e-16 ***
## regionnorthwest    -0.081679   0.029685  -2.752  0.00602 **
## regionsoutheast    -0.150057   0.029946  -5.011 6.20e-07 ***
## regionsouthwest    -0.126912   0.029753  -4.266 2.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3669 on 1243 degrees of freedom
## Multiple R-squared:  0.8275, Adjusted R-squared:  0.8264
## F-statistic: 745.3 on 8 and 1243 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m6)
```



```
Anova(m6) # all variables are important
```

```
## Anova Table (Type II tests)
##
## Response: log(charges)
##                      Sum Sq   Df   F value    Pr(>F)    
## sqrt(age)            366.40    1 2722.0097 < 2.2e-16 ***
## bmi                    2.50    1   18.5658 1.771e-05 ***
## I(children + 0.5)     19.39    1  144.0122 < 2.2e-16 ***
## sex                    1.14    1    8.4881  0.003639 ** 
## smoker               300.58    1 2233.0348 < 2.2e-16 ***
## region                 3.92    3    9.7112 2.456e-06 ***
## Residuals            167.32 1243                        
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m7 <- step( m6, k=log(nrow(df))) # Step doesn't show any variable that can be deleted
```

```
# But we try a model without region, because it is a multiclass
m6reg <- lm(log(charges) ~sqrt(age) + bmi + I(children+0.5) + sex + smoker,
            data = df[-resout,])
# summary(m6reg)

anova(m6reg, m6)
```

```
## Analysis of Variance Table
## 
## Model 1: log(charges) ~ sqrt(age) + bmi + I(children + 0.5) + sex + smoker
## Model 2: log(charges) ~ sqrt(age) + bmi + I(children + 0.5) + sex + smoker + 
##     region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
## 1   1246 171.24                                  
## 2   1243 167.32  3    3.9216 9.7112 2.456e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

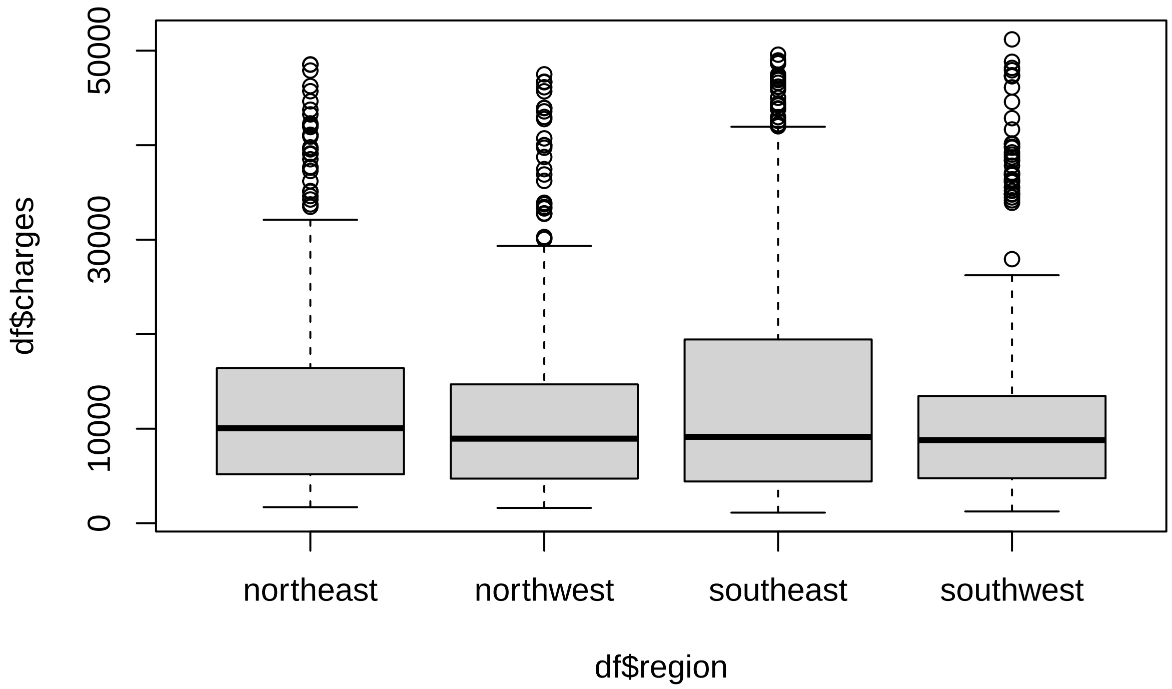
```
# Ho rejected, so models are not equivalent, so we can not delete region
```

