Assignment 1 Report

Eliya Tiram & Achraf Hmimou

Statistical Inference and Modeling

Index

```
library(GGally)
library(data.table)
library(car)

library(rpart)
library(chemometrics)
library(mvoutlier)

library(sgeostat)
library(lmtest)
```

Preparing the data in the environment

```
# Clear plots
if(!is.null(dev.list())) dev.off()

# Clean workspace
rm(list=ls())
#load data
df <- read.csv("insurance.csv")
```

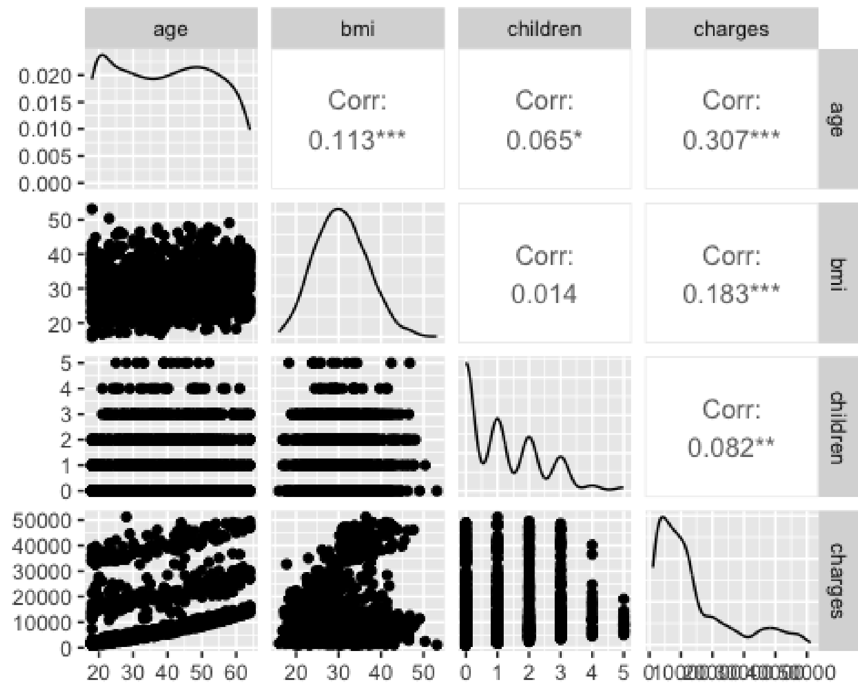
Explanatory data analysis

```
summary(df)

##       age            sex                 bmi           children
##  Min.   :18.00   Length:1323        Min.   :15.96   Min.   :0.00
##  1st Qu.:27.00   Class :character   1st Qu.:26.22   1st Qu.:0.00
##  Median :39.00   Mode  :character   Median :30.30   Median :1.00
##  Mean   :39.31                      Mean   :30.62   Mean   :1.08
##  3rd Qu.:51.00                      3rd Qu.:34.60   3rd Qu.:2.00
##  Max.   :64.00                      Max.   :53.13   Max.   :5.00
##     smoker              region             charges          f.sex
f.smoker
##  Length:1323        Length:1323        Min.   : 1122   female:654   no
:1058
##  Class :character   Class :character   1st Qu.: 4729   male  :669   yes:
265
##  Mode  :character   Mode  :character   Median : 9305
##                                        Mean   :13047
##                                        3rd Qu.:16265
##                                        Max.   :51195
##       f.region
##  northeast:320
##  northwest:321
##  southeast:360
##  southwest:322
```

```r
#numeric variables
summary(df[,c(1,3,4,7)])
```

```
##       age             bmi           children        charges     
##  Min.   :18.00   Min.   :15.96   Min.   :0.00   Min.   : 1122  
##  1st Qu.:27.00   1st Qu.:26.22   1st Qu.:0.00   1st Qu.: 4729  
##  Median :39.00   Median :30.30   Median :1.00   Median : 9305  
##  Mean   :39.31   Mean   :30.62   Mean   :1.08   Mean   :13047  
##  3rd Qu.:51.00   3rd Qu.:34.60   3rd Qu.:2.00   3rd Qu.:16265  
##  Max.   :64.00   Max.   :53.13   Max.   :5.00   Max.   :51195
```

```r
#plot(df[,c(1,3,4,7)])
ggpairs(df[,c(1,3,4,7)])
```

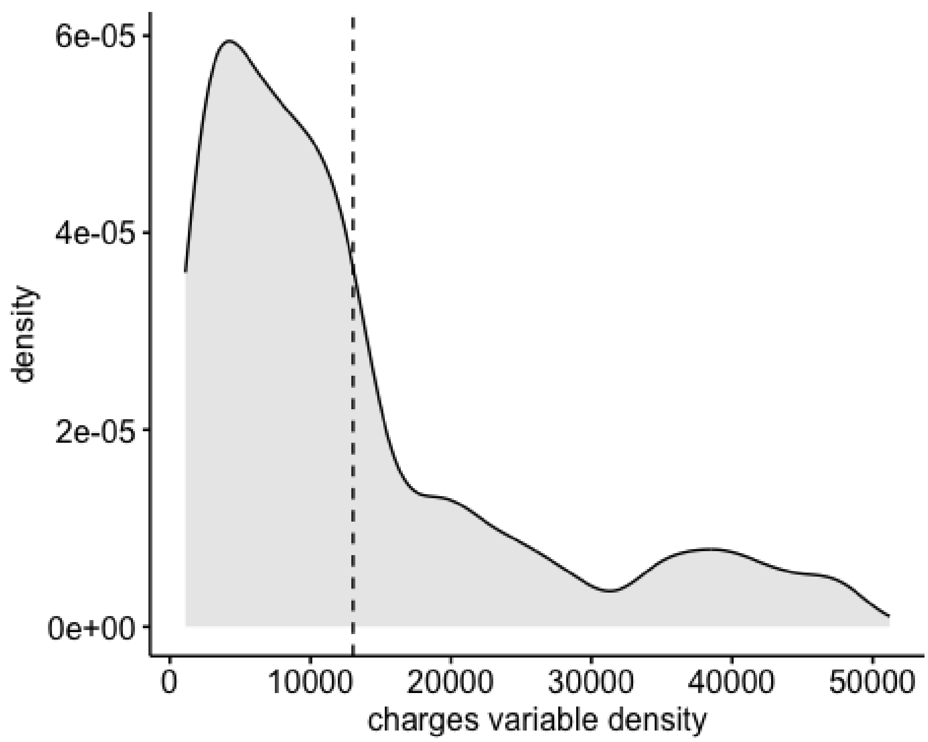



```r
#categorical variables
summary(df[,c(1,4,8:10)])
```

```
##       age           children         f.sex       f.smoker        f.region  
##  Min.   :18.00   Min.   :0.00   female:654   no :1058   northeast:320  
##  1st Qu.:27.00   1st Qu.:0.00   male  :669   yes: 265   northwest:321  
##  Median :39.00   Median :1.00                           southeast:360  
##  Mean   :39.31   Mean   :1.08                           southwest:322  
##  3rd Qu.:51.00   3rd Qu.:2.00                                          
##  Max.   :64.00   Max.   :5.00
```

From the summary we can see the factor values, it seems that sex and region are distributed equally and not much smokers compare to the non smokers. age and number of children looks about right and there are values in a range that makes sense. In addition, we see low correlation (0.198) between the target variable and the other numeric explanatory variable bmi. We don't see any pattern in the relation between the two variables. We see a number of extreme values with high bmi and/or charges.

```
# Density plot to check the distribution
ggpubr::ggdensity(df$charges,  fill = "lightgray", add = "mean",  xlab =
"charges variable density")

## Warning: `geom_vline()`: Ignoring `mapping` because `xintercept` was
provided.

## Warning: `geom_vline()`: Ignoring `data` because `xintercept` was
provided.

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2
3.4.0.
## i Please use `after_stat(density)` instead.
## i The deprecated feature was likely used in the ggpubr package.
##   Please report the issue at
<]8;;https://github.com/kassambara/ggpubr/issueshttps://github.com/kassambara
/ggpubr/issues]8;;>.
```




```
# Shapiro Test to asses that data on response variable is normaly
distribution
# H0 = Data is normally distributed
# H1 = Data is not normally distributed
# alfa = 0.05
shapiro.test(df$charges)
```

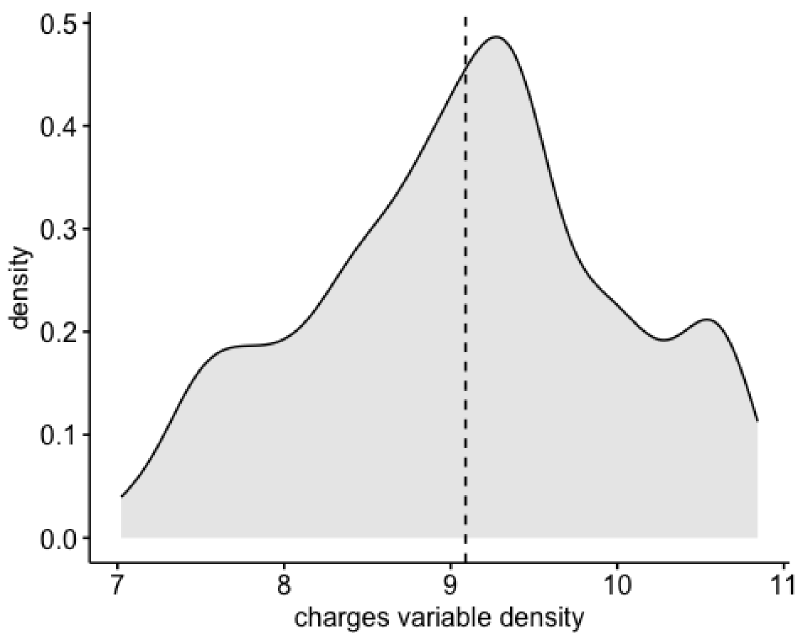
```
##
##  Shapiro-Wilk normality test
##
## data:  df$charges
## W = 0.81754, p-value < 2.2e-16
```

As we can see, the density plot shows that data is not normally distributed. To asses that, we can use one of many statistical tests that check normality on data. In this case, we use Shapiro test.

