ST301_A1_S17_394

Janani Shashinika

2022-11-03

INTRODUCTION -

An insurance company wants to develop a model to predict the annual medical claims made by its customers. The data analytic team of the company assumes that the following variables may be useful in predicting the annual medical claims made by a given policyholder.

1. age : age of the policyholder
2. gender : the policyholder's gender - female, male
3. bmi : body mass index of the policyholder
4. num dependents : number of dependents covered by the health insurance (spouse and children below age 18)
5. is smoker : smoking status of the policyholder - yes, no
6. working env : working environment of the policyholder - construction site, factory, office
7. tot claims : total amount of claims made by the policyholder

We have given six independent variables (including 3 categorical variables- sex, is_smoker, working_env) and one response variable (tot_claims) since only one response variable and more than one independent variables, we have to use Multiple Linear Regression to find the predictions for this problem.

We use 'insurance_claims' dataset to explore the relationships between the response variable and the other 06 variables.

```
ins_claims = read.csv("insurance_claims.csv")
head(ins_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
```

EXPLORATORY ANALYSIS -

Here, we are going to convert the Categorical data(sex, is_smoker and working_env) into Numerical form to make the predictive models.

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

ins_claims$is_smoker = as.numeric(factor(ins_claims$is_smoker,labels = c("no","yes")))
```

```
ins_claims$working_env = as.numeric(factor(ins_claims$working_env,labels = c("factory","office","constr

head(ins_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
```

```
dim(ins_claims)
```

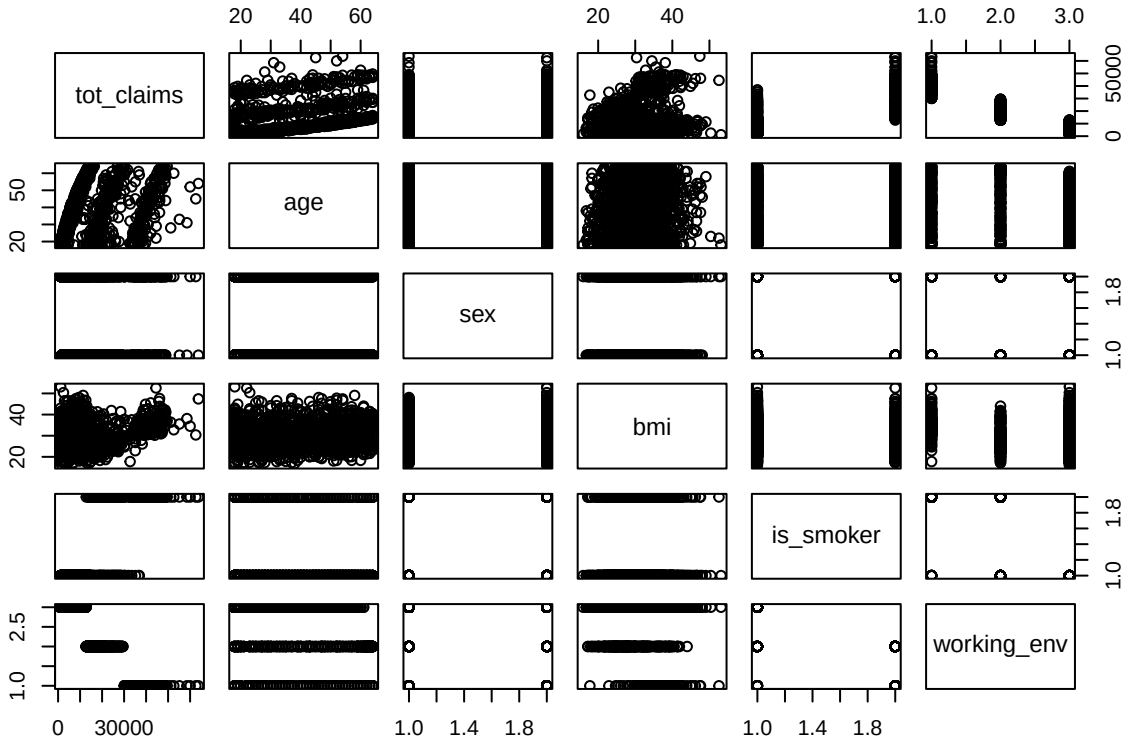
```
## [1] 1338    7
```

Number of observations in the data set = 1338