Redefining factors

The variables region and bmi were candidates to be redefined, region because it has 4 categories, so it was wanted to be reduced to 2, and bmi because there was a pattern with it:

- region: According to the distributions by each group the south seems to have a common distribution by checking the min value, the median and the max, and the same with the north, so we will group them by north and south
- bmi: In the distribution of charges, the 3rd Qu. IS 16390, so above that we can consider expensive charges, so we decided to draw a line in charges of 30000, almost the double of the 3rd Qu. and it can be seen that approximately in bmi's higher than 30 it starts to be more frequent, so it was finally decided to create a factor that indicates if the individual has a bmi higher or equal than 30 or not

```
#region
par(mfrow=c(1,1))
plot(df$charges ~ df$region)
```



```
tapply(df$charges, df$region, summary)
```

```
## $northeast
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1695    5179   10043   13267   16398   48549
##
## $northwest
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1621    4724    8948   12171   14626   47496
##
## $southeast
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4415    9141   14385   19444   49578
##
## $southwest
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1242    4750    8791   12223   13454   51195
```

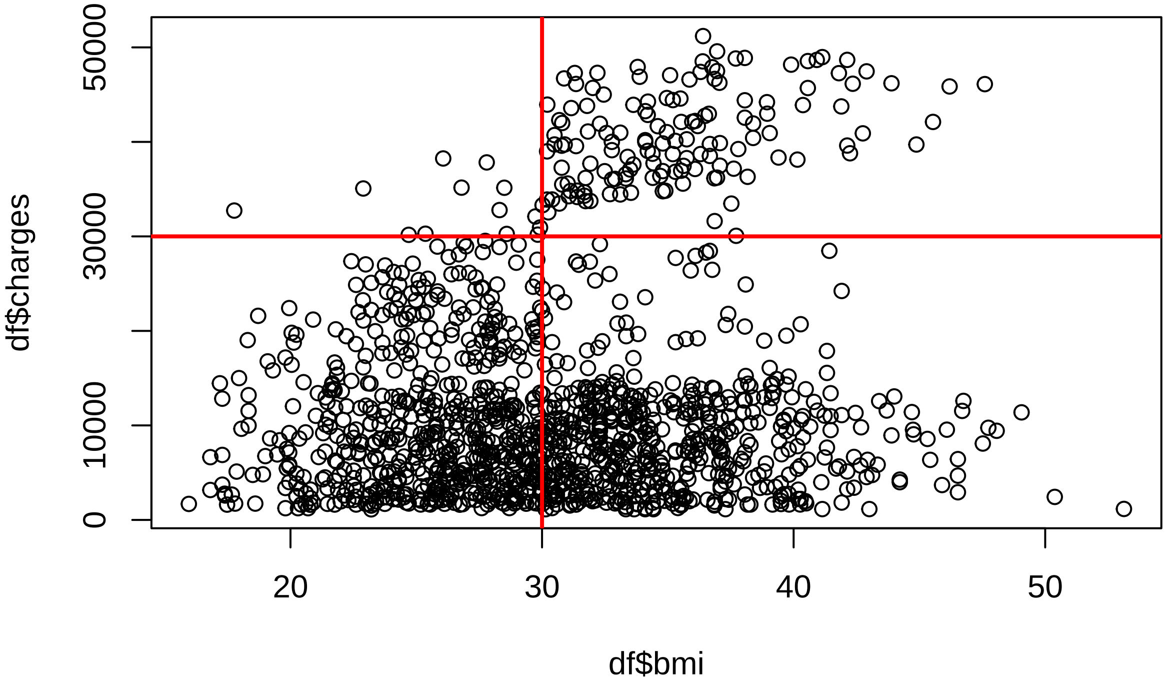
```
df$f.reg<-0
ll<-which(df$region %in% c("northeast","northwest"))
df$f.reg[ll]<-1
df$f.reg <- factor( df$f.reg, labels=c("south","north"))

#bmi
summary(df$bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.96   26.22   30.30   30.62   34.60   53.13
```

```
summary(df$charges)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4724    9303   13050   16390   51195
```

```
plot(df$charges ~ df$bmi)
abline( h=30000, lwd=2, col="red")
abline( v=30, lwd=2, col="red")
```



```
df$f.bmi<-0
ll<-which(df$bmi >= 30)
df$f.bmi[ll]<-1
df$f.bmi <- factor( df$f.bmi, labels=c("<30",">=30"))
```