The result of the Shapiro test shows that data in variable **charges** is not normally distributed since *p-value* is less than the significance level (0.05) so we reject the null hypothesis (data is normally distributed) and we conclude that data is not normally distributed (alternative hypothesis)

Let's try to apply the log transformation

```
# Density plot to check the distribution
ggpubr::ggdensity(log(df$charges),  fill = "lightgray", add = "mean",  xlab =
"charges variable density")

## Warning: `geom_vline()`: Ignoring `mapping` because `xintercept` was
provided.

## Warning: `geom_vline()`: Ignoring `data` because `xintercept` was
provided.
```

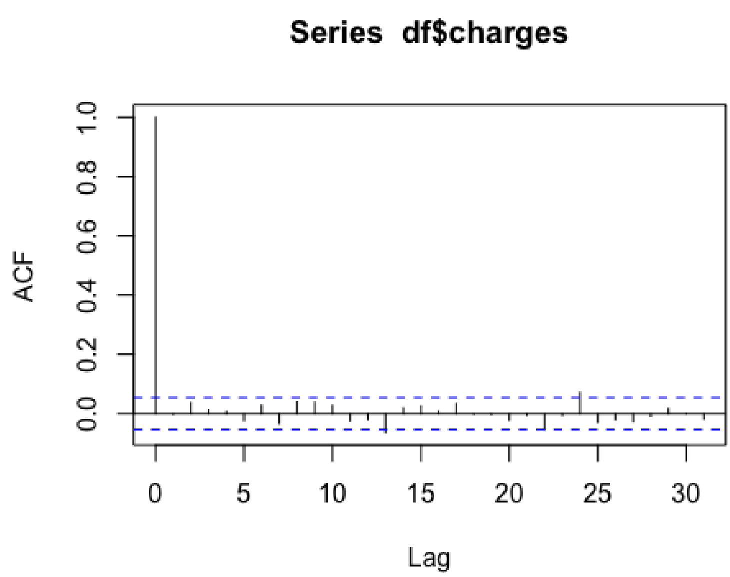



```
# Shapiro Test to asses that data on response variable is normaly
distribution
# H0 = Data is normally distributed
# H1 = Data is not normally distributed
```

```r
# alfa = 0.05
shapiro.test(log(df$charges))

##
##  Shapiro-Wilk normality test
##
## data:  log(df$charges)
## W = 0.98152, p-value = 5.679e-12
```

The null hypothesis can be still rejected so data still not being normally distributed.

```r
par(mfrow=c(1,1))
acf(df$charges)
```




```r
dwtest(df$charges~1)

##
##  Durbin-Watson test
##
## data:  df$charges ~ 1
## DW = 2.0054, p-value = 0.5394
## alternative hypothesis: true autocorrelation is greater than 0
```

Address tests to discard serial correlation: In the acf (auto correlation function) we can see from the graph that the data is not correlated where we have the blue threshold and all lines are within the threshold, we do see that there is one or two lines that crosses the threshold but just in a little bit so we leave it as it is without random the order of the observations. In addition we address Durbin-Watson test to check whether true autocorrelation is greater or not than 0. We see p-value 0.5183, thus we don't reject the null hypothesis and say that true autocorrelation is not greater than 0.

```r
#library(DataExplorer)
#create_report(df, y= "charges")

library(FactoMineR)
res.con <- condes(df[,c(1,3,4,7,8:10)], num.var = 4 , proba = 0.01 )
res.con$quanti

##          correlation      p.value
## age       0.30679657 3.128392e-30
## bmi       0.18280602 2.091908e-11
## children  0.08239851 2.705520e-03

res.con$quali

##                 R2       p.value
## f.smoker 0.6169962 1.418037e-277

res.con$category

##                Estimate       p.value
## f.smoker=yes  11493.36 1.418037e-277
## f.smoker=no  -11493.36 1.418037e-277
```

Association to the target variable, we see the numeric variable age 0.301 which is the most associated but the number is quite low and it is not strong association. Following age, we have bmi and then children.

For categorical variables we see that f.smoker is globally associated to charges, in particular, f.smoker=yes is very remarkable. Let's check the case of smoker category.

```r
res.cat <- catdes(df[,c(9,3,4,7,8, 1, 10)], num.var = 1 , proba = 0.01 )

res.cat$quanti

## $no
##            v.test Mean in category Overall mean sd in category Overall sd
## charges -28.55992         8443.049     13047.35       5993.179   11712.29
##               p.value
## charges 2.115342e-179
##
## $yes
##           v.test Mean in category Overall mean sd in category Overall sd
## charges 28.55992         31429.78     13047.35       10904.12   11712.29
##               p.value
## charges 2.115342e-179

res.cat$category

## $no
##                Cla/Mod  Mod/Cla   Global     p.value    v.test
## f.sex=female 83.02752 51.32325 49.43311 0.006035893  2.745825
## f.sex=male   76.98057 48.67675 50.56689 0.006035893 -2.745825
```

```
##
## $yes
##                  Cla/Mod  Mod/Cla   Global     p.value    v.test
## f.sex=male     23.01943 58.11321 50.56689 0.006035893  2.745825
## f.sex=female 16.97248 41.88679 49.43311 0.006035893 -2.745825
```

```
res.cat$quanti.var
```

```
##                Eta2       P-value
## charges 0.6169962 1.418037e-277
```

We can see that the mean of charges for smokers is much more higher than people who don't smoke. Smoking seems a very important influence in the price for having high insurance charges.

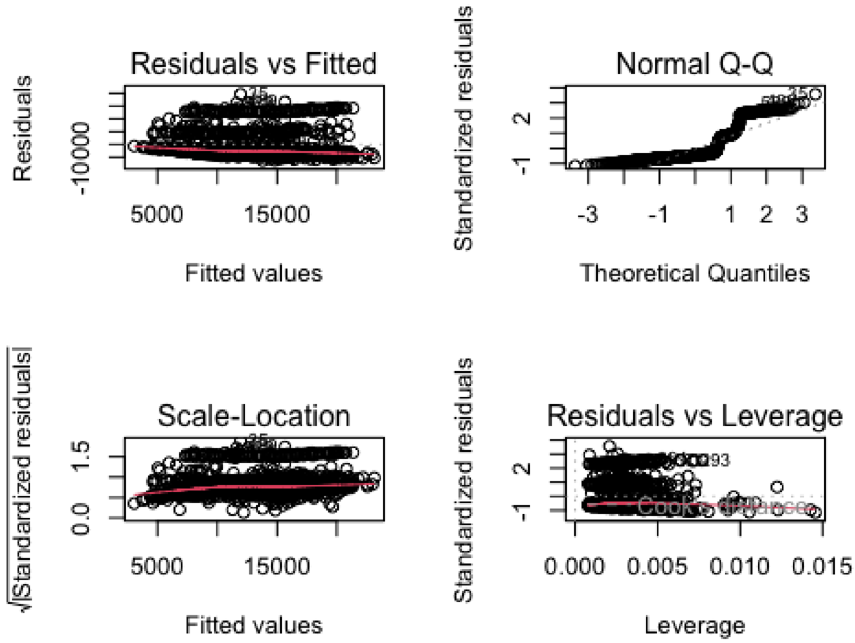
Building the model

First model

```
m1<-lm(charges~bmi+age+children, data = df)
summary(m1)
```

```
##
## Call:
## lm(formula = charges ~ bmi + age + children, data = df)
##
## Residuals:
##    Min      1Q Median     3Q    Max
## -12628  -6735  -5057   5894  39232
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5853.05    1710.99  -3.421 0.000643 ***
## bmi           288.72      50.14   5.758 1.06e-08 ***
## age           239.14      21.79  10.977  < 2e-16 ***
## children      610.04     255.28   2.390 0.017003 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11000 on 1319 degrees of freedom
## Multiple R-squared:  0.1202, Adjusted R-squared:  0.1182
## F-statistic: 60.05 on 3 and 1319 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m1)
```




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

Looking at the summary of the model, the RSquared is very low and there is a lot of residual standard error.

If we study the residual error looking at the plots we can see that the data is not following a normal distribution since there are deviations of the line (Normal Q-Q plot). Also there are a lot of sparsity in the variance (Scale-Location plot).

Asses multicollinearity

Maybe there is multicollinearity that is causing bad results

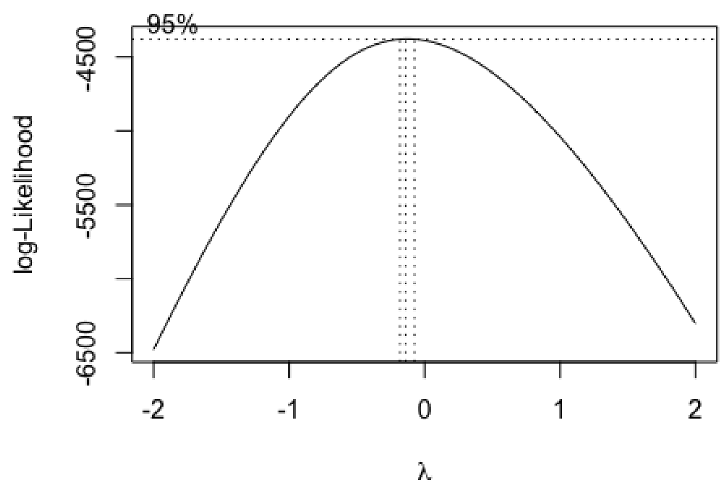
```
car::vif(m1)

##      bmi      age children
## 1.012957 1.017013 1.004239
```

The vif values are low (less than 5) so there aren't problems of multidisciplinary.

Let's try to do some transformations to the data.

Transformation

```
library(MASS)

boxcox(charges~bmi+age+children, data = df)
```



The boxplots shows that the lambda values are close to 0 so a logarithmic transformation to the target variable should help to improve the results

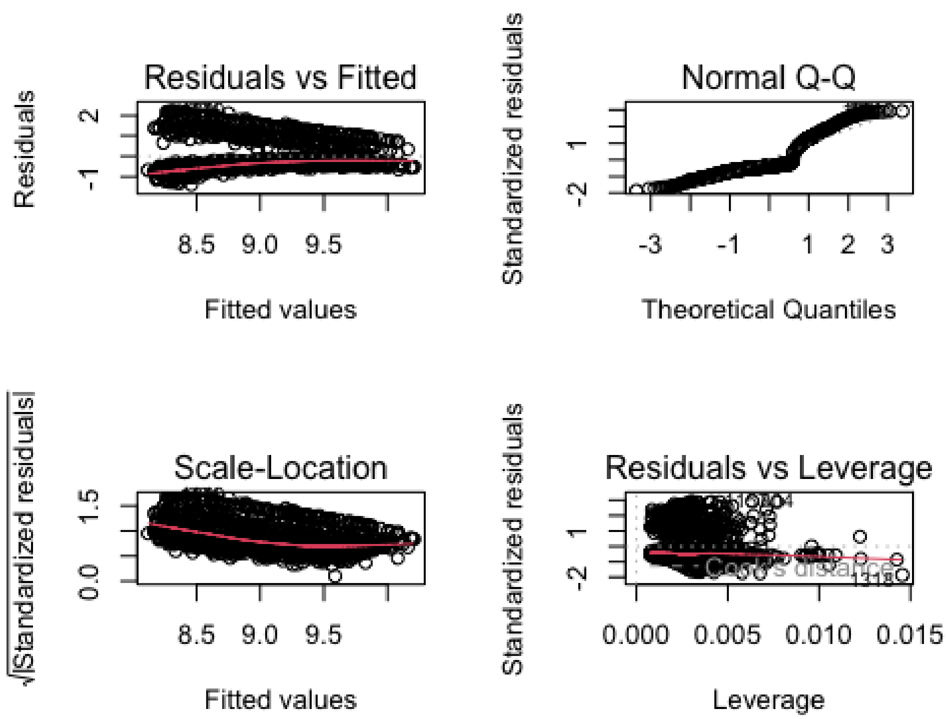
```
# (only for numerical variables)

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

##                     MLE of lambda Score Statistic (z) Pr(>|z|)
## bmi                      -1.07828             -1.4110  0.15824
## age                       0.42692             -1.7687  0.07694 .
## I(children + 0.5)         0.25004             -1.7969  0.07235 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## iterations =  16

# poly(age,3) for adding ortogonal polynomial
```

