ECONOMETRICS PROJECT

INDIAN INSTITUTE OF FOREIGN TRADE

भारतीय विदेश व्यापार संस्थान

SOMEWRIT SEKHAR MAITI Roll No. 47

Introduction:

With this project we are analyzing the effects of smoking habits in individuals and whether gender plays a role in determining the insurance premiums.

We have taken this dataset from archives of University of Florida, USA. The study includes the age, BMI, smoking habits, gender, region and the premium paid by each individual.
The dataset includes data of 1338 individuals living in a certain region in Florida.

Research Questions:

1) Does smoking affect health premium payments?
2) Does gender play a role in premium charges paid by the customer?

Methodology:

We have performed an OLS regression analysis to verify our research questions.

We have considered premium paid as our dependent variable (Y)

Our explanatory variables include:

1) Age
2) BMI
3) Smoking Habits (Dummy Variable, 1 for "Smoker" and 0 for "Non-Smoker")
4) Gender (Dummy Variable, 1 for "Male", 0 for "Female")
5) Children (Discrete Variable takes values between 0 to 5)

Model Building:

$$\text{Premium Paid} = \beta0 + \beta1(\text{Age}) + \beta2(\text{BMI}) + \beta3(\text{Smoking\_dummy}) + \beta4(\text{Gender\_dummy}) + \beta5(\text{Children}) + \varepsilon$$

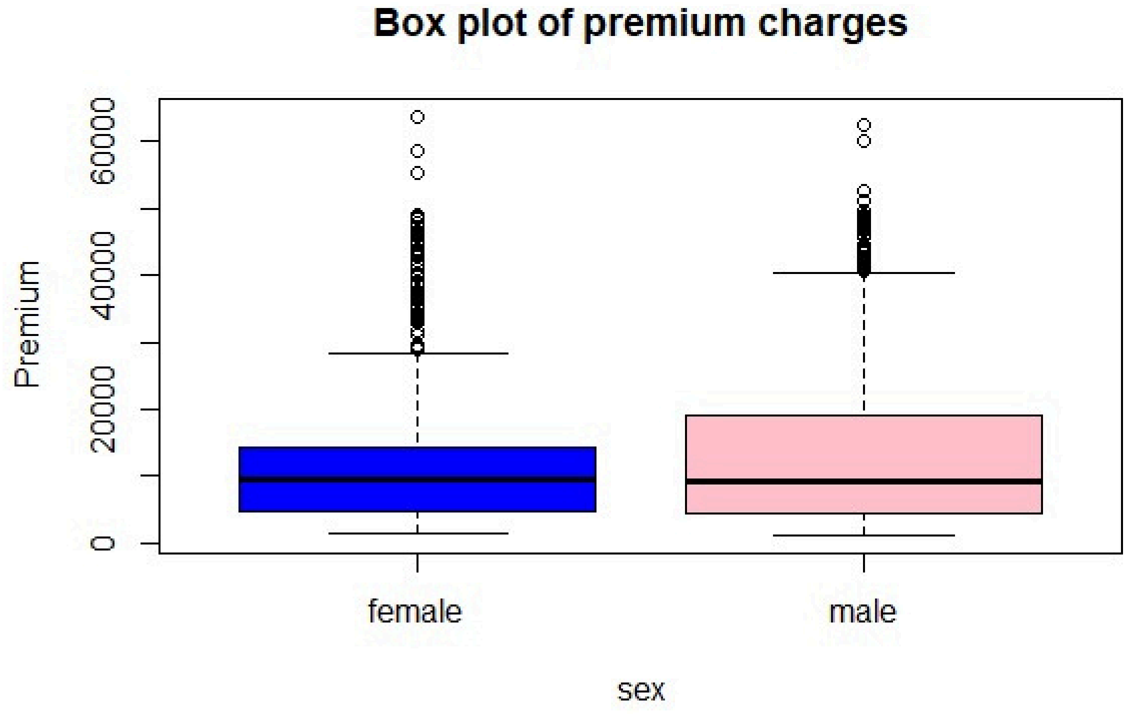
DESCRIPTIVE STATISTICS

Mean of Premium – 13270.42

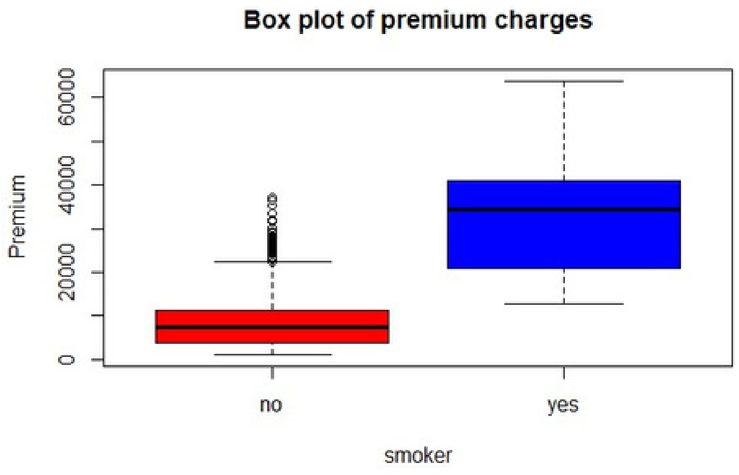
Mean Age – 39.21

Mean BMI – 30.66

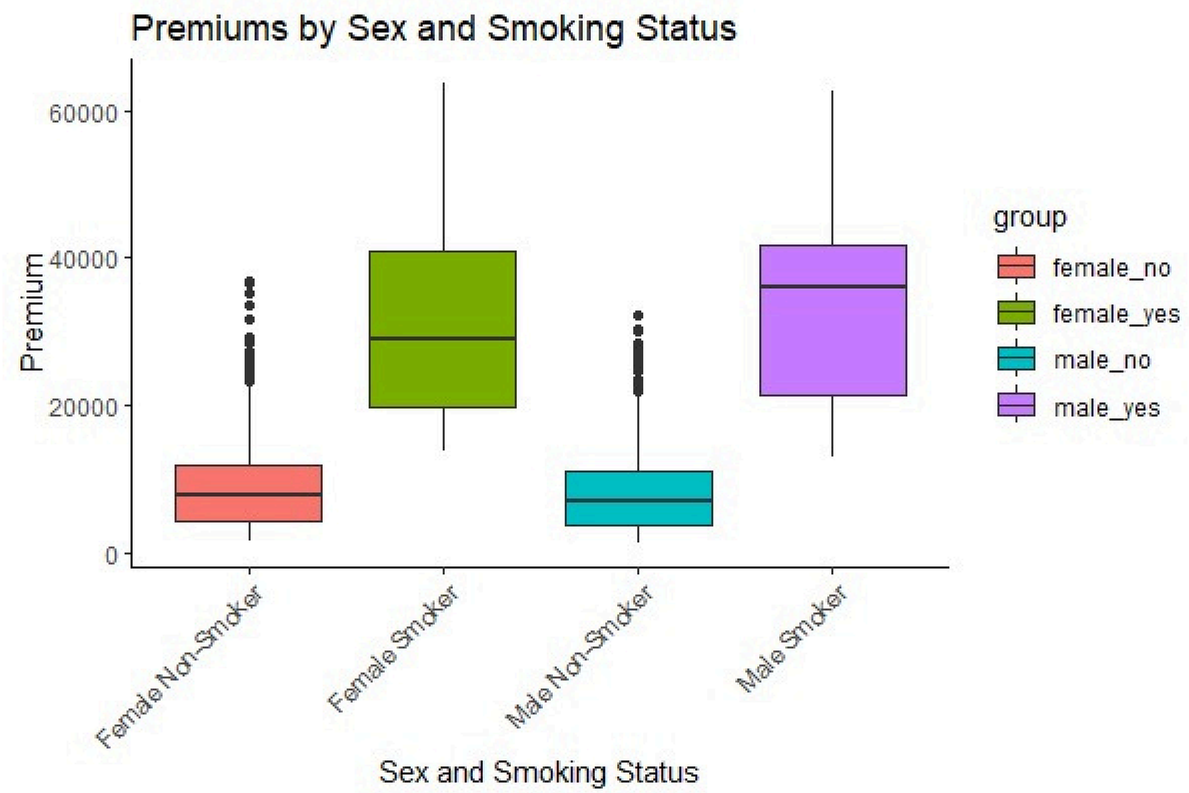
BOX PLOTS



The given diagram is a box plot comparing premium charges based on gender (female and male). The box represents the inter-quartile range i.e. the middle 50% of the data. The line inside the box is the median showing the central tendency of the data. The whiskers extend to the smallest and largest values within 1.5 times the interquartile range. The dots above the whiskers represents outliers which are values significantly higher than the rest. So, we can draw a conclusion from the data that males on average pay a higher premium than females, males also exhibit greater variability in premium charges, with some individuals paying extremely high premiums (as shown by the outliers) the disparity could result from the differences in risk profiles, healthcare habits or socio-economic factors between genders.



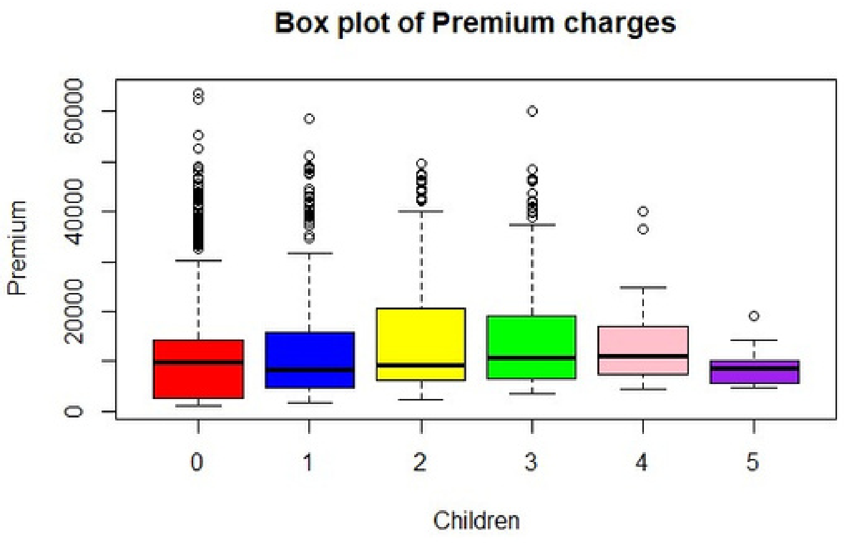
The boxplots shows that smokers pay a higher premium as compared to the non smokers. In the next box plot we will show how premium payments vary between males and females who are smokers and non-smokers so that we can get a better picture to explain the variability in the data.



Female non-smokers have the lowest median premium, whereas male smokers have the highest median premium, followed by female smokers, showing the trend that smokers indeed have to pay a higher premium.

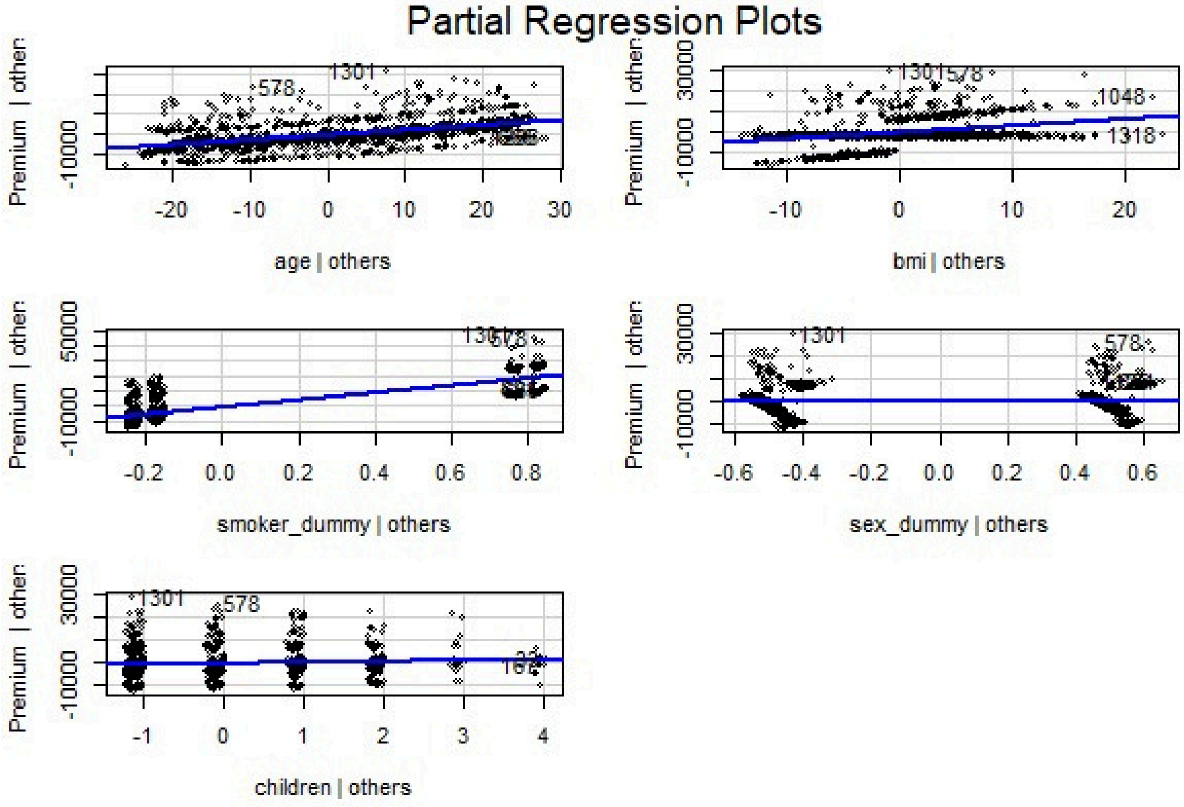
Non-smokers have fewer outliers compared to smokers indicating more consistency in their premiums.