Recalculate the model

The impact of the new factors was tested in two new models, one using only the new region and other using the new region and bmi, and comparing the BIC'S the result was that the best model was the one with only the new region.

```
# Checking the change only in region
m8 <- lm(log(charges) ~  sqrt(age) + bmi + I(children+0.5) + sex + smoker + f.reg,
data = df[-resout,])
# summary(m8) #R2 = 0.8264

# Checking the change in region and in bmi
m9 <- lm(log(charges) ~  sqrt(age) + f.bmi + I(children+0.5) + sex + smoker + f.reg,
         data = df[-resout,])
# summary(m9) #R2 = 0.8261
AIC(m9,m8,m7, k=log(nrow(df)))

# Comparing the models the m8 has the lower BIC(1099.018), so we decide to keep only the
# region factor transformation for the model
m10<- step(m9, k=log(nrow(df))) # No changes
```

Adding interactions

After adding the interactions and applying and step to delete the non significant interactions a final model was obtained. In that model's summary it was checked that the p-value of bmi was 0.57, so it was tested if it could be deleted, but by comparing the R2 and the BIC it didn't improved, so it was kept.

```
m11 <- lm(log(charges) ~  (sqrt(age) +bmi+ I(children+0.5)) * (sex + smoker + f.reg),
          data = df[-resout,])
#summary(m11) # R2 = 0.8643
m12 <- step(m11, k=log(nrow(df)))
```

```
summary(m12) # R2 = 0.8632
```

```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5) +
##     sex + smoker + f.reg + sqrt(age):sex + sqrt(age):smoker +
##     sqrt(age):f.reg + bmi:smoker + I(children + 0.5):smoker,
##     data = df[-resout, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36812 -0.12703 -0.06791 -0.02083  2.11532
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 5.248590   0.104597  50.179  < 2e-16 ***
## sqrt(age)                   0.533745   0.014562  36.654  < 2e-16 ***
## bmi                         0.000993   0.001751   0.567 0.570724
## I(children + 0.5)           0.114657   0.008333  13.760  < 2e-16 ***
## sexmale                    -0.464135   0.105186  -4.412 1.11e-05 ***
## smokeryes                   2.548418   0.195503  13.035  < 2e-16 ***
## f.regnorth                  0.475025   0.104862   4.530 6.47e-06 ***
## sqrt(age):sexmale           0.061649   0.016581   3.718 0.000210 ***
## sqrt(age):smokeryes        -0.405780   0.026646 -15.229  < 2e-16 ***
## sqrt(age):f.regnorth       -0.059080   0.016523  -3.576 0.000363 ***
## bmi:smokeryes               0.053314   0.004524  11.785  < 2e-16 ***
## I(children + 0.5):smokeryes -0.098324   0.021788  -4.513 7.00e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3271 on 1240 degrees of freedom
## Multiple R-squared:  0.8632, Adjusted R-squared:  0.862
## F-statistic: 711.6 on 11 and 1240 DF,  p-value: < 2.2e-16
```