```
str(ins_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 ...
```

```
pairs(~tot_claims+age+sex+bmi+is_smoker+working_env,data = ins_claims)
```



We can assume that there are no missing values in the dataset according to the summary statistics.

MODEL FITTING -

Here let's use Forward selection method that based on adjusted R squared value as the variable selection method.

Iteration 01 :

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

```
## [1] 0.002536334
```

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

```
## [1] 0.08872432
```

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

```
## [1] 0.03862008
```

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

```
## [1] 0.003878717
```

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

```
## [1] 0.6194802
```

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

```
## [1] 0.8614734
```

working_env is added - R squared value = 0.8614734

Iteration 02 -

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

```
## [1] 0.886051
```

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

```
## [1] 0.8614025
```

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

```
## [1] 0.865276
```

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

```
## [1] 0.8643907
```

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

```
## [1] 0.8679371
```

Age is added - R squared value= 0.886051

Iteration 03 -

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

```
## [1] 0.8859661
```

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

```
## [1] 0.888348
```

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

```
## [1] 0.8883291
```

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

```
## [1] 0.9013036
```

is_smoker is added - R squared value = 0.9013036

Iteration 04 -

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

```
## [1] 0.9012491
```

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

```
## [1] 0.9063182
```

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

```
## [1] 0.903542
```

bmi is added - R squared value = 0.903542

Iteration 05 -

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

```
## [1] 0.9063071
```

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

```
## [1] 0.9085069
```

children is added - R squared value = 0.9085069

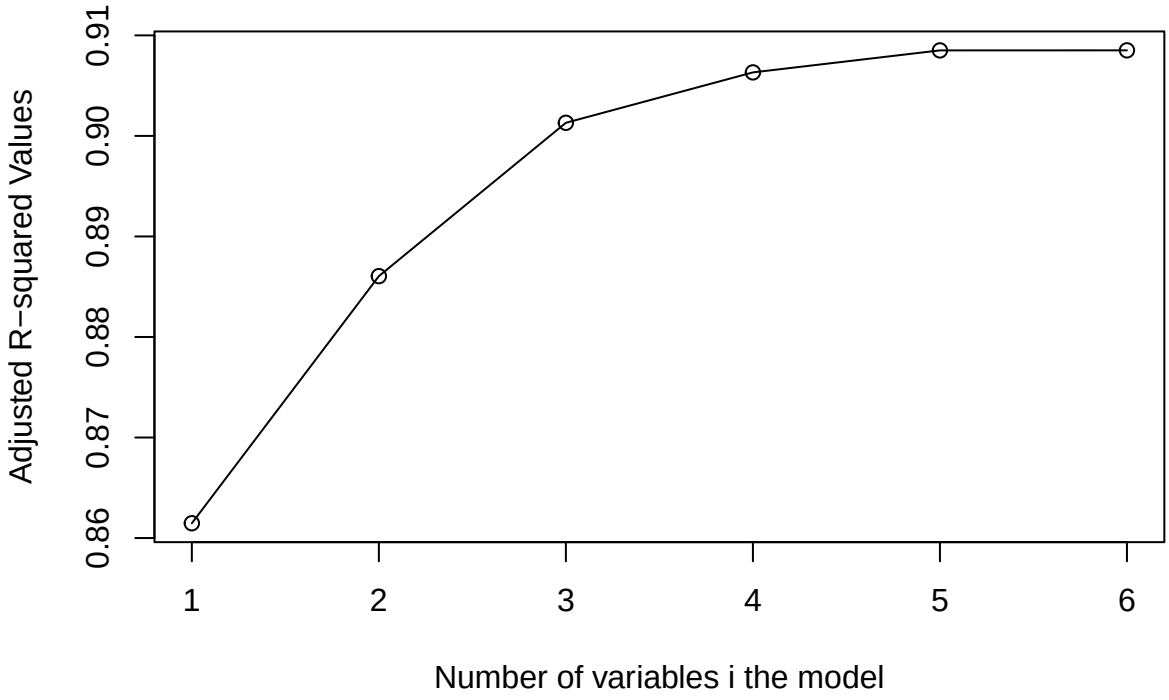
Iteration 06 -

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

```
## [1] 0.9085106
```

Graph of the Adjusted R squared Values -

```
plot(c(1,2,3,4,5,6),c(0.8614734,0.886051,0.9013036,0.9063182,0.9085069,0.9085106), xlab = "Number of va
```



Here, We can remove the sex variable since there is no significant increment of Adjusted R squared value in Iteration 06.

Forward Selection based on F-test -

```
con.model = lm(tot_claims ~ 1, data = ins_claims)
add1(con.model,scope = tot_claims~age+sex+bmi+children+is_smoker+working_env,test="F")
```

```
## Single term additions
##
## Model:
## tot_claims ~ 1
##             Df  Sum of Sq        RSS   AIC   F value    Pr(>F)
## <none>                   1.9607e+11 25160
## age          1 1.7530e+10 1.7854e+11 25037  131.1740 < 2.2e-16 ***
## sex          1 6.4359e+08 1.9543e+11 25158    4.3997   0.03613 *
## bmi          1 7.7134e+09 1.8836e+11 25109   54.7093 2.459e-13 ***
## children     1 9.0660e+08 1.9517e+11 25156    6.2060   0.01285 *
## is_smoker    1 1.2152e+11 7.4554e+10 23868 2177.6149 < 2.2e-16 ***
## working_env  1 1.6893e+11 2.7141e+10 22516 8315.5757 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the p-values age, is_smoker and working_env are significant. First, let's add working_env to the model.

```
add1(update(con.model, ~ . +working_env), scope = tot_claims~age+sex+bmi+children+is_smoker+working_env
```

```
## Single term additions
##
## Model:
## tot_claims ~ working_env
##           Df  Sum of Sq        RSS   AIC  F value    Pr(>F)
## <none>                  2.7141e+10 22516
## age        1 4832144066 2.2309e+10 22256 289.1614 < 2.2e-16 ***
## sex        1    6437096 2.7135e+10 22518   0.3167    0.5737
## bmi        1  764784429 2.6376e+10 22480  38.7084 6.573e-10 ***
## children   1  591471755 2.6550e+10 22489  29.7410 5.877e-08 ***
## is_smoker  1 1285786035 2.5855e+10 22453  66.3894 8.429e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can add age to the model since it has most significant p value.

```
add1(update(con.model, ~ . +working_env+age), scope = tot_claims~age+sex+bmi+children+is_smoker+working
```

```
## Single term additions
##
## Model:
## tot_claims ~ working_env + age
##           Df  Sum of Sq        RSS   AIC  F value    Pr(>F)
## <none>                  2.2309e+10 22256
## sex        1     105032 2.2309e+10 22258   0.0063    0.9368
## bmi        1  466076521 2.1843e+10 22230  28.4644 1.120e-07 ***
## children   1  462388978 2.1847e+10 22230  28.2344 1.258e-07 ***
## is_smoker  1 3000631878 1.9308e+10 22065 207.3109 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can add is_smoker to the model since it has the p value now.

```
add1(update(con.model, ~ . +working_env+age+is_smoker), scope = tot_claims~age+sex+bmi+children+is_smok
```

```
## Single term additions
##
## Model:
## tot_claims ~ working_env + age + is_smoker
##          Df Sum of Sq        RSS   AIC F value    Pr(>F)
## <none>                1.9308e+10 22065
## sex       1   3823000 1.9305e+10 22067   0.264    0.6075
## bmi       1 994780328 1.8314e+10 21996  72.407 < 2.2e-16 ***
## children  1 452064148 1.8856e+10 22035  31.957 1.925e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we can add bmi to the model as it has the most significant p value.

```
add1(update(con.model, ~ . +working_env+age+is_smoker+bmi), scope = tot_claims~age+sex+bmi+children+is_
```

```
## Single term additions
##
## Model:
## tot_claims ~ working_env + age + is_smoker + bmi
##          Df Sum of Sq        RSS   AIC F value    Pr(>F)
## <none>                1.8314e+10 21996
## sex       1  11565653 1.8302e+10 21997  0.8417    0.3591
## children  1 441273934 1.7872e+10 21965 32.8875 1.207e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now children can be added to the model as it has the significant p value.

```
add1(update(con.model, ~ . +working_env+age+is_smoker+bmi+children), scope = tot_claims~age+sex+bmi+chil
```

```
## Single term additions
##
## Model:
## tot_claims ~ working_env + age + is_smoker + bmi + children
##        Df Sum of Sq        RSS   AIC F value Pr(>F)
## <none>              1.7872e+10 21965
## sex     1  14140693 1.7858e+10 21966  1.0539 0.3048
```

Here we can not add the sex variable to the model,because it's p value is greater than 0.05. Therefore we can remove the sex variable from the model, according to the forward selection method selection method based on F test.

Reduced Model -

```
reduced_model = lm(tot_claims ~ working_env+age+is_smoker+bmi+children,data = ins_claims)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = tot_claims ~ working_env + age + is_smoker + bmi +
##     children, data = ins_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 ***
## working_env -12170.053    252.355 -48.226  < 2e-16 ***
## age            160.017      7.463  21.442  < 2e-16 ***
## is_smoker     6942.301    428.930  16.185  < 2e-16 ***
## bmi            144.977     16.929   8.564  < 2e-16 ***
## children       477.031     83.182   5.735 1.21e-08 ***
## ---
```

```
## 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
```

We can say that the model is significant because the p value of this reduced model is 2.2e-16.as concluded from the adjusted R squared value of 0.9085 ,there is a strong relationship between the variables.

```
full_model = lm(tot_claims ~ . , data = ins_claims)
summary(full_model)
```

```
##
## Call:
## lm(formula = tot_claims ~ ., data = ins_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
```

VALIDATION -

Using the partial f test we can check whether the reduced model is adequate or not. Null hypothesis(Ho) : Reduced model is not adequate Alternative hypothesis : Reduced model is not adequate

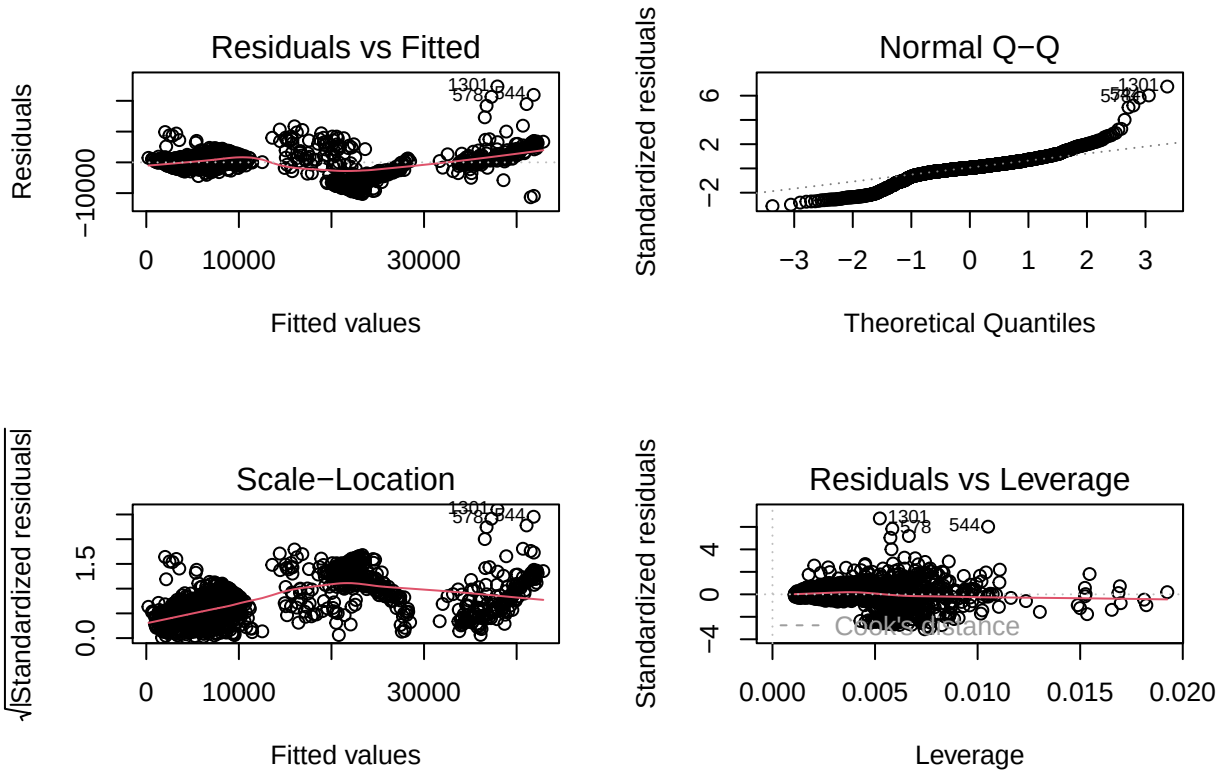
```
anova(reduced_model,full_model)
```

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

Considering the anova table, the p-value of this fitted regression model is 0.3048. Since it is greater than 0.05,we do not have enough evidence to reject the null hypothesis(Ho). Therefore we can say, there is enough evidence to say that the reduced model is adequate.