The transformations of the explanatory variables are not performed since all p-values are above 0.05 significance level.

```
m2 <- lm(log(charges)~bmi+ age+children, data = df)

summary(m2)

##
## Call:
## lm(formula = log(charges) ~ bmi + age + children, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3991 -0.4339 -0.3051  0.4777  2.2134
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.360187   0.117968  62.392  < 2e-16 ***
## bmi         0.009188   0.003457   2.657  0.00797 **
## age         0.033885   0.001502  22.558  < 2e-16 ***
## children    0.107190   0.017601   6.090 1.48e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7586 on 1319 degrees of freedom
## Multiple R-squared:  0.3105, Adjusted R-squared:  0.309
## F-statistic:   198 on 3 and 1319 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m2)
```

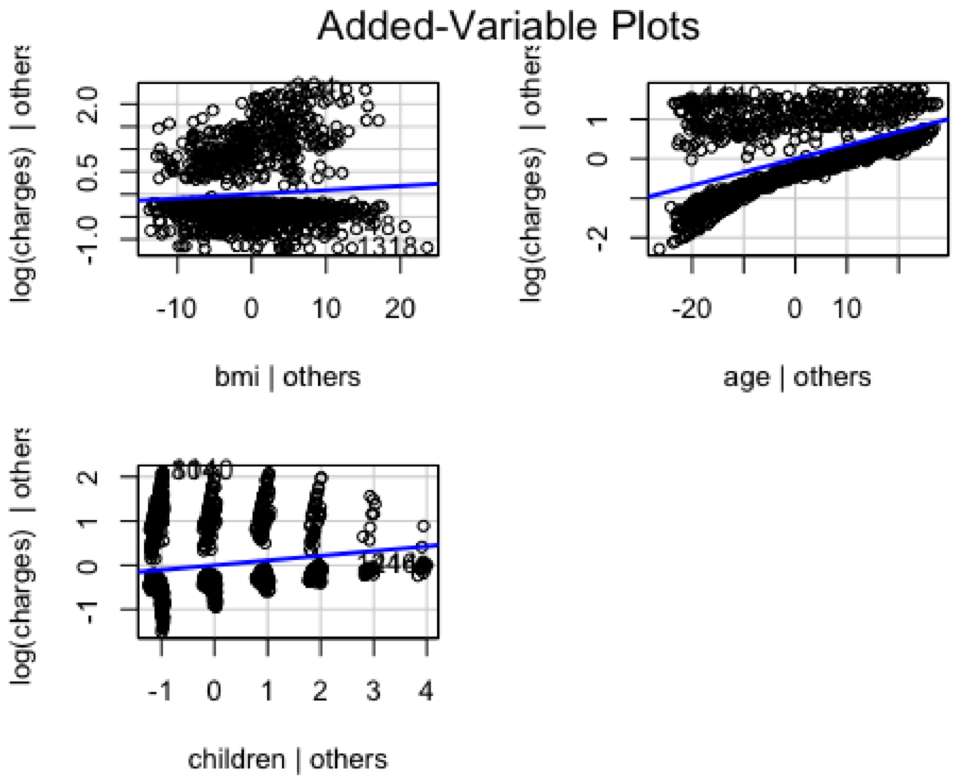



```
par(mfrow=c(1,1))
```

The model is still not performing very well. However if we check the study of residuals we can see that it results in an improvement.

The normal Q-Q plot still have a deviation but is that big as the m1 and if we check the Scale-Location of the standard residuals the variance is better.

```
avPlots(m2)
```



The partial regressions plots shows that all regresors have two big clusters of data.

```
AIC(m1,m2)

##    df       AIC
## m1  5 28383.903
## m2  5  3029.469
```

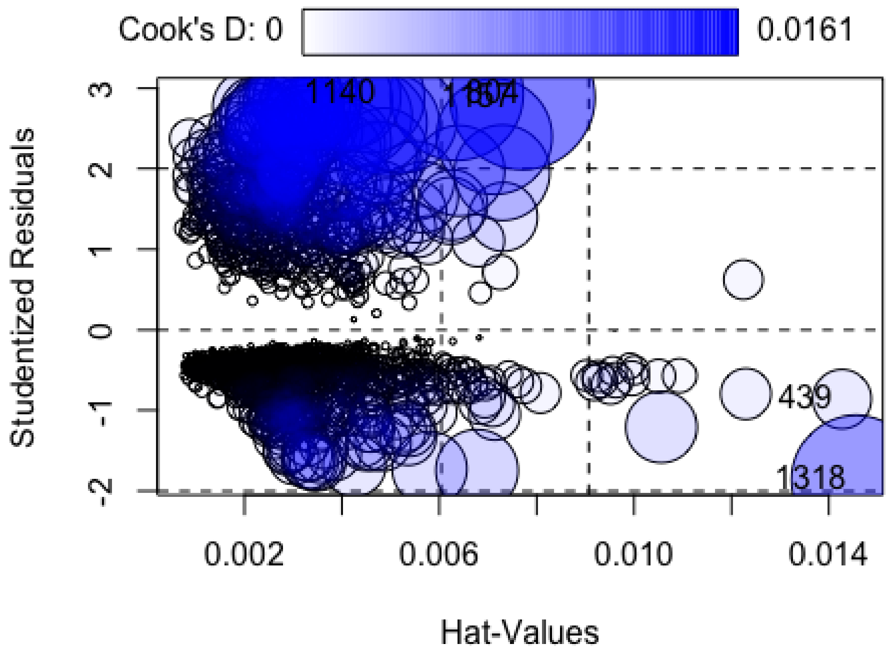
The AIC test shows that model 2 is performing much better than model 1 so we will continue with it.

Influential data

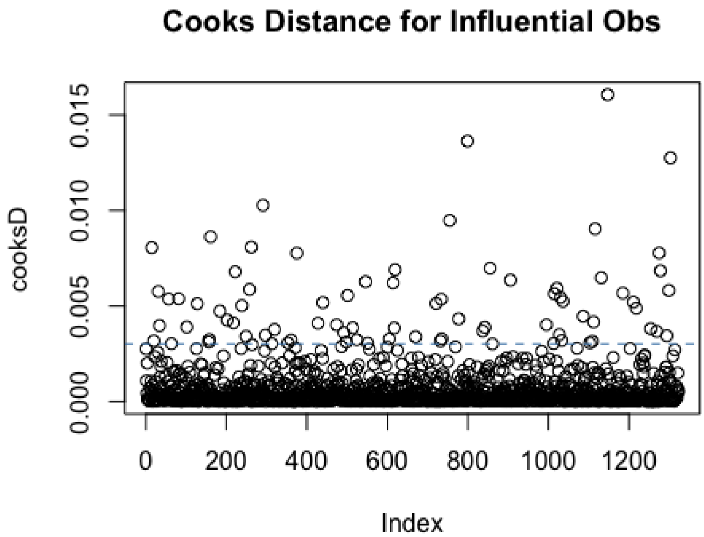
Maybe, by removing influential data the results can be improved.

- Residual outliers
- Influential values

```
library(car)

influencePlot(m2)
```

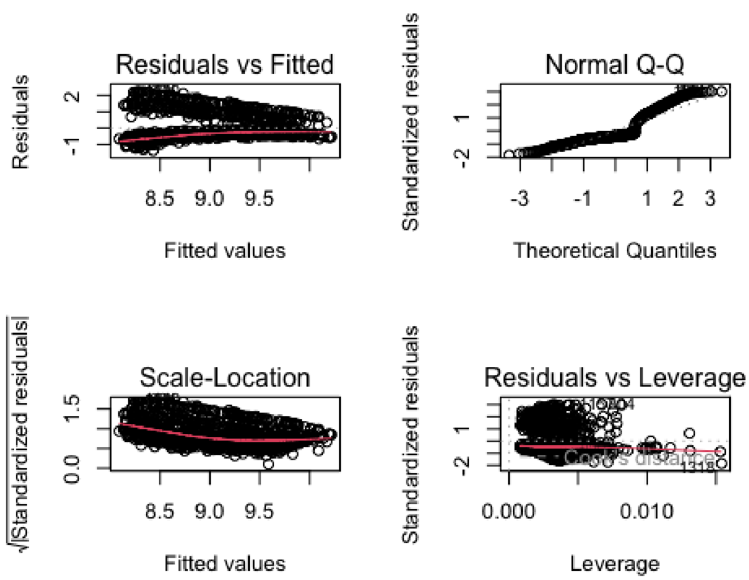



```
##           StudRes         Hat       CookD
## 439  -0.8586974 0.014270736 0.002669288
## 804   2.9280088 0.006355073 0.013629744
## 1140  2.9305989 0.003026175 0.006479950
## 1157  2.8840078 0.007706514 0.016060091
## 1318 -1.8596140 0.014562223 0.012751918

# there are a lot of influential data


# With cooks distance
cooksD <- cooks.distance(m2)
n <- nrow(df)
plot(cooksD, main = "Cooks Distance for Influential Obs")
abline(h = 4/n, lty = 2, col = "steelblue") # add cutoff line
```

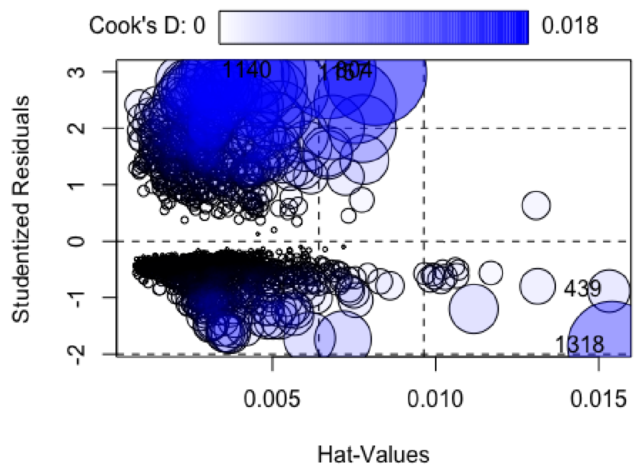



```
influential_obs <- as.numeric(names(cooksD)[(cooksD > (4/n))])
influential_obs

##  [1]   15   20   31   35   58   65   83  103  129  158  159  162  186  204
220
## [16]  224  241  251  260  264  293  299  315  322  355  363  378  431  443
477
## [31]  495  501  504  517  527  550  555  610  619  622  624  675  726  737
739
## [46]  740  760  782  804  843  848  861  912 1002 1020 1022 1028 1034 1037
1040
## [61] 1043 1094 1112 1118 1121 1125 1140 1157 1197 1224 1232 1268 1283 1289
1292
## [76] 1309 1314 1318

length(influential_obs)

## [1] 78

m3 <- lm(log(charges)~bmi+age+children, data=df[-influential_obs,])
summary(m3)

##
## Call:
## lm(formula = log(charges) ~ bmi + age + children, data =
df[-influential_obs,
##     ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3732 -0.4262 -0.3001  0.4490  2.2392
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.320468   0.120584  60.708  < 2e-16 ***
## bmi         0.009217   0.003512   2.624  0.00879 **
## age         0.034569   0.001524  22.691  < 2e-16 ***
## children    0.109175   0.017948   6.083 1.57e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.747 on 1241 degrees of freedom
## Multiple R-squared:  0.3241, Adjusted R-squared:  0.3225
## F-statistic: 198.4 on 3 and 1241 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(m3)
```



