STAT 632 Final Project

Raksha Ramaraj and Yogesh Gupta

2022-04-21

```r
insurance= read.csv("insurance (1).csv")
head(insurance)
```

```
##   age sex    bmi children smoker    region   charges
## 1  19   0 27.900        0      1 southwest 16884.924
## 2  18   1 33.770        1      0 southeast  1725.552
## 3  28   1 33.000        3      0 southeast  4449.462
## 4  33   1 22.705        0      0 northwest 21984.471
## 5  32   1 28.880        0      0 northwest  3866.855
## 6  31   0 25.740        0      0 southeast  3756.622
```

```r
summary(insurance)
```

```
##       age             sex              bmi           children    
##  Min.   :18.00   Min.   :0.0000   Min.   :15.96   Min.   :0.000  
##  1st Qu.:27.00   1st Qu.:0.0000   1st Qu.:26.30   1st Qu.:0.000  
##  Median :39.00   Median :1.0000   Median :30.40   Median :1.000  
##  Mean   :39.21   Mean   :0.5052   Mean   :30.66   Mean   :1.095  
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:34.69   3rd Qu.:2.000  
##  Max.   :64.00   Max.   :1.0000   Max.   :53.13   Max.   :5.000  
##      smoker          region             charges     
##  Min.   :0.0000   Length:1338        Min.   : 1122  
##  1st Qu.:0.0000   Class :character   1st Qu.: 4740  
##  Median :0.0000   Mode  :character   Median : 9382  
##  Mean   :0.2048                      Mean   :13270  
##  3rd Qu.:0.0000                      3rd Qu.:16640  
##  Max.   :1.0000                      Max.   :63770  
```

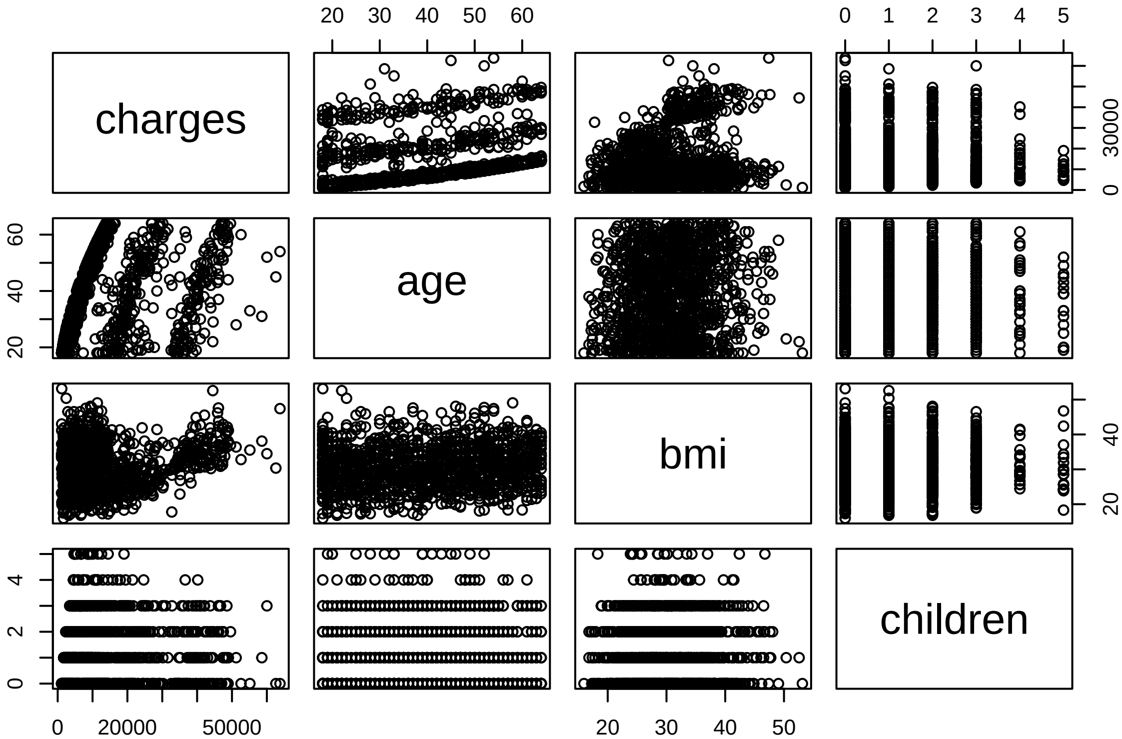
```r
dim(insurance)
```

```
## [1] 1338    7
```

```r
names(insurance)
```

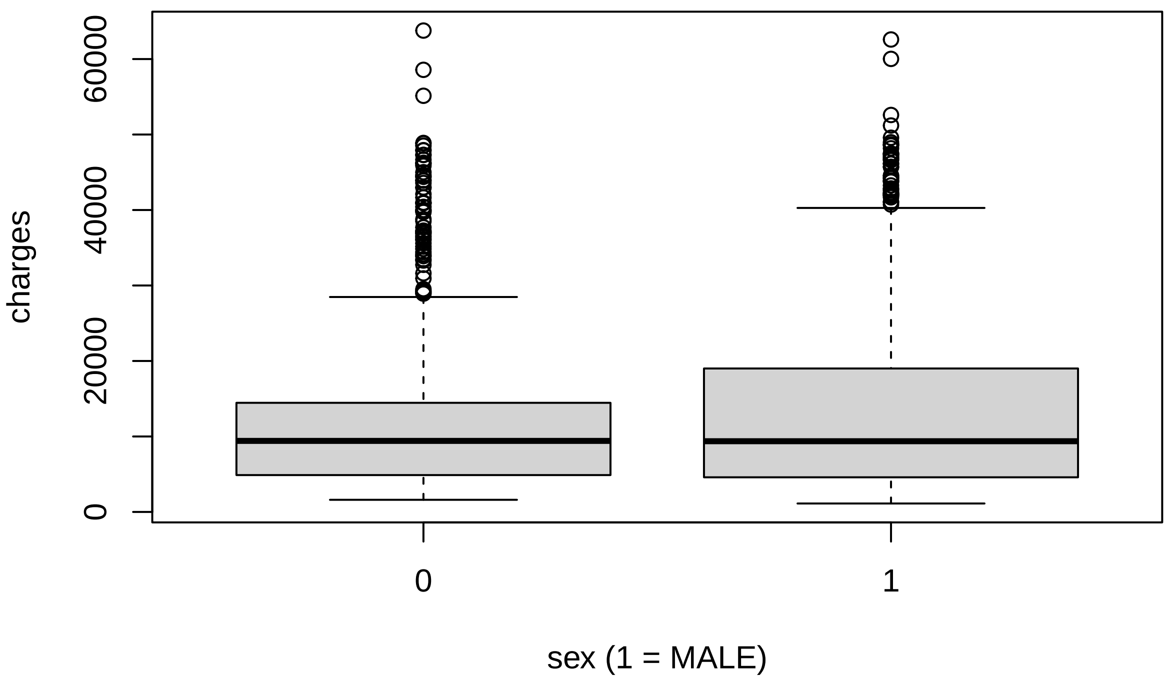
```
## [1] "age"      "sex"      "bmi"      "children" "smoker"   "region"   "charges"
```

```r
pairs(charges~age+bmi+ children,data= insurance)
```



Box plot for gender.

```
boxplot(charges ~ sex, data= insurance,
ylab="charges", xlab="sex (1 = MALE)")
```