Residual Analysis -

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




```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(performance)
check_normality(reduced_model)
```

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

The points on the Normal Q-Q plot(the plot of the standardized residuals vs the theoretical quantiles) provide an indication of the normality of the residuals.If the error terms are normally distributed, the points will fall on the 45-degree reference line.But in here, it is slightly deviated in the bottom end and the upper end(two tails).Only the upper Middle part is aligned with the 45-degree reference line.So they are not normally distributed and the normality assumption is violated.

```
library(tidyverse)
library(performance)
check_heteroscedasticity(reduced_model)
```

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

The bottom left plot shows that the assumption of constant variance is violated in this dataset, as the line is not horizontal but shows a some different pattern.

```
library(tidyverse)
library(performance)
check_autocorrelation(reduced_model)
```

```
## OK: Residuals appear to be independent and not autocorrelated (p = 0.938).
```

The error terms are uncorrelated.

```
library(tidyverse)
library(performance)
check_outliers(reduced_model)
```

```
## OK: No outliers detected.
```

Therefore,No outliers were detected in this model.

```
cor(ins_claims)
```

```
##                     age         sex          bmi    children    is_smoker
## age          1.00000000 -0.02085587  0.109271882  0.04246900 -0.025018752
## sex         -0.02085587  1.00000000  0.046371151  0.01716298  0.076184817
## bmi          0.10927188  0.04637115  1.000000000  0.01275890  0.003750426
## children     0.04246900  0.01716298  0.012758901  1.00000000  0.007673120
## is_smoker   -0.02501875  0.07618482  0.003750426  0.00767312  1.000000000
## working_env -0.15505210 -0.06788169 -0.147128907 -0.01409200 -0.795243654
## tot_claims   0.29900819  0.05729206  0.198340969  0.06799823  0.787251430
##             working_env  tot_claims
## age         -0.15505210  0.29900819
## sex         -0.06788169  0.05729206
## bmi         -0.14712891  0.19834097
## children    -0.01409200  0.06799823
## is_smoker   -0.79524365  0.78725143
## working_env  1.00000000 -0.92821172
## tot_claims  -0.92821172  1.00000000
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.2
```

```
## 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(MASS)
```

```
##
## Attaching package: 'MASS'
```

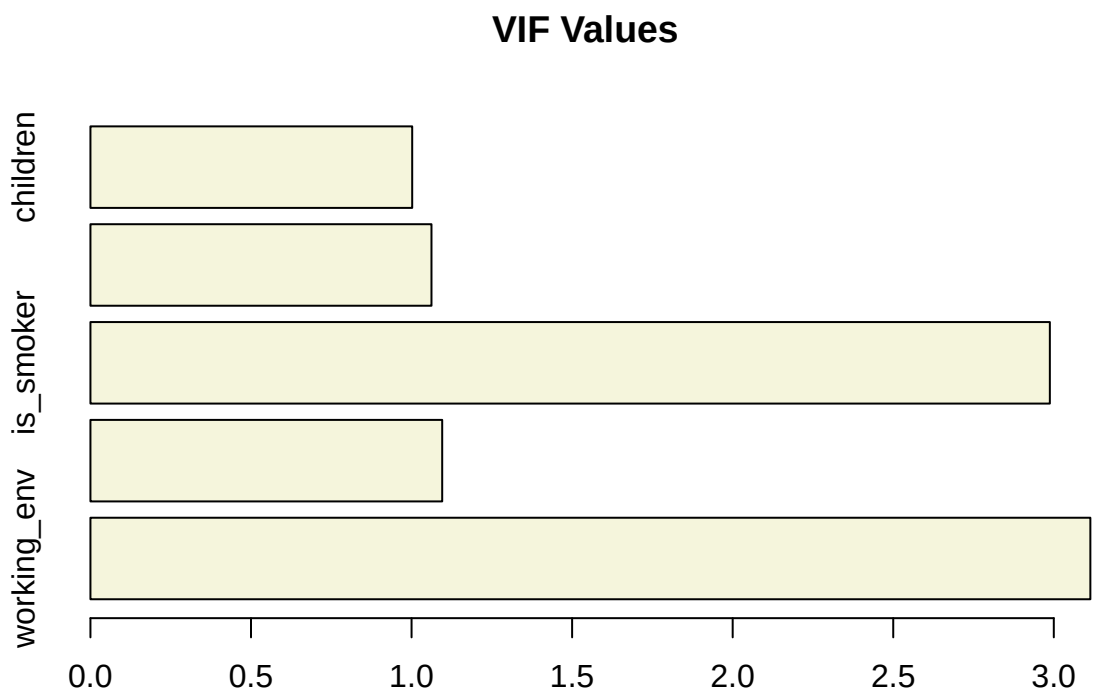
```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
vif(reduced_model)
```

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

There is no any violation in multicollinearity in the model, since the vif scores are far below to 5.

```
vifval = vif(reduced_model)
barplot(vifval,main = "VIF Values", horiz = TRUE, col = "beige")
abline(v=4, lwd=3,lty=2)
```



Considering the 04 assumptions two are violated- Normality and Heteroscedasticity

DISCUSSION AND CONCLUSION -

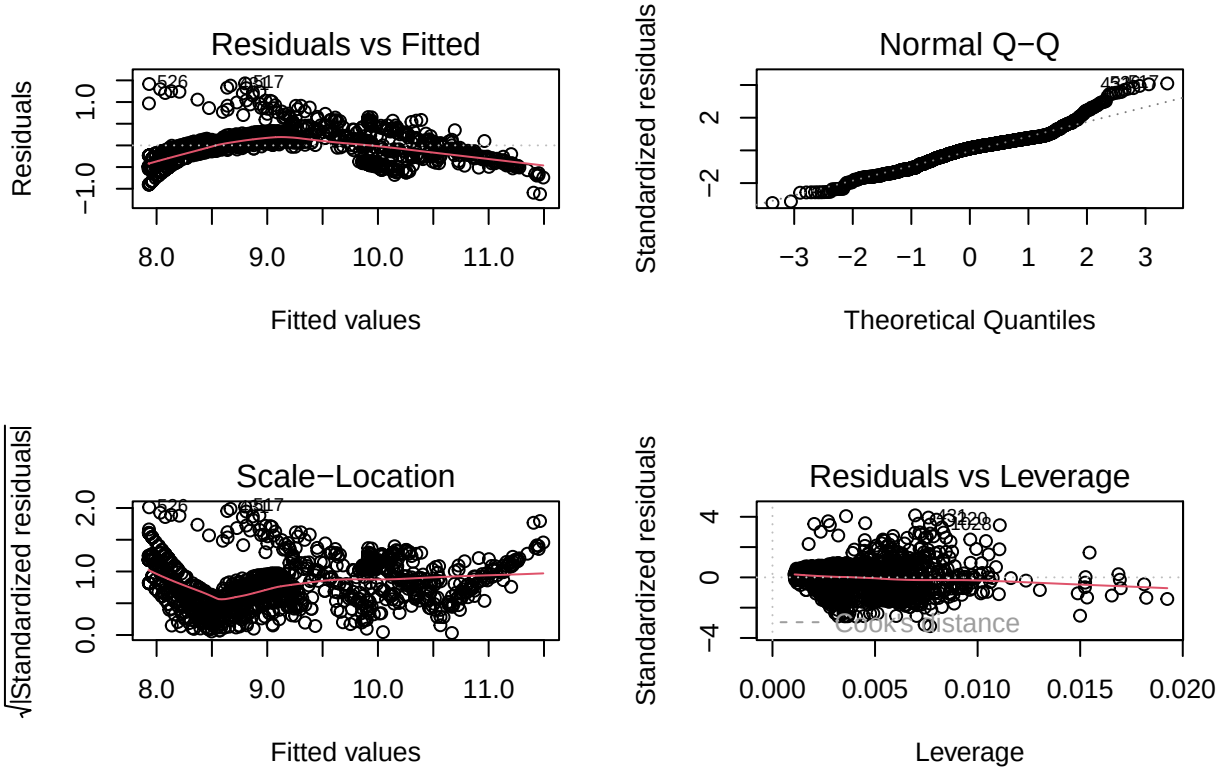
As the normality assumption and the heteroscedasticity are violated let's use log transformation method to fix the violations here.

```
ins_claims2 = lm(log(tot_claims) ~ age+working_env+is_smoker+bmi+children,data = ins_claims)
summary(ins_claims2)
```

```
##
## Call:
## lm(formula = log(tot_claims) ~ age + working_env + is_smoker +
##     bmi + children, data = ins_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 ***
## working_env -0.7063503  0.0241555 -29.242   <2e-16 ***
## is_smoker    0.5641608  0.0410574  13.741   <2e-16 ***
## bmi          0.0003439  0.0016205   0.212    0.832
## children     0.1014024  0.0079623  12.735   <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
```

```
par(mfrow = c(2,2))
plot(ins_claims2)
```




```
library(tidyverse)
library(performance)
check_heteroscedasticity(ins_claims2)
```

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

The constant variance(homoscedastic) among the residuals is detected now and the violation has fixed.

```
library(tidyverse)
library(performance)
check_autocorrelation(ins_claims2)
```

```
## OK: Residuals appear to be independent and not autocorrelated (p = 0.498).
```

Here, the autocorrelation assumption is not violated as previous.

```
library(tidyverse)
library(performance)
check_outliers(ins_claims2)
```

```
## OK: No outliers detected.
```

Here, we can not define any outliers.

```
library(tidyverse)
library(performance)
check_normality(ins_claims2)
```

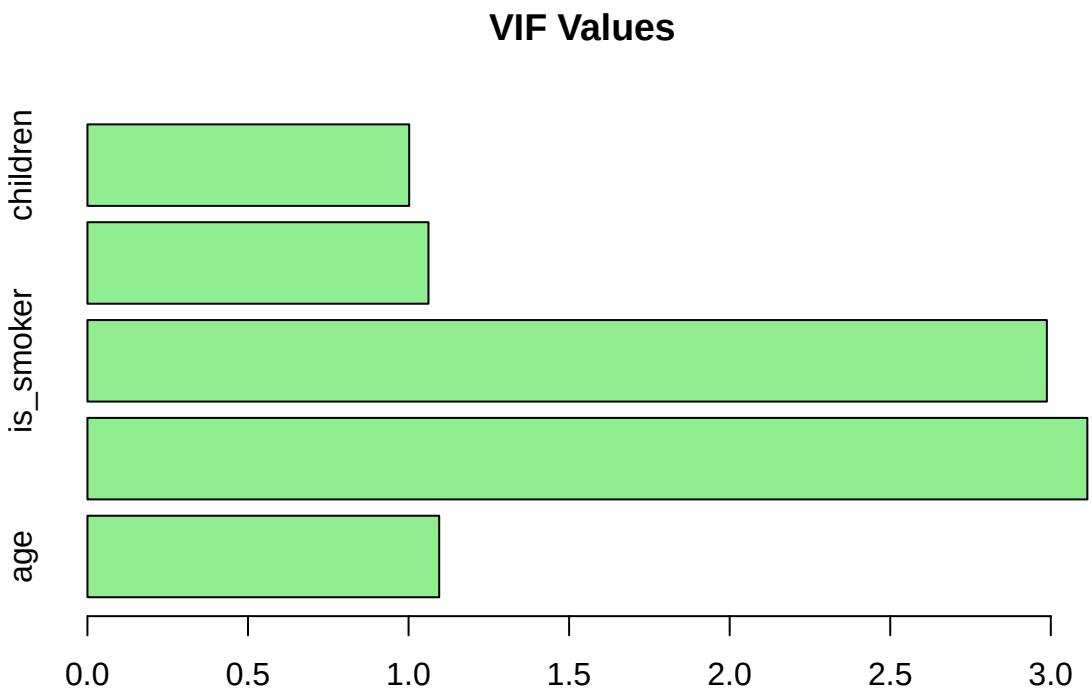
```
## Warning: Non-normality of residuals detected (p < .001).
```

Normality assumption is still violated.

```
library(car)
library(quantmod)
library(MASS)
vif(ins_claims2)
```

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

```
vifval = vif(ins_claims2)
barplot(vifval,main = "VIF Values", horiz = TRUE, col = "light green")
abline(v=4, lwd=3,lty=2)
```



The vif scores are far below 4.Therefore, the independent variables are not highly correlated and multicollinearity assumption is not violated. Considering the Central Limit Theorem,that tha distrbution of residuals will approximately normal. Finally, as we have a large sample of data in this case,we can approximate the normality here. Assuming all factors and we can get final output like as follows.

```
coef(ins_claims2)
```

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

Therefore the final model is,

log(tot_claims) = 8.9522670887 + (-0.7063503004)working_env + (0.0291043745)age + (0.5641607659)is_smoker + (0.0003438572)bmi + (0.1014024291)children