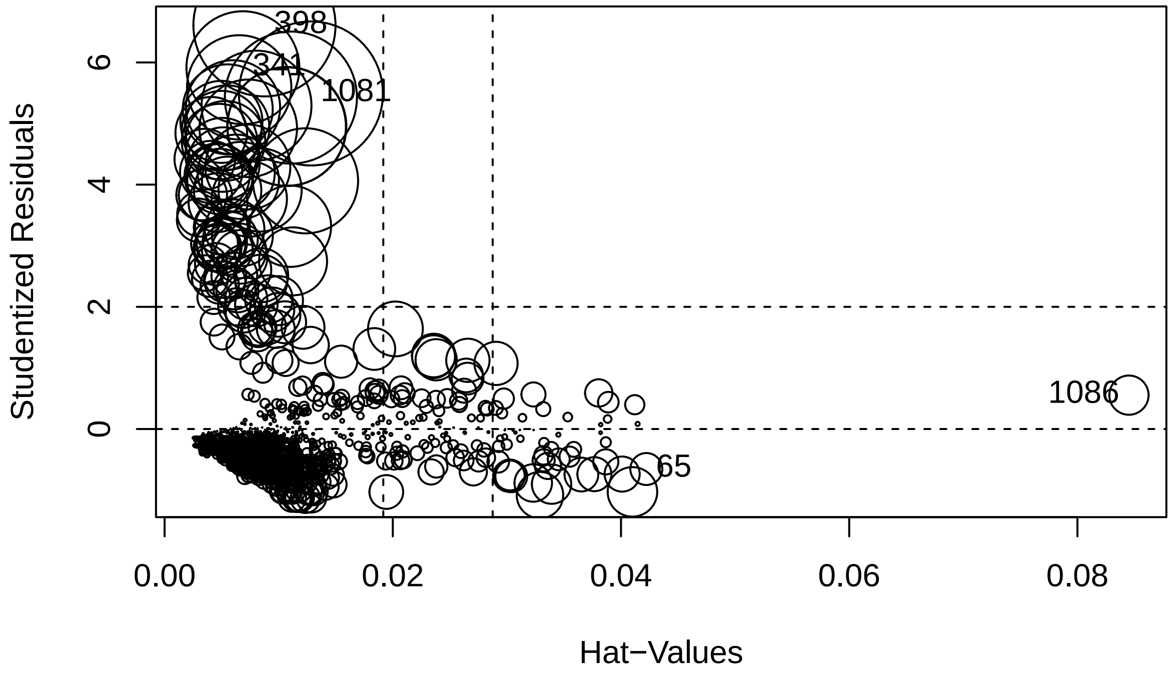
```
m13 <-lm(formula = log(charges) ~ sqrt(age)  + I(children + 0.5) + sex + smoker +
           f.reg + sqrt(age):sex + sqrt(age):smoker + sqrt(age):f.reg + bmi:smoker +
           I(children + 0.5):smoker, data = df[-resout, ])
# summary(m13)
AIC(m13,m12,m11, k=log(nrow(df)))
```

```
##     df      AIC
## m13 13 836.0310
## m12 13 836.0310
## m11 17 854.9863
```

Residual Analysis and Validation

To validate the final model the influencePlot was checked, and there were two values with hatvalues higher than the threshold, so a model was generated excluding them, and getting no real improvement

```
influencePlot(m12)
```



```
##         StudRes         Hat       CookD
## 65   -0.6541387 0.042221000 0.001572614
## 341   5.9138504 0.006864097 0.019606288
## 398   6.6067905 0.008752762 0.031051071
## 1081  5.4923432 0.012824765 0.031907485
## 1086  0.5546353 0.084516427 0.002367919
```

```
2*(length(coef(m12))/nrow(df)) # Threshold for Hatvalues
```

```
## [1] 0.01804511
```

```
#There are two values with hat values higher than the threshold so a model without them
# was decided to be tested
hh <- which(rownames(df) %in% c("65", "1086"));
m13<- lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5) + sex + smoker +
           f.reg + sqrt(age):sex + sqrt(age):smoker + sqrt(age):f.reg + bmi:smoker +
           I(children + 0.5):smoker, data = df[- c(resout,hh), ])
m14 <- step(m13, k=log(nrow(df)))
#summary(m14)
```

Final results and conclusions

- The final model was the one obtained after executing the step after adding the interactions and the final dataset was trimmed, deleting the influential data based on its cook's distance. This model has a coeficient of determination of 0.8632 and an BIC of 836.031.

- Age negatively affects the model because of how it is distributed in the three patterns previously discussed

- The final model does not achieve the get a fully Normal Q-Q plot, this could happen because in the residual vs fitted graph there is still a pattern, with that in mind it can be concluded that there is still a small pattern in the residuals that could not be identified

- Smoker is an important variable that highly affects the charges variable, if the person smokes, the charges tends to increase, despite the other characteristics, this can be confirmed in the allEffects plots

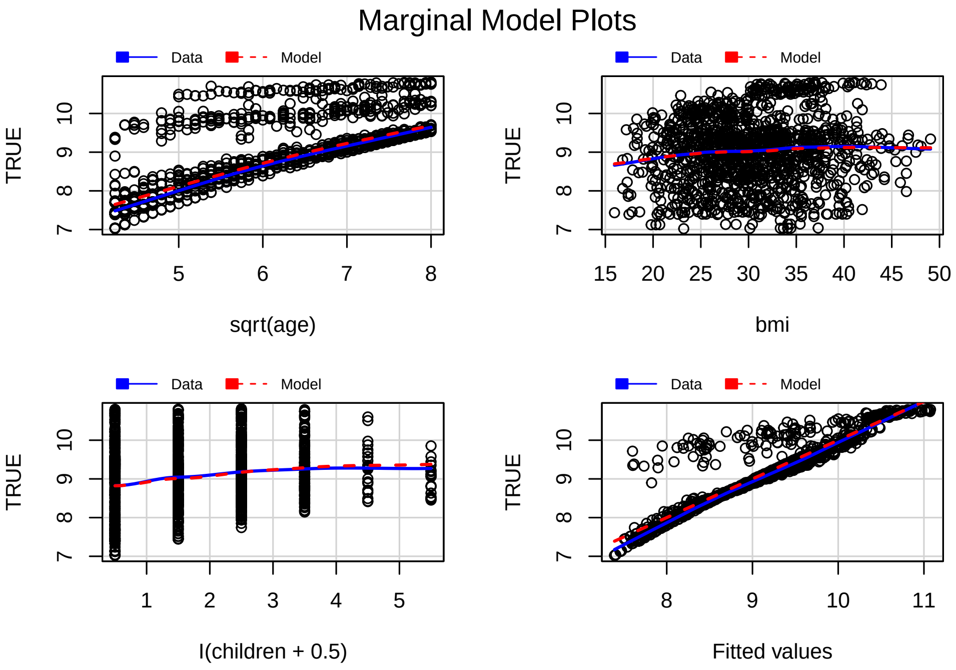
- In the Marginal Model Plots the patterns are very accurate in all the numeric variables having only an slightly difference in age, this tells us that there might be the problem and there can be a patter in its residuals

```
model<- lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5) + sex + smoker
           + f.reg + sqrt(age):sex + sqrt(age):smoker + sqrt(age):f.reg + bmi:smoker +
             I(children + 0.5):smoker, data = df[- resout, ])
summary(model)
```