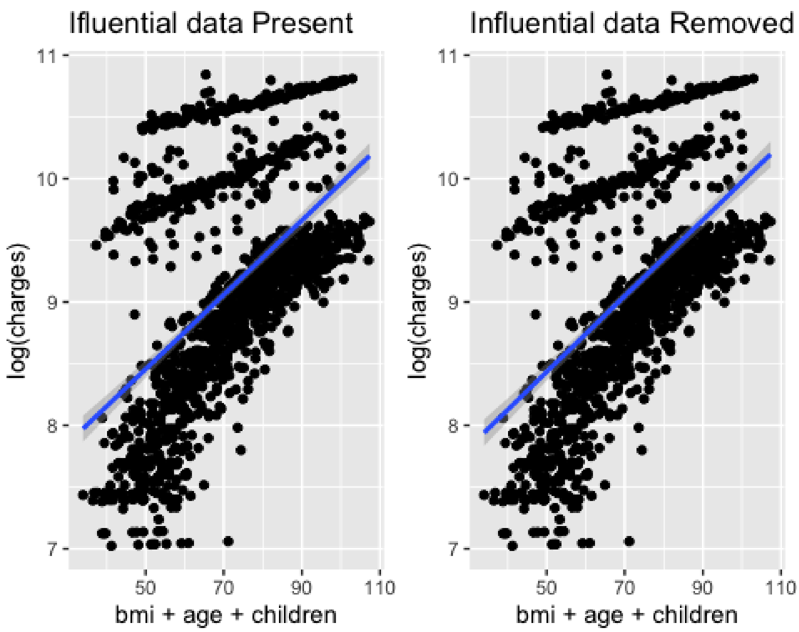

```
par(mfrow=c(1,1))

influencePlot(m3)
```




```
##          StudRes         Hat       CookD
## 439  -0.8822072 0.015318123 0.003027391
## 804   3.0105579 0.006725594 0.015243428
## 1140  3.0123385 0.003233415 0.007311376
## 1157  2.9649173 0.008153223 0.017952793
## 1318 -1.8545683 0.015404544 0.013426531
```

```
#create scatterplot with influential data present
outliers_present <- ggplot(data = df, aes(x = bmi + age + children, y =
log(charges))) +
  geom_point() +
  geom_smooth(method = lm) +
#  ylim(0, 200) +
  ggtitle("Ifluential data Present")

#create scatterplot with influential data removed
outliers_removed <- ggplot(data = df[-influential_obs,], aes(x = bmi + age +
children, y = log(charges))) +
  geom_point() +
  geom_smooth(method = lm) +
#  ylim(0, 200) +
  ggtitle("Influential data Removed")

#plot both scatterplots side by side
gridExtra::grid.arrange(outliers_present, outliers_removed, ncol = 2)
```

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



Adding factors

- Check that meaning of a factor could not be related to the numerical variables so one should be used.
- AIC test to compare

```
summary(df)

##       age            sex                   bmi           children   
##  Min.   :18.00   Length:1323        Min.   :15.96   Min.   :0.00  
##  1st Qu.:27.00   Class :character   1st Qu.:26.22   1st Qu.:0.00  
##  Median :39.00   Mode  :character   Median :30.30   Median :1.00  
##  Mean   :39.31                      Mean   :30.62   Mean   :1.08  
##  3rd Qu.:51.00                      3rd Qu.:34.60   3rd Qu.:2.00  
##  Max.   :64.00                      Max.   :53.13   Max.   :5.00  
##     smoker             region             charges          f.sex    
f.smoker
##  Length:1323        Length:1323        Min.   : 1122   female:654   no 
:1058  
##  Class :character   Class :character   1st Qu.: 4729   male  :669   yes:
265  
##  Mode  :character   Mode  :character   Median : 9305                
##                                        Mean   :13047                
##                                        3rd Qu.:16265                
##                                        Max.   :51195                
##       f.region  
##  northeast:320  
##  northwest:321  
##  southeast:360  
##  southwest:322  
```

```
##
##
```

```r
m4 <- lm(log(charges)~bmi+age+children+f.sex+f.smoker+f.region,
data=df[-influential_obs,])
summary(m4)
```

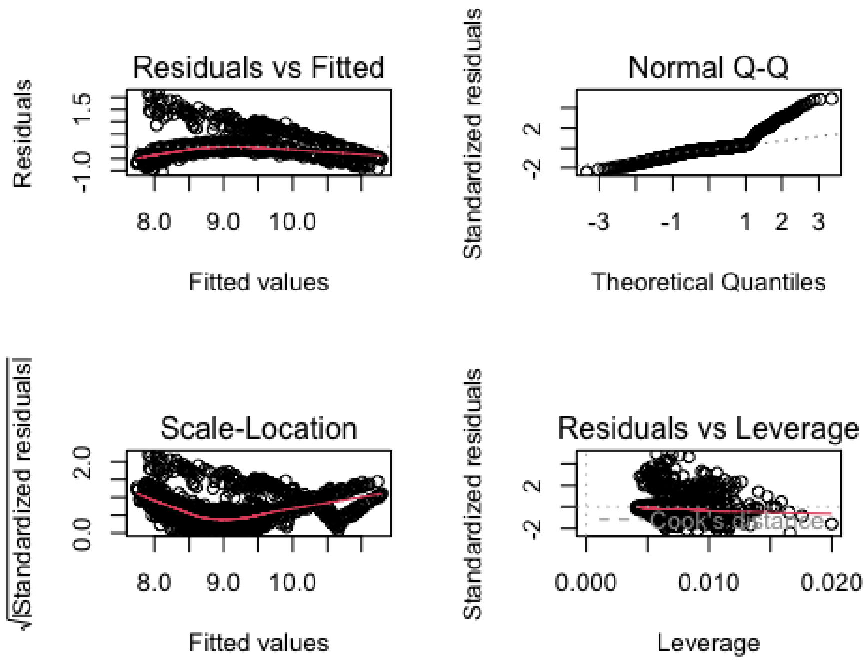
```
##
## Call:
## lm(formula = log(charges) ~ bmi + age + children + f.sex + f.smoker +
##     f.region, data = df[-influential_obs, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05304 -0.19908 -0.05181  0.05959  2.16809
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.0366548  0.0752441  93.518  < 2e-16 ***
## bmi                 0.0124863  0.0021630   5.773 9.86e-09 ***
## age                 0.0349941  0.0008989  38.929  < 2e-16 ***
## children            0.1023007  0.0105952   9.655  < 2e-16 ***
## f.sexmale          -0.0746327  0.0250596  -2.978 0.002956 **
## f.smokeryes         1.5244571  0.0316950  48.098  < 2e-16 ***
## f.regionnorthwest -0.0613902  0.0357542  -1.717 0.086229 .
## f.regionsoutheast -0.1431207  0.0359860  -3.977 7.38e-05 ***
## f.regionsouthwest -0.1267455  0.0360563  -3.515 0.000455 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4402 on 1236 degrees of freedom
## Multiple R-squared:  0.7662, Adjusted R-squared:  0.7647
## F-statistic: 506.4 on 8 and 1236 DF,  p-value: < 2.2e-16
```

Let's try to check if there are factors that could be removed

```r
Anova(m4)
```

```
## Anova Table (Type II tests)
##
## Response: log(charges)
##           Sum Sq   Df   F value    Pr(>F)
## bmi         6.46    1   33.3225 9.863e-09 ***
## age       293.68    1 1515.5052 < 2.2e-16 ***
## children   18.07    1   93.2266 < 2.2e-16 ***
## f.sex       1.72    1    8.8698 0.0029559 **
## f.smoker  448.30    1 2313.3955 < 2.2e-16 ***
## f.region    3.80    3    6.5320 0.0002206 ***
## Residuals 239.52 1236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
m5 <- step( m4 )

## Start:  AIC=-2034.08
## log(charges) ~ bmi + age + children + f.sex + f.smoker + f.region
##
##              Df Sum of Sq    RSS      AIC
## <none>                    239.52 -2034.08
## - f.sex     1      1.72 241.24 -2027.17
## - f.region  3      3.80 243.32 -2020.49
## - bmi       1      6.46 245.98 -2002.96
## - children  1     18.07 257.59 -1945.54
## - age       1    293.68 533.20 -1039.74
## - f.smoker  1    448.30 687.82  -722.73

summary(m5)

##
## Call:
## lm(formula = log(charges) ~ bmi + age + children + f.sex + f.smoker +
##     f.region, data = df[-influential_obs, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.05304 -0.19908 -0.05181  0.05959  2.16809
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.0366548  0.0752441  93.518  < 2e-16 ***
## bmi                0.0124863  0.0021630   5.773 9.86e-09 ***
## age                0.0349941  0.0008989  38.929  < 2e-16 ***
## children           0.1023007  0.0105952   9.655  < 2e-16 ***
## f.sexmale         -0.0746327  0.0250596  -2.978 0.002956 **
## f.smokeryes        1.5244571  0.0316950  48.098  < 2e-16 ***
## f.regionnorthwest -0.0613902  0.0357542  -1.717 0.086229 .
## f.regionsoutheast -0.1431207  0.0359860  -3.977 7.38e-05 ***
## f.regionsouthwest -0.1267455  0.0360563  -3.515 0.000455 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4402 on 1236 degrees of freedom
## Multiple R-squared:  0.7662, Adjusted R-squared:  0.7647
## F-statistic: 506.4 on 8 and 1236 DF,  p-value: < 2.2e-16

par( mfrow = c(2,2))
plot( m5, id.n=0 )
```




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

Le's try to transform age into a factor

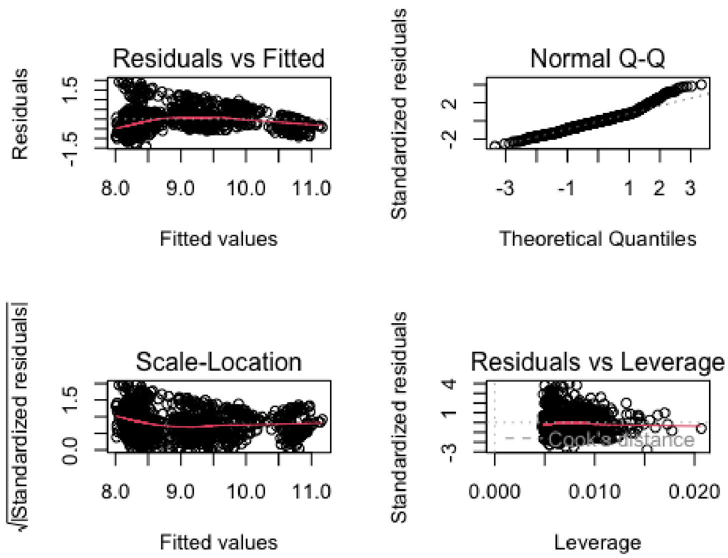
```
df$age_range <- cut(df$age, breaks = quantile(df$age,probs = c(0,0.5,1)),
include.lowest = T)
summary(df)
```

```
##       age            sex                  bmi           children
##  Min.   :18.00   Length:1323        Min.   :15.96   Min.   :0.00
##  1st Qu.:27.00   Class :character   1st Qu.:26.22   1st Qu.:0.00
##  Median :39.00   Mode  :character   Median :30.30   Median :1.00
##  Mean   :39.31                      Mean   :30.62   Mean   :1.08
##  3rd Qu.:51.00                      3rd Qu.:34.60   3rd Qu.:2.00
##  Max.   :64.00                      Max.   :53.13   Max.   :5.00
##     smoker             region             charges          f.sex
f.smoker
##  Length:1323        Length:1323        Min.   : 1122   female:654   no
:1058
##  Class :character   Class :character   1st Qu.: 4729   male  :669   yes:
265
##  Mode  :character   Mode  :character   Median : 9305
##                                        Mean   :13047
##                                        3rd Qu.:16265
##                                        Max.   :51195
##       f.region     age_range
##  northeast:320   [18,39]:663
##  northwest:321   (39,64]:660
##  southeast:360
```

```
##  southwest:322
##
```

