|  |
| --- |
| title: “SIM - Assignment 1” |
| author: “Míriam Méndez, Gabriel Zarate” |
| date: “” |
| output: |
| word\_document: default |
| pdf\_document: default |
| subtitle: Medical cost |
| editor\_options: |
| chunk\_output\_type: console |

# 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.

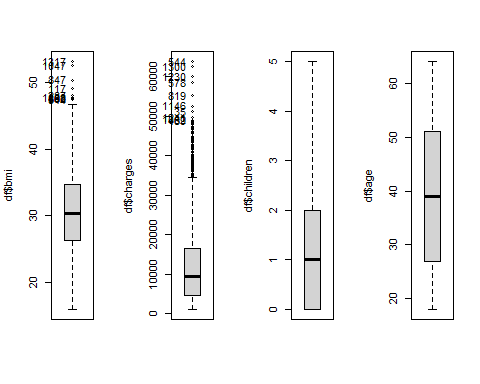
par(mfrow=c(1,4))  
Boxplot(df$bmi) #bmi seem to have outliers

## [1] 117 287 402 544 847 860 1047 1088 1317

Boxplot(df$charges) #charges seem to have outliers

## [1] 544 1300 1230 578 819 1146 35 1241 1062 489

Boxplot(df$children) #Children no outliers  
Boxplot(df$age) #Age no outliers



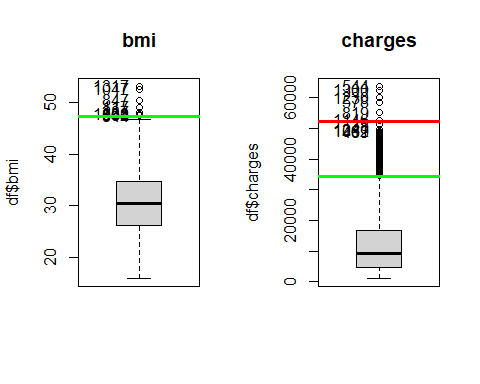
par(mfrow=c(1,2))  
#Checking outliers for bmi  
ss<-summary(df$bmi);  
# Upper/lower severe threshold  
utso2<-ss[5]+3\*(ss[5]-ss[2]);  
utsi2<-ss[2]-3\*(ss[5]-ss[2]);  
# Upper/lower mild threshold  
utmo2<-ss[5]+1.5\*(ss[5]-ss[2]);   
utmi2<-ss[2]-1.5\*(ss[5]-ss[2]);   
Boxplot(df$bmi, main="bmi")

## [1] 117 287 402 544 847 860 1047 1088 1317

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" )

## [1] 544 1300 1230 578 819 1146 35 1241 1062 489

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));md

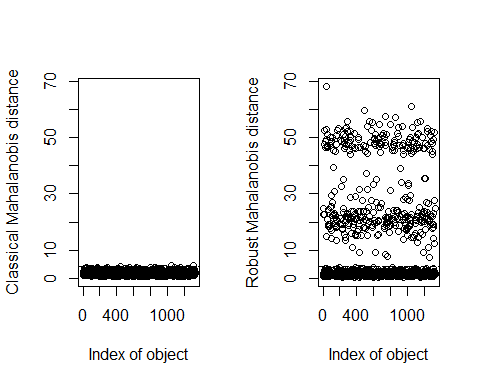
## [1] 544 578 819 1146 1230 1300

df <- df[-md,]

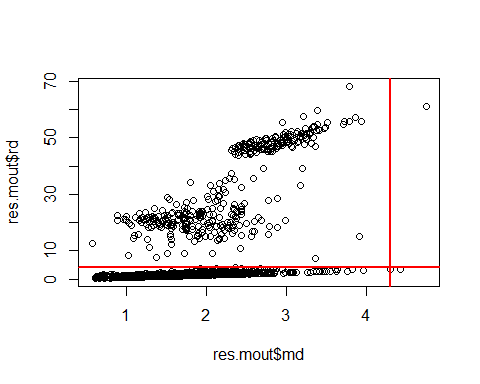
## Multivariate Outliers

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.

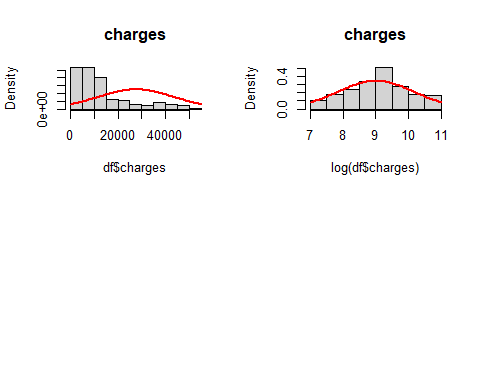
Additionally, the normality of the logarithmic transformation of charge was tested, getting the same results by checking the histogram, comparing it with the curve, and by applying the Shapiro Test, rejecting the H0, so it does not has a log-normal distribution.

par(mfrow=c(2,2))  
  
#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="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))

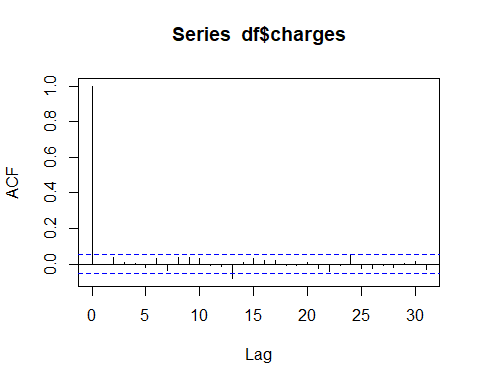
##   
## Shapiro-Wilk normality test  
##   
## data: log(df$charges)  
## W = 0.98185, p-value = 7.028e-12