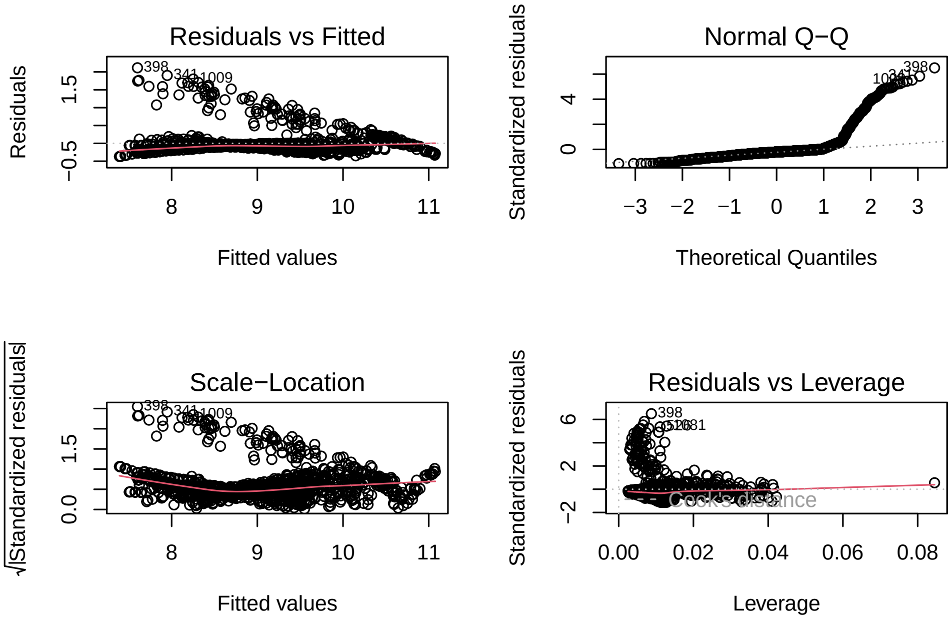
```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5) +
##     sex + smoker + f.reg + sqrt(age):sex + sqrt(age):smoker +
##     sqrt(age):f.reg + bmi:smoker + I(children + 0.5):smoker,
##     data = df[-resout, ])
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.36812 -0.12703 -0.06791 -0.02083  2.11532 
## 
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                  5.248590   0.104597  50.179  < 2e-16 ***
## sqrt(age)                    0.533745   0.014562  36.654  < 2e-16 ***
## bmi                          0.000993   0.001751   0.567 0.570724    
## I(children + 0.5)            0.114657   0.008333  13.760  < 2e-16 ***
## sexmale                     -0.464135   0.105186  -4.412 1.11e-05 ***
## smokeryes                    2.548418   0.195503  13.035  < 2e-16 ***
## f.regnorth                   0.475025   0.104862   4.530 6.47e-06 ***
## sqrt(age):sexmale            0.061649   0.016581   3.718 0.000210 ***
## sqrt(age):smokeryes         -0.405780   0.026646 -15.229  < 2e-16 ***
## sqrt(age):f.regnorth        -0.059080   0.016523  -3.576 0.000363 ***
## bmi:smokeryes                0.053314   0.004524  11.785  < 2e-16 ***
## I(children + 0.5):smokeryes -0.098324   0.021788  -4.513 7.00e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3271 on 1240 degrees of freedom
## Multiple R-squared:  0.8632, Adjusted R-squared:  0.862 
## F-statistic: 711.6 on 11 and 1240 DF,  p-value: < 2.2e-16
```

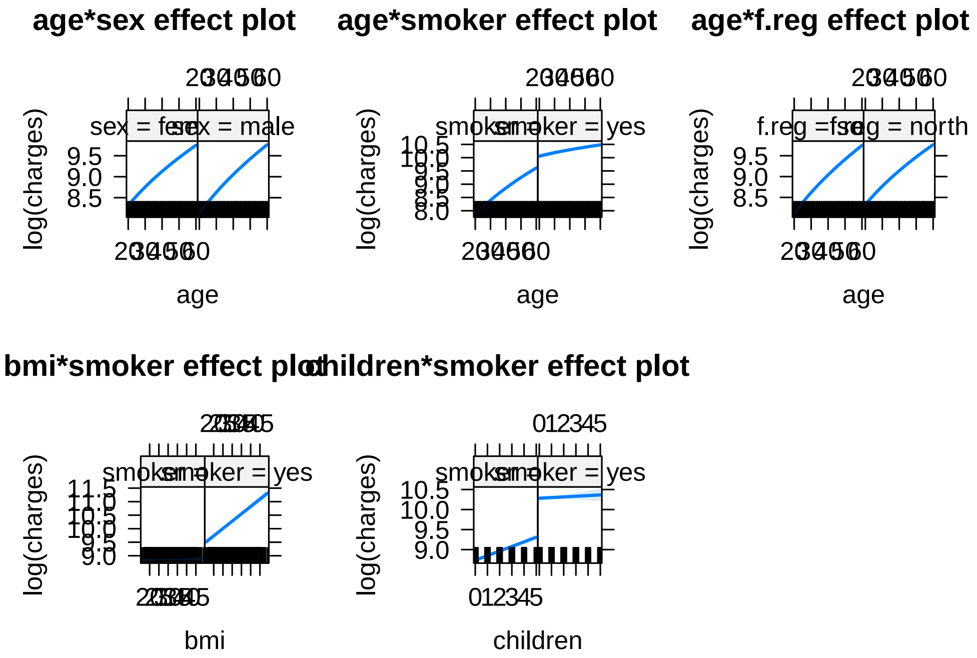
```
par(mfrow=c(2,2))
marginalModelPlots(model)
```



```
plot(model)
```



```
plot(allEffects(model))
```


age*sex effect plot
age*smoker effect plot
age*f.reg effect plot
bmi*smoker effect plot
children*smoker effect plot
log(charges)
age
bmi
children
sex = male
smoker = yes
f.reg = north