We have created a new variable called age_range where we divide the ages into 4 groups according to the 4 quantiles. From the summary (and the new column in the data set) we see 4 groups of ages and how many observations fit into each age group. The results do not change a lot with 4 quantiles and we tried with 2 groups and this got a more interesting result.

```
m6 <- lm(log(charges)~bmi+children+f.sex+f.smoker+f.region+age_range,
data=df[-influential_obs,])
summary(m6)

##
## Call:
## lm(formula = log(charges) ~ bmi + children + f.sex + f.smoker +
##     f.region + age_range, data = df[-influential_obs, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40598 -0.32131 -0.02742  0.25982  2.02237
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.923742   0.080374  98.586  < 2e-16 ***
## bmi                0.014752   0.002484   5.939 3.73e-09 ***
## children           0.113214   0.012172   9.301  < 2e-16 ***
## f.sexmale         -0.083762   0.028814  -2.907 0.003715 **
## f.smokeryes        1.520719   0.036448  41.722  < 2e-16 ***
## f.regionnorthwest -0.060609   0.041120  -1.474 0.140747
## f.regionsoutheast -0.158606   0.041374  -3.833 0.000133 ***
## f.regionsouthwest -0.125455   0.041467  -3.025 0.002534 **
## age_range(39,64]   0.838466   0.028852  29.061  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5062 on 1236 degrees of freedom
## Multiple R-squared:  0.6908, Adjusted R-squared:  0.6888
## F-statistic: 345.2 on 8 and 1236 DF,  p-value: < 2.2e-16

par( mfrow = c(2,2))
plot( m6, id.n=0 )
```




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

AIC(m5,m6)

##    df      AIC
## m5 10 1501.080
## m6 10 1849.079
```

Removing age as a numerical explanatory variable and adding it as a factor does not improve things in general. However we can see that the normal Q-Q plot from the residuals is better since we reduced the impact of the age variable. We will continue with model 5.

Adding interactions

With the model added factors will try to check adding double interactions between all numerical and factors.

```
m7 <- lm(log(charges)~bmi+age
         +children * (f.sex+f.smoker+f.region), data=df[-influential_obs,])

summary(m7)

##
## Call:
## lm(formula = log(charges) ~ bmi + age + children * (f.sex + f.smoker +
##     f.region), data = df[-influential_obs, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99576 -0.21193 -0.05748  0.06408  2.17375
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                  7.0207859  0.0783886  89.564  < 2e-16 ***
## bmi                          0.0125351  0.0021465   5.840 6.68e-09 ***
## age                          0.0350238  0.0008924  39.246  < 2e-16 ***
## children                     0.1122317  0.0247948   4.526 6.58e-06 ***
## f.sexmale                   -0.0968541  0.0336195  -2.881 0.004034 **
## f.smokeryes                  1.6711212  0.0435590  38.365  < 2e-16 ***
## f.regionnorthwest           -0.0724783  0.0484149  -1.497 0.134643
## f.regionsoutheast           -0.1654342  0.0470327  -3.517 0.000452 ***
## f.regionsouthwest           -0.0973501  0.0479806  -2.029 0.042679 *
## children:f.sexmale           0.0250942  0.0210920   1.190 0.234374
## children:f.smokeryes        -0.1319506  0.0272271  -4.846 1.42e-06 ***
## children:f.regionnorthwest   0.0101619  0.0304850   0.333 0.738934
## children:f.regionsoutheast   0.0174022  0.0296315   0.587 0.557118
## children:f.regionsouthwest  -0.0230416  0.0297476  -0.775 0.438741
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4364 on 1231 degrees of freedom
## Multiple R-squared:  0.7711, Adjusted R-squared:  0.7687
## F-statistic: 319.1 on 13 and 1231 DF,  p-value: < 2.2e-16
```

```
m8 <- step(m7) # see which is the best combination
```

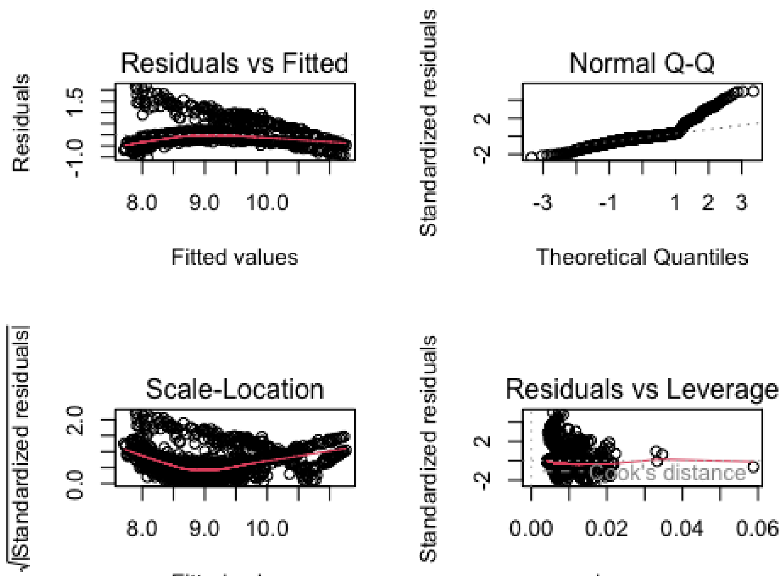
```
## Start:  AIC=-2050.6
## log(charges) ~ bmi + age + children * (f.sex + f.smoker + f.region)
##
##                     Df Sum of Sq    RSS     AIC
## - children:f.region  3     0.418 234.89 -2054.4
## - children:f.sex     1     0.270 234.74 -2051.2
## <none>                           234.47 -2050.6
## - children:f.smoker  1     4.474 238.94 -2029.1
## - bmi                1     6.496 240.97 -2018.6
## - age                1   293.369 527.84 -1042.3
##
## Step:  AIC=-2054.38
## log(charges) ~ bmi + age + children + f.sex + f.smoker + f.region +
##     children:f.sex + children:f.smoker
##
##                     Df Sum of Sq    RSS     AIC
## - children:f.sex     1     0.247 235.14 -2055.1
## <none>                           234.89 -2054.4
## - f.region           3     3.864 238.75 -2040.1
## - children:f.smoker  1     4.514 239.40 -2032.7
## - bmi                1     6.407 241.30 -2022.9
## - age                1   294.200 529.09 -1045.4
##
## Step:  AIC=-2055.08
## log(charges) ~ bmi + age + children + f.sex + f.smoker + f.region +
##     children:f.smoker
##
```

```
##                        Df Sum of Sq    RSS     AIC
## <none>                              235.14 -2055.1
## - f.sex                 1     1.503 236.64 -2049.1
## - f.region              3     3.887 239.02 -2040.7
## - children:f.smoker     1     4.384 239.52 -2034.1
## - bmi                   1     6.435 241.57 -2023.5
## - age                   1   294.017 529.15 -1047.2
```

```
summary(m8)
```

```
##
## Call:
## lm(formula = log(charges) ~ bmi + age + children + f.sex + f.smoker +
##     f.region + children:f.smoker, data = df[-influential_obs,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02519 -0.20790 -0.05672  0.06422  2.17187
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.009093   0.074803  93.700  < 2e-16 ***
## bmi                     0.012465   0.002144   5.814 7.76e-09 ***
## age                     0.035014   0.000891  39.297  < 2e-16 ***
## children                0.125946   0.011601  10.857  < 2e-16 ***
## f.sexmale              -0.069849   0.024859  -2.810 0.005036 **
## f.smokeryes             1.669037   0.043529  38.343  < 2e-16 ***
## f.regionnorthwest      -0.060998   0.035440  -1.721 0.085472 .
## f.regionsoutheast      -0.147420   0.035681  -4.132 3.84e-05 ***
## f.regionsouthwest      -0.123750   0.035745  -3.462 0.000554 ***
## children:f.smokeryes -0.130210   0.027135  -4.799 1.79e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4363 on 1235 degrees of freedom
## Multiple R-squared:  0.7705, Adjusted R-squared:  0.7688
## F-statistic: 460.7 on 9 and 1235 DF,  p-value: < 2.2e-16
```

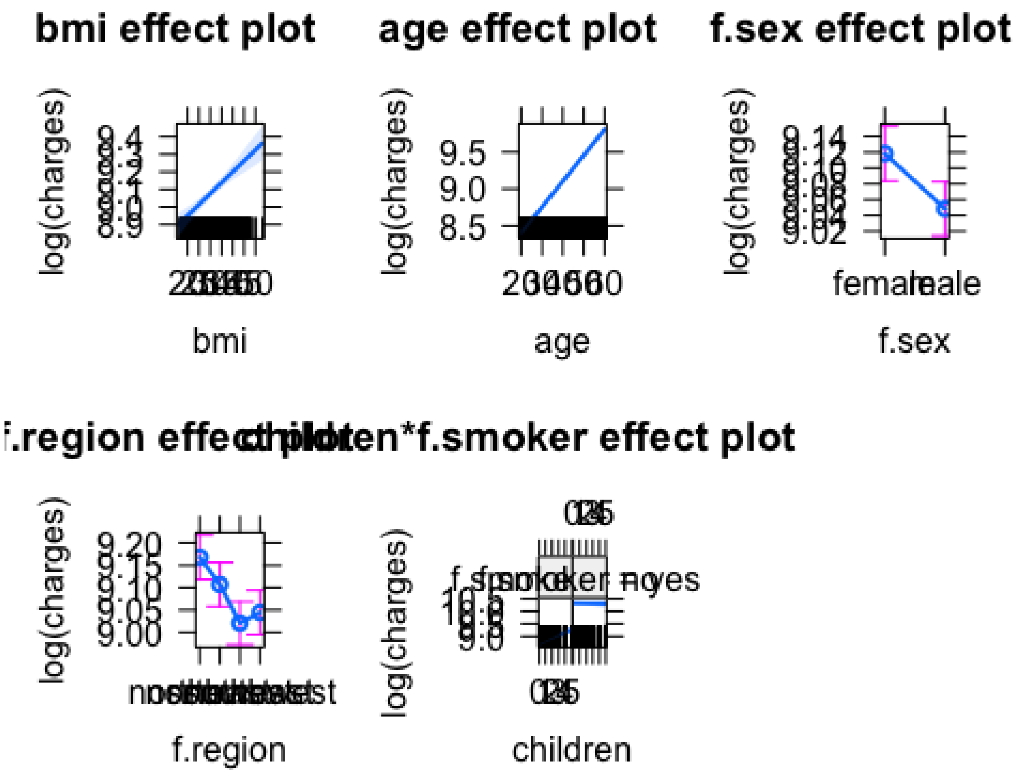
```
par( mfrow = c(2,2))
plot( m8, id.n=0 )
```




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

This will be our final model after several iterations.

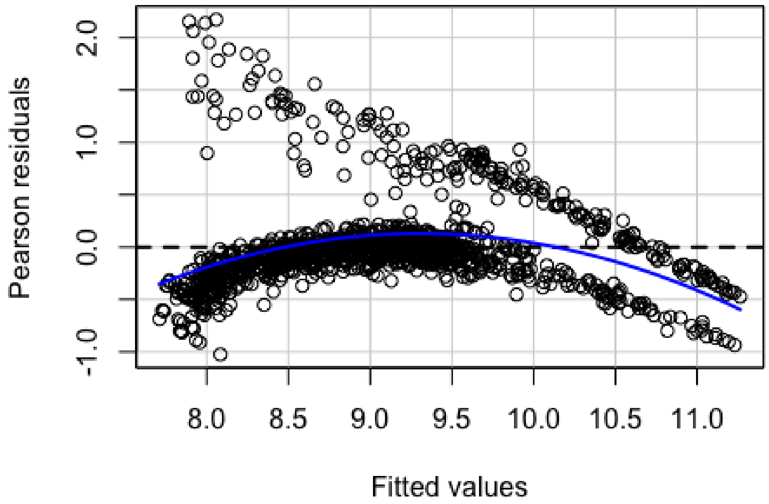
```
library(effects)

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.

plot(allEffects(m8))
```



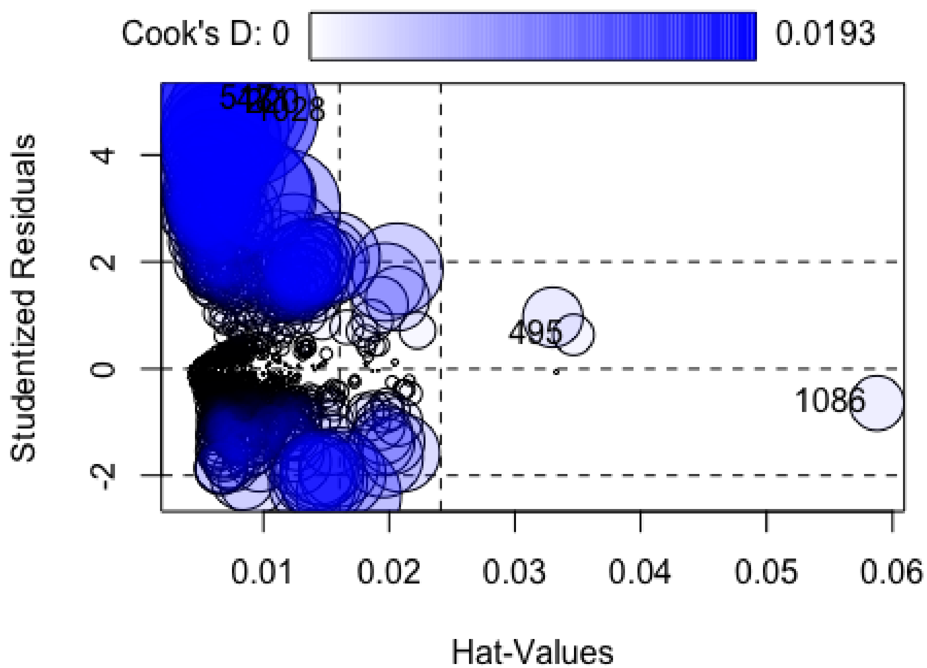
The allEffects plot shows that being a female have an effect of increasing the charges. In addition, we can see that having more children has an effect of increasing the charges on no smokers. On the other hand, smokers seem to have to pay much more regardless to the number of children .

Validation of the model

```
library(car)

residualPlot(m8)
```

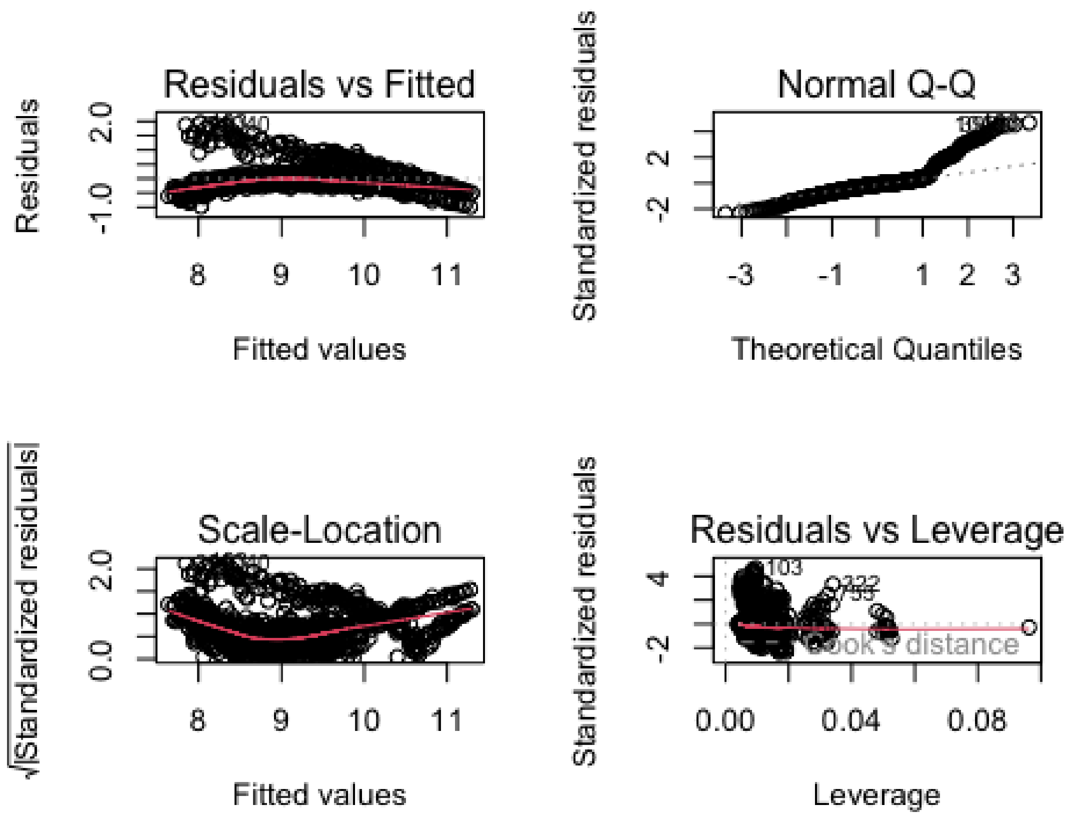



```
influencePlot(m8)
```




```
##           StudRes          Hat       CookD
## 220    4.9672541 0.007782677 0.018989292
## 431    5.0007994 0.007074228 0.017477513
## 495    0.6337794 0.034659237 0.001442863
## 517    5.0408291 0.005715592 0.014323673
## 1028   4.7909844 0.008465434 0.019254782
## 1086  -0.6454875 0.058775305 0.002603050

# there are a lot of influential data
```

```
influential_after_iteractions <- which(rownames(df) %in% c("517","1028",
"220", "431"))

influential_after_iteractions

## [1]  218  428  514 1021

influential_obs <- c(influential_obs, influential_after_iteractions)

m9 <- lm(log(charges)~bmi+age+children+children * (f.sex+f.smoker+f.region),
data=df[-influential_obs,])
summary(m9)

##
## Call:
## lm(formula = log(charges) ~ bmi + age + children + children *
##     (f.sex + f.smoker + f.region), data = df[-influential_obs,
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98467 -0.20570 -0.05223  0.06808  1.94903
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                6.9808247  0.0755475  92.403  < 2e-16 ***
## bmi                        0.0132065  0.0020684   6.385 2.43e-10 ***
## age                        0.0356108  0.0008598  41.416  < 2e-16 ***
## children                   0.1109512  0.0238384   4.654 3.60e-06 ***
## f.sexmale                 -0.1081588  0.0324006  -3.338 0.000869 ***
## f.smokeryes                1.6869411  0.0418988  40.262  < 2e-16 ***
## f.regionnorthwest         -0.0851715  0.0466150  -1.827 0.067924 .
## f.regionsoutheast         -0.1863705  0.0453029  -4.114 4.15e-05 ***
## f.regionsouthwest         -0.1109497  0.0461916  -2.402 0.016456 *
## children:f.sexmale         0.0287765  0.0202966   1.418 0.156503
## children:f.smokeryes      -0.1374206  0.0261779  -5.249 1.80e-07 ***
## children:f.regionnorthwest  0.0159379  0.0293268   0.543 0.586913
## children:f.regionsoutheast  0.0225097  0.0284977   0.790 0.429752
## children:f.regionsouthwest -0.0177469  0.0286148  -0.620 0.535240
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4195 on 1227 degrees of freedom
## Multiple R-squared:  0.7884, Adjusted R-squared:  0.7861
## F-statistic: 351.6 on 13 and 1227 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(m9)
```




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

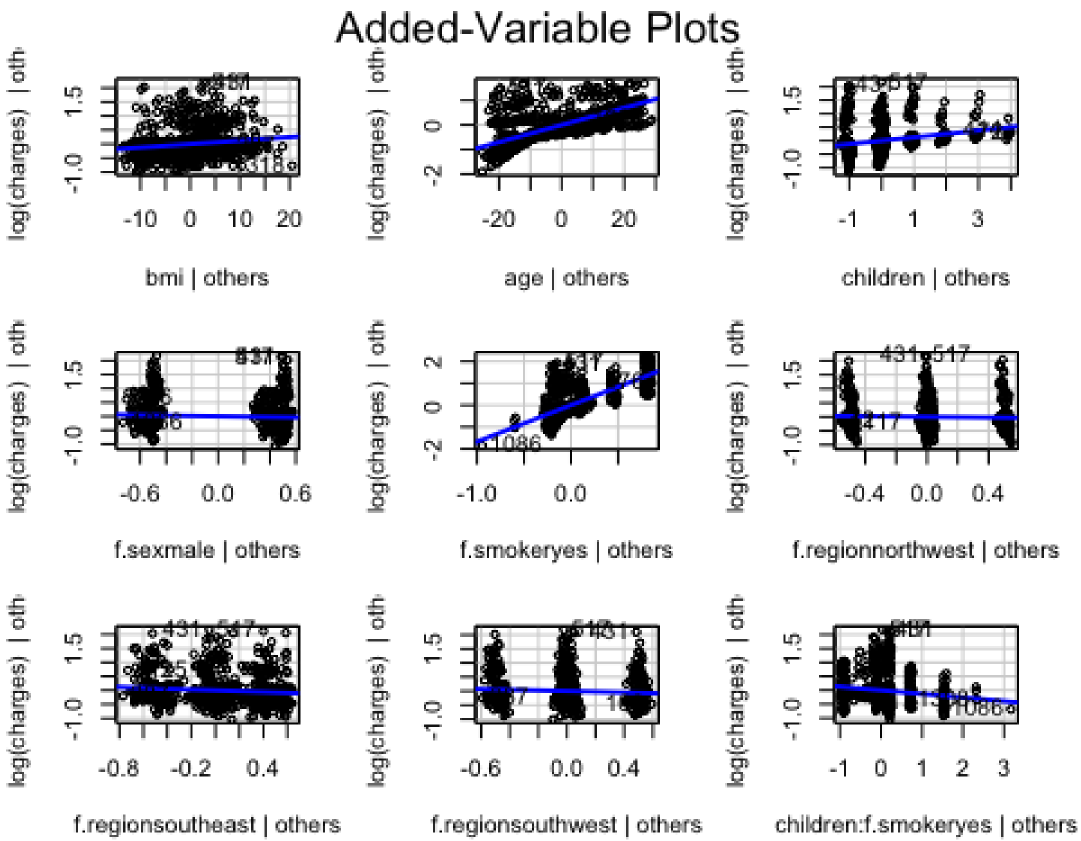
We addressed again influential data after adding interactions and we removed some observations.

The final model created has an adjusted R-squared score of 0.78 which is good. However, studying the residual plots there are patterns that are producing a deviation in the normal Q-Q.

This pattern is mainly introduced by the **age** variable which we tried to reduce the impact transforming it into a factor with an age-range variable and removing it from the numerical explanatory variables. This transformation helped us to have a better normal Q-Q plot but reduced significantly the R-squared score.

Our decision is that we keep the age variable since we consider it an important variable and it has a positive impact on the r-squared score so the model will be better explained.

```
avPlots(m8)
```



We can see in the partial regression plots that people who are older are paying more charges and also people who smoke are significantly paying more. Also having more childrens and having a high bmi is affecting paying more.

We managed to reach a good R-Squared which explains a lot of the variable charges and could help to make an prediction of what a person would be paying.

ANNEX 1: Data cleaning

Data format

```
is.null(df) #no nulls in the data

## [1] FALSE

replace(df,which(df %like% " "), '') #close all blank space

which(df=="") #no blanks found in the data

## integer(0)

#check for distinct values and whether there are differences in them
unique(df$sex) #expecting 2 values

## [1] "female" "male"

unique(df$smoker) #expecting 2 values

## [1] "yes" "no"

unique(df$region) #expecting 4 values

## [1] "southwest" "southeast" "northwest" "northeast"

#we can see that data is consistent for categorical variables
df$f.sex <- factor(df$sex,labels = c("female","male"));
df$f.smoker <- factor(df$smoker,labels = c("no","yes"))
df$f.region <- factor(df$region,labels =
c("northeast","northwest","southeast","southwest"))
summary(df) #from the summary we can see the factor values, it seems that sex
and region are distributed equally and not much smokers compare to the non
smokers.

##       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          f.sex
f.smoker
##  Length:1338        Length:1338        Min.   : 1122   female:662   no
:1064
##  Class :character   Class :character   1st Qu.: 4740   male  :676   yes:
274
```

```
##  Mode  :character   Mode  :character   Median : 9382
##                                         Mean   :13270
##                                         3rd Qu.:16640
##                                         Max.   :63770
##       f.region
##  northeast:324
##  northwest:325
##  southeast:364
##  southwest:325
##
##
```

```
dim(df)
```

```
## [1] 1338   10
```

```
unique(df)

#There is only one observation which repeat twice, it makes sense that a
person with the same properties will have the same charge and since it's only
one we decide to leave it there.
#outliers
```

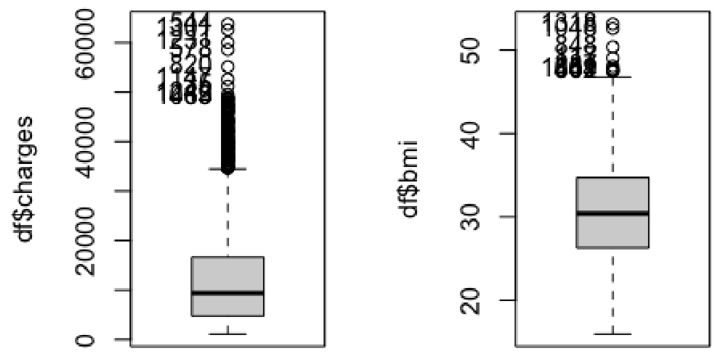
Outlier detection

Univariate

```
par(mfrow=c(1,2))
Boxplot(df$charges)
```

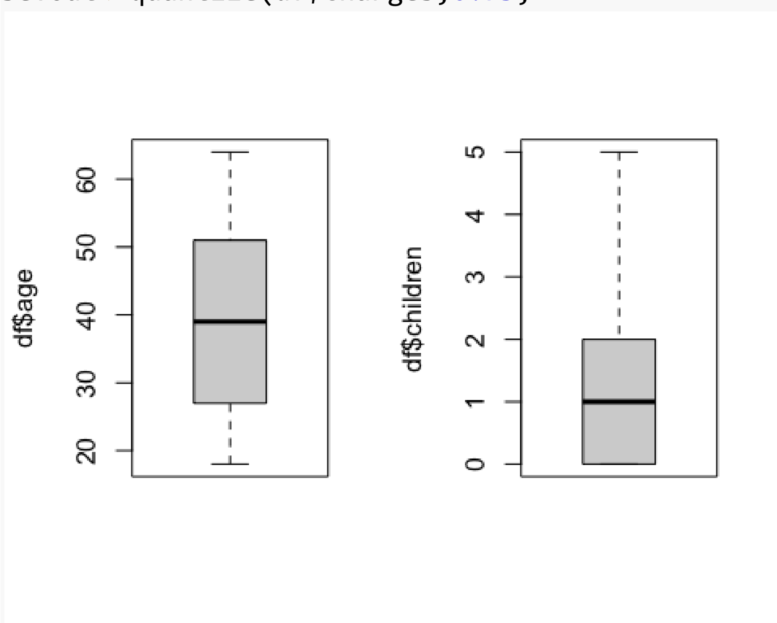
```
##  [1]  544 1301 1231  578  820 1147   35 1242 1063  489
```

```
Boxplot(df$bmi)
```




```
## [1]  117  287  402  544  848  861 1048 1089 1318
```

```
Boxplot(df$age)
Boxplot(df$children)

# treat outliers for charges variable
sevout<-quantile(df$charges,0.75,
```

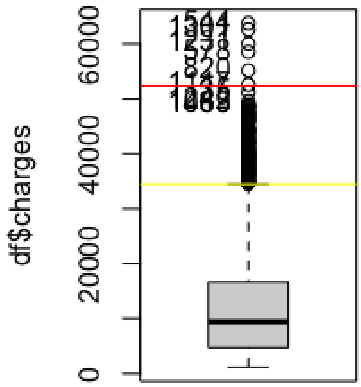



```
na.rm=TRUE)+3*(quantile(df$charges,0.75,na.rm=TRUE)-quantile(df$charges,0.25,
na.rm=TRUE))
sevout
```

```
##      75%
## 52338.79
```

```
sev_out_lower <-
quantile(df$charges,0.25,na.rm=TRUE)-3*(quantile(df$charges,0.75,na.rm=TRUE)-
quantile(df$charges,0.25,na.rm=TRUE))

mist<-quantile(df$charges,0.75,na.rm=TRUE)+1.5*(quantile(df$charges,0.75,na.r
m=TRUE)-quantile(df$charges,0.25,na.rm=TRUE))
mist

##      75%
## 34489.35

mist_out_lower <-
quantile(df$charges,0.25,na.rm=TRUE)-1.5*(quantile(df$charges,0.75,na.rm=TRUE
)-quantile(df$charges,0.25,na.rm=TRUE))

# get list of outliers
loutse<-which(df$charges>sevout);length(loutse)

## [1] 6

loutmist <-which(df$charges>mist);length(loutmist)

## [1] 139

low_out_sever <- which(df$charges<sev_out_lower);low_out_sever

## integer(0)

low_out_mild <- which(df$charges<mist_out_lower);low_out_mild

## integer(0)

# see outliers
Boxplot(df$charges)

##  [1]  544 1301 1231  578  820 1147   35 1242 1063  489

abline(h=sevout,col="red")
abline(h=mist,col="yellow")

# Since there are only 6 severe outliers, we will remove them from the
dataset,
df <- df[-which(df$charges >= sevout),]

# check severe outliers for bmi atrribute
sevout_bmi<-quantile(df$bmi,0.75,na.rm=TRUE)+3*(quantile(df$bmi,0.75,na.rm=TR
UE)-quantile(df$bmi,0.25,na.rm=TRUE));sevout_bmi

##    75%
## 59.815
```

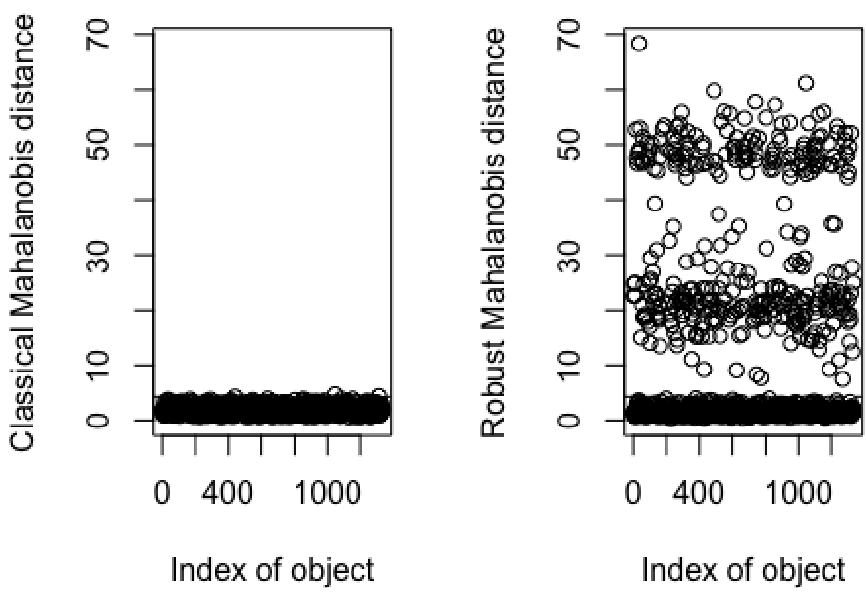
```
mist_bmi <-
quantile(df$bmi,0.75,na.rm=TRUE)+1.5*(quantile(df$bmi,0.75,na.rm=TRUE)-quanti
le(df$bmi,0.25,na.rm=TRUE))
loutse_bmi<-which(df$bmi>sevout_bmi);length(loutse_bmi) # no severe outliers
for bmi

## [1] 0

colSums(is.na(df))

##       age       sex       bmi  children    smoker    region   charges     f.sex
##         0         0         0         0         0         0         0         0
## f.smoker f.region
##         0         0

serout_lower_bmi <-
quantile(df$bmi,0.25,na.rm=TRUE)-3*(quantile(df$bmi,0.75,na.rm=TRUE)-quantile
(df$bmi,0.25,na.rm=TRUE));serout_lower_bmi

##     25%
## 1.02375

mist_lower_bmi <-
quantile(df$bmi,0.25,na.rm=TRUE)-1.5*(quantile(df$bmi,0.75,na.rm=TRUE)-quanti
le(df$bmi,0.25,na.rm=TRUE));mist_lower_bmi

##      25%
## 13.62187

up_sever_bmi <- which(df$bmi > sevout_bmi); up_sever_bmi

## integer(0)

up_mild_bmi <- which(df$bmi > mist_bmi); up_mild_bmi

## [1]  117  287  402  845  858 1045 1086 1312

low_sever_bmi <- which(df$bmi < serout_lower_bmi); low_sever_bmi

## integer(0)

low_mild_bmi <- which(df$bmi < mist_lower_bmi); low_mild_bmi

## integer(0)
```



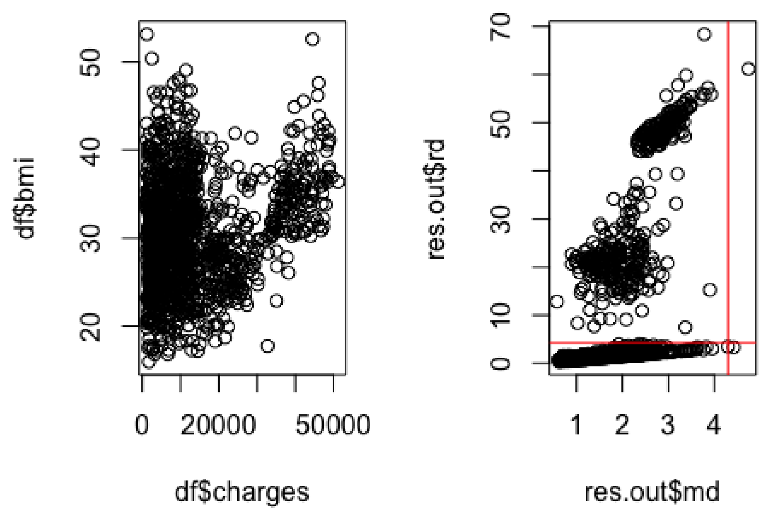
We can see extreme outliers for both charges and bmi, since it's just several observation it might be the case that for a certain bmi, age or smokers the charge value is raising by a lot compare to the rest. from looking at the high value of column charges it can be seen that all are smokers and mid-high bmi, also some of the ages I see are relatively high. For the target variable we can see there is no lower bound for extreme and mild outliers, it's also can be seen on the Boxplot(). For variable bmi, mild outliers on the upper bound and no sever upper bound outliers and not lower bound outliers. We decided to delete the 6 univariate outliers since the charges are very high, even though all 6 observation are smokers, there are 274 smokers in the dataset and their charges values are not as high as the extreme outliers observations

Multivariate

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



```
#str(res.out)
plot(df$charges,df$bmi)
res.out$cutoff

## [1] 4.297305
which((res.out$md > res.out$cutoff) & (res.out$rd > res.out$cutoff))

## 1048
## 1045

plot( res.out$md, res.out$rd )
abline(h=res.out$cutoff, col="red")
abline(v=res.out$cutoff, col="red")
```




```
df <- df[-which(res.out$md > res.out$cutoff & res.out$rd > res.out$cutoff),]
```

For the multivariate outliers, we have chosen the quantile to be a very high value so outliers we get are very extreme compared to our values in the dataset. Observation number 1048 is the multivariate outlier we have got and it's indeed a very high value of charge and bmi. Since this observation is so extreme we will remove it from the dataset. We see from the plot of classical Mahalanobis distance vs robust Mahalanobis distance that there is one observation (1048) that is behind the cutoff value, in addition we can indicate 3 clusters and number of observations that are a bit far from the clusters, it can be suspected as influential data. We also plot charges vs bmi and we can see on the top right corner of the graph there is one observation which has high charge and bmi.

Missing data

```
##       age      sex      bmi  children   smoker   region  charges    f.sex
##         0        0        0         0        0        0        0        0
```

```
## f.smoker f.region
##        0        0
```

There is no missing data in the dataframe so no further imputation is needed

Data Validation

After doing the pre-processing steps where we detected and removed outliers, we will check if data makes sense using common sense and domain knowledge.

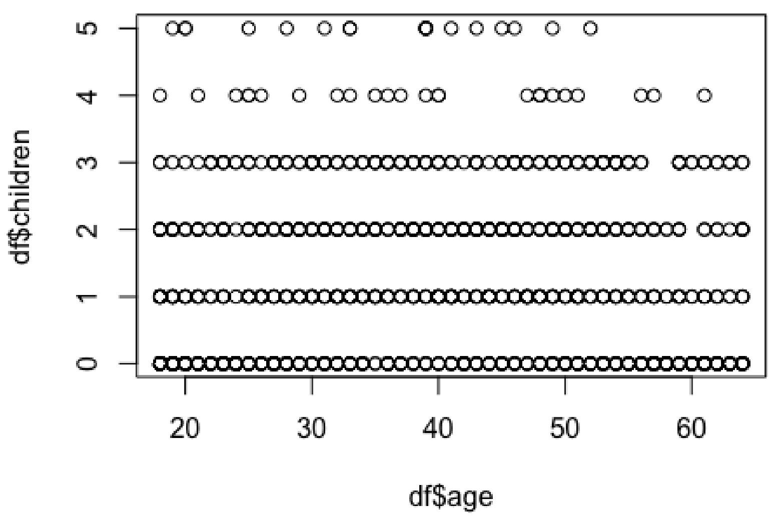
```
summary(df)

##       age            sex                  bmi           children
##  Min.   :18.00   Length:1331        Min.   :15.96   Min.   :0.000
##  1st Qu.:26.50   Class :character   1st Qu.:26.22   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.30   Median :1.000
##  Mean   :39.19                      Mean   :30.62   Mean   :1.097
##  3rd Qu.:51.00                      3rd Qu.:34.60   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker              region             charges          f.sex
f.smoker
##  Length:1331        Length:1331        Min.   : 1122   female:659   no
:1064
##  Class :character   Class :character   1st Qu.: 4720   male  :672   yes:
267
##  Mode  :character   Mode  :character   Median : 9302
##                                        Mean   :13042
##                                        3rd Qu.:16359
##                                        Max.   :51195
##       f.region
##  northeast:323
##  northwest:323
##  southeast:361
##  southwest:324
```

We have ages ranging from 18 to 64, and which bmi ranging from 16 to 53 which are values that are in the following table. The balance between factor variable is really good. However, only 20% of the sample are smokers.

WEIGHT	lbs	90	100	110	120	130	140	150	160	170	180	190	200	210	220	230	240	250	260	270	280	290
	kgs	41	45	50	54	59	64	68	73	77	82	86	91	95	100	104	109	113	118	122	127	132
HEIGHT ft/in	cm		Underweight					Healthy				Overweight				Obese				Extremely Obese		
4'8"	142.2	20	22	25	27	29	31	34	36	38	40	43	45	47	49	52	54	56	58	61	63	65
4'9"	144.7	19	22	24	26	28	30	32	35	37	39	41	43	45	48	50	52	54	56	58	61	63
4'10"	147.3	19	21	23	25	27	29	31	33	36	38	40	42	44	46	48	50	52	54	56	59	61
4'11"	149.8	18	20	22	24	26	28	30	32	34	36	38	40	42	44	46	48	51	53	55	57	59
4'12"	152.4	18	20	21	23	25	27	29	31	33	35	37	39	41	43	45	47	49	51	53	55	57
5'1"	154.9	17	19	21	23	25	26	28	30	32	34	36	38	40	42	43	45	47	49	51	53	55
5'2"	157.4	16	18	20	22	24	26	27	29	31	33	35	37	38	40	42	44	46	48	49	51	53
5'3"	160.0	16	18	19	21	23	25	27	28	30	32	34	35	37	39	41	43	44	46	48	50	51
5'4"	162.5	15	17	19	21	22	24	26	27	29	31	33	34	36	38	39	41	43	45	46	48	50
5'5"	165.1	15	17	18	20	22	23	25	27	28	30	32	33	35	37	38	40	42	43	45	47	48
5'6"	167.6	15	16	18	19	21	23	24	26	27	29	31	32	34	36	37	39	40	42	44	45	47
5'7"	170.1	14	16	17	19	20	22	24	25	27	28	30	31	33	34	36	38	39	41	42	44	45
5'8"	172.7	14	15	17	18	20	21	23	24	26	27	29	30	32	33	35	37	38	40	41	43	44
5'9"	175.2	13	15	16	18	19	21	22	24	25	27	28	30	31	33	34	35	37	38	40	41	43
5'10"	177.8	13	14	16	17	19	20	22	23	24	26	27	29	30	32	33	34	36	37	39	40	42
5'11"	180.3	13	14	15	17	18	20	21	22	24	25	27	28	29	31	32	33	35	36	38	39	40
6'0"	182.8	12	14	15	16	18	19	20	22	23	24	26	27	28	30	31	33	34	35	37	38	39
6'1"	185.4	12	13	15	16	17	18	20	21	22	24	25	26	28	29	30	32	33	34	36	37	38
6'2"	187.9	12	13	14	15	17	18	19	21	22	23	24	26	27	28	30	31	32	33	35	36	37
6'3"	190.5	11	13	14	15	16	18	19	20	21	23	24	25	26	28	29	30	31	33	34	35	36
6'4"	193.0	11	12	13	15	16	17	18	19	21	22	23	24	26	27	28	29	30	32	33	34	35
6'5"	195.5	11	12	13	14	15	17	18	19	20	21	23	24	25	26	27	28	30	31	32	33	34
6'6"	198.1	10	12	13	14	15	16	17	18	20	21	22	23	24	25	27	28	29	30	31	32	34
6'7"	200.6	10	11	12	14	15	16	17	18	19	20	21	23	24	25	26	27	28	29	30	32	33
6'8"	203.2	10	11	12	13	14	15	16	18	19	20	21	22	23	24	25	26	27	29	30	31	32
6'9"	205.7	10	11	12	13	14	15	16	17	18	19	20	21	23	24	25	26	27	28	29	30	31
6'10"	208.2	9	10	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
6'11"	210.8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	27	28	29	30

Let's see how the relationship between children per age.

```
plot(df$children~df$age)
```



As we can see in the plot, there are individuals with age 20 that have from 3 to 5 children which is really strange.

```
thr2five_children <- which(df$age <= 20 & df$children>2)

thr2five_children
```

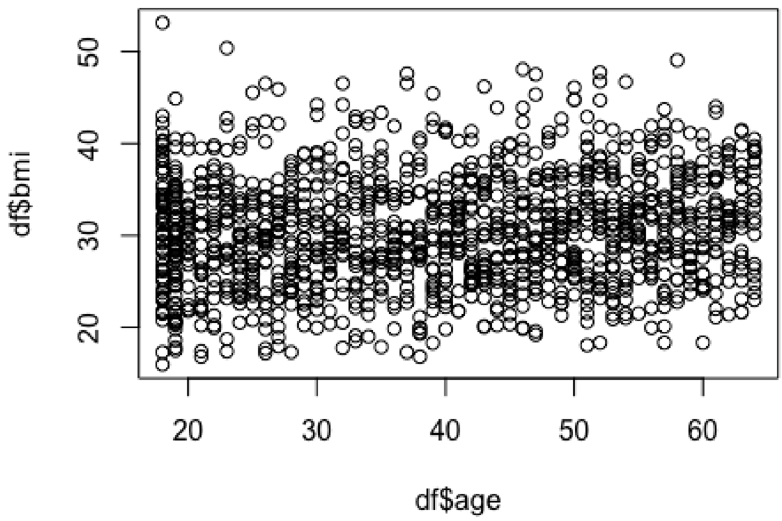
```
## [1]   33  167  370  982 1092 1182 1191 1200
```

These observations will be removed since it's something very unlikely.

```
df <- df[-thr2five_children,]
```

Let's check now the bmi values per age to see if there is any weird case:

```
plot(df$bmi~df$age)
```



In this case the plot shows there are young people who have a really high bmi. Since data is from EEUU, and there are a lot of obesity problems, we decide that these observations are not going to be removed.