Therefore, smoking habits are a major factor when determining the premiums, the same cannot be said of gender yet.

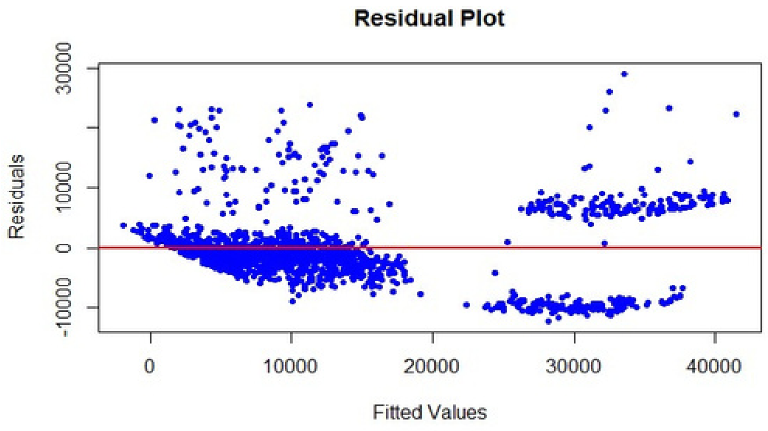


Similarly, the boxplots for premium payments vs number of children shows a peculiar result that median is consistent with the fact that having more children results in higher premium payments. Although to verify this we need to perform the regression analysis to be fully sure of its impact.

MODEL DIAGNOSTICS



The partial regression plots show that the variables all have a linear relation with the dependent variable (premium paid by the individual). All the variables show a positive trendline as opposed to the sex_dummy variable which may indicate that the variable has little to explain the variability in the premium paid.



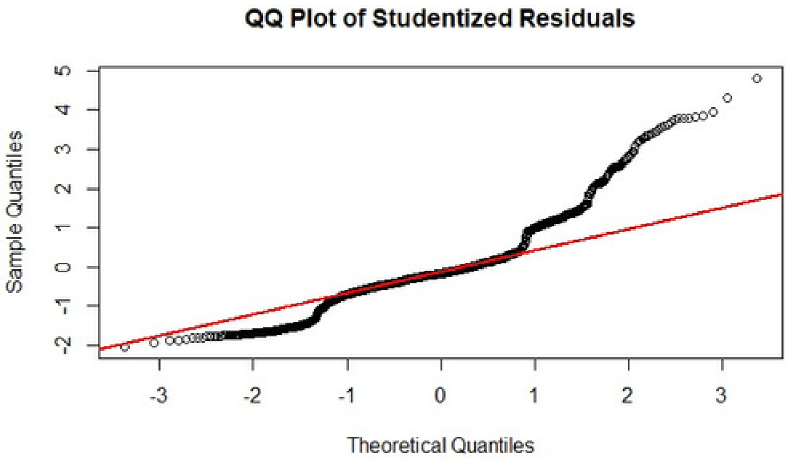
The residual plot shows that the residuals are not evenly scattered showing a discernible pattern. The residual plot suggests that the relation between predictors and the response variables is not strictly linear and also shows heteroscedasticity (the spread of the residuals increases as fitted values grow, there is a funnel shaped pattern this is a sign of heteroscedasticity) which is incompatible with the OLS assumptions.

We conducted the Breusch Pagan test to check for Heteroscedasticity

STUDENTIZED BREUSCH-PAGAN TEST

BP = 113.48, df = 5, p-value < 2.2e-16

The statistic shows there is heteroscedasticity present.



We can see from the QQ plot that the data points significantly deviate from the red line specially at both tails. This suggests that the residuals are not normally distributed. The data shows significant departures from normality as evidenced by the QQ plot. It may be necessary to address this issue by various transformations.

After conducting the Shapiro-Wilk test we get the following results.

SHAPIRO-WILK NORMALITY TEST

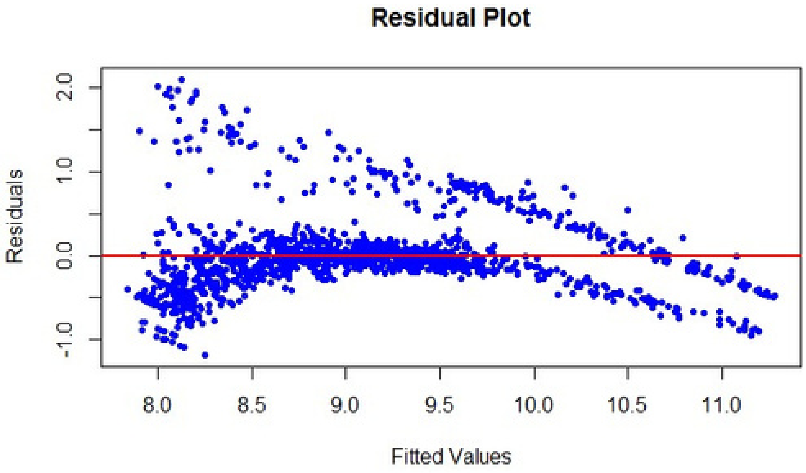
W = 0.98094, p-value = 2.629e-12

From this statistic we can see that the residuals are not normally distributed.

To remedy this assumption we have tried to transform the dependent variable using various robust methods such as :

1) Log Transformation of the Dependent variable
2) Box-Cox Transformation

TESTING FOR Log MODEL



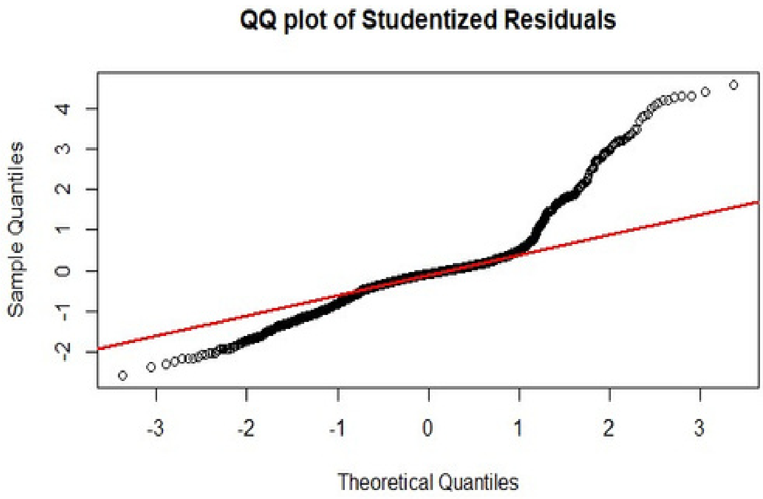
The residual plot for the log transformed model shows residuals are not randomly scattered around 0. There is a clear pattern or trend in the data which indicates that the variance of the residuals changes with fitted values suggesting heteroscedasticity.

We checked for heteroscedasticity using the Breusch Pagan test and got the following results.

STUDENTIZED BREUSCH-PAGAN TEST

BP = 90.862, df = 5, p-value < 2.2e-16

This statistic shows that heteroscedasticity is present.



We can see from the plot that the points in the middle roughly follow the red line, indicating some normality in the central region. However, at both ends or tails the points deviate significantly suggesting heavy tails or outliers in the data. This indicates that the residuals are not perfectly normal, as the distribution has extreme values. We checked for normality using the Shapiro Wilk test and got the following results.

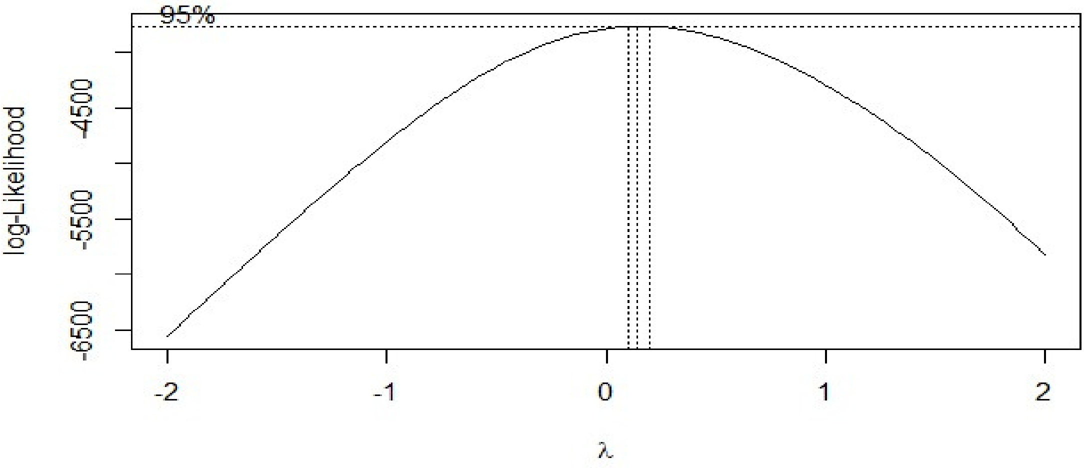
SHAPIRO-WILK NORMALITY TEST

W = 0.98316, p-value = 2.291e-11

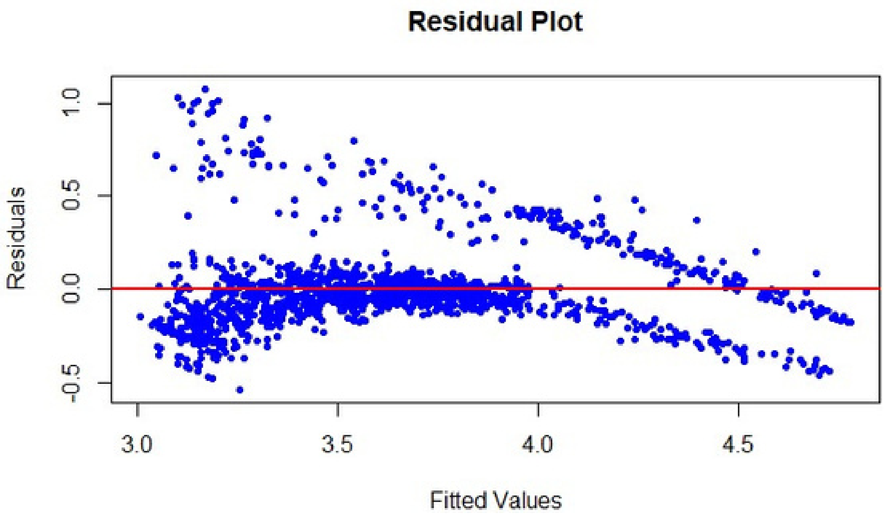
From this statistic we can see that the residuals are not normally distributed.

Since, log transformation did not provide the necessary results we proceed with the Box-Cox Transformation.

Testing for Box-Cox Transformed Model



The best lambda value has been found out to be 0.1414141 using the maximum likelihood estimation. This value is used as the exponent to the dependent variable while conducting the box-cox diagnostic tests.



The residual plot for the log transformed model shows residuals are not randomly scattered around 0. There is a clear pattern or trend in the data which indicates that the variance of the residuals changes with fitted values suggesting heteroscedasticity.

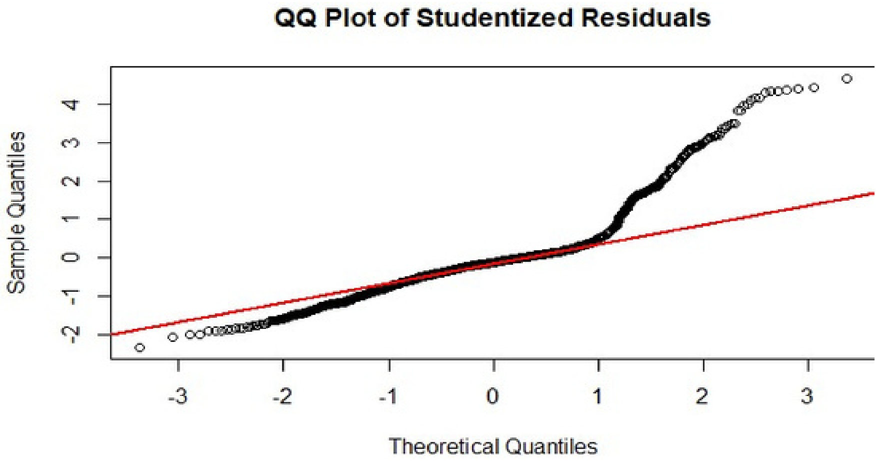
We checked for

heteroscedasticity using the Breusch Pagan test and got the following results.

STUDENTIZED BREUSCH-PAGAN TEST

BP = 62.91, df = 5, p-value = 7.088e-13

The test shows that heteroscedasticity is present.



We can see from the plot that the points till 1 on the Y-axis, roughly follow the red line, indicating some normality in this region. However, at the other right tail the points deviate significantly suggesting outliers in the data. This indicates that the residuals are not perfectly normal, as the distribution has extreme values. We checked for normality using the Shapiro Wilk test and got the following results.

SHAPIRO-WILK NORMALITY TEST

W = 0.98075, p-value = 2.271e-12

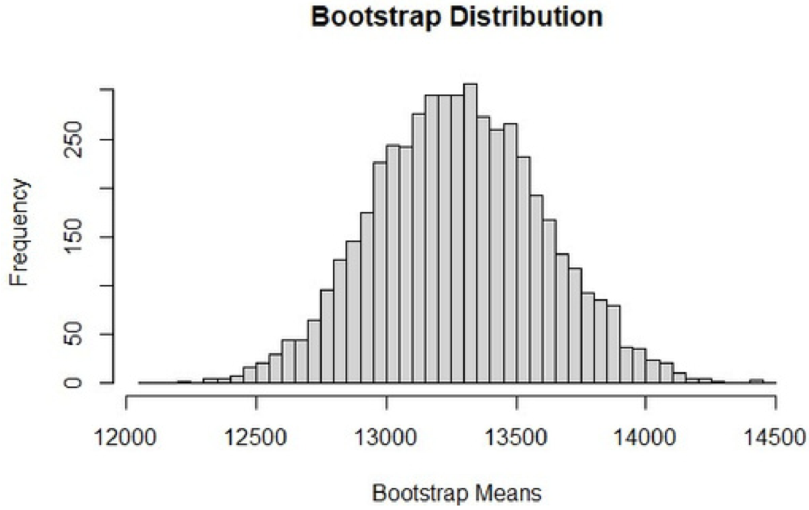
From this statistic we can see that the residuals are not normally distributed.

We have stopped our transformation process after the Box-Cox transformation as none of the transformations yielded the necessary results, so we are left to consider that the model neither normal nor homoscedastic.

To remedy this we consider bootstrapping to cure normality to check if bootstrapping leads to any changes in our estimates.

BOOTSTRAPPING

We ran the bootstrapping method in R to get 5000 samples of the dataset and checked for normality.



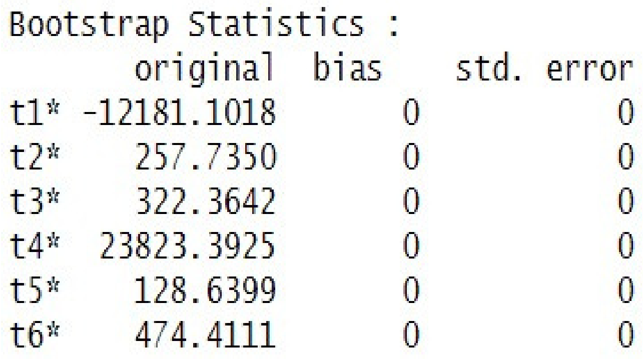
From the Histogram we can see that, the data has achieved normality. The mean of the bootstrap samples represents an accurate estimate of the population mean. Additionally the spread (width) of the distribution can help determine the variability of the estimate and construct the confidence intervals. The conclusion is that the population mean likely falls within of the range of the observed bootstrap means (around 12000 to 14500).

Bootstrap Mean = 13278.31   Population Mean (Premium) = 13270.43

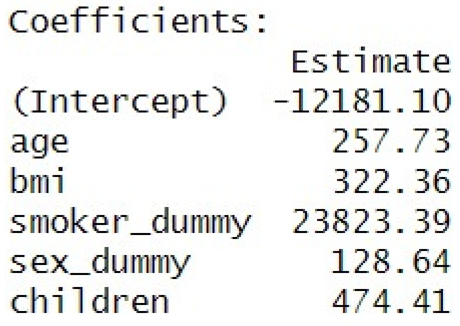
SHAPIRO-WILK NORMALITY TEST

W = 0.9995, p-value = 0.2133

Since, the p-value is larger than 0.05, we cannot reject the null hypothesis that data is normal. Therefore, we fail to reject the null hypothesis, thereby proving that after bootstrapping the distribution has attained normality.



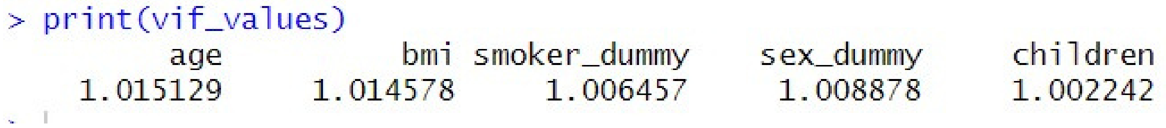
```
Bootstrap Statistics :
      original  bias    std. error
t1* -12181.1018     0            0
t2*    257.7350     0            0
t3*    322.3642     0            0
t4*  23823.3925     0            0
t5*    128.6399     0            0
t6*    474.4111     0            0
```



```
Coefficients:
                Estimate
(Intercept)   -12181.10
age              257.73
bmi              322.36
smoker_dummy   23823.39
sex_dummy        128.64
children         474.41
```

From the coefficients achieved after bootstrapping and coefficients after regression are same with the same sign, we can conclude that there is strong consistency. It also states that sex_dummy is irrelevant as the p-value has found out to be 0.7.

MULTICOLLINEARITY

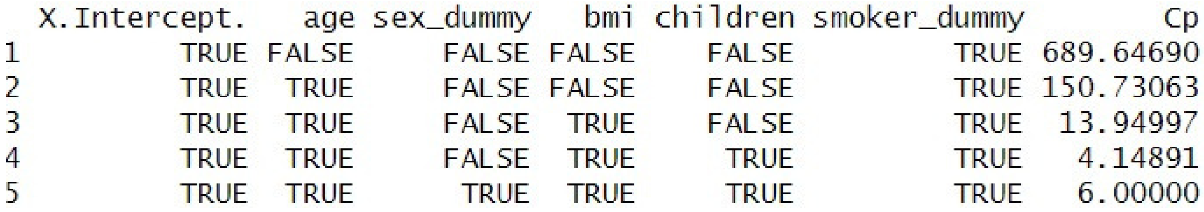


```
> print(vif_values)
         age          bmi smoker_dummy    sex_dummy     children 
    1.015129     1.014578     1.006457     1.008878     1.002242 
```

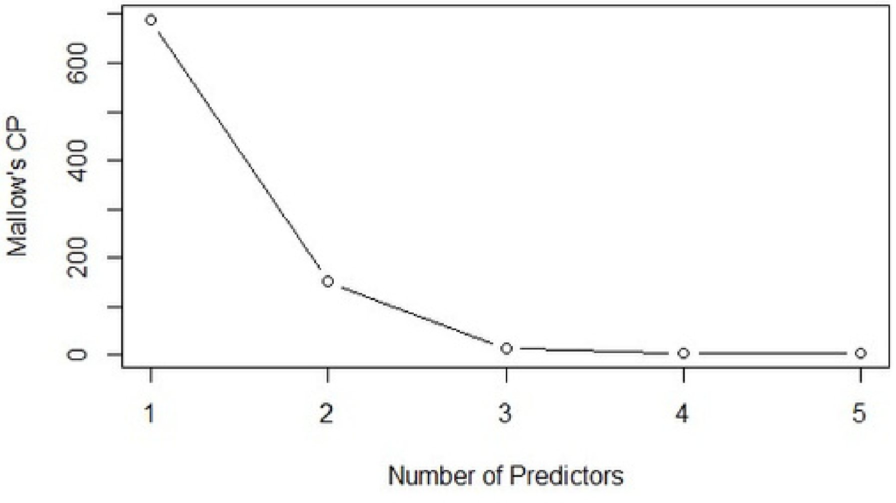
VIF measures how much the variance of the regression coefficients is inflated due to multicollinearity. So if a variance inflation factor is 1 it indicates that the variable is not multicollinear with other variables in the model. It means that, its inclusion in the model does not cause instability due to correlation with other predictors. It is safe to keep in the model.

NOTE: We have not included the correlation matrix as 2 of our variables are dummy variables and "children" and "age" is a discrete variable as it only takes positive integer values.

VARIABLE SELECTION



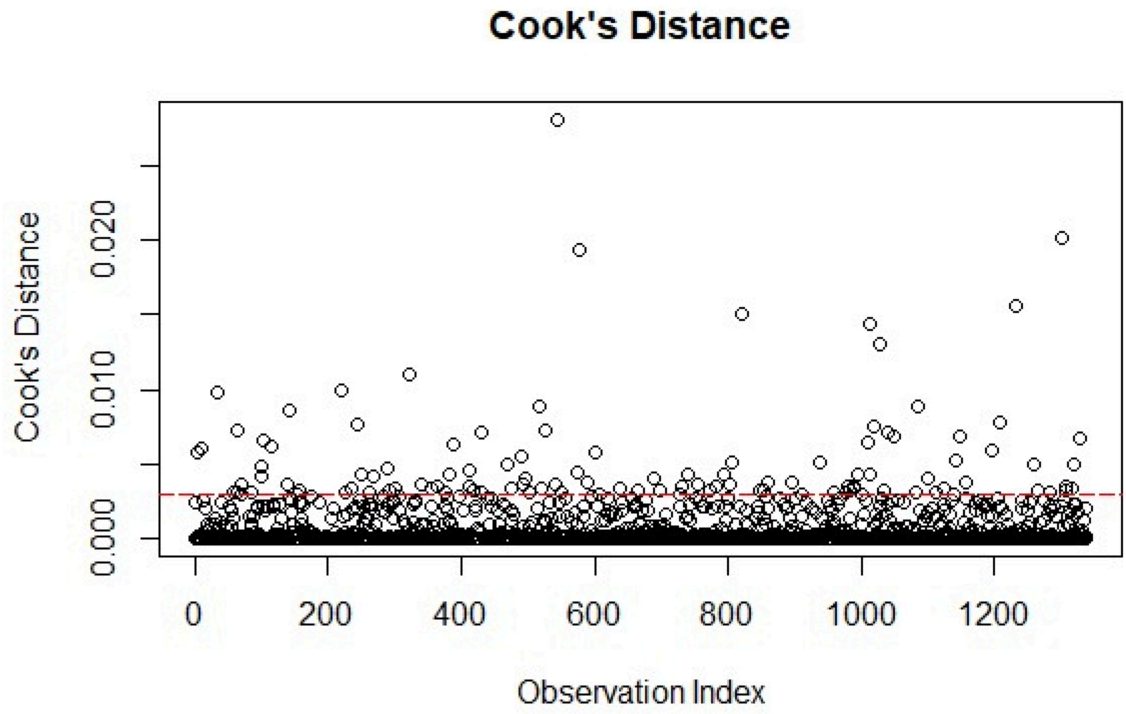
```
  X.Intercept.   age sex_dummy   bmi children smoker_dummy        Cp
1         TRUE FALSE     FALSE FALSE    FALSE         TRUE 689.64690
2         TRUE  TRUE     FALSE FALSE    FALSE         TRUE 150.73063
3         TRUE  TRUE     FALSE  TRUE    FALSE         TRUE  13.94997
4         TRUE  TRUE     FALSE  TRUE     TRUE         TRUE   4.14891
5         TRUE  TRUE      TRUE  TRUE     TRUE         TRUE   6.00000
```



Mallow's CP helps to select the model with an appropriate number of predictors. A good model typically has a CP value close to the number of predictors plus 1 (P + 1). Adding sex_dummy as a predictor increases the CP value to exactly to 6. This corresponds to the total predictors now being 5 plus intercept, suggesting that including sex_dummy balances the model's complexity and fit. This implies that including sex_dummy contributes meaningfully to the model's explanatory power making model 5 a better overall choice compared to model 4.

INFLUENCE ANALYSIS

BY COOK'S DISTANCE

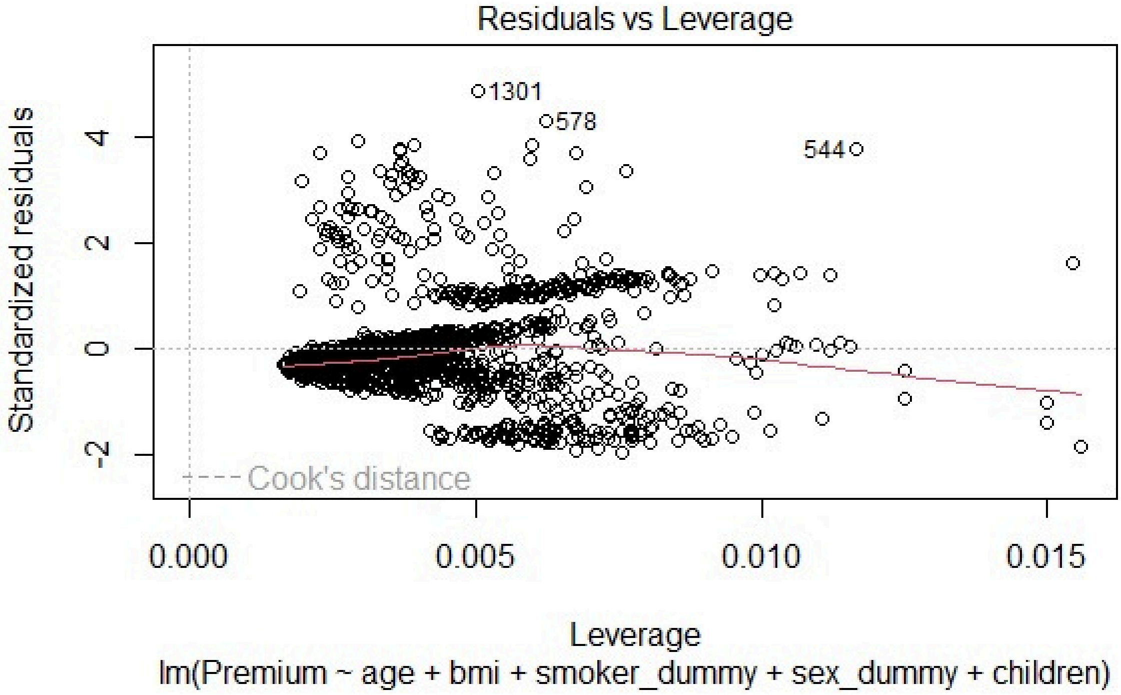


Cook's Distance measures the combined impact of an observation on both the fitted values and the regression coefficients. Its particularly useful for identifying influential data points that might unduly affect the regression model's results. So, we can see from the above data that the majority of the points are tightly clustered near 0 indicating a well-behaved data set with minimal influence from individual observations. So no points appear to be significantly beyond the threshold meaning the model isn't unduly influenced by any particular observation.

Model's Robustness: The absence of high Cook's Distance value suggests that the regression model is robust and reliable with no single data point disproportionately driving the results.

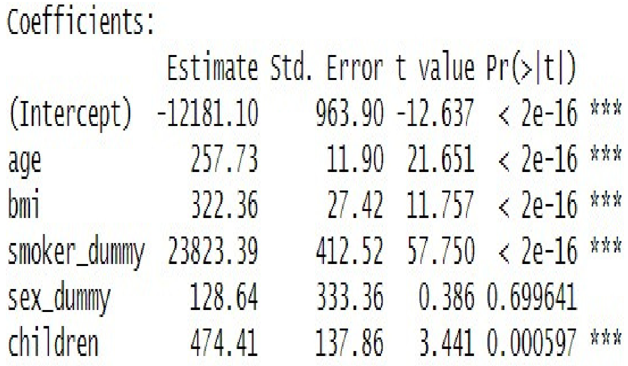
Influential Points in the Data – 109 observations

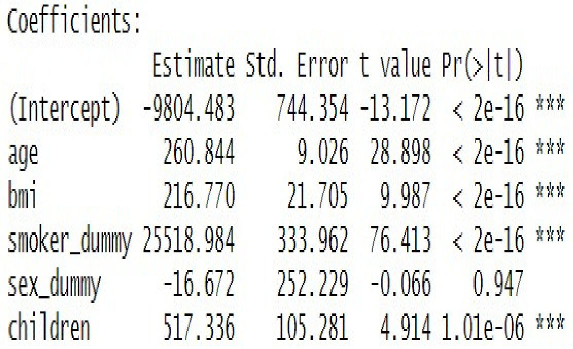
Threshold level - 4/nrow(insurance_data)



Based on the diagram of the residual vs leverage plot the following conclusions we have drawn is that:

1) Good Model fit for most data – The majority of the points are scattered near the center which standardized residuals close to zero and low leverage, suggesting that the model fits most of the data well.
2) Presence of Influential points – A few points (example 1301, 578, 544) are labelled as potential influential observations these points have either high leverage or high residuals and they may significantly affect the regression model.


```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12181.10     963.90 -12.637  < 2e-16 ***
age            257.73      11.90  21.651  < 2e-16 ***
bmi            322.36      27.42  11.757  < 2e-16 ***
smoker_dummy 23823.39     412.52  57.750  < 2e-16 ***
sex_dummy      128.64     333.36   0.386 0.699641
children       474.41     137.86   3.441 0.000597 ***
```


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9804.483    744.354 -13.172  < 2e-16 ***
age            260.844      9.026  28.898  < 2e-16 ***
bmi            216.770     21.705   9.987  < 2e-16 ***
smoker_dummy 25518.984    333.962  76.413  < 2e-16 ***
sex_dummy      -16.672    252.229  -0.066    0.947
children       517.336    105.281   4.914 1.01e-06 ***
```

Original Model R-squared – 0.7497 Adjusted R-squared – 0.7488

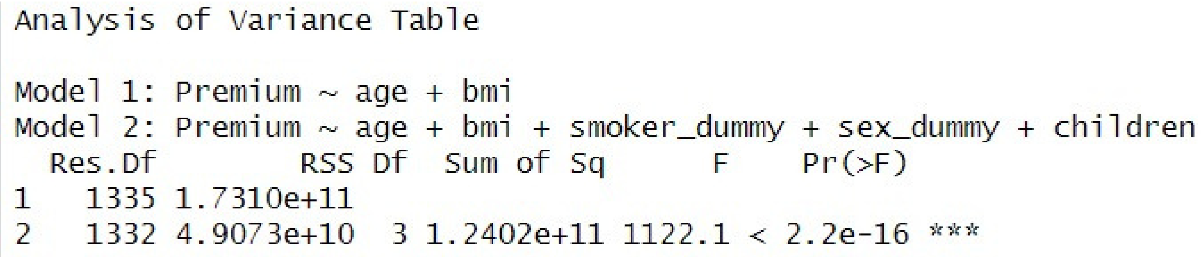
Residual Standard Error - 6070

Outlier removed R-squared – 0.852 Adjusted R-squared - 0.8514

Residual Standard Error – 4394

After cleaning the model by removing outliers and influential points the model has a significantly lower standard error and higher R-squared showing that the influential points had a considerably high leverage and influential points could be reason for fat tails observed in the QQ plot and funnel shape as seen in the residual plot.

HYPOTHESIS TESTING

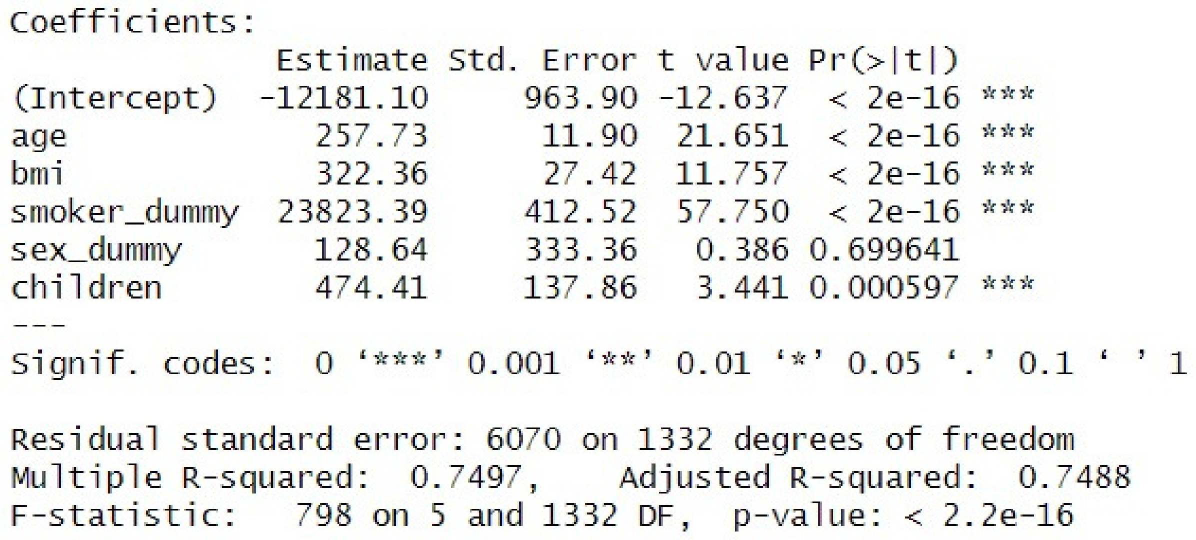


```
Analysis of Variance Table

Model 1: Premium ~ age + bmi
Model 2: Premium ~ age + bmi + smoker_dummy + sex_dummy + children
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
1   1335 1.7310e+11                                   
2   1332 4.9073e+10  3 1.2402e+11 1122.1 < 2.2e-16 ***
```

From the low p-value we can see that with the introduction of additional variables it improves the quality of model. So we consider the second for our analysis and drawing conclusions.

CONCLUSIONS



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -12181.10     963.90 -12.637  < 2e-16 ***
age             257.73      11.90  21.651  < 2e-16 ***
bmi             322.36      27.42  11.757  < 2e-16 ***
smoker_dummy  23823.39     412.52  57.750  < 2e-16 ***
sex_dummy       128.64     333.36   0.386 0.699641    
children        474.41     137.86   3.441 0.000597 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6070 on 1332 degrees of freedom
Multiple R-squared:  0.7497,	Adjusted R-squared:  0.7488 
F-statistic:   798 on 5 and 1332 DF,  p-value: < 2.2e-16
```

To answer our research questions and check the impact of each variable we will employ this regression summary for our analysis.

1) The model has a high R-squared and adjusted R-squared showing that the model answers most of the variability in the data.
2) Smoking Habits are significant in this study showing that a smoker has to pay a higher premium as compared to a non-smoker.
3) The sex of the individual is insignificant in this model showing that there is variation in the amount paid by a female and male. Although the initial boxplots showed a disparity in the raw data, regression analysis accounts for multiple factors simultaneously, which overshadowed the independent effect of gender on premium payments.
4) Also the premium amount depends positively on the variables such as age, bmi, smoking habits and number of children and each of them have a significant effect on the premium paid by the individual.