PREDICTING AND ANALYZING FACTORS INFLUENCING HEALTH INSURANCE PREMIUM

A PROJECT REPORT

ABSTRACT

This project investigates different demographic and lifestyle factors that affect medical charges. The factors we investigated are age, body mass index (BMI), number of children, smoking status, region, and sex. We used regression analysis to determine the most significant predictors of medical expenses and the strength and direction of their relationship with these charges.

We want to find out how each of these factors affects the variation in expense, so that individuals and insurers can understand and estimate medical health insurance premiums pricing better. For example, age and BMI are thought to be correlated with higher premiums because of higher health risks, while smoking status also matters. By looking into these factors systematically, we want to see which has the most impact and provide insights for health insurance pricing policy.

The result of this study is not only for understanding how health insurance premiums are priced but also aims at predicting premiums based on these factors enabling consumers and insurers to make more informed decisions.

INTRODUCTION

Health insurance helps manage medical bills and financial risk. But the premiums are determined by many factors that reflect an individual's risk and lifestyle. Knowing these factors is important for consumers looking for affordable cover and insurers looking for fair and transparent pricing.

Here we are looking at the relationship between health insurance premiums and a bunch of demographic and lifestyle variables: age, body mass index (BMI), number of kids, smoker or not, region, and medical expenses. We chose these because we assume they impact health risks and therefore insurance premiums. For example, age and BMI are often linked to higher health risks, smoking is a known contributor to higher health costs, and the region because healthcare is available and costly in different areas.

We will use regression to see which of these factors matter and how much. We will look at the strength and direction of those relationships to see what is going on under the hood.

This report shows what affects premiums and the key findings for consumers and insurers. Consumers can see how their personal and lifestyle choices impact their premiums and insurers can use this to design fairer pricing. This helps us understand what drives health insurance costs and get actionable for transparency, affordability, and fairness in health insurance.

PROBLEM STATEMENT

Insurance premiums are calculated on many factors like age, BMI, smoking status, region, and medical expenses. But consumers are unclear how these factors impact the premium and insurers struggle to get fair and accurate pricing models.

Our challenge is to build a model that calculates health insurance premiums based on these factors. The model should identify the top predictors, quantify the impact, and show the relationships between these variables. To make premium pricing more transparent so consumers can make informed decisions and insurers can design fair policies.

DATA SOURCE

https://www.kaggle.com/datasets/mirichoi0218/insurance/data

PROPOSED METHODOLOGY

The methodology to analyze factors affecting health insurance premiums and build a predictive model starts with data preprocessing. First, the dataset will be checked for null or missing values and necessary actions will be taken to fill the gaps. Categorical variables like smoking status and region will be converted into numerical format using one-hot or label encoding. This will make the data ready for analysis and modeling.

After the data is prepared, it will be split into training (70%) and testing (30%) sets. The training data will be used to build a Linear Regression model, and the testing data will evaluate the model on unseen data. The training phase will involve fitting the regression model and checking the fit to see which factors are significant in predicting health insurance premiums.

After the model is built, Exploratory Data Analysis (EDA) will be done to see the relationships between independent variables and the response variable. Basic plots will be used to check if the relationships are linear. Heteroscedasticity (non-constant variance of residuals) will also be checked to make sure the assumptions of linear regression are met.

Further diagnostic checks will include checking for multicollinearity among predictors using Variance Inflation Factor (VIF) to identify highly correlated variables that can affect the model’s reliability. The presence of influential points like outliers or high-leverage points will also be checked using diagnostic plots and influence measures like Cook’s Distance.

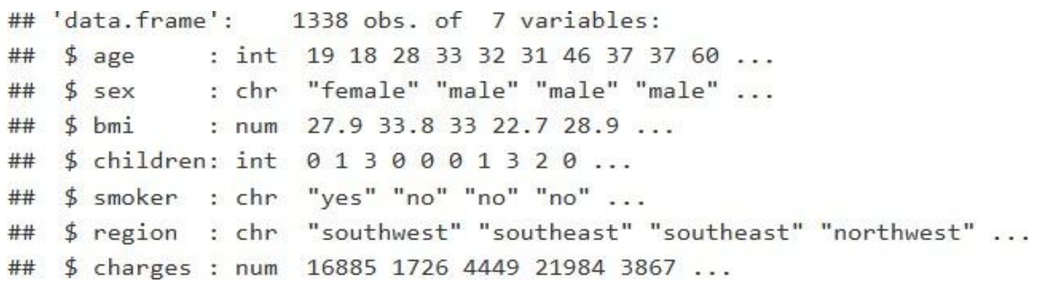
If heteroscedasticity, multicollinearity, or influential points are found, remediation methods will be applied. This may include data transformations, variable selection, or regularization techniques to make the model more robust and accurate.

Finally, the trained model will be tested using the 30% testing dataset. Predictions will be compared with actual values and the model will be evaluated using metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE). Insights from the model will help to identify the most significant factors affecting health insurance premiums. The results will be interpreted and presented in a clear and concise manner.

This will give us a complete picture of the factors affecting premiums and solve the problems during model building.

ANALYSIS AND RESULTS

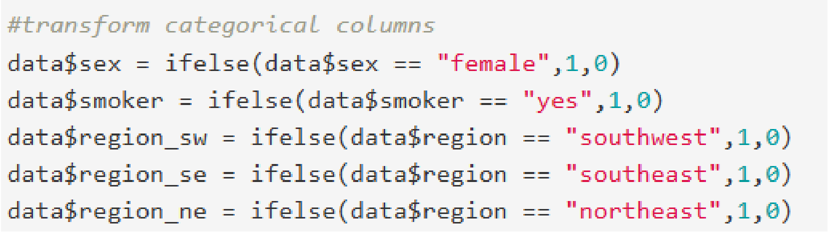
Description Of Data Set:



```
## '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 ...
##  $ smoker  : chr  "yes" "no" "no" "no" ...
##  $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num  16885 1726 4449 21984 3867 ...
```

- The dataset consists of around 1338 observations with the response variable as charges and predictors such as age, sex, BMI, children, smoker, and region. Out of which sex and smoker are binary qualitative variables and the region predictor has 4 labels: northwest, northeast, southwest and southeast.

Transforming categorical columns:

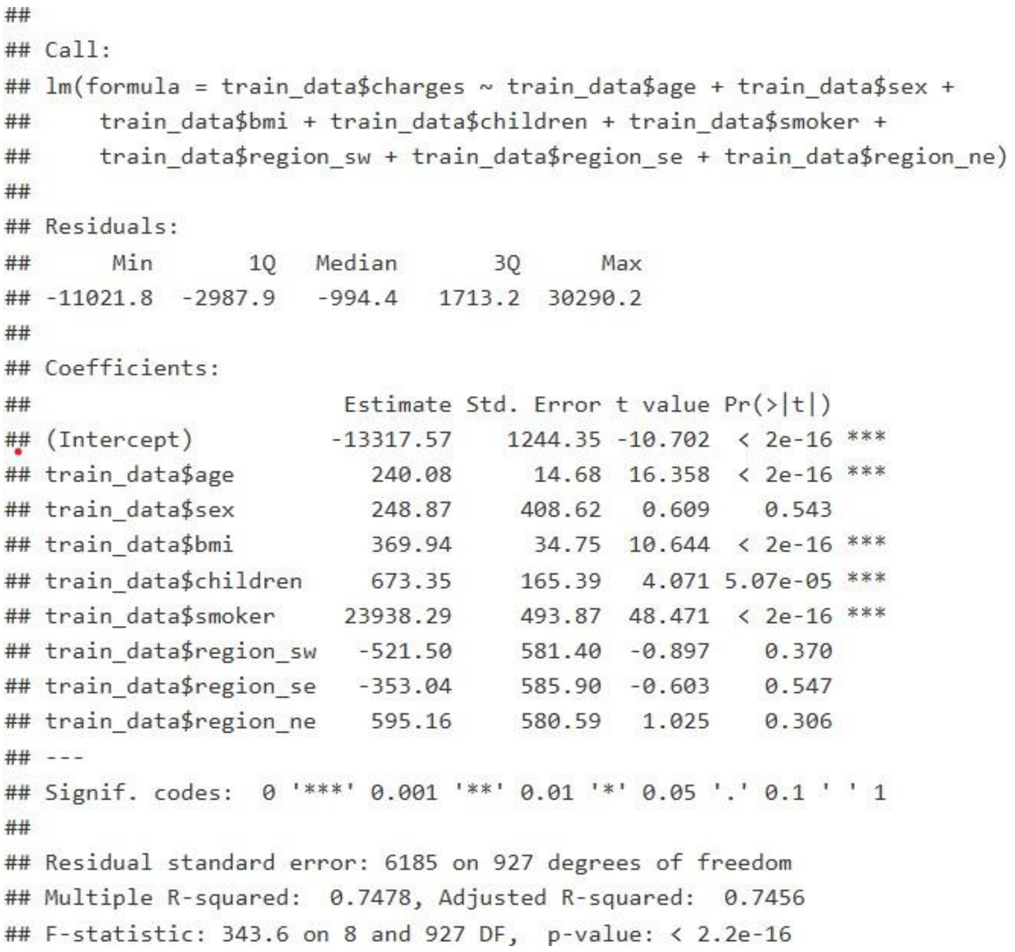


```
#transform categorical columns
data$sex = ifelse(data$sex == "female",1,0)
data$smoker = ifelse(data$smoker == "yes",1,0)
data$region_sw = ifelse(data$region == "southwest",1,0)
data$region_se = ifelse(data$region == "southeast",1,0)
data$region_ne = ifelse(data$region == "northeast",1,0)
```

- We have transformed the categorical predictors to numeric predictors using dummy variables. We have used on-hot coding to identify a female candidate as 1 and a male candidate as 0, a smoker as 1 and a nonsmoker as 0. Created 3 dummy variables to indicate the regions.

- We have sliced the data into training (70% of the data set) and testing set (30% of the data set) and then the model has been created using only the 70% of the data set.

Observations from the initial model:



```
##
## Call:
## lm(formula = train_data$charges ~ train_data$age + train_data$sex +
##     train_data$bmi + train_data$children + train_data$smoker +
##     train_data$region_sw + train_data$region_se + train_data$region_ne)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11021.8  -2987.9   -994.4   1713.2  30290.2
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -13317.57    1244.35 -10.702  < 2e-16 ***
## train_data$age          240.08      14.68  16.358  < 2e-16 ***
## train_data$sex          248.87     408.62   0.609    0.543
## train_data$bmi          369.94      34.75  10.644  < 2e-16 ***
## train_data$children     673.35     165.39   4.071 5.07e-05 ***
## train_data$smoker     23938.29     493.87  48.471  < 2e-16 ***
## train_data$region_sw   -521.50     581.40  -0.897    0.370
## train_data$region_se   -353.04     585.90  -0.603    0.547
## train_data$region_ne    595.16     580.59   1.025    0.306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6185 on 927 degrees of freedom
## Multiple R-squared:  0.7478, Adjusted R-squared:  0.7456
## F-statistic: 343.6 on 8 and 927 DF,  p-value: < 2.2e-16
```

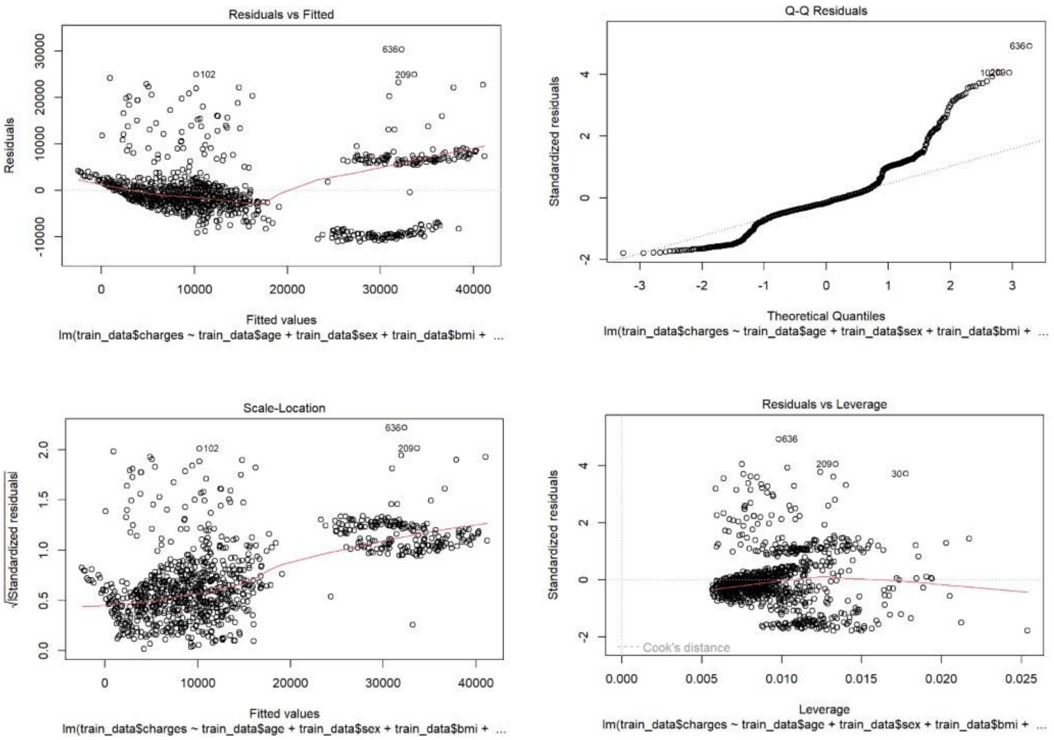
- Initially there were 1338 observations but right now when we investigate the summary, we can find only 936 observations. This implies only 70 percent of the data have been used to train the model.
- The R-squared value and the adjusted R-squared value of the initial model which is formed using all the predictors is 0.7478 indicating that the linear model with all the predictors was able to capture around 74% of the variance or the pattern in predicting charges.

- The f-statistic value is high (343.6) indicating that there is a relationship between charges and at least one of the predictors used in the model. The p-value associated with the f-statistic is small and below the threshold confirming that this observation is statistically significant.
- On initial analysis of the p values associated with the coefficients of the predictors we can say that the coefficients of the factors sex and region are not statistically significant as their p values are above the threshold.

The significant predictors are:

- **Age:** Medical charges increase by approximately $240.08 for each additional year of age holding all other factors constant.
- **BMI:** An increase of one unit in BMI corresponds to an average increase of $369.94 in charges holding all other factors constant.
- **Number of children:** Each additional child increases medical charges by $673.35 on average holding all other factors constant.
- **Smoking:** Smokers incur significantly higher charges, with an average increase of $23,938.29 compared to non-smokers holding all other factors constant.

Observations on the basic plots of the initial model:



- **Residual vs Fitted:** The U shape in this plot indicates the presence of non-linearity. The spread of the residuals with the increase of the fitted values indicates the presence of heteroscedasticity
- **Q-Q Residuals:** From the graph we can see that the residuals are deviating from the diagonal line specifically in the upper tail indicating that the normal distribution assumption has been violated. This also tells that there are outliers, and influential points present, and the model might have a difficulty in capturing extreme values
- **Scale Location:** The increasing spread and trend in the residuals suggest that the model may not satisfy the assumption of homoscedasticity. The red smooth line has a slight upward trend, especially noticeable at higher fitted values, further supporting the presence of heteroscedasticity.

- **Residual vs Leverage:** this plot highlights the outliers present (636,209,30).

Identifying potential issues:

Multicollinearity check:

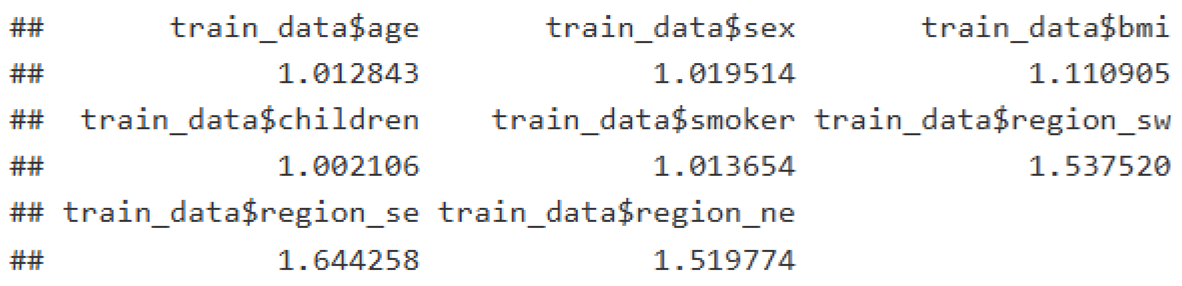
One of the assumptions for a linear regression model is that the predictors are not correlated with each other making it difficult to determine the individual effect of each predictor on the dependent variable. Variance Inflation Factor (VIF) measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. The VIF result can be read as

- VIF = 1, then there is no correlation between the predictor and the other variables, indicating no multicollinearity.

- 1< VIF <5 indicates Moderate correlation; multicollinearity is not a severe issue.

- VIF > = 5 indicates that there is high correlation, suggesting the presence of multicollinearity. The variable may be problematic and its inclusion in the model could lead to inflated standard errors and unstable coefficient estimates.

- VIF > = 10 indicates severe multicollinearity. In such cases, the predictor is highly correlated with one or more other variables, and it is advisable to reconsider its inclusion or apply remedial measures.



```
##      train_data$age      train_data$sex      train_data$bmi
##            1.012843            1.019514            1.110905
##  train_data$children   train_data$smoker train_data$region_sw
##            1.002106            1.013654            1.537520
## train_data$region_se train_data$region_ne
##            1.644258            1.519774
```

- Based on the result, the VIF values associated for the predictors are all low (in the range of 1 to 1.6) indicating that there is **no multicollinearity present**.

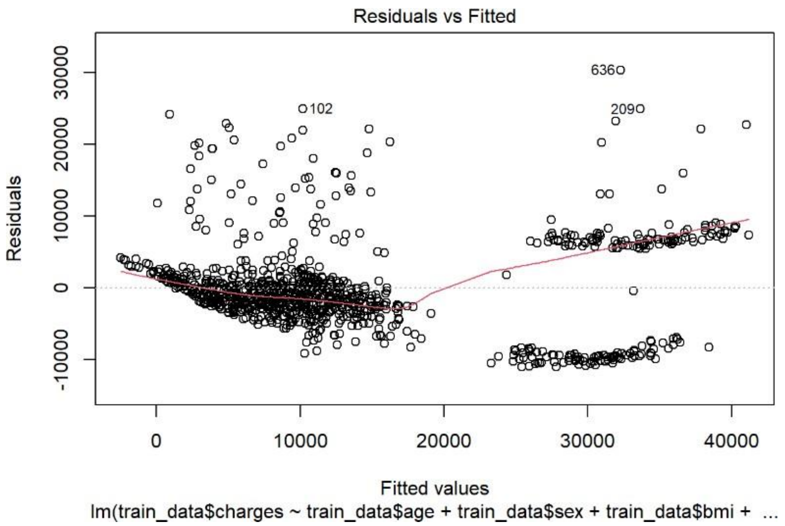
Autocorrelation check:

- Autocorrelation in residuals indicates that the errors are not independent, which can bias standard errors and reduce the validity of statistical tests. The Durbin-Watson (DW) test assesses the presence of autocorrelation, particularly first-order (lag-1) autocorrelation.
- A DW statistic close to 2 suggests no autocorrelation, values < 2 indicate positive autocorrelation, and values > 2 suggest negative autocorrelation.
- Autocorrelation can undermine the validity of regression assumptions and lead to inefficient estimators.

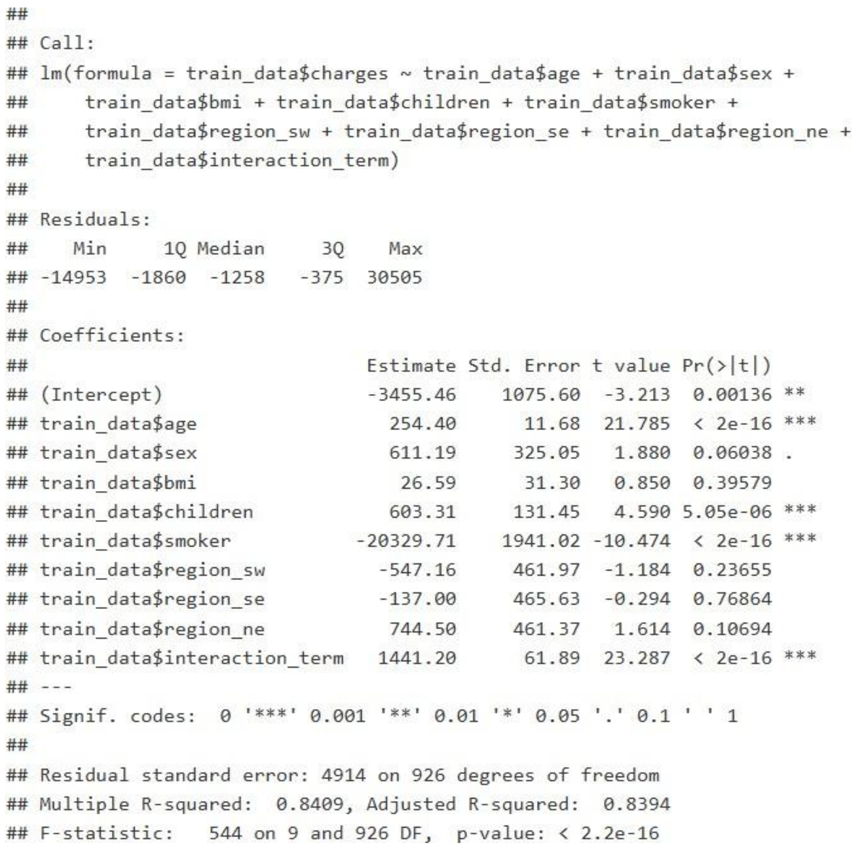
```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.9762, p-value = 0.3582
## alternative hypothesis: true autocorrelation is greater than 0
```

- The result of the DW test is 1.97 which is remarkably close to 2, indicating that there is no autocorrelation. The p value associated is high, meaning the model's residuals are independent and **do not exhibit notable correlation patterns**.

Non-linearity check:



- Another assumption associated with linear models is that the relationship between the predictors and the response variable is linear. As observed by the graph Residual vs Fitted, we identified the presence of non-linearity.
- To improve the model, we have **introduced** an **interaction term** BMI * SMOKER and as a result.



```
##
## Call:
## lm(formula = train_data$charges ~ train_data$age + train_data$sex +
##     train_data$bmi + train_data$children + train_data$smoker +
##     train_data$region_sw + train_data$region_se + train_data$region_ne +
##     train_data$interaction_term)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14953  -1860  -1258   -375  30505
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -3455.46    1075.60  -3.213  0.00136 **
## train_data$age                 254.40      11.68  21.785  < 2e-16 ***
## train_data$sex                 611.19     325.05   1.880  0.06038 .
## train_data$bmi                  26.59      31.30   0.850  0.39579
## train_data$children            603.31     131.45   4.590 5.05e-06 ***
## train_data$smoker           -20329.71    1941.02 -10.474  < 2e-16 ***
## train_data$region_sw          -547.16     461.97  -1.184  0.23655
## train_data$region_se          -137.00     465.63  -0.294  0.76864
## train_data$region_ne           744.50     461.37   1.614  0.10694
## train_data$interaction_term   1441.20      61.89  23.287  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4914 on 926 degrees of freedom
## Multiple R-squared:  0.8409, Adjusted R-squared:  0.8394
## F-statistic:   544 on 9 and 926 DF,  p-value: < 2.2e-16
```

- We observe that the adjusted R-squared of the model has improved from 0.74 to 0.83 as soon as the interaction term has been added.

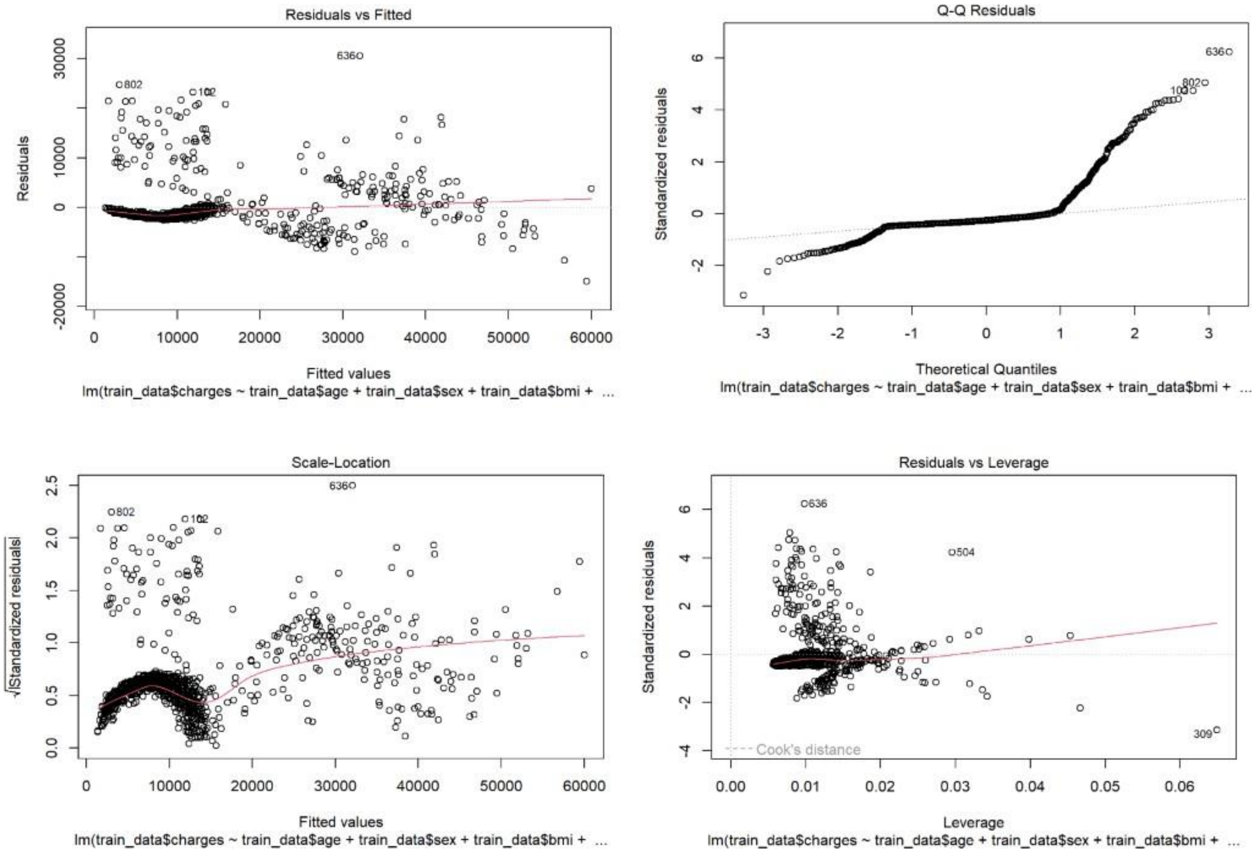
From the summary of the corrected model the significant predictors are

- **Age:** With one unit increase of age the charges are expected to increase by $254.40 holding all other factors constant

- **Number of children:** with the increase of one child the charges increase on an average by $603.31 holding all other factors constant

- **Interaction term (BMI * Smoker)**: The interaction term between BMI and smoking status suggests that the relationship between BMI and medical charges is different for smokers compared to non-smokers. Specifically, smokers with higher BMI may experience a charge of $1441.20 more than non-smokers with similar BMI values keeping other factors constant, reflecting compounded health risks associated with both high BMI and smoking.

Basic Plots Including the Correlation Term:

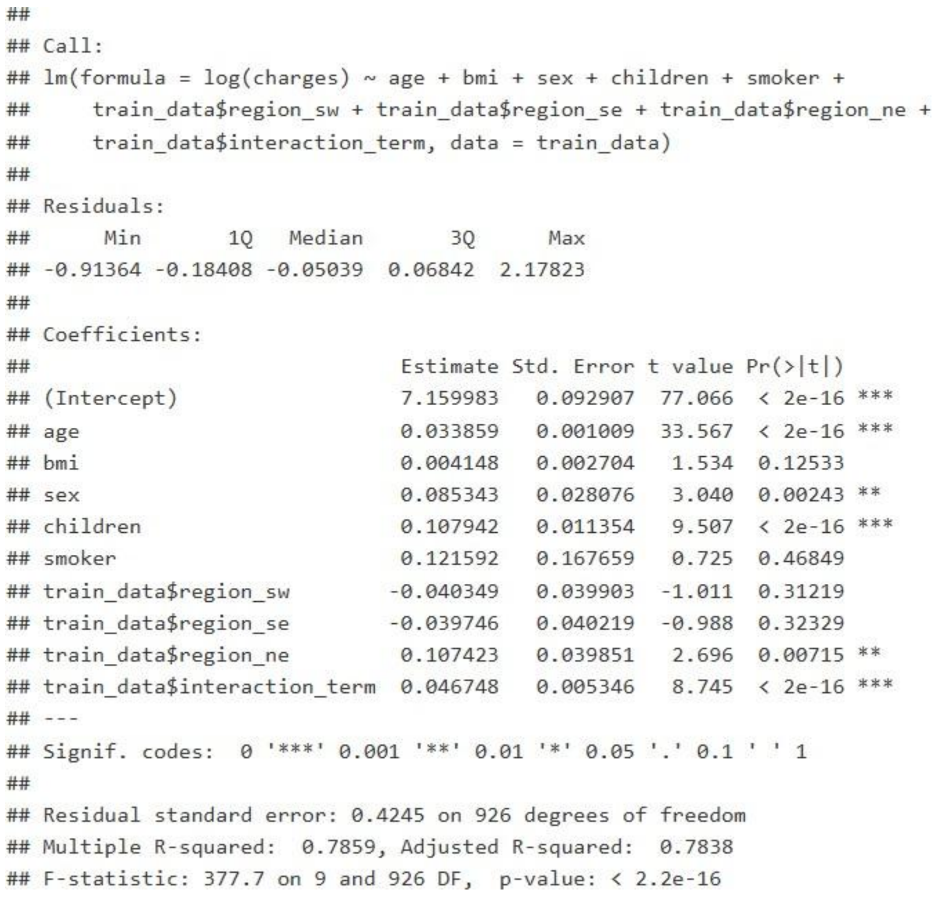


- From the Residuals vs Fitted plot of the new modified model we can see that the red smooth line no longer has a significant U shape, capturing the nonlinear relationship.

Resolution for heteroscedasticity:

- We have used logarithmic transformations on the response variable to reduce the heteroscedasticity.

The Summary of our Model after adding the Logarithmic transformation:

