ST-301_Assignment_01

S/17/436

2022-11-07

Install Packages

The following packages are used to evaluate the fitted model for the given data set.

```
library(MASS)
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.2 ──
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ✕ dplyr::select() masks MASS::select()

library(gridExtra)

##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine

library(corrplot)

## corrplot 0.92 loaded

library(performance)
library(tinytex)
library(sp)
```

Import the data set

```
insurance_claims <- read.csv("E:/3rd year 1st sem/ST-301 (Regression
Analysis)/insurance_claims.csv")
```

```
attach(insurance_claims)
head(insurance_claims)
```

```
##   age    sex    bmi children is_smoker working_env tot_claims
## 1  19 female 27.900        0       yes     factory  16884.924
## 2  18   male 33.770        1        no      office   1725.552
## 3  28   male 33.000        3        no      office   4449.462
## 4  33   male 22.705        0        no     factory  21984.471
## 5  32   male 28.880        0        no      office   3866.855
## 6  31 female 25.740        0        no      office   3756.622
```

```
summary(insurance_claims)
```

```
##       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  
##   is_smoker          working_env          tot_claims   
##  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
```

```
str(insurance_claims)
```

```
## 'data.frame':    1338 obs. of  7 variables:
##  $ age        : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex        : chr  "female" "male" "male" "male" ...
##  $ bmi        : num  27.9 33.8 33 22.7 28.9 ...
##  $ children   : int  0 1 3 0 0 0 1 3 2 0 ...
##  $ is_smoker  : chr  "yes" "no" "no" "no" ...
##  $ working_env: chr  "factory" "office" "office" "factory" ...
##  $ tot_claims : num  16885 1726 4449 21984 3867 ...
```

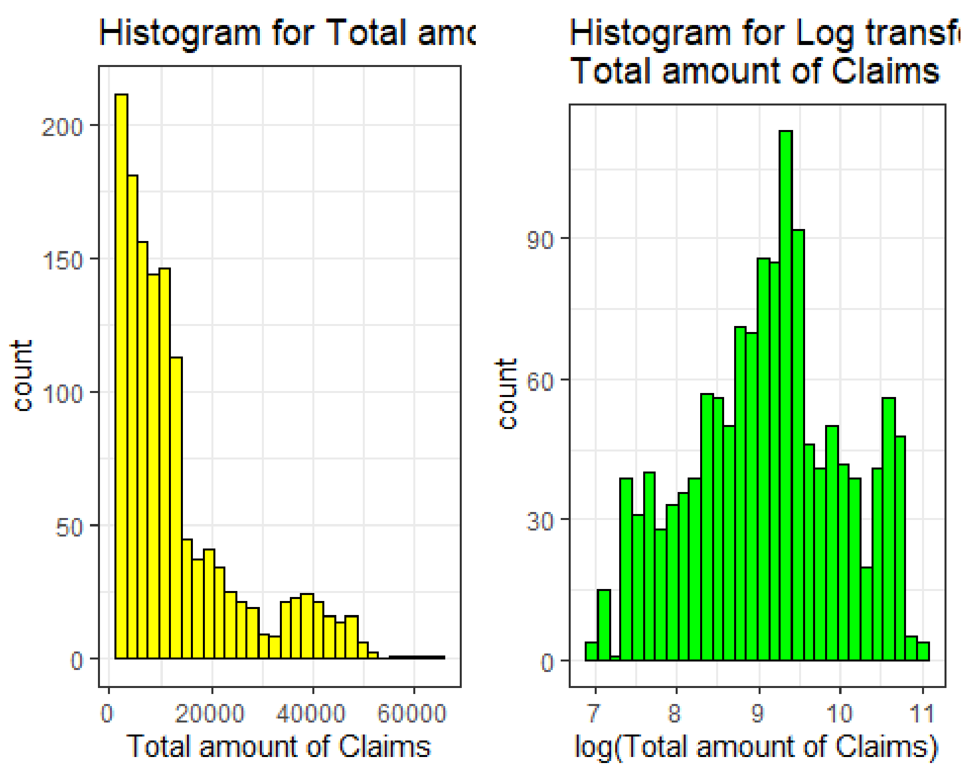
Note that, there are three categorical variables in the given data set. Such as, ‘sex’ , ‘is_smoker’ and ‘working_env’

Exploratory Analysis

Under the exploratory analysis, We have to look at the relationship between categorical variables and numerical variables. Also, want to look at the relationship between each and every variables with the response variable.

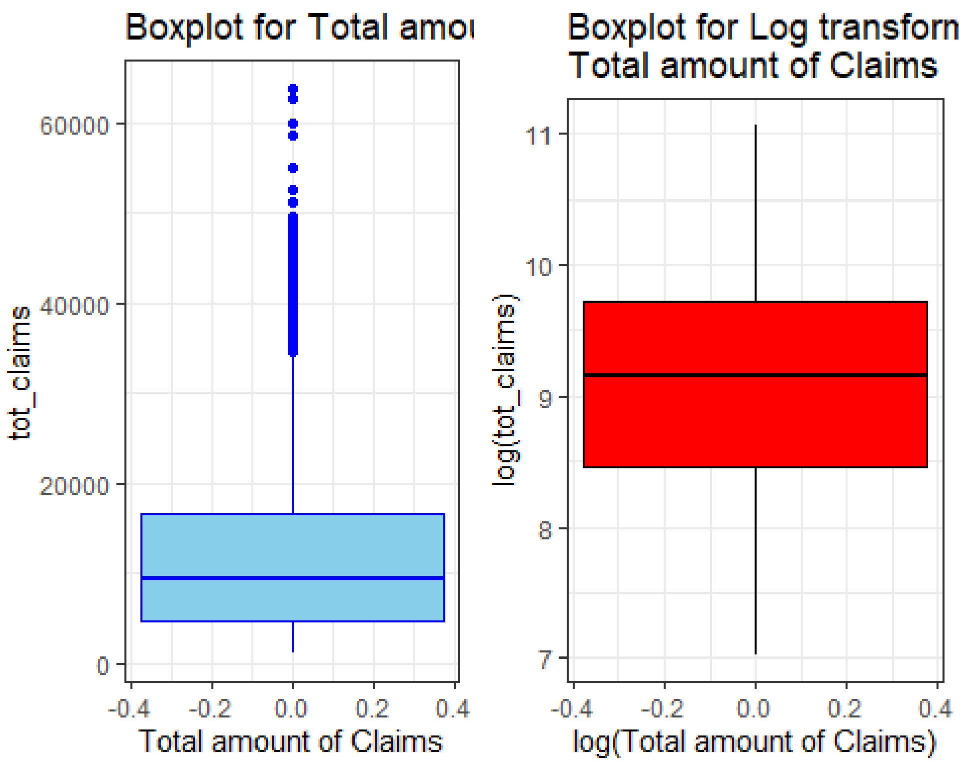
1. Total amount of claims made by the policyholder

```
p1 <- insurance_claims %>%
  ggplot(aes(x = tot_claims)) +
  geom_histogram(col = "black", fill = "yellow") +
  labs(x = "Total amount of Claims",
       title = "Histogram for Total amount of Claims") +
  theme_bw()

p2 <- insurance_claims %>%
  ggplot(aes(x = log(tot_claims))) +
  geom_histogram(col = "black", fill = "green") +
  labs(x = "log(Total amount of Claims)",
       title = "Histogram for Log transformed\nTotal amount of Claims") +
  theme_bw()

grid.arrange(p1,p2, ncol = 2)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
p01 <- insurance_claims %>%
  ggplot(aes(y = tot_claims)) +
  geom_boxplot(col = "blue", fill = "skyblue") +
  labs(x = "Total amount of Claims",
       title = "Boxplot for Total amount of Claims") +
```

```
  theme_bw()

p02 <- insurance_claims %>%
  ggplot(aes(y = log(tot_claims))) +
  geom_boxplot(col = "black", fill = "red") +
  labs(x = "log(Total amount of Claims)",
       title = "Boxplot for Log transformed\nTotal amount of Claims") +
  theme_bw()

grid.arrange(p01,p02, ncol = 2)
```

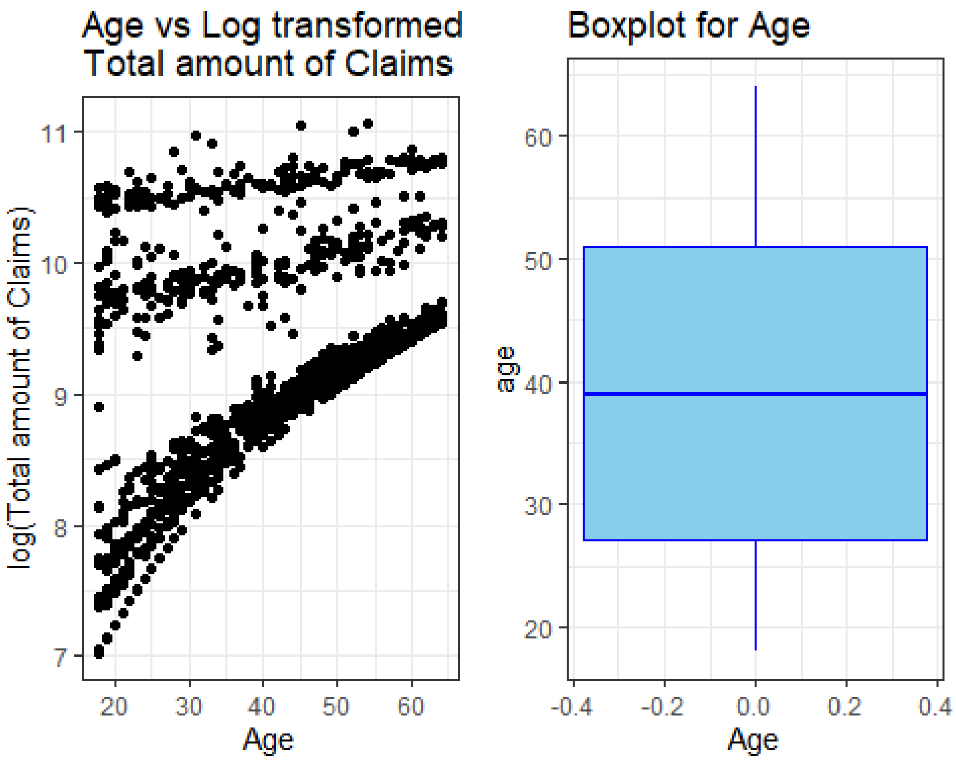


'tot_claims' variable is not normally distributed. Then I applied log transformation to the variable and it seems that, transformed variable is fairly normally distributed. So, we used log transformed variable for further analysis.

2. Relationship between Age and Total amount of claims

```
p3 <- insurance_claims %>%
  ggplot(aes(x = age,y = log(tot_claims))) +
  geom_point(col = "black") +
  labs(x = "Age",y = "log(Total amount of Claims)",
       title = "Age vs Log transformed\nTotal amount of Claims") +
  theme_bw()

p4 <- insurance_claims %>%
  ggplot(aes(y = age)) +
  geom_boxplot(col = "blue", fill = "skyblue") +
```

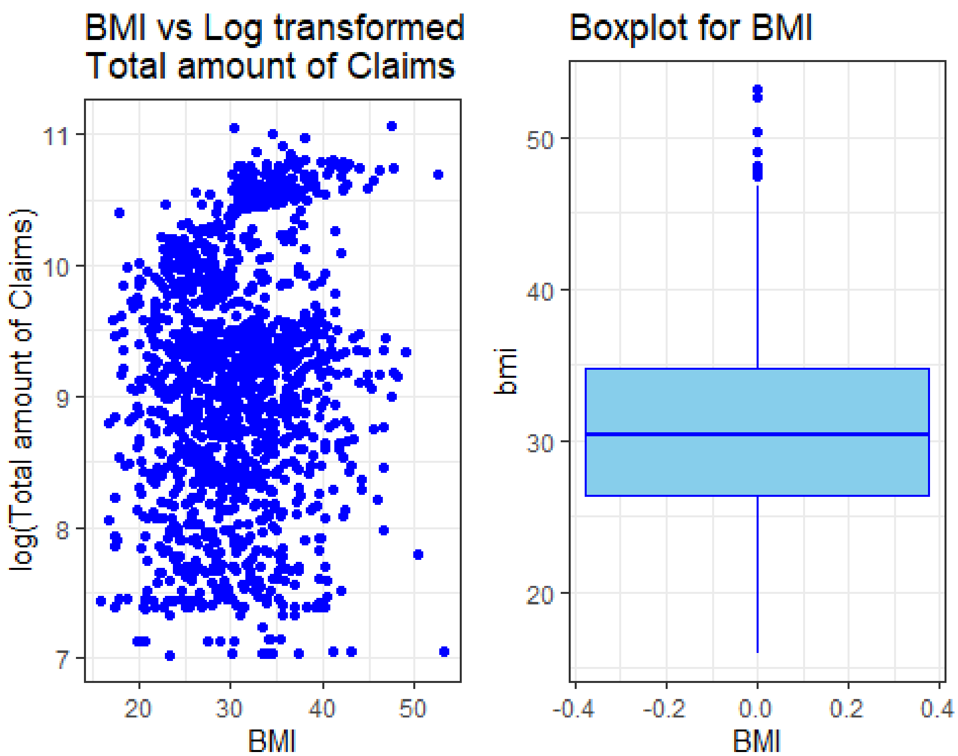
```
  labs(x = "Age",
       title = "Boxplot for Age") +
  theme_bw()

grid.arrange(p3,p4, ncol = 2)
```



There is a a moderately positive relationship between Age and log transformed variable. Age variable does not contain any outliers.

3. Relationship between BMI and Total amount of claims

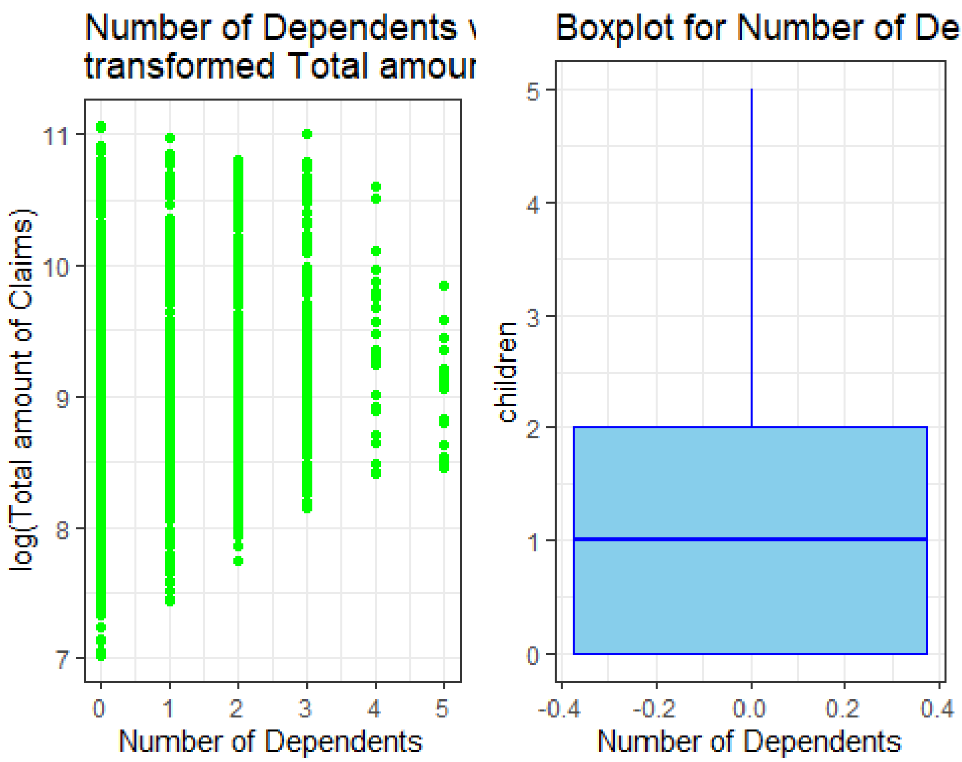
```
p5 <- insurance_claims %>%
  ggplot(aes(x = bmi,y = log(tot_claims))) +
  geom_point(col = "blue") +
  labs(x = "BMI",y = "log(Total amount of Claims)",
       title = "BMI vs Log transformed\nTotal amount of Claims") +
  theme_bw()

p6 <- insurance_claims %>%
  ggplot(aes(y = bmi)) +
  geom_boxplot(col = "blue", fill = "skyblue") +
  labs(x = "BMI",
       title = "Boxplot for BMI") +
  theme_bw()

grid.arrange(p5,p6, ncol = 2)
```



It seems that there exists a very poor relationship between 'bmi' and log transformed variable. There are some outlires in ``bmi`` variable.

4. Relationship between Number of Dependents and Total amount of claims

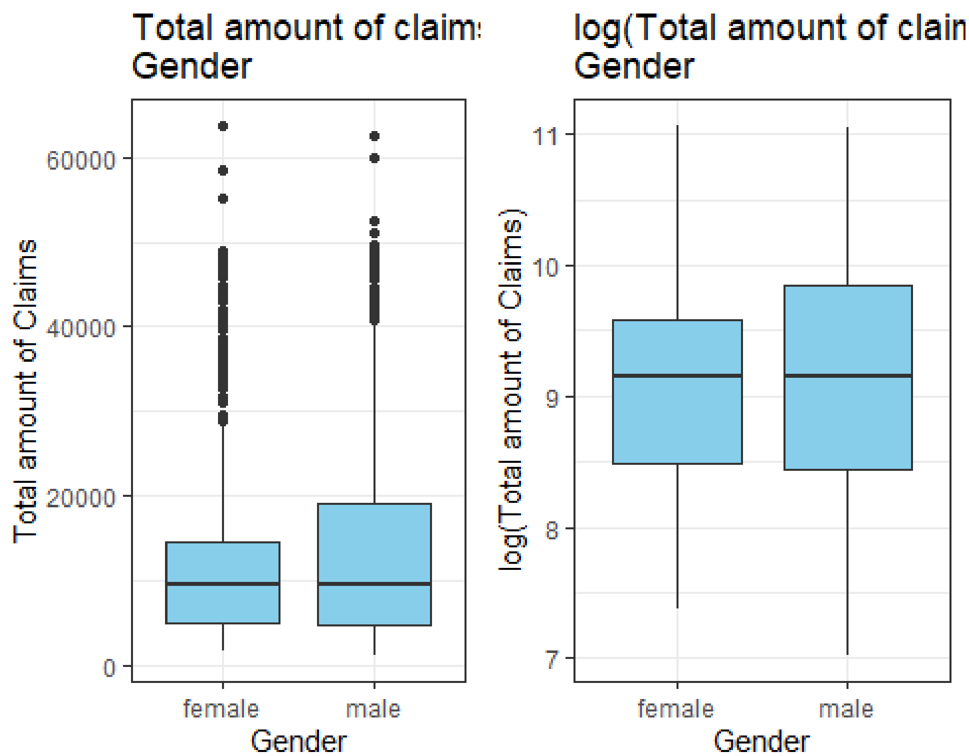
```
p7 <- insurance_claims %>%
  ggplot(aes(x = children,y = log(tot_claims))) +
  geom_point(col = "green") +
  labs(x = "Number of Dependents",y = "log(Total amount of Claims)",
       title = "Number of Dependents vs Log\ntransformed Total amount of
Claims") +
  theme_bw()

p8 <- insurance_claims %>%
  ggplot(aes(y = children)) +
  geom_boxplot(col = "blue", fill = "skyblue") +
  labs(x = "Number of Dependents",
       title = "Boxplot for Number of Dependents") +
  theme_bw()

grid.arrange(p7,p8, ncol = 2)
```



There is a poor relationship with 'children' and log transformed variable. But we can not detect any outlires in this variable.

5. Total amount of claims across Gender

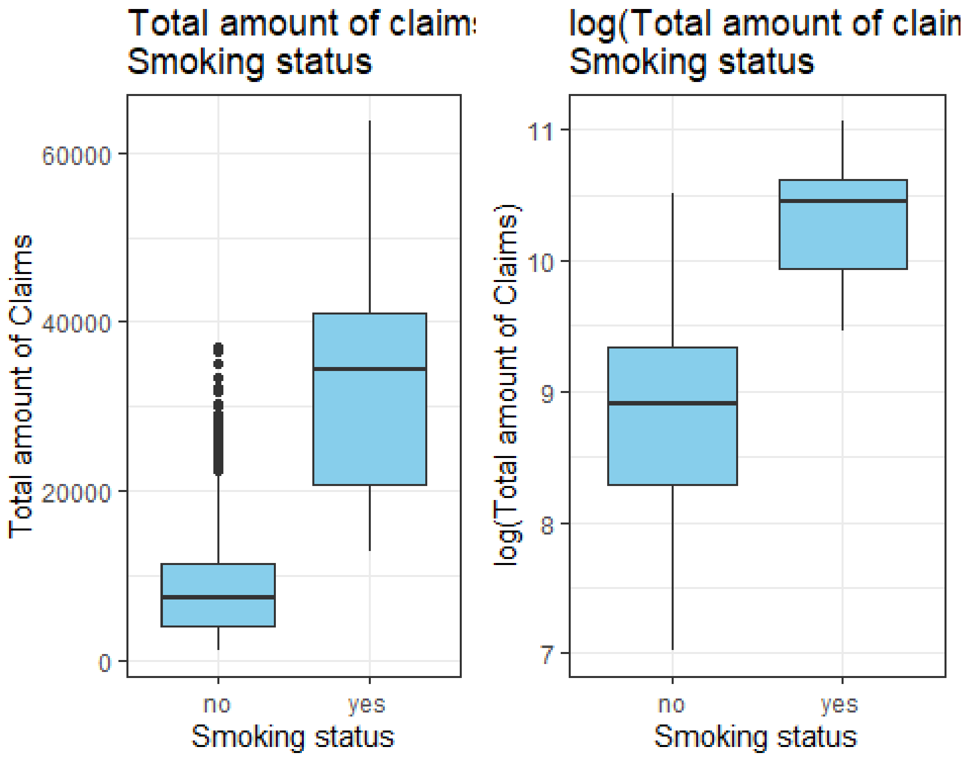
```
p9 <- insurance_claims %>%
  ggplot(aes(x = sex, y = tot_claims)) +
  geom_boxplot(fill = "skyblue") +
  labs(x = "Gender",y = "Total amount of Claims",
       title = "Total amount of claims across\nGender") +
  theme_bw()

p10 <- insurance_claims %>%
  ggplot(aes(x = sex,y = log(tot_claims))) +
  geom_boxplot(fill = "skyblue") +
  labs(x = "Gender",y = "log(Total amount of Claims)",
       title = "log(Total amount of claims) across\nGender") +
  theme_bw()

grid.arrange(p9,p10, ncol = 2)
```



We can see that there are outlires in the right side boxplot. But after applied the log transformation, we can not detect any outlires. Further, both distributions are fairly normally distributed because of both medians are approximately equal.

6. Total amount of claims across Smoking status

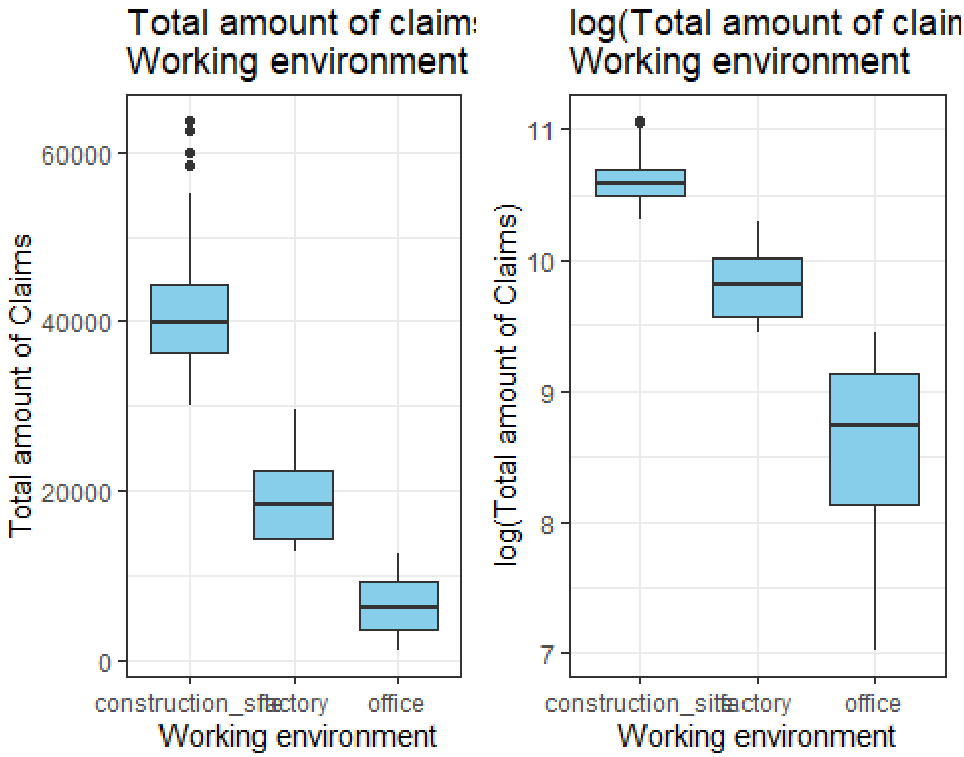
```
p11 <- insurance_claims %>%
  ggplot(aes(x = is_smoker, y = tot_claims)) +
  geom_boxplot(fill = "skyblue") +
  labs(x = "Smoking status",y = "Total amount of Claims",
       title = "Total amount of claims across\nSmoking status") +
  theme_bw()

p12 <- insurance_claims %>%
  ggplot(aes(x = is_smoker,y = log(tot_claims))) +
  geom_boxplot(fill = "skyblue") +
  labs(x = "Smoking status",y = "log(Total amount of Claims)",
       title = "log(Total amount of claims) across\nSmoking status") +
  theme_bw()

grid.arrange(p11,p12, ncol = 2)
```



We can see that there are outlires in the right side boxplot. But after transformed, we can not detect any outliers and both distributions are fairly normally distributed. In 'yes' category has significantly higher median than 'no' category. So, it seems that 'is_smoker' variable has an effect on 'tot_claims' variable.

7. Total amount of claims across working environment

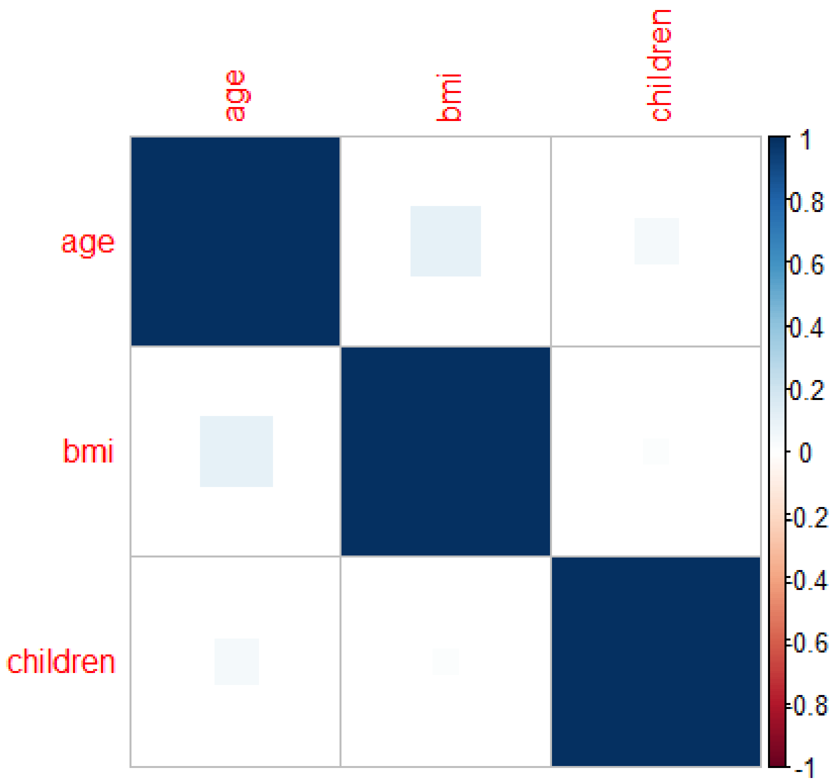
```
p13 <- insurance_claims %>%
  ggplot(aes(x = working_env, y = tot_claims)) +
  geom_boxplot(fill = "skyblue") +
  labs(x = " Working environment",y = "Total amount of Claims",
       title = "Total amount of claims across\nWorking environment") +
  theme_bw()

p14 <- insurance_claims %>%
  ggplot(aes(x = working_env,y = log(tot_claims))) +
  geom_boxplot(fill = "skyblue") +
  labs(x = "Working environment",y = "log(Total amount of Claims)",
       title = "log(Total amount of claims) across\nWorking environment") +
  theme_bw()

grid.arrange(p13,p14, ncol = 2)
```



We can see that there are outlires in the right side boxplots. In 'construction_site' category has significantly higher median than other categories. So, it seems 'working_env' variable has an effect on 'tot_claims' variable.

Correlation between Age, BMI and Number of Dependents

```
corrplot(cor(insurance_claims[,c("age","bmi","children")]),method = 'square')
```



There is a very poor correlation between 'bmi and 'age'. Also, there is no relationship between other pairs in the plot.

Handle the categorical variables

Here, we have to covert all the categorical variables as numeric.

```
insurance_claims$sex <- as.numeric(factor(insurance_claims$sex , labels =
c("male" , "female")))

insurance_claims$is_smoker <- as.numeric(factor(insurance_claims$is_smoker ,
labels = c("yes" , "no")))

insurance_claims$working_env <-
as.numeric(factor(insurance_claims$working_env , labels = c("factory" ,
"office" , "construction_site")))

str(insurance_claims)

## 'data.frame':    1338 obs. of  7 variables:
##  $ age        : int  19 18 28 33 32 31 46 37 37 60 ...
##  $ sex        : num  1 2 2 2 2 1 1 1 2 1 ...
##  $ bmi        : num  27.9 33.8 33 22.7 28.9 ...
##  $ children   : int  0 1 3 0 0 0 1 3 2 0 ...
##  $ is_smoker  : num  2 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ working_env: num  2 3 3 2 3 3 3 3 3 2 ...
##  $ tot_claims : num  16885 1726 4449 21984 3867 ...
```

```
head(insurance_claims)
```

```
##   age sex    bmi children is_smoker working_env tot_claims
## 1  19   1 27.900        0         2           2  16884.924
## 2  18   2 33.770        1         1           3   1725.552
## 3  28   2 33.000        3         1           3   4449.462
## 4  33   2 22.705        0         1           2  21984.471
## 5  32   2 28.880        0         1           3   3866.855
## 6  31   1 25.740        0         1           3   3756.622
```

Model Fitting

For the purpose of model fitting, I have used the forward selection method based on Adjusted R-squared values to select the significance variables.

Iteration 01

```
summary(lm(tot_claims ~ age, data = insurance_claims))$adj.r.squared

## [1] 0.08872432

summary(lm(tot_claims ~ bmi, data = insurance_claims))$adj.r.squared

## [1] 0.03862008

summary(lm(tot_claims ~ sex, data = insurance_claims))$adj.r.squared

## [1] 0.002536334

summary(lm(tot_claims ~ children, data = insurance_claims))$adj.r.squared

## [1] 0.003878717

summary(lm(tot_claims ~ is_smoker, data = insurance_claims))$adj.r.squared

## [1] 0.6194802

summary(lm(tot_claims ~ working_env, data = insurance_claims))$adj.r.squared

## [1] 0.8614734
```

Since 'working_env' variable has the largest adjusted R-squared value as 0.8614734, that variable is included to the model

Iteration 02

```
summary(lm(tot_claims ~ working_env + age, data =
insurance_claims))$adj.r.squared

## [1] 0.886051
```

```
summary(lm(tot_claims ~ working_env + sex, data =
insurance_claims))$adj.r.squared

## [1] 0.8614025

summary(lm(tot_claims ~ working_env + bmi, data =
insurance_claims))$adj.r.squared

## [1] 0.865276

summary(lm(tot_claims ~ working_env + children, data =
insurance_claims))$adj.r.squared

## [1] 0.8643907

summary(lm(tot_claims ~ working_env + is_smoker, data =
insurance_claims))$adj.r.squared

## [1] 0.8679371
```

Here 'age' variable has the highest adjusted R-squared value as 0.886051. Therefore, 'age' is added to the model.

Iteration 03

```
summary(lm(tot_claims ~ working_env + age + sex, data =
insurance_claims))$adj.r.squared

## [1] 0.8859661

summary(lm(tot_claims ~ working_env + age + bmi, data =
insurance_claims))$adj.r.squared

## [1] 0.888348

summary(lm(tot_claims ~ working_env + age + children, data =
insurance_claims))$adj.r.squared

## [1] 0.8883291

summary(lm(tot_claims ~ working_env + age + is_smoker, data =
insurance_claims))$adj.r.squared

## [1] 0.9013036
```

Note that, 'is_smoker' variable has largest adjusted R-squared value as 0.9013036. So, this variable is also added to the model.

Iteration 04

```
summary(lm(tot_claims ~ working_env + age + is_smoker + sex, data =
insurance_claims))$adj.r.squared

## [1] 0.9012491
```

```r
summary(lm(tot_claims ~ working_env + age + is_smoker + bmi, data =
insurance_claims))$adj.r.squared

## [1] 0.9063182

summary(lm(tot_claims ~ working_env + age + is_smoker + children, data =
insurance_claims))$adj.r.squared

## [1] 0.903542
```

Since 'bmi' variable has largest adjusted R-squared as 0.9063182. 'bmi' variable is included to the model.

Iteration 05

```r
summary(lm(tot_claims ~ working_env + age + is_smoker + bmi + sex, data =
insurance_claims))$adj.r.squared

## [1] 0.9063071

summary(lm(tot_claims ~ working_env + age + is_smoker + bmi + children, data
= insurance_claims))$adj.r.squared

## [1] 0.9085069
```

Here 'children' variable has highest adjusted R-squared value as 0.9085069. So, this variable is also in the model.

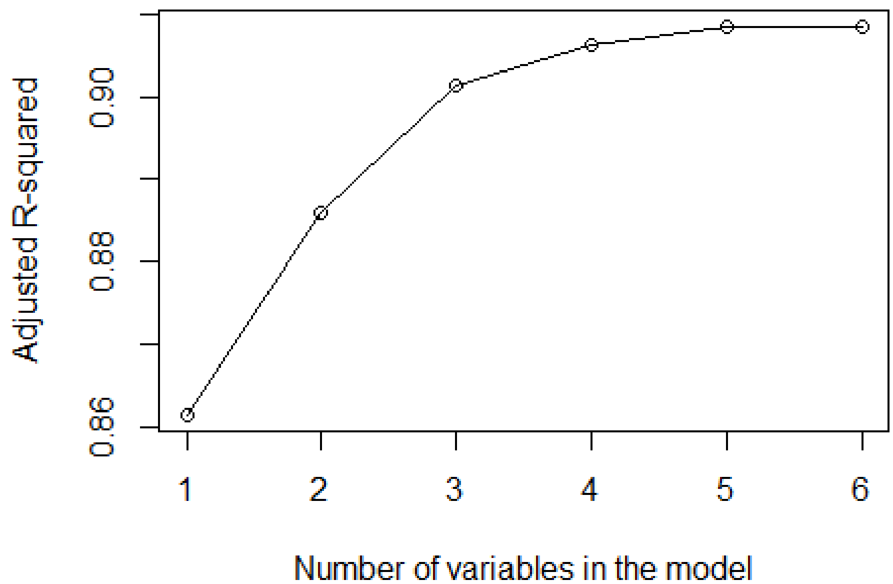
Iteration 06

```r
summary(lm(tot_claims ~ working_env + age + is_smoker + bmi + children + sex,
data = insurance_claims))$adj.r.squared

## [1] 0.9085106
```

Note that, when the 'sex' variable is added to the model there is no any significance change in adjusted R-squared value. Based on that reason, we cannot include the 'sex' variable for the above fitted model.

Plot all the iteration

```r
plot(c(1,2,3,4,5,6),c(0.8614734,0.886051,0.9013036,0.9063182,0.9085069,
0.9085106),
xlab = "Number of variables in the model", ylab ="Adjusted R-squared",
type="o")
```



According to the above plot, we have to include the following variables in order to obtain the best fitted model.

- age
- bmi
- children
- is_smoker
- working_env

Full model

Obtained the full model by including all the variables as follows.

```
full_model <- lm(tot_claims ~ . , data = insurance_claims)
drop1(full_model, test = "F")

## Single term deletions
##
## Model:
## tot_claims ~ age + sex + bmi + children + is_smoker + working_env
##              Df  Sum of Sq        RSS   AIC   F value    Pr(>F)
## <none>                     1.7858e+10 21966
```

```
## age          1 6.1495e+09 2.4008e+10 22360  458.3293 < 2.2e-16 ***
## sex          1 1.4141e+07 1.7872e+10 21965    1.0539    0.3048
## bmi          1 9.9272e+08 1.8851e+10 22037   73.9888 < 2.2e-16 ***
## children     1 4.4385e+08 1.8302e+10 21997   33.0808 1.095e-08 ***
## is_smoker    1 3.5266e+09 2.1385e+10 22205  262.8435 < 2.2e-16 ***
## working_env  1 3.1215e+10 4.9073e+10 23317 2326.4839 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(full_model)
```

```
##
## Call:
## lm(formula = tot_claims ~ ., data = insurance_claims)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11247.0  -1187.3    184.6   1669.3  24756.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24886.835   1435.366  17.338  < 2e-16 ***
## age            159.817      7.465  21.409  < 2e-16 ***
## sex           -206.535    201.182  -1.027    0.305
## bmi            145.770     16.947   8.602  < 2e-16 ***
## children       478.491     83.193   5.752  1.1e-08 ***
## is_smoker     6958.673    429.218  16.212  < 2e-16 ***
## working_env -12172.133    252.358 -48.234  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3663 on 1331 degrees of freedom
## Multiple R-squared:  0.9089, Adjusted R-squared:  0.9085
## F-statistic:  2214 on 6 and 1331 DF,  p-value: < 2.2e-16
```

According to the above results, we have to exclude the 'sex' variable as it is not significant to the fitted model. Further, it has high p value of 0.305 (>0.05) than the other variables.

Reduced model

Obtained the reduced model by dropping 'sex' variable from the full_model.

```
red_model <- lm(tot_claims ~ age + bmi + children + is_smoker + working_env ,
data = insurance_claims)
summary(red_model)
```

```
##
## Call:
## lm(formula = tot_claims ~ age + bmi + children + is_smoker +
##     working_env, data = insurance_claims)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11334.8  -1162.1    182.1   1684.2  24667.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24608.441   1409.546  17.458  < 2e-16 ***
## age            160.017      7.463  21.442  < 2e-16 ***
## bmi            144.977     16.929   8.564  < 2e-16 ***
## children       477.031     83.182   5.735 1.21e-08 ***
## is_smoker     6942.301    428.930  16.185  < 2e-16 ***
## working_env -12170.053    252.355 -48.226  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3663 on 1332 degrees of freedom
## Multiple R-squared:  0.9088, Adjusted R-squared:  0.9085
## F-statistic:  2656 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Validation of the model

Here, I have used Partial F Test to check the adequacy of the reduced model.

- **null hypothesis** : Reduced model is adequate

  vs

- **alternative** : Reduced model is not adequate

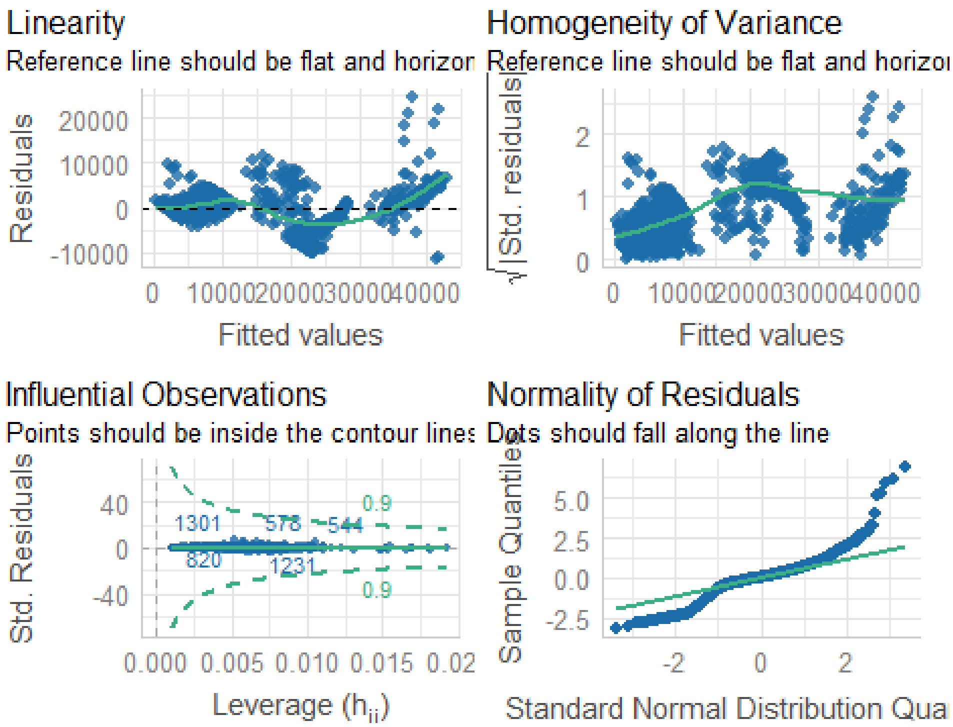
```
anova(full_model,red_model)
```

```
## Analysis of Variance Table
##
## Model 1: tot_claims ~ age + sex + bmi + children + is_smoker + working_env
## Model 2: tot_claims ~ age + bmi + children + is_smoker + working_env
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1   1331 1.7858e+10
## 2   1332 1.7872e+10 -1 -14140693 1.0539 0.3048
```

By looking at the ANOVA table, we can detect that the p-value (0.3048) is greater than 0.05 at 5% significance level. That means we don't have enough evidence to reject null hypothesis at 5% significance level. Moreover, we can conclude that the reduced model is adequate.

Residual analysis

```
check_model(red_model, check = c("linearity","homogeneity","qq","outliers"))
```

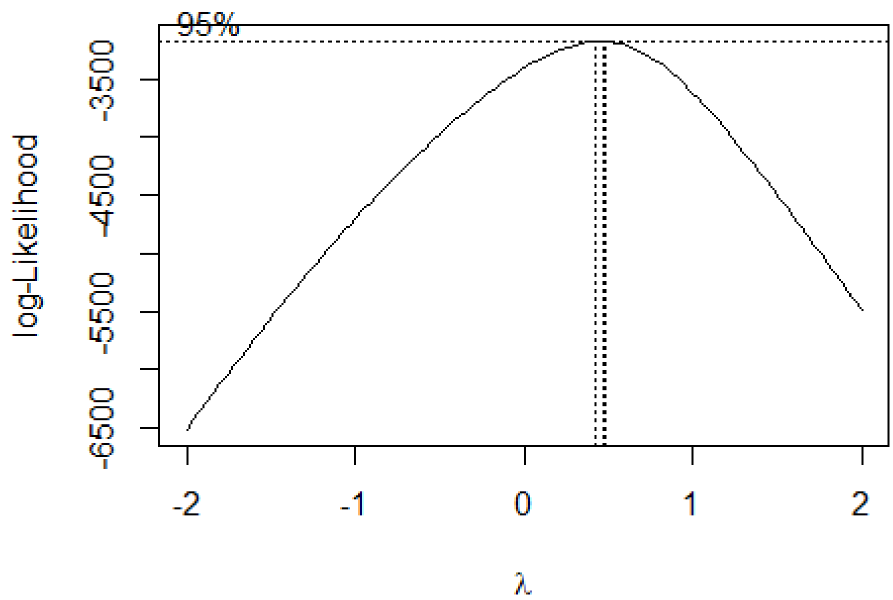



```
check_normality(red_model)

## Warning: Non-normality of residuals detected (p < .001).

check_heteroskedasticity(red_model)

## Warning: Heteroscedasticity (non-constant error variance) detected (p <
.001).

check_outliers(red_model)

## OK: No outliers detected.

check_autocorrelation(red_model)

## OK: Residuals appear to be independent and not autocorrelated (p = 0.998).
```

By looking at the above plot and results that we obtained, we can detect that the normality of residuals and heteroskedasticity is violated. So, we have to use the transformation method to correct those violation.

Box-cox transformation

```
box_trans <- boxcox(red_model)
```

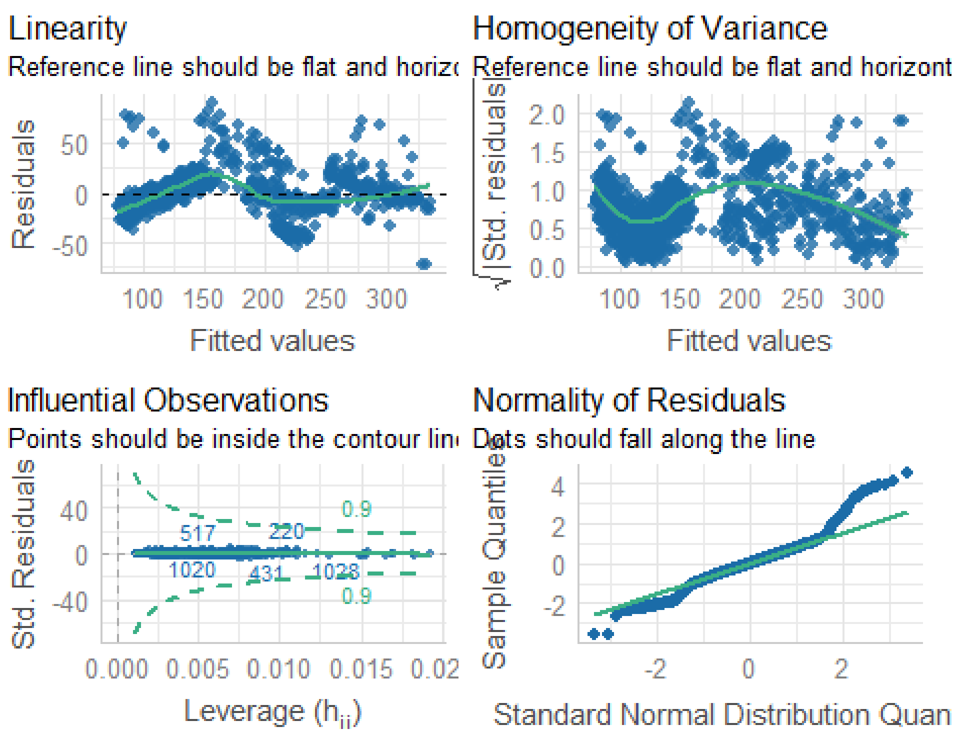



```
(lambda <- box_trans$x[which.max(box_trans$y)])

## [1] 0.4646465

fit_model <- lm(((tot_claims^lambda-1)/lambda) ~ working_env + is_smoker +
bmi + children + age, data = insurance_claims)

summary(fit_model)

##
## Call:
## lm(formula = ((tot_claims^lambda - 1)/lambda) ~ working_env +
##     is_smoker + bmi + children + age, data = insurance_claims)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.549 -10.617  -0.124  10.492  90.610
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 193.97208    7.75948  24.998  < 2e-16 ***
## working_env -63.40573    1.38920 -45.642  < 2e-16 ***
## is_smoker    40.89734    2.36124  17.320  < 2e-16 ***
## bmi           0.37064    0.09320   3.977 7.36e-05 ***
## children      4.84877    0.45792  10.589  < 2e-16 ***
## age           1.54199    0.04108  37.535  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.16 on 1332 degrees of freedom
## Multiple R-squared:  0.9132, Adjusted R-squared:  0.9129
## F-statistic:  2802 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Check the model assumption

```
check_model(fit_model, check = c("qq", "linearity", "homogeneity",
"outliers"))
```




```
check_normality(fit_model)

## Warning: Non-normality of residuals detected (p < .001).

check_heteroskedasticity(fit_model)

## Warning: Heteroscedasticity (non-constant error variance) detected (p <
.001).

check_outliers(fit_model)

## OK: No outliers detected.

check_autocorrelation(fit_model)

## OK: Residuals appear to be independent and not autocorrelated (p = 0.244).
```

In order to correct the non constant error of variance, we can use log transformation.

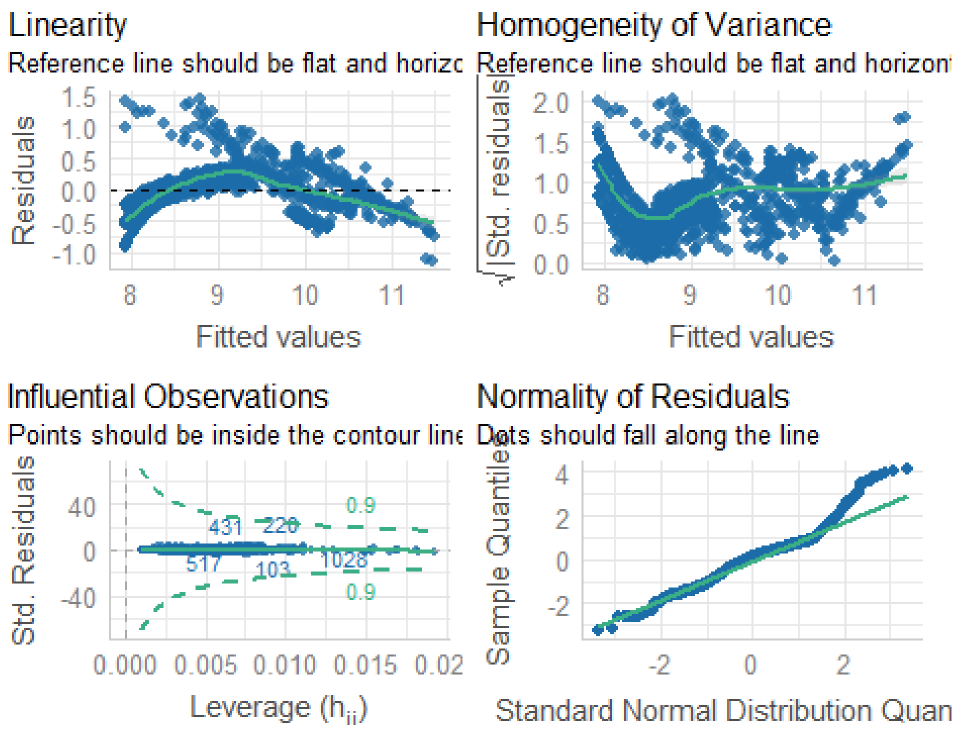
Log transformation

```
log_model <- lm(log(tot_claims) ~ age + bmi + children + is_smoker +
working_env, data = insurance_claims )
summary(log_model)
```

```
##
## Call:
## lm(formula = log(tot_claims) ~ age + bmi + children + is_smoker +
##     working_env, data = insurance_claims)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1242 -0.2302  0.0434  0.1943  1.4307
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.9522671  0.1349223  66.351   <2e-16 ***
## age          0.0291044  0.0007143  40.744   <2e-16 ***
## bmi          0.0003439  0.0016205   0.212    0.832
## children     0.1014024  0.0079623  12.735   <2e-16 ***
## is_smoker    0.5641608  0.0410574  13.741   <2e-16 ***
## working_env -0.7063503  0.0241555 -29.242   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3506 on 1332 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8546
## F-statistic:  1573 on 5 and 1332 DF,  p-value: < 2.2e-16
```

Check the model assumption

```
check_model(log_model, check = c("qq", "linearity", "homogeneity",
"outliers"))
```



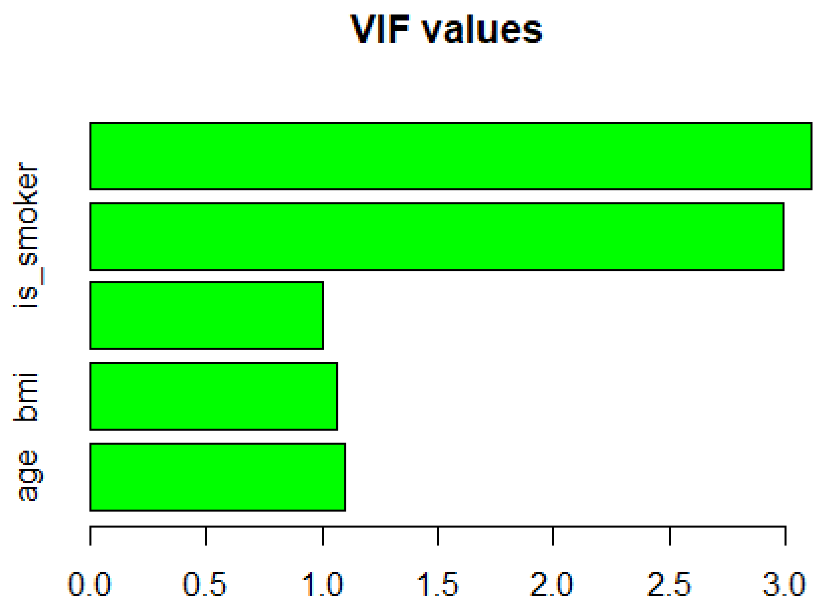

```
check_normality(log_model)

## Warning: Non-normality of residuals detected (p < .001).

check_heteroskedasticity(log_model)

## OK: Error variance appears to be homoscedastic (p = 0.232).

check_outliers(log_model)

## OK: No outliers detected.

check_autocorrelation(log_model)

## OK: Residuals appear to be independent and not autocorrelated (p = 0.494).
```

Still normality assumption is violated.

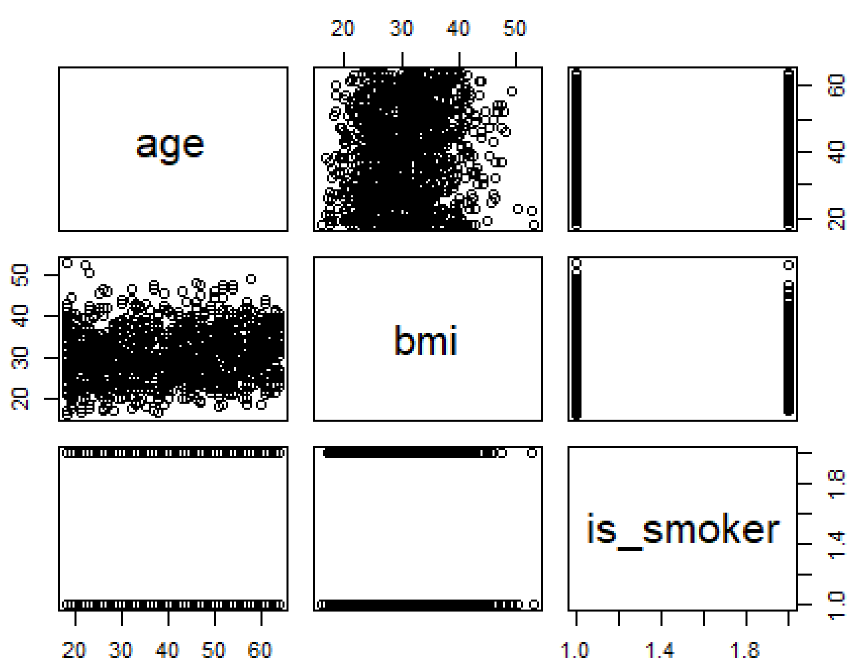
Multicolinearity

```
library(caTools)
library(car)

## Loading required package: carData

##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

library(quantmod)

## Loading required package: xts

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##
## Attaching package: 'xts'

## The following objects are masked from 'package:dplyr':
##
##     first, last

## Loading required package: TTR

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

library(xts)
library(zoo)
vif(red_model)

##         age         bmi    children   is_smoker working_env
##    1.095446    1.062040    1.001950    2.987664    3.113844

vif_values <- vif(red_model)
barplot(vif_values, main = "VIF values", horiz = TRUE, col = "green")
abline(v = 4, lwd = 3, lty = 2)
```




```
insurance_claims %>% select(age, bmi, is_smoker) %>% pairs()
```



According to the above plot, we can conclude that the variables are uncorrelated. Therefore, the multicolinearity does not effect when predict the annual claims.

Discussion

In the best fitted model, each and every exploratory variables should be uncorrelated. If we detect the multicolinearity of the fitted model, It would be directly effected when predict the response variable. So, in this multiple linear regression analysis, we didn't detect the multicolinearity.

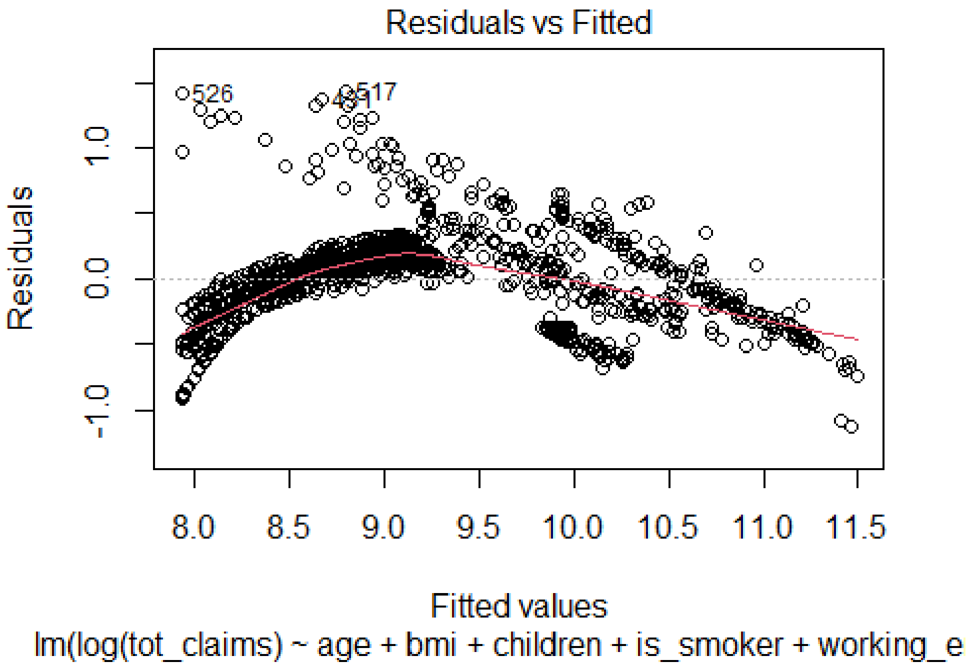
When checking the assumption, normality assumption was violated even use the log and boxcox transformation.

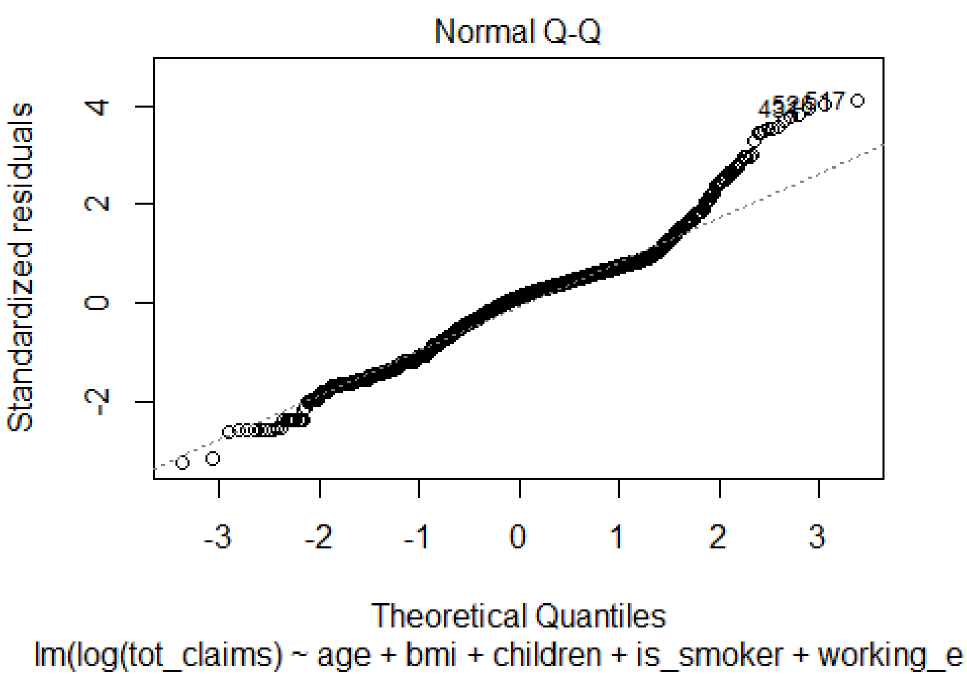
```
dim(insurance_claims)

## [1] 1338    7
```

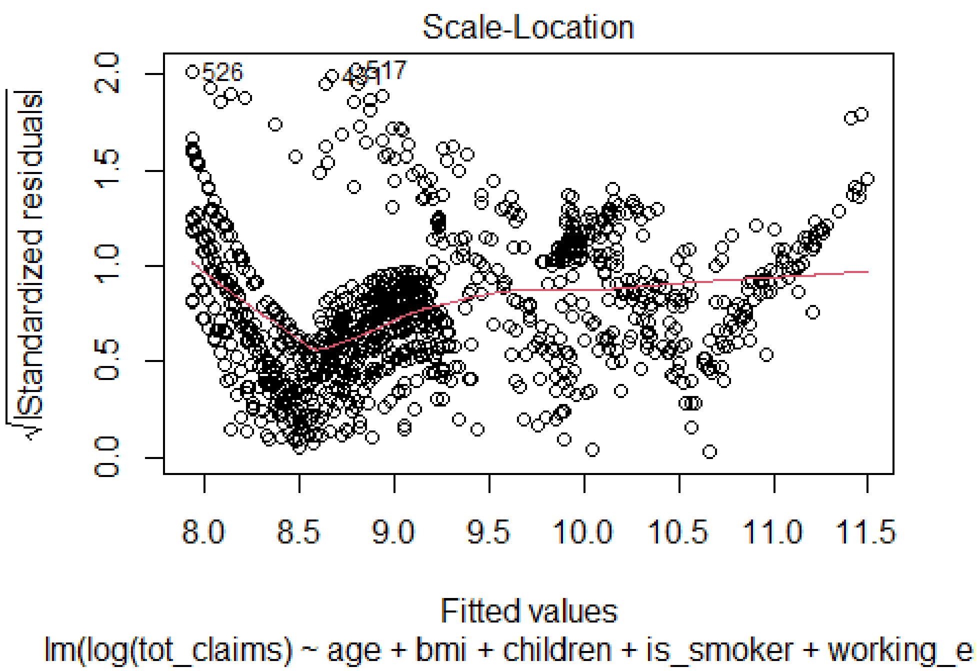
This data set contains 1338 observations. By Central Limit Therom for sufficiently large sample we can conclude that the residual will approximately normal.

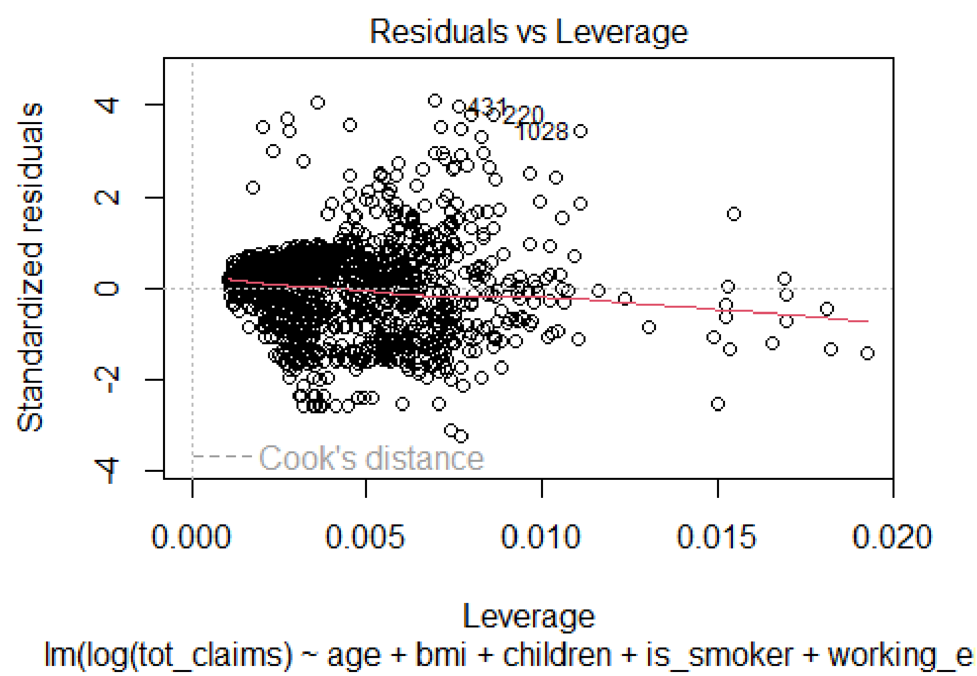
Conclusion

```
coef_log_model <- coef(log_model)
coef_log_model

##   (Intercept)           age           bmi      children     is_smoker 
##  8.9522670887  0.0291043745  0.0003438572  0.1014024291  0.5641607659 
##   working_env 
## -0.7063503004

plot(log_model) + geom_abline(intercept = coef_log_model[1], slope =
coef_log_model[2], color = "red")
```


Residuals vs Fitted
Residuals
Fitted values
lm(log(tot_claims) ~ age + bmi + children + is_smoker + working_e
526
517
-1.0
0.0
1.0
8.0
8.5
9.0
9.5
10.0
10.5
11.0
11.5


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(log(tot_claims) ~ age + bmi + children + is_smoker + working_e
517
-2
0
2
4
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
526
431
517
8.0
8.5
9.0
9.5
10.0
10.5
11.0
11.5
Fitted values
lm(log(tot_claims) ~ age + bmi + children + is_smoker + working_e


Residuals vs Leverage
Standardized residuals
-4
-2
0
2
4
431
220
1028
Cook's distance
0.000
0.005
0.010
0.015
0.020
Leverage
lm(log(tot_claims) ~ age + bmi + children + is_smoker + working_e

```
## NULL
```

Best fitted model

log(tot_claims) = 8.95226 + (0.0291)age + (0.00034)bmi + (0.1014)children + (0.56416)is_smoker