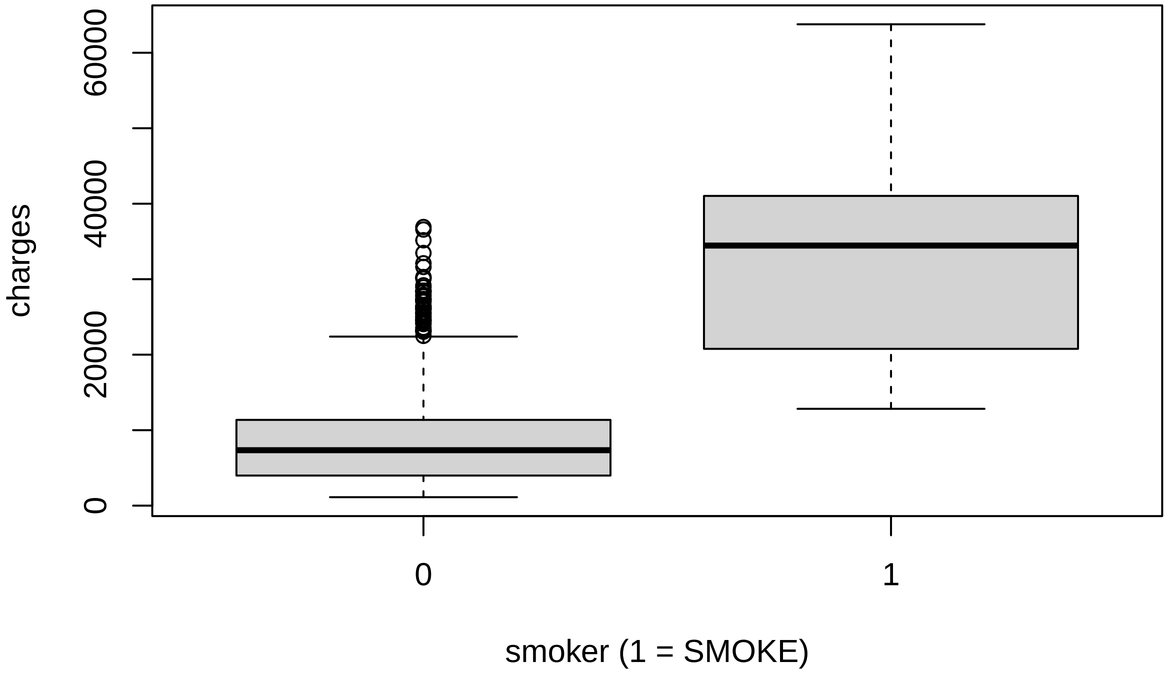


Inference

Both male and female seems to be having almost the same insurance charges. For females the threshold for the third quartile is nearly 16K dollars and for males it is nearly 20k dollars. From the above plot it seems like the gender might not have lot of impact on the response variable.

Box plot for smoker

```
boxplot(charges ~ smoker, data= insurance,
ylab="charges", xlab="smoker (1 = SMOKE)")
```

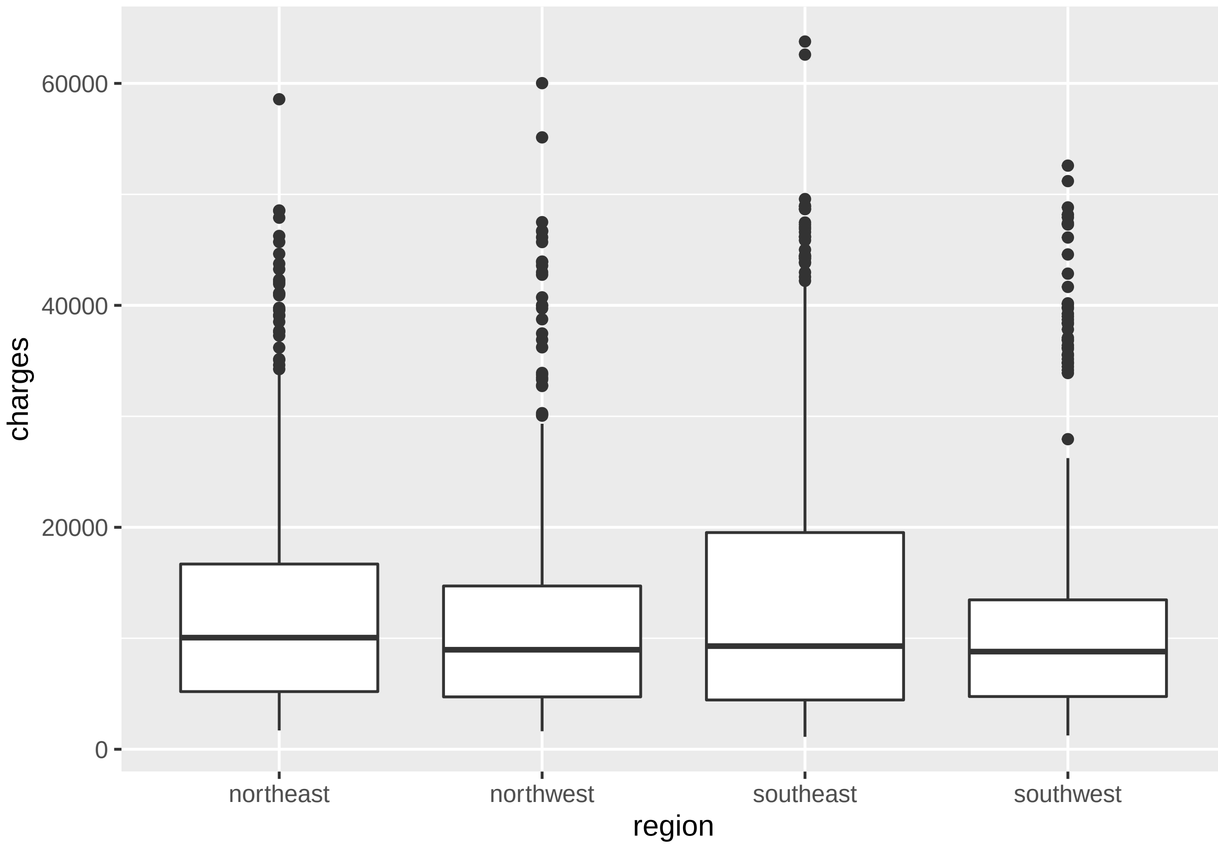


Inference:

From the plot we can see that the smokers have more charges compared to the non smokers. Hence this variable might be of importance.

Box plot for various regions.

```
library(ggplot2)
ggplot(insurance, aes(y= charges, x= region)) +
geom_boxplot()
```



Inference:

The third quartile for the southeast seems to be a little higher compared to the rest of the regions.But overall there might not be lot of impact on the model due to this variable.

Fitting a multi linear regression model with all the variables

Initially we can fit a multi linear regression with all the variables. since variable region is a categorical variable with four categories, North east, North west, South east, South west. We will have to factor it before including in the model.

```
model1= lm(charges~ age + sex + bmi + children + smoker + factor(region) , data= insurance)
summary(model1)
```

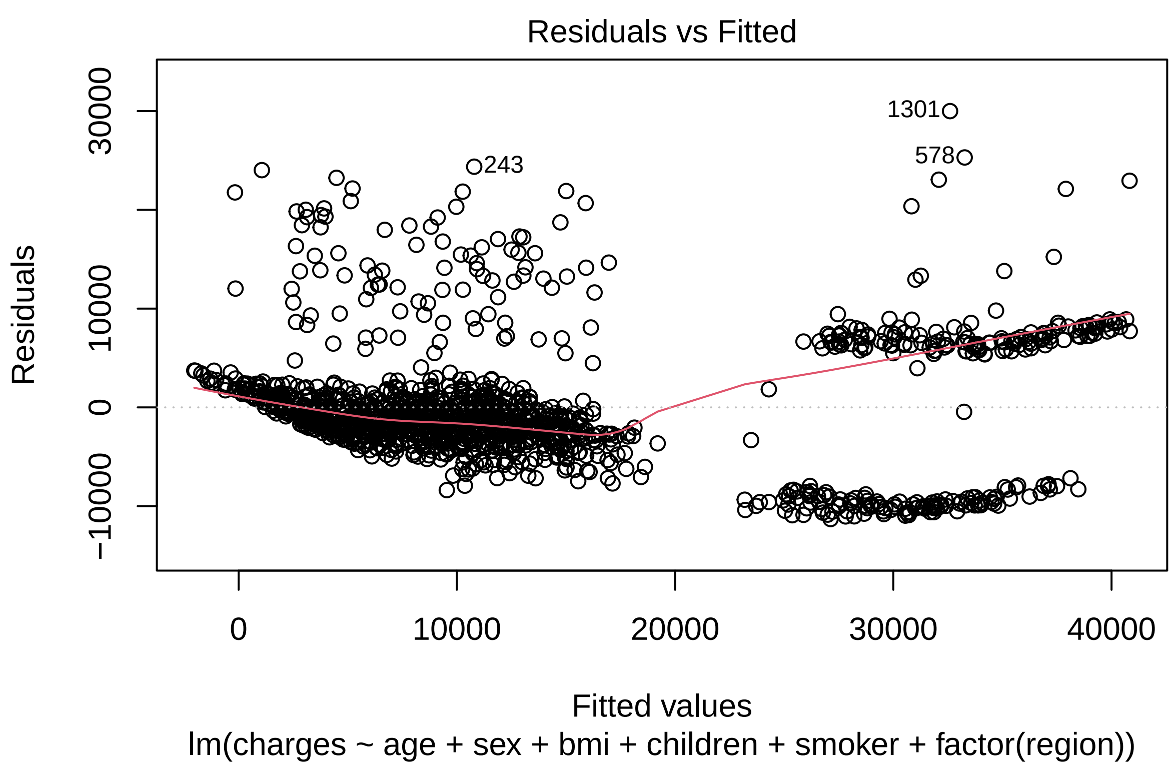
```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     factor(region), data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -11938.5      987.8 -12.086  < 2e-16 ***
## age                        256.9       11.9  21.587  < 2e-16 ***
```

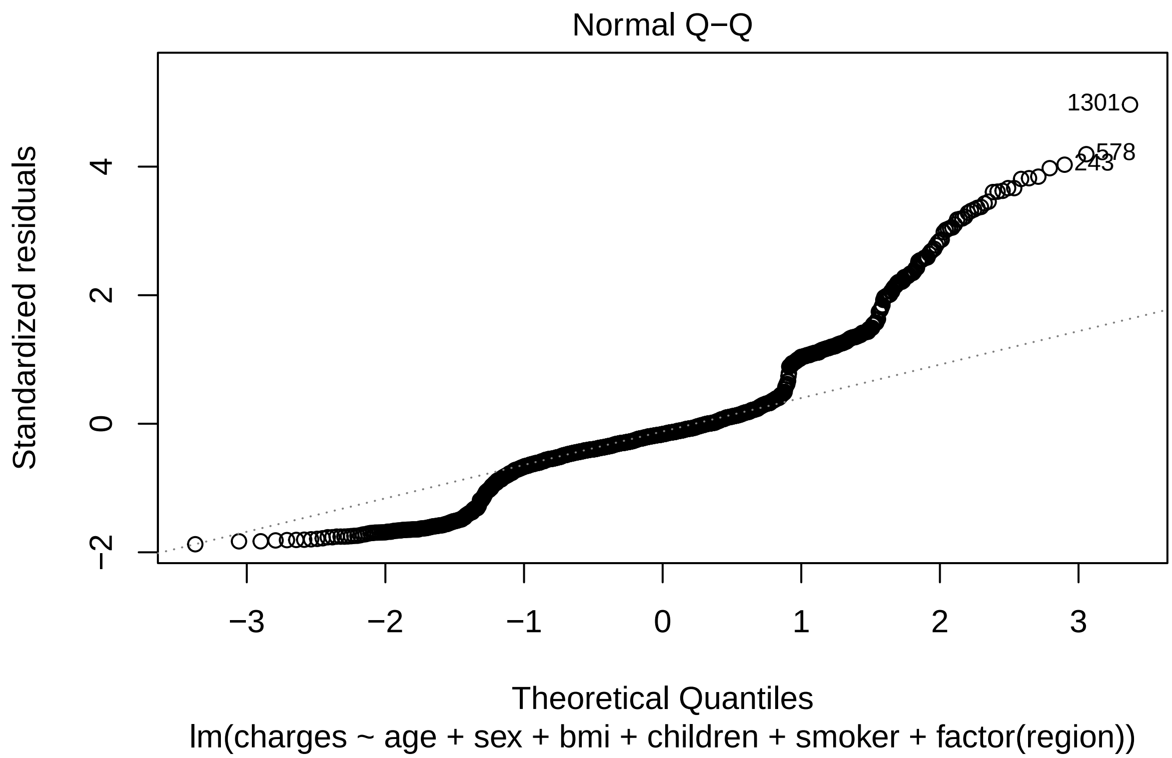
```
## sex                          -131.3      332.9  -0.394 0.693348    
## bmi                           339.2       28.6  11.860  < 2e-16 ***
## children                      475.5      137.8   3.451 0.000577 ***
## smoker                      23848.5      413.1  57.723  < 2e-16 ***
## factor(region)northwest      -353.0      476.3  -0.741 0.458769    
## factor(region)southeast     -1035.0      478.7  -2.162 0.030782 *  
## factor(region)southwest      -960.0      477.9  -2.009 0.044765 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494 
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Inference:

From the above model we can see that most of the co efficiency are significant but for the variable sex. Using $\alpha = 0.05$.

```
plot(model1, 1:2)
```



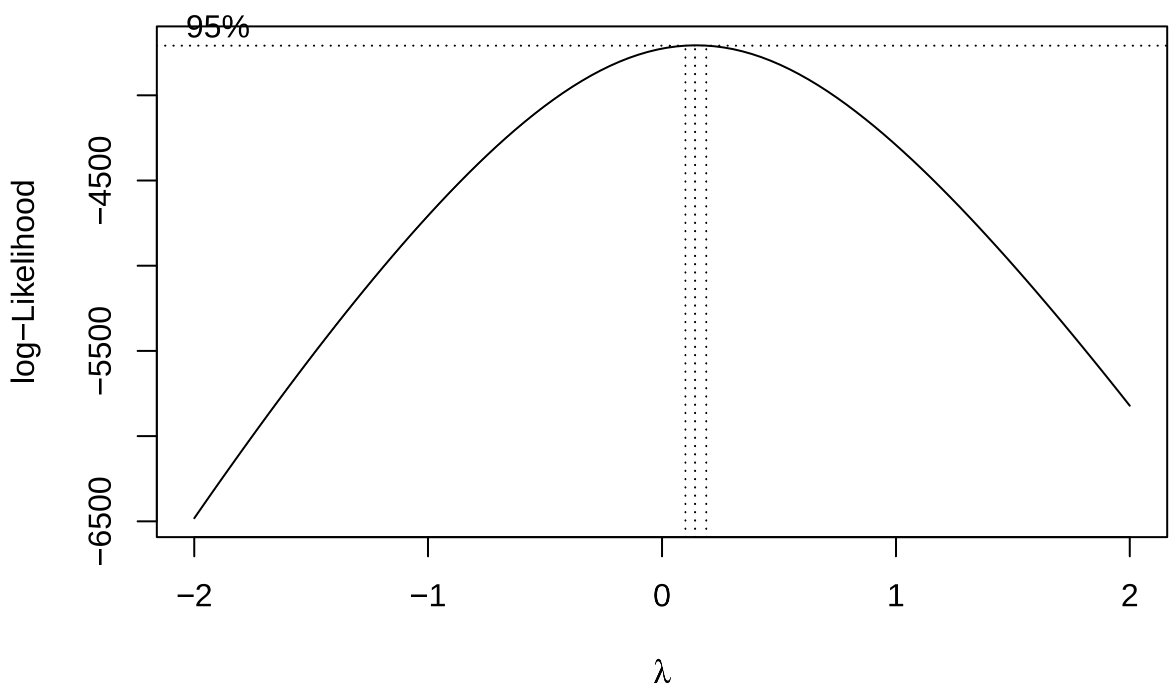


Box Cox Transformation

```
library(MASS)
library(car)
```

```
## Loading required package: carData
```

```
boxcox(model1)
```



```r
summary(powerTransform(model1))
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.1462        0.15       0.1002       0.1921
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                              LRT df       pval
## LR test, lambda = (0) 38.75118  1 4.8142e-10
##
## Likelihood ratio test that no transformation is needed
##                              LRT df       pval
## LR test, lambda = (1) 1169.213  1 < 2.22e-16
```

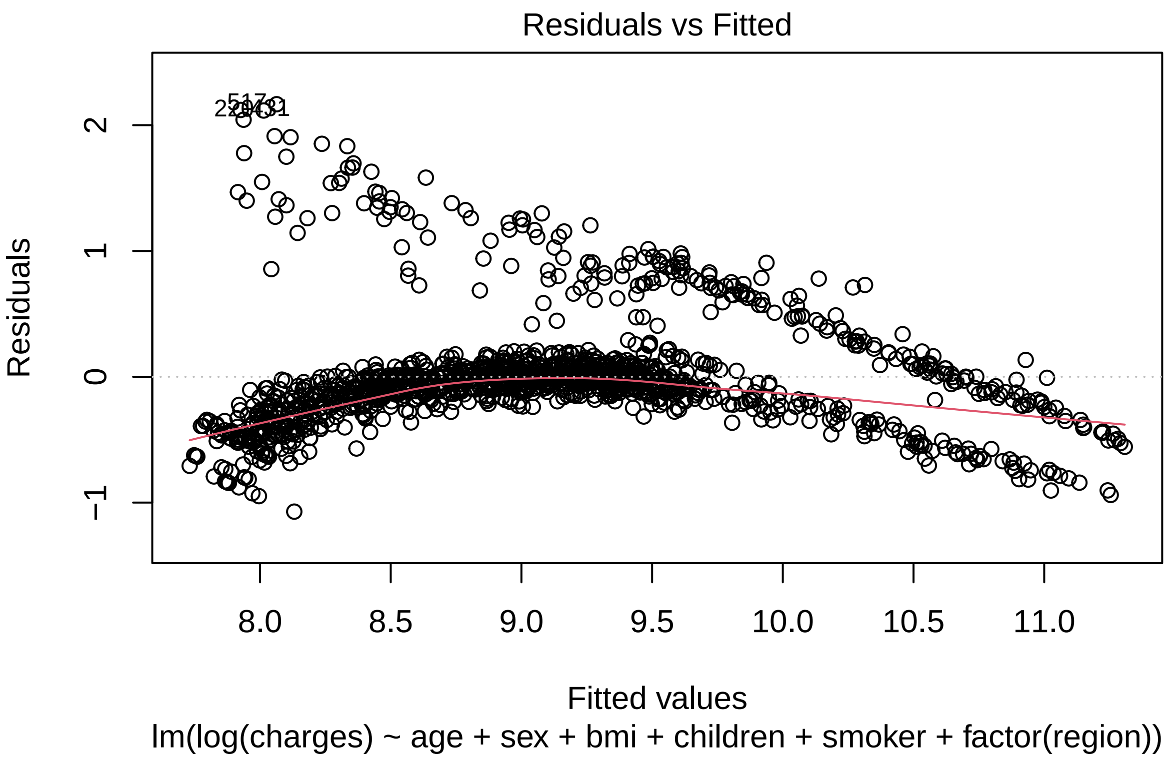
Since $\lambda = 0.15$. We can consider log transformation.

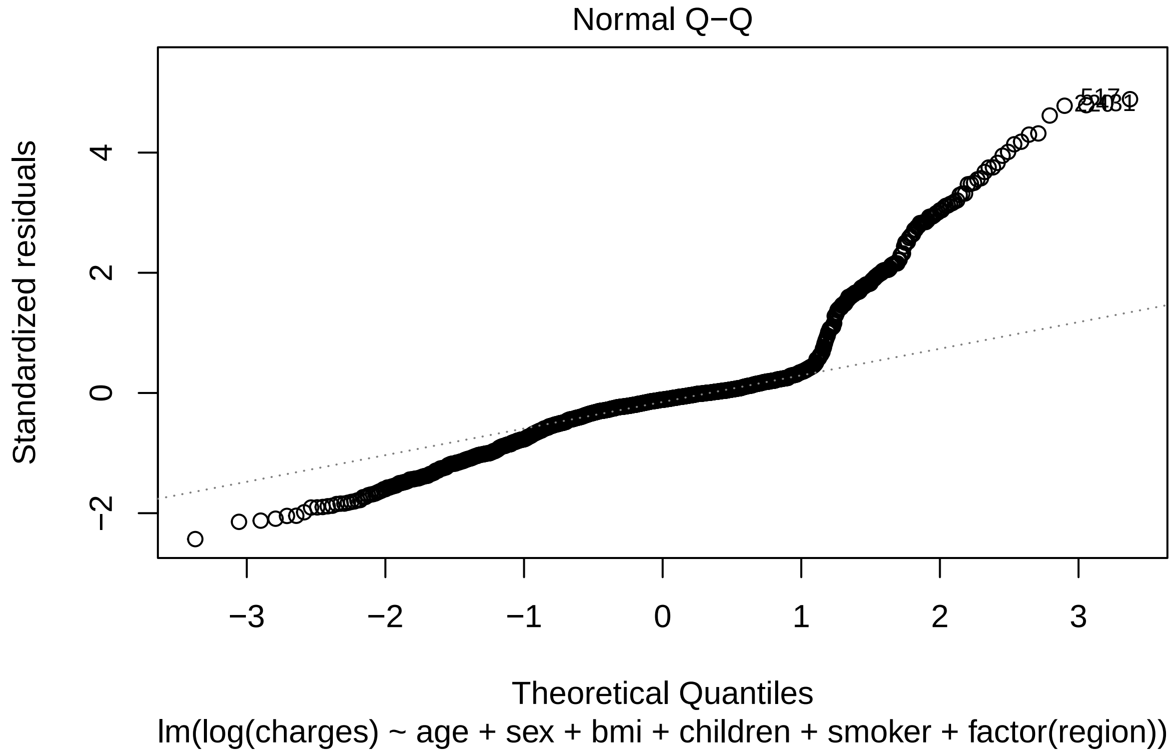
Fitting the model including transformation on response variable

```r
model2= lm(log(charges)~  age + sex + bmi + children + smoker + factor(region) , data= insurance)
summary(model2)
```

```
##
## Call:
## lm(formula = log(charges) ~ age + sex + bmi + children + smoker +
##     factor(region), data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07186 -0.19835 -0.04917  0.06598  2.16636
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.0305581  0.0723960  97.112  < 2e-16 ***
## age                      0.0345816  0.0008721  39.655  < 2e-16 ***
## sex                     -0.0754164  0.0244012  -3.091 0.002038 **
## bmi                      0.0133748  0.0020960   6.381 2.42e-10 ***
## children                 0.1018568  0.0100995  10.085  < 2e-16 ***
## smoker                   1.5543228  0.0302795  51.333  < 2e-16 ***
## factor(region)northwest -0.0637876  0.0349057  -1.827 0.067860 .
## factor(region)southeast -0.1571967  0.0350828  -4.481 8.08e-06 ***
## factor(region)southwest -0.1289522  0.0350271  -3.681 0.000241 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4443 on 1329 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7666
## F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```r
plot(model2, 1:2)
```



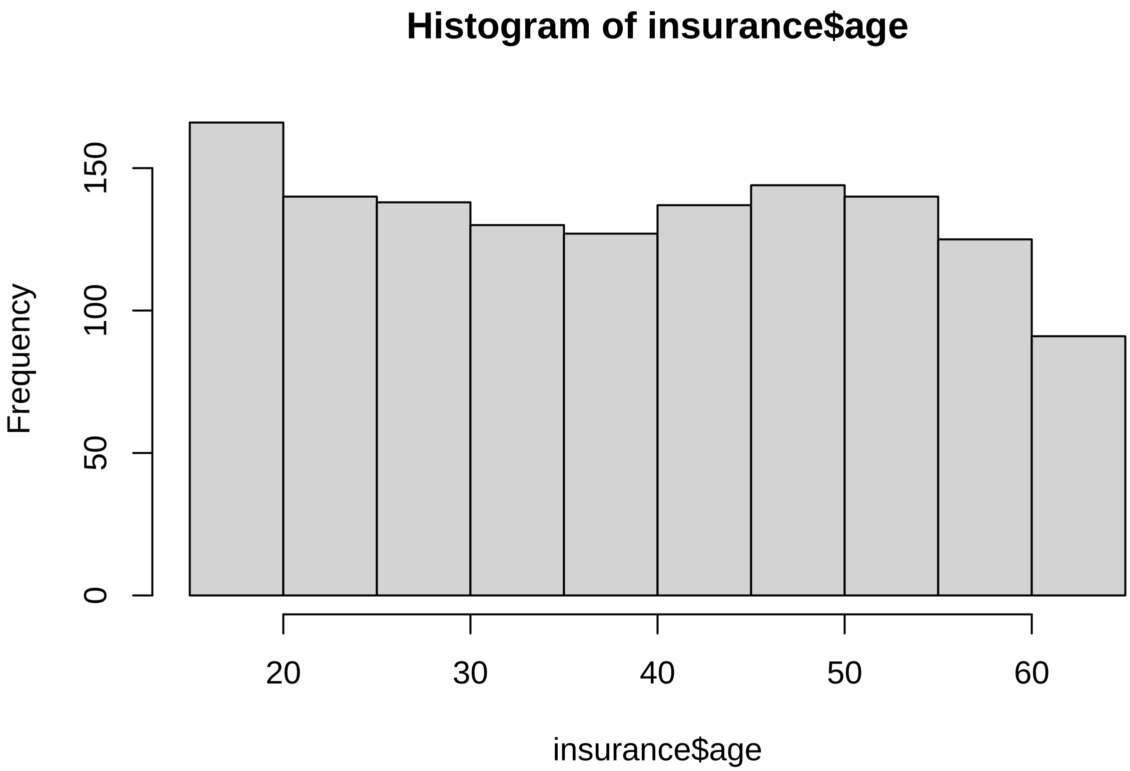


Inference:

Even after the transformation on the response variable the assumptions of normality and constant variance are not satisfied. We can check the distribution of the other variables

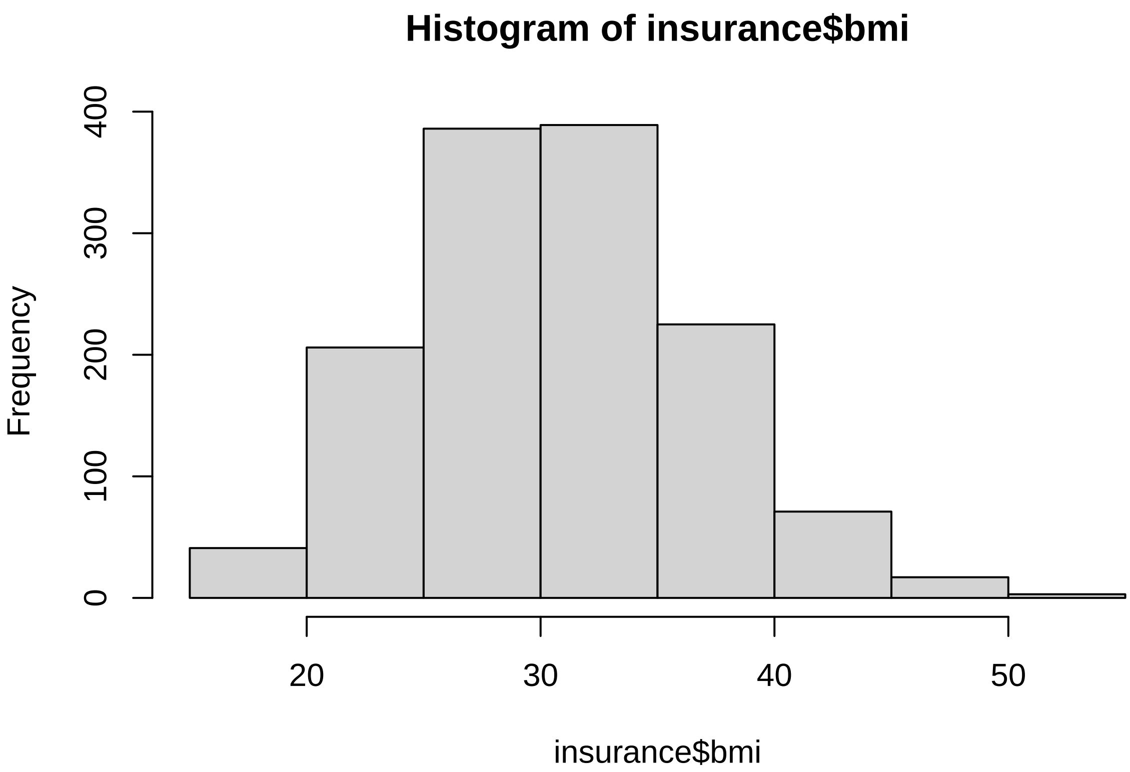
Let us check the distribution of the age.

```
hist(insurance$age)
```



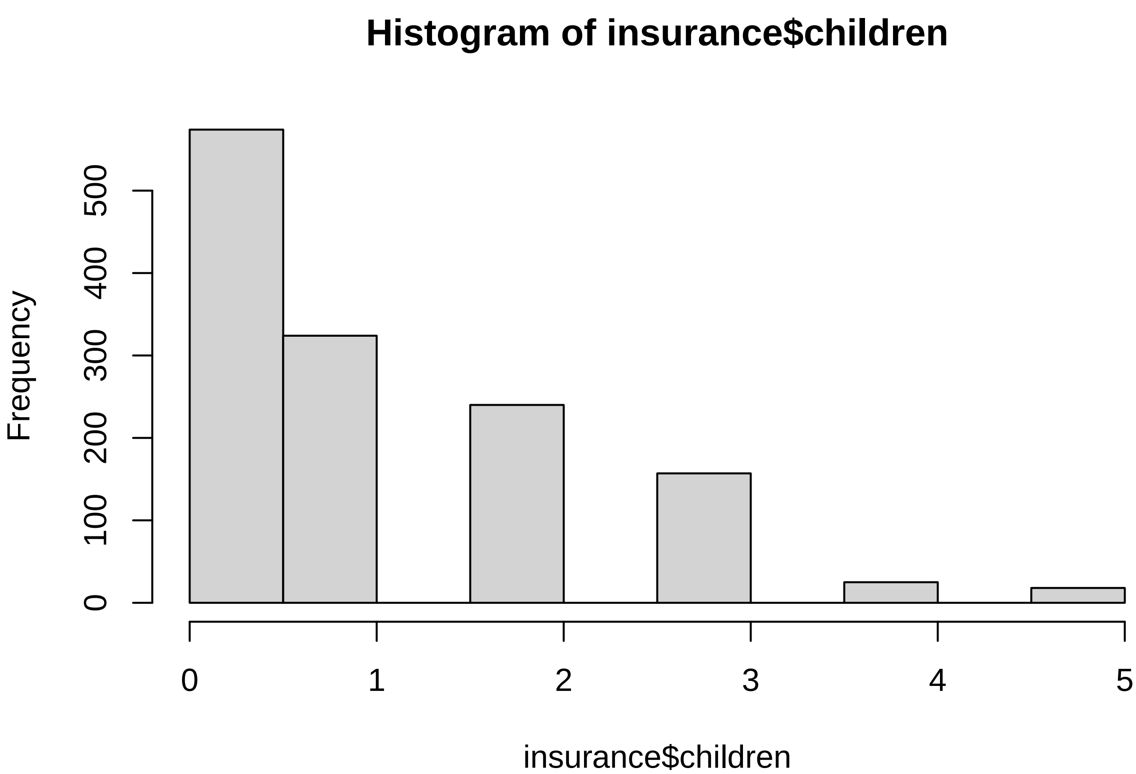
Distribution of BMI variable:

```
hist(insurance$bmi)
```



Distribution of no. of dependants:

```
hist(insurance$children)
```

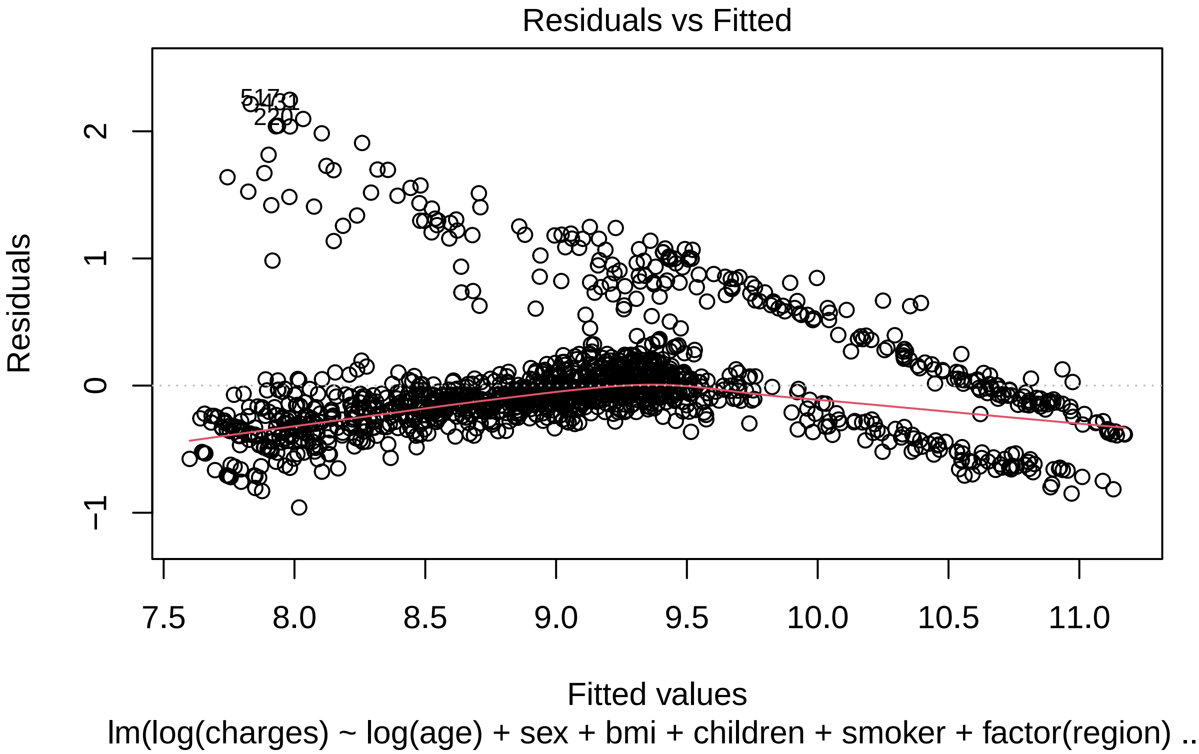


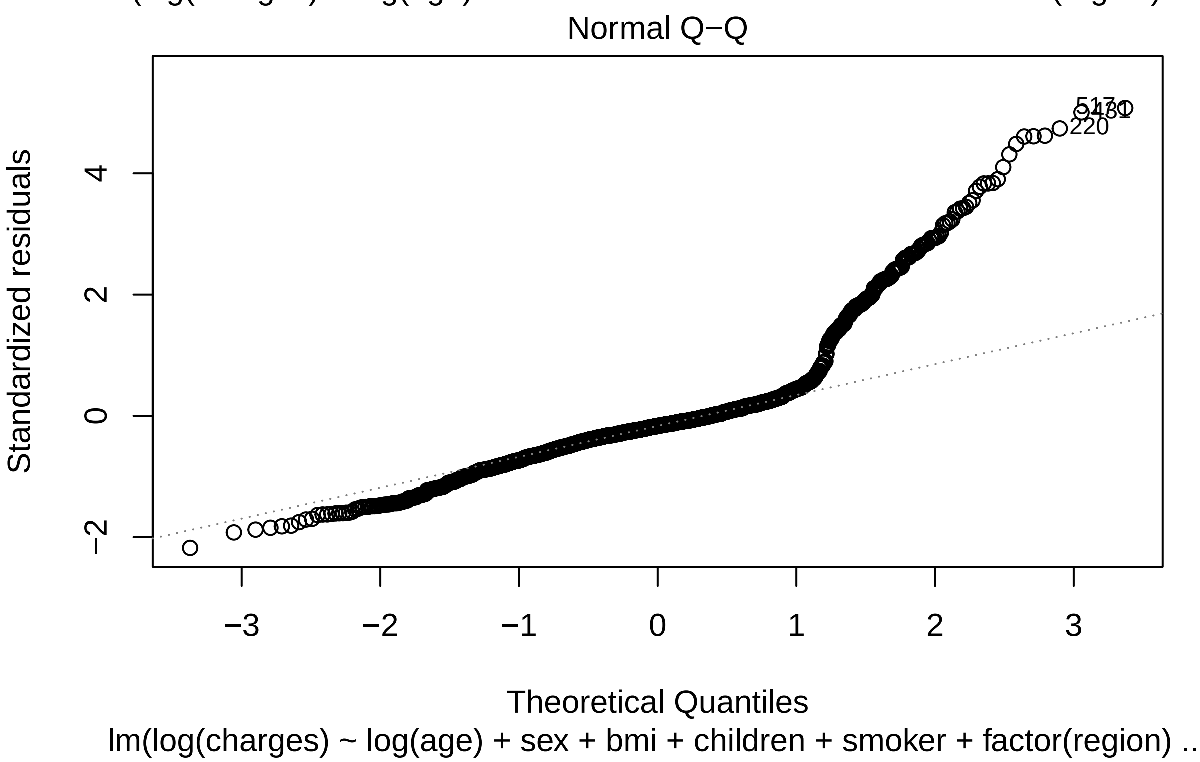
Fitting a multilinear regression model with transformation on response and age variable.

```
model3= lm(log(charges)~  log(age) + sex + bmi + children + smoker + factor(region) , data= insurance)
summary(model3)
```

```
##
## Call:
## lm(formula = log(charges) ~ log(age) + sex + bmi + children +
##     smoker + factor(region), data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95870 -0.22594 -0.07411  0.07838  2.24765
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.901494   0.125294  31.139  < 2e-16 ***
## log(age)                 1.247780   0.031411  39.724  < 2e-16 ***
## sex                     -0.074875   0.024379  -3.071 0.002174 **
## bmi                      0.013985   0.002092   6.684 3.41e-11 ***
## children                 0.082306   0.010122   8.131 9.64e-16 ***
## smoker                   1.551896   0.030250  51.303  < 2e-16 ***
## factor(region)northwest -0.063589   0.034873  -1.823 0.068461 .
## factor(region)southeast -0.158353   0.035049  -4.518 6.79e-06 ***
## factor(region)southwest -0.130858   0.034994  -3.739 0.000192 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4439 on 1329 degrees of freedom
## Multiple R-squared:  0.7684, Adjusted R-squared:  0.767
## F-statistic: 551.1 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```
plot(model3,1:2)
```






```r
s1 = summary(model2)
s2 = summary(model3)
```

```r
s1$adj.r.squared
```

```
## [1] 0.7665509
```

```
s2$adj.r.squared
```

```
## [1] 0.766989
```

Inference:

We see that the adj. R square value for model with transformation on response and age variable is 76.69 which performs better that the model with transformation only on response variable where the adj. r square value is 76.65.

Random Forest:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x dplyr::select() masks MASS::select()
## x purrr::some()   masks car::some()
```

```
library(ISLR2)
```

```
##
## Attaching package: 'ISLR2'
## The following object is masked from 'package:MASS':
##
##     Boston
```

```
library(randomForest)
```

```
## randomForest 4.7-1
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
## The following object is masked from 'package:dplyr':
##
##     combine
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(vip)
```

```
##
## Attaching package: 'vip'
```

```
## The following object is masked from 'package:utils':
##
##     vi
```

```r
region= as.factor(insurance$region)
```

```r
set.seed(652) # make results reproducible
rf1 <- randomForest(charges~ age + sex + bmi + children + smoker + region , data= insurance, importance
rf1
```

```
##
## Call:
##  randomForest(formula = charges ~ age + sex + bmi + children +      smoker + region, data = insuranc
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 21971071
##                     % Var explained: 85.01
```

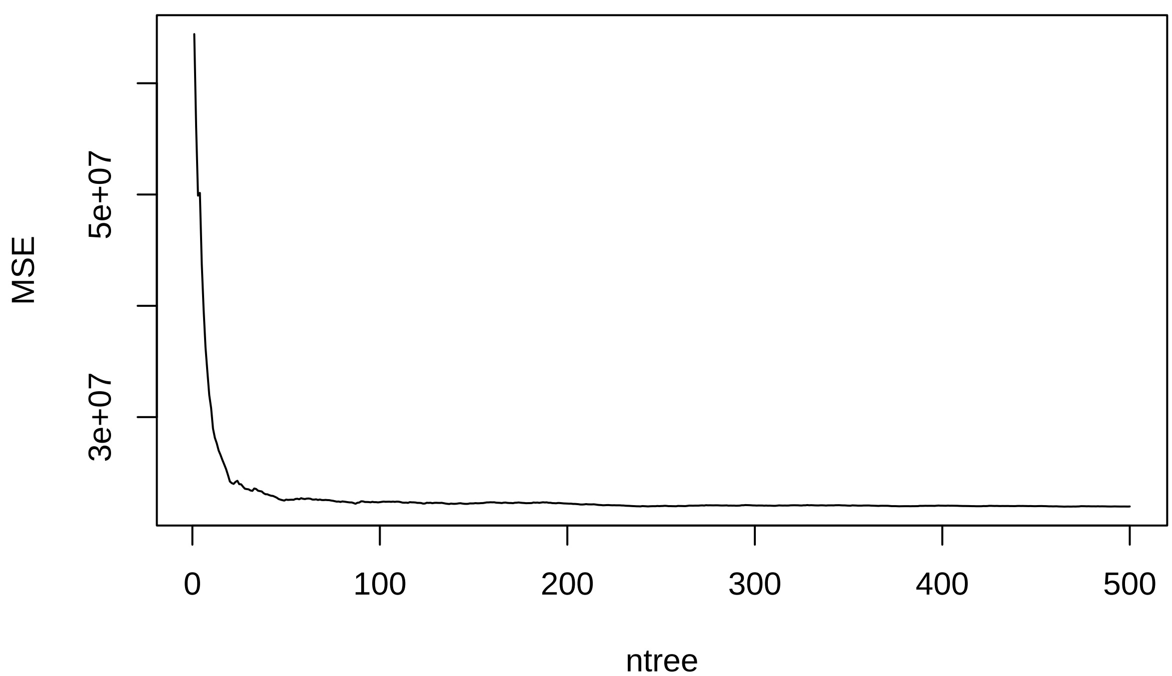
```r
sqrt(21971071)
```

```
## [1] 4687.331
```

The RMSE on the OOB data is $\sqrt{21971071} = 4687.331$, and the R2 on the OOB data is about 85.01%.

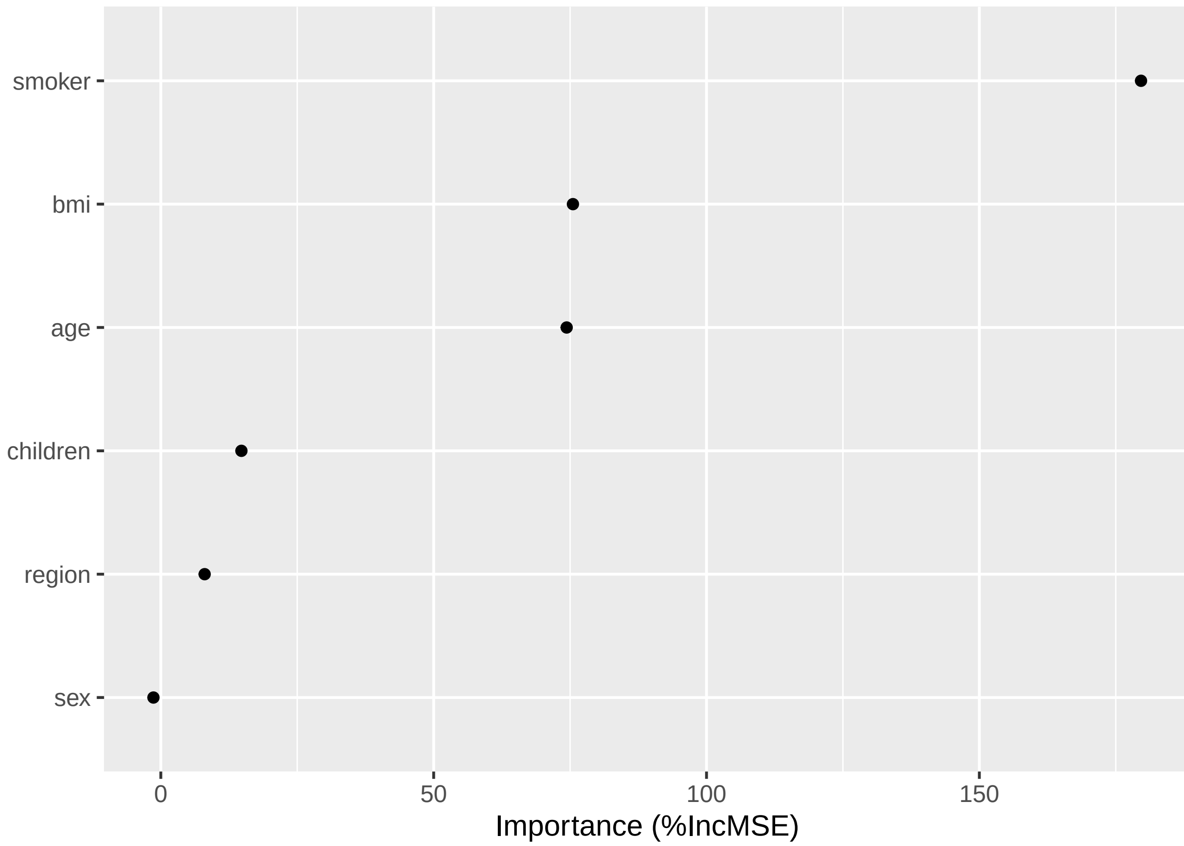
This plot below show the MSE as the number of trees in the model increases. We see that the MSE stabilizes by about 100 trees.

```r
plot(c(1: 500), rf1$mse, xlab="ntree", ylab="MSE", type="l")
```



Below is a variable importance plot which gives a ranking of the predictors in the model from most important (smoker) to least important (sex). The variable importance measure here is based on the increase in MSE when permuting each variable in the OOB data.

```r
vip(rf1, num_features = 6, geom = "point", include_type = TRUE)
```



CROSS VALIDATION

Here we use cross-validation (hold-out method) to check the performance of random forests.

```
# split data into 70% training and 30% test set
set.seed(652)
n <- nrow(insurance)
train_index <- sample(1:n, round(0.7*n))
insurance_train <- insurance[train_index, ]
insurance_test <- insurance[-train_index, ]
```

```
rf2 <- randomForest(charges~ age + sex + bmi + children + smoker + region , data= insurance, importance
```

```
# make predictions on test set and compute RMSE
pred_rf2 <- predict(rf2, newdata = insurance_test)
```

```
RMSE <- function(y, y_hat) {
  sqrt(mean((y - y_hat)^2))
}
```

```
RMSE(insurance_test$charges, pred_rf2)
```

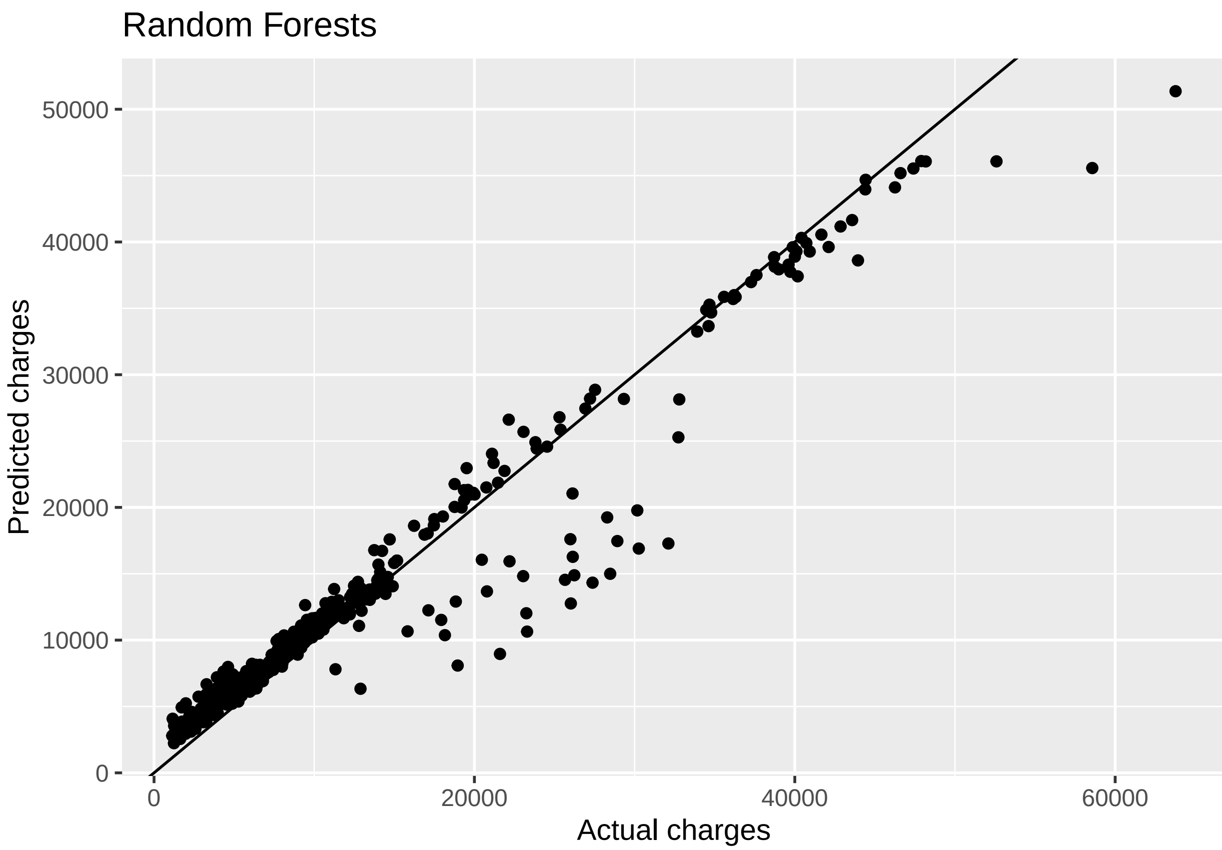
```
## [1] 3102.828
```

Inference:

In terms of RMSE, the Random Forest model performed better than multiple linear regression.

```r
pred_df <- data.frame(
  Actual = insurance_test$charges,
  Pred_RF = pred_rf2
)
```

```r
ggplot(pred_df, aes(x = Actual, y = Pred_RF)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1) +
  xlab("Actual charges") + ylab("Predicted charges") +
  ggtitle("Random Forests")
```



Inference:

Looking at the actual Vs predicted plot the model seems to be working fine since most of the points are lying on the straight line.