## 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.

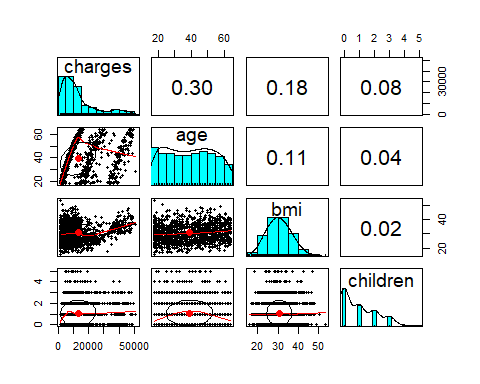
Secondly, in these scatter plots we found interesting relationships between:

* *age* and *charges* : displays several relatively straight lines.
* *bmi* and *charges*: has two distinct groups of points.

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

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



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

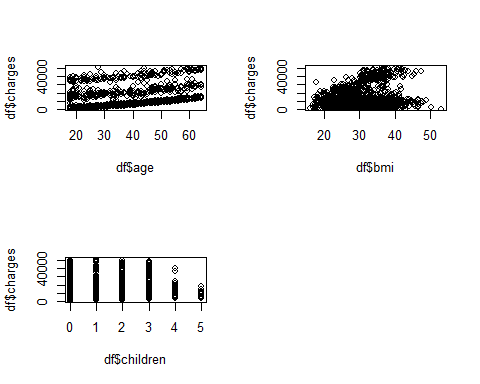
res.con$category

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

Checking the relation of charges with the other numerical variables by using graphics:

* With 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.
* With 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
* With 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

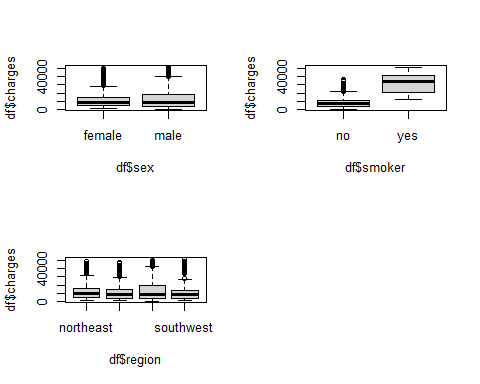
par( mfrow = c(2,2))  
plot(df$charges ~ df$age)  
plot(df$charges ~ df$bmi)  
plot(df$charges ~ df$children)



Checking the relation of charges with the other categorical variables by using graphics:

* With 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.
* With 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.
* With region eems to not be any significative differences between them

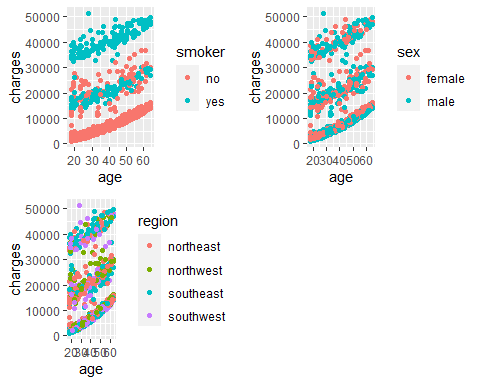
par( mfrow = c(2,2))  
plot(df$charges ~ df$sex)  
tapply(df$charges, df$sex, summary)  
plot(df$charges ~ df$smoker)  
tapply(df$charges, df$smoker, summary)  
plot(df$charges ~ df$region)  
tapply(df$charges, df$region, summary)



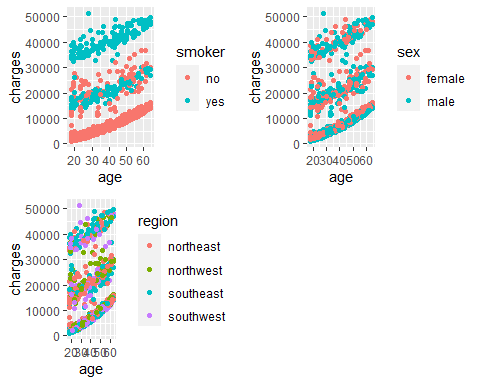
Checking the relation of charges with the interaction of other variables:

* 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

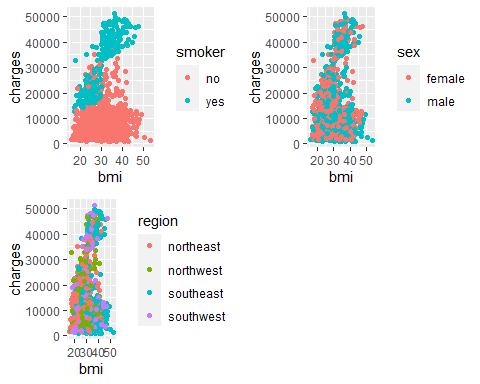
plot1 <- ggplot(df, aes(x=age , y = charges, color = smoker)) + geom\_point()  
plot2 <- ggplot(df, aes(x=age , y = charges, color = sex)) + geom\_point()  
plot3 <- ggplot(df, aes(x=age , y = charges, color = region)) + geom\_point()  
grid.arrange( plot1, plot2,plot3 , ncol=2)



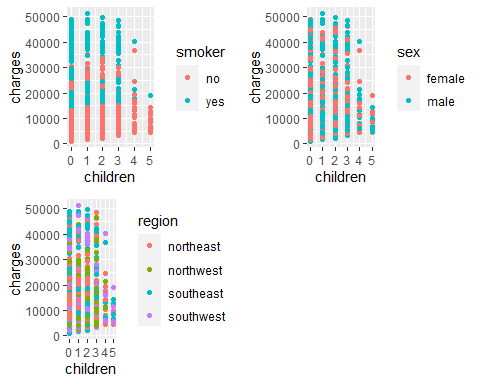
plot1 <-ggplot(df, aes(x=age , y = charges, color = smoker)) + geom\_point()  
plot2 <-ggplot(df, aes(x=age , y = charges, color = sex)) + geom\_point()  
plot3 <-ggplot(df, aes(x=age , y = charges, color = region)) + geom\_point()  
grid.arrange( plot1, plot2,plot3 , ncol=2)



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

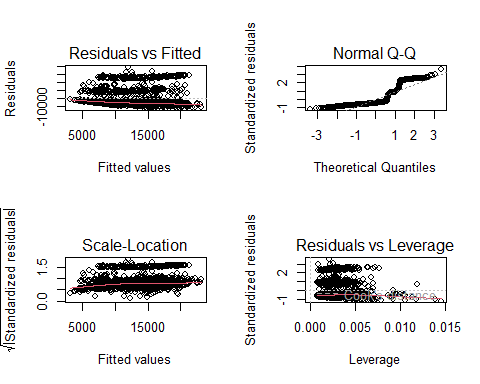
## Model with only numerical variables

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.

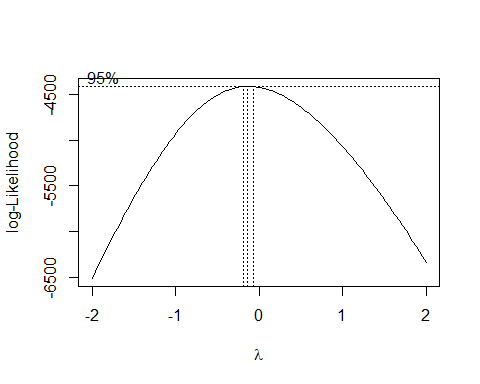
m0 <- lm(charges ~ age + bmi + children , data = df)  
summary(m0)

##   
## 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



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

# 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

##   
## Call:  
## lm(formula = log(charges) ~ age + I(age^2) + bmi + I(children +   
## 0.5), data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.3607 -0.4542 -0.3123 0.5249 2.2493   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.0064349 0.2081155 33.666 < 2e-16 \*\*\*  
## age 0.0520133 0.0102215 5.089 4.13e-07 \*\*\*  
## I(age^2) -0.0002320 0.0001275 -1.819 0.06907 .   
## bmi 0.0094726 0.0034585 2.739 0.00625 \*\*   
## I(children + 0.5) 0.0981836 0.0181102 5.421 7.02e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7591 on 1325 degrees of freedom  
## Multiple R-squared: 0.3081, Adjusted R-squared: 0.306   
## F-statistic: 147.5 on 4 and 1325 DF, p-value: < 2.2e-16

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

##   
## Call:  
## lm(formula = log(charges) ~ I(age^2) + bmi + I(children + 0.5),   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.5204 -0.4345 -0.2725 0.4549 2.1097   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.8985063 0.1132052 69.772 < 2e-16 \*\*\*  
## I(age^2) 0.0004099 0.0000188 21.805 < 2e-16 \*\*\*  
## bmi 0.0091332 0.0034902 2.617 0.00898 \*\*   
## I(children + 0.5) 0.1259642 0.0174290 7.227 8.29e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7662 on 1326 degrees of freedom  
## Multiple R-squared: 0.2946, Adjusted R-squared: 0.293   
## F-statistic: 184.6 on 3 and 1326 DF, p-value: < 2.2e-16

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

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.

#we use 2 because it is a small dataset  
llev <- which( hatvalues(m1) > 2\*(length(coef(m1))/nrow(df)));llev

## 27 33 37 63 72 95 104 117 167 171 173 200 251 287 345 370   
## 27 33 37 63 72 95 104 117 167 171 173 200 251 287 345 370   
## 379 399 414 419 426 439 443 481 495 535 550 553 569 587 604 636   
## 379 399 414 419 426 439 443 481 495 535 549 552 568 584 601 633   
## 641 660 661 665 753 769 771 802 804 826 848 861 878 891 896 909   
## 638 657 658 662 750 766 768 799 801 822 844 857 874 887 892 905   
## 933 938 970 985 1013 1052 1069 1075 1080 1086 1096 1117 1131 1157 1205 1237   
## 929 934 966 981 1009 1047 1064 1070 1075 1081 1091 1112 1126 1151 1199 1230   
## 1242 1246 1266 1273 1302 1318   
## 1235 1239 1259 1266 1294 1310

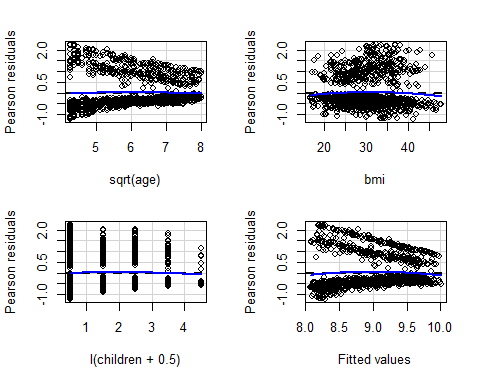
length(llev)

## [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

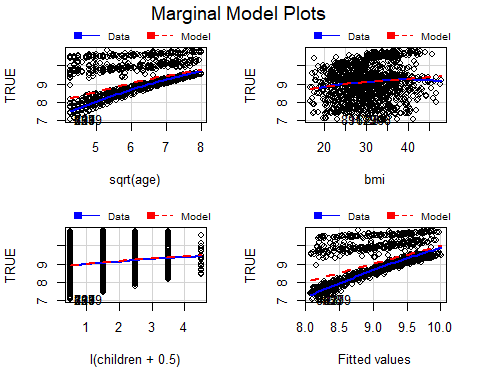
##   
## Call:  
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5),   
## data = df[-llev, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2454 -0.4633 -0.3208 0.5370 2.2694   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.043210 0.156227 38.682 < 2e-16 \*\*\*  
## sqrt(age) 0.413428 0.019207 21.524 < 2e-16 \*\*\*  
## bmi 0.010595 0.003711 2.855 0.00437 \*\*   
## I(children + 0.5) 0.114024 0.019556 5.831 7.01e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.762 on 1256 degrees of freedom  
## Multiple R-squared: 0.3051, Adjusted R-squared: 0.3034   
## F-statistic: 183.8 on 3 and 1256 DF, p-value: < 2.2e-16

residualPlots(m3)



## Test stat Pr(>|Test stat|)   
## sqrt(age) -0.1561 0.87597   
## bmi -1.4209 0.15560   
## I(children + 0.5) -0.8708 0.38401   
## Tukey test -1.8726 0.06113 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

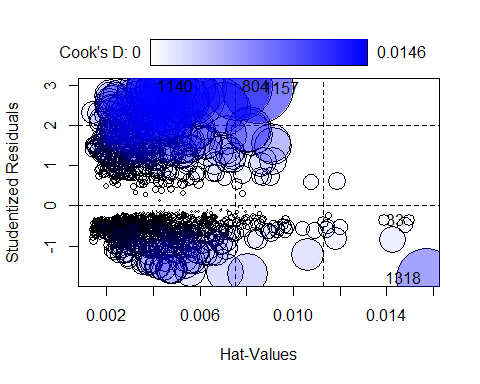
marginalModelPlots(m3, id=list(n=5, labels=rownames(df)))



### 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

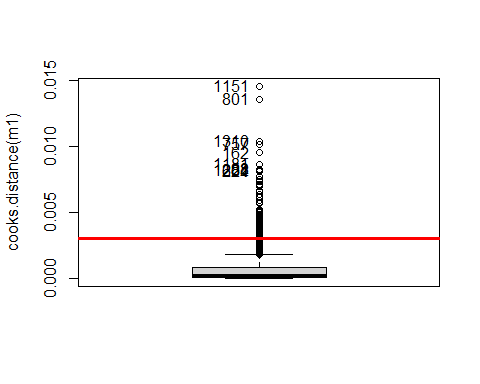
# Threshold Chatterjee-Hadi  
thChH <- 4/ (nrow(df) - length(coef(m1)));thChH

## [1] 0.003018868

#Actual influent data Cook´s distance: outliers in cook´s distance  
Boxplot(cooks.distance(m1))

## [1] 1151 801 1310 757 162 1181 58 1282 621 264

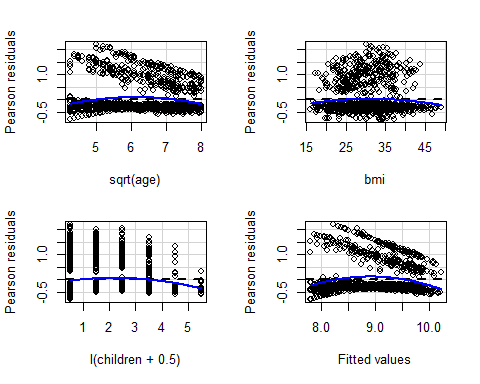
abline(h=thChH,col="red",lwd=3)



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

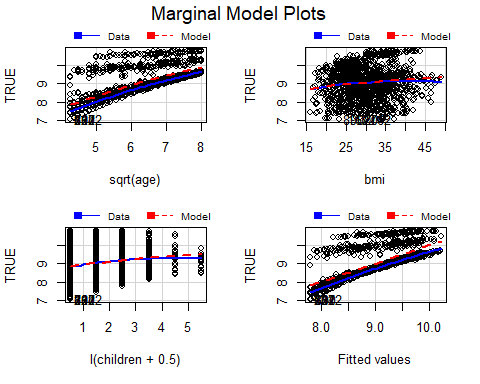
##   
## Call:  
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5),   
## data = df[-resout, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.7879 -0.3448 -0.2632 -0.1008 2.2137   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.4626087 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.

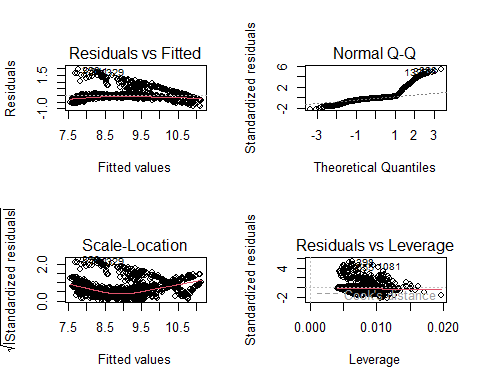
names(df)

## [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)



par(mfrow=c(1,1))  
  
Anova(m6)

## 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

#all variables are important  
  
m7 <- step( m6, k=log(nrow(df)))

## Start: AIC=-2455.04  
## log(charges) ~ sqrt(age) + bmi + I(children + 0.5) + sex + smoker +   
## region  
##   
## Df Sum of Sq RSS AIC  
## <none> 167.32 -2455.0  
## - sex 1 1.14 168.46 -2453.7  
## - region 3 3.92 171.24 -2447.6  
## - bmi 1 2.50 169.82 -2443.7  
## - I(children + 0.5) 1 19.39 186.70 -2325.0  
## - smoker 1 300.58 467.90 -1174.7  
## - sqrt(age) 1 366.40 533.72 -1009.9

#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)

##   
## Call:  
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5) +   
## sex + smoker, data = df[-resout, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.80331 -0.14354 -0.06038 0.01831 1.87382   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.404926 0.076055 71.066 < 2e-16 \*\*\*  
## sqrt(age) 0.494545 0.009531 51.889 < 2e-16 \*\*\*  
## bmi 0.005833 0.001794 3.251 0.00118 \*\*   
## I(children + 0.5) 0.102915 0.008711 11.814 < 2e-16 \*\*\*  
## sexmale -0.061075 0.021054 -2.901 0.00379 \*\*   
## smokeryes 1.341022 0.028657 46.796 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3707 on 1246 degrees of freedom  
## Multiple R-squared: 0.8235, Adjusted R-squared: 0.8227   
## F-statistic: 1162 on 5 and 1246 DF, p-value: < 2.2e-16

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

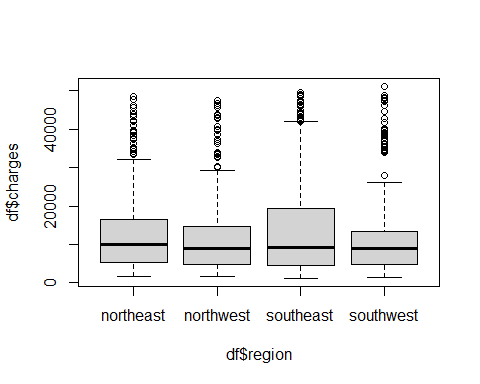
#Ho rejected, so models are not equivalent, so we can not delete region

### 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  
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

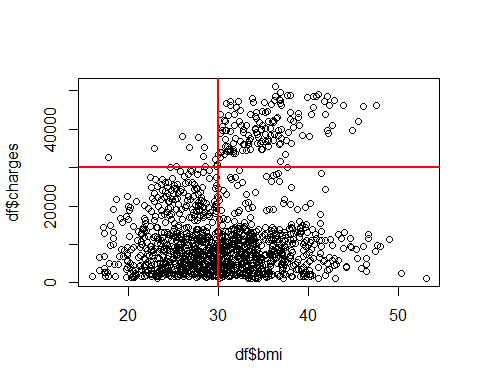
df$f.reg<-0  
ll<-which(df$region %in% c("northeast","northwest"))  
df$f.reg[ll]<-1  
df$f.reg <- factor( df$f.reg, labels=c("south","north"))  
  
#bmi  
summary(df$bmi)

## 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

##   
## Call:  
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5) +   
## sex + smoker + f.reg, data = df[-resout, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.75574 -0.14606 -0.06148 0.00882 1.91852   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.308899 0.078335 67.772 < 2e-16 \*\*\*  
## sqrt(age) 0.492988 0.009462 52.102 < 2e-16 \*\*\*  
## bmi 0.007717 0.001827 4.223 2.58e-05 \*\*\*  
## I(children + 0.5) 0.102986 0.008643 11.916 < 2e-16 \*\*\*  
## sexmale -0.061095 0.020888 -2.925 0.00351 \*\*   
## smokeryes 1.345018 0.028445 47.285 < 2e-16 \*\*\*  
## f.regnorth 0.097449 0.021350 4.564 5.51e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3678 on 1245 degrees of freedom  
## Multiple R-squared: 0.8264, Adjusted R-squared: 0.8255   
## F-statistic: 987.5 on 6 and 1245 DF, p-value: < 2.2e-16

#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

##   
## Call:  
## lm(formula = log(charges) ~ sqrt(age) + f.bmi + I(children +   
## 0.5) + sex + smoker + f.reg, data = df[-resout, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.76707 -0.13819 -0.05895 0.00715 1.88181   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.492335 0.062088 88.460 < 2e-16 \*\*\*  
## sqrt(age) 0.494828 0.009415 52.558 < 2e-16 \*\*\*  
## f.bmi>=30 0.085303 0.021260 4.012 6.37e-05 \*\*\*  
## I(children + 0.5) 0.103085 0.008648 11.920 < 2e-16 \*\*\*  
## sexmale -0.060959 0.020903 -2.916 0.00361 \*\*   
## smokeryes 1.343368 0.028442 47.231 < 2e-16 \*\*\*  
## f.regnorth 0.089717 0.021049 4.262 2.18e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3681 on 1245 degrees of freedom  
## Multiple R-squared: 0.8261, Adjusted R-squared: 0.8253   
## F-statistic: 985.9 on 6 and 1245 DF, p-value: < 2.2e-16

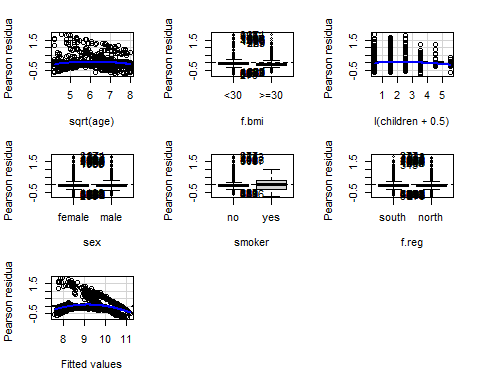
AIC(m9,m8,m7, k=log(nrow(df)))

## df AIC  
## m9 8 1100.742  
## m8 8 1099.018  
## m7 10 1105.175

#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)))

## Start: AIC=-2459.47  
## log(charges) ~ sqrt(age) + f.bmi + I(children + 0.5) + sex +   
## smoker + f.reg  
##   
## Df Sum of Sq RSS AIC  
## <none> 168.65 -2459.5  
## - sex 1 1.15 169.81 -2458.1  
## - f.bmi 1 2.18 170.83 -2450.6  
## - f.reg 1 2.46 171.11 -2448.5  
## - I(children + 0.5) 1 19.25 187.90 -2331.4  
## - smoker 1 302.19 470.85 -1181.2  
## - sqrt(age) 1 374.20 542.86 -1003.1

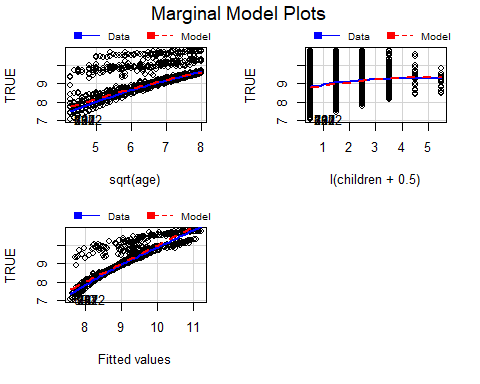
#No changes  
  
residualPlots( m9)



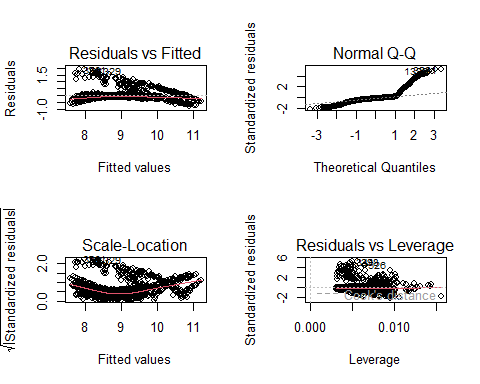
## Test stat Pr(>|Test stat|)   
## sqrt(age) -5.1668 2.772e-07 \*\*\*  
## f.bmi   
## I(children + 0.5) -2.6534 0.008069 \*\*   
## sex   
## smoker   
## f.reg   
## Tukey test -13.5130 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

marginalModelPlots(m9, id=list(n=5, labels=rownames(df)))

## Warning in mmps(...): Interactions and/or factors skipped



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



par(mfrow=c(1,1))

## 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

##   
## Call:  
## lm(formula = log(charges) ~ (sqrt(age) + bmi + I(children + 0.5)) \*   
## (sex + smoker + f.reg), data = df[-resout, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.39569 -0.12829 -0.06861 -0.01996 2.12606   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.3694117 0.1193109 45.004 < 2e-16 \*\*\*  
## sqrt(age) 0.5394042 0.0146670 36.777 < 2e-16 \*\*\*  
## bmi -0.0031224 0.0028222 -1.106 0.268779   
## I(children + 0.5) 0.0976739 0.0134076 7.285 5.72e-13 \*\*\*  
## sexmale -0.5151069 0.1335636 -3.857 0.000121 \*\*\*  
## smokeryes 2.5740695 0.1953059 13.180 < 2e-16 \*\*\*  
## f.regnorth 0.2534481 0.1346085 1.883 0.059955 .   
## sqrt(age):sexmale 0.0587936 0.0167682 3.506 0.000471 \*\*\*  
## sqrt(age):smokeryes -0.4075396 0.0266233 -15.308 < 2e-16 \*\*\*  
## sqrt(age):f.regnorth -0.0662422 0.0167077 -3.965 7.77e-05 \*\*\*  
## bmi:sexmale 0.0009588 0.0031900 0.301 0.763805   
## bmi:smokeryes 0.0529949 0.0045269 11.707 < 2e-16 \*\*\*  
## bmi:f.regnorth 0.0083426 0.0032746 2.548 0.010966 \*   
## I(children + 0.5):sexmale 0.0259561 0.0154166 1.684 0.092503 .   
## I(children + 0.5):smokeryes -0.1018900 0.0218051 -4.673 3.30e-06 \*\*\*  
## I(children + 0.5):f.regnorth 0.0085054 0.0153568 0.554 0.579781   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3263 on 1236 degrees of freedom  
## Multiple R-squared: 0.8643, Adjusted R-squared: 0.8627   
## F-statistic: 524.9 on 15 and 1236 DF, p-value: < 2.2e-16

m12 <- step(m11, k=log(nrow(df)))

## Start: AIC=-2705.23  
## log(charges) ~ (sqrt(age) + bmi + I(children + 0.5)) \* (sex +   
## smoker + f.reg)  
##   
## Df Sum of Sq RSS AIC  
## - bmi:sex 1 0.0096 131.62 -2712.3  
## - I(children + 0.5):f.reg 1 0.0327 131.64 -2712.1  
## - I(children + 0.5):sex 1 0.3018 131.91 -2709.6  
## - bmi:f.reg 1 0.6911 132.30 -2705.9  
## <none> 131.61 -2705.2  
## - sqrt(age):sex 1 1.3091 132.92 -2700.0  
## - sqrt(age):f.reg 1 1.6738 133.28 -2696.6  
## - I(children + 0.5):smoker 1 2.3250 133.94 -2690.5  
## - bmi:smoker 1 14.5928 146.20 -2580.8  
## - sqrt(age):smoker 1 24.9511 156.56 -2495.1  
##   
## Step: AIC=-2712.33  
## 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 + bmi:f.reg + I(children + 0.5):sex + I(children +   
## 0.5):smoker + I(children + 0.5):f.reg  
##   
## Df Sum of Sq RSS AIC  
## - I(children + 0.5):f.reg 1 0.0326 131.65 -2719.2  
## - I(children + 0.5):sex 1 0.3037 131.92 -2716.6  
## - bmi:f.reg 1 0.6816 132.30 -2713.1  
## <none> 131.62 -2712.3  
## - sqrt(age):sex 1 1.3775 133.00 -2706.5  
## - sqrt(age):f.reg 1 1.6675 133.29 -2703.8  
## - I(children + 0.5):smoker 1 2.3167 133.94 -2697.7  
## - bmi:smoker 1 14.6754 146.30 -2587.2  
## - sqrt(age):smoker 1 25.0413 156.66 -2501.5  
##   
## Step: AIC=-2719.21  
## 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 + bmi:f.reg + I(children + 0.5):sex + I(children +   
## 0.5):smoker  
##   
## Df Sum of Sq RSS AIC  
## - I(children + 0.5):sex 1 0.3058 131.96 -2723.5  
## - bmi:f.reg 1 0.6879 132.34 -2719.9  
## <none> 131.65 -2719.2  
## - sqrt(age):sex 1 1.3770 133.03 -2713.4  
## - sqrt(age):f.reg 1 1.6506 133.30 -2710.8  
## - I(children + 0.5):smoker 1 2.3125 133.97 -2704.6  
## - bmi:smoker 1 14.6735 146.33 -2594.1  
## - sqrt(age):smoker 1 25.0182 156.67 -2508.6  
##   
## Step: AIC=-2723.5  
## 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 + bmi:f.reg + I(children + 0.5):smoker  
##   
## Df Sum of Sq RSS AIC  
## - bmi:f.reg 1 0.6879 132.65 -2724.2  
## <none> 131.96 -2723.5  
## - sqrt(age):sex 1 1.4425 133.40 -2717.1  
## - sqrt(age):f.reg 1 1.6470 133.61 -2715.2  
## - I(children + 0.5):smoker 1 2.2131 134.17 -2709.9  
## - bmi:smoker 1 14.9438 146.90 -2596.4  
## - sqrt(age):smoker 1 25.0185 156.98 -2513.3  
##   
## Step: AIC=-2724.18  
## 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  
##   
## Df Sum of Sq RSS AIC  
## <none> 132.65 -2724.2  
## - sqrt(age):f.reg 1 1.3677 134.01 -2718.5  
## - sqrt(age):sex 1 1.4788 134.13 -2717.5  
## - I(children + 0.5):smoker 1 2.1786 134.82 -2711.0  
## - bmi:smoker 1 14.8574 147.50 -2598.5  
## - sqrt(age):smoker 1 24.8089 157.46 -2516.7

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

#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, ])  
  
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)

##   
## Call:  
## 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,   
## ])  
##   
## 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 \*\*\*  
## 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 \*\*\*  
## smokerno:bmi 0.000993 0.001751 0.567 0.570724   
## smokeryes:bmi 0.054307 0.004211 12.897 < 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

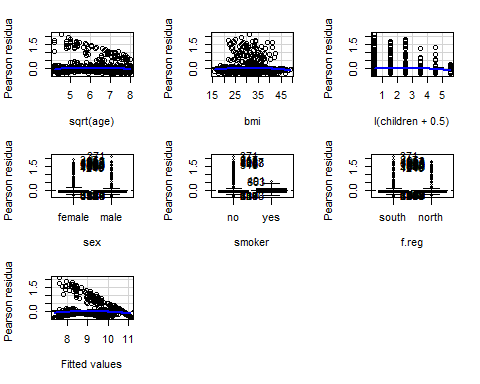
AIC(m13,m12,m11, k=log(nrow(df)))

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

### Residual Analysis and Validation

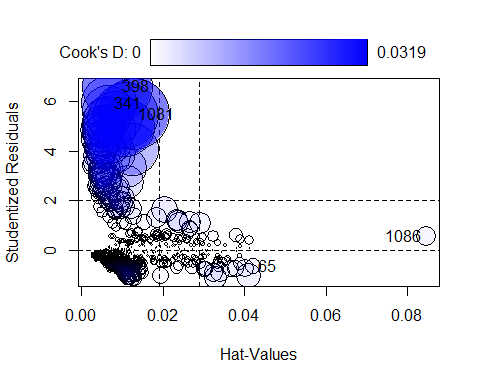
To validate the final model the influencePlot was checked, and there were two values with hatvalues higher than the threshold, so a model was generated excluding them, and getting no real improvement

residualPlots(m12)



## Test stat Pr(>|Test stat|)   
## sqrt(age) -2.9557 0.003178 \*\*   
## bmi -2.2296 0.025953 \*   
## I(children + 0.5) -2.5203 0.011849 \*   
## sex   
## smoker   
## f.reg   
## Tukey test -6.9523 3.595e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

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

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

## [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)))

## Start: AIC=-2718.44  
## 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  
##   
## Df Sum of Sq RSS AIC  
## <none> 132.57 -2718.4  
## - sqrt(age):f.reg 1 1.3867 133.95 -2712.6  
## - sqrt(age):sex 1 1.5024 134.07 -2711.6  
## - I(children + 0.5):smoker 1 2.2217 134.79 -2704.9  
## - bmi:smoker 1 14.6585 147.23 -2594.5  
## - sqrt(age):smoker 1 24.6892 157.26 -2512.2

summary(m14)

##   
## 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[-c(resout, hh), ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.36652 -0.12770 -0.06783 -0.02037 2.11664   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.248171 0.104654 50.148 < 2e-16 \*\*\*  
## sqrt(age) 0.533689 0.014569 36.631 < 2e-16 \*\*\*  
## bmi 0.001005 0.001752 0.574 0.566212   
## I(children + 0.5) 0.114670 0.008337 13.754 < 2e-16 \*\*\*  
## sexmale -0.467799 0.105406 -4.438 9.88e-06 \*\*\*  
## smokeryes 2.565796 0.198507 12.925 < 2e-16 \*\*\*  
## f.regnorth 0.478848 0.105058 4.558 5.68e-06 \*\*\*  
## sqrt(age):sexmale 0.062223 0.016612 3.746 0.000188 \*\*\*  
## sqrt(age):smokeryes -0.408522 0.026904 -15.184 < 2e-16 \*\*\*  
## sqrt(age):f.regnorth -0.059557 0.016550 -3.599 0.000333 \*\*\*  
## bmi:smokeryes 0.053560 0.004578 11.700 < 2e-16 \*\*\*  
## I(children + 0.5):smokeryes -0.102188 0.022434 -4.555 5.76e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3272 on 1238 degrees of freedom  
## Multiple R-squared: 0.8632, Adjusted R-squared: 0.862   
## F-statistic: 710 on 11 and 1238 DF, p-value: < 2.2e-16

## Final results

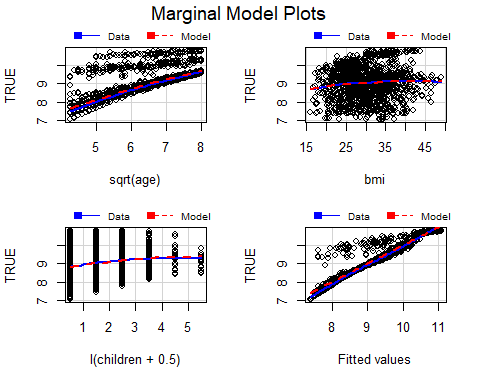
The final model was the one obtained after excecuting the step after adding the interactions and the final dataset was trimmed, deleting the influential data based on its cook’s distance.

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)

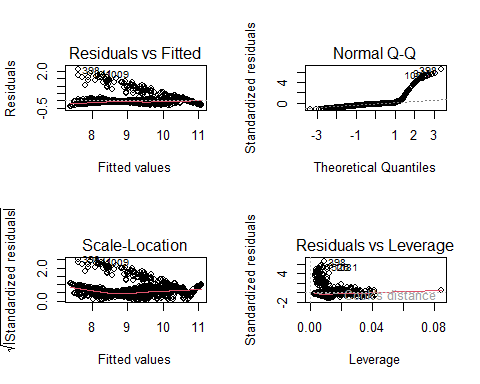
##   
## Call:  
## lm(formula = log(charges) ~ sqrt(age) + bmi + I(children + 0.5) +   
## sex + smoker + f.reg + sqrt(age):sex + sqrt(age):smoker +   
## sqrt(age):f.reg + bmi:smoker + I(children + 0.5):smoker,   
## data = df[-resout, ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.36812 -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

marginalModelPlots(model)

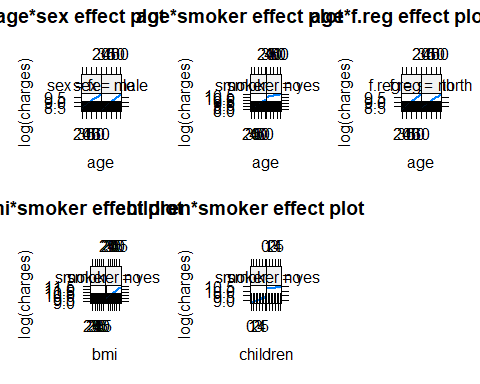
## Warning in mmps(...): Interactions and/or factors skipped



par(mfrow=c(2,2))  
plot(model)



par(mfrow=c(1,1))  
plot(allEffects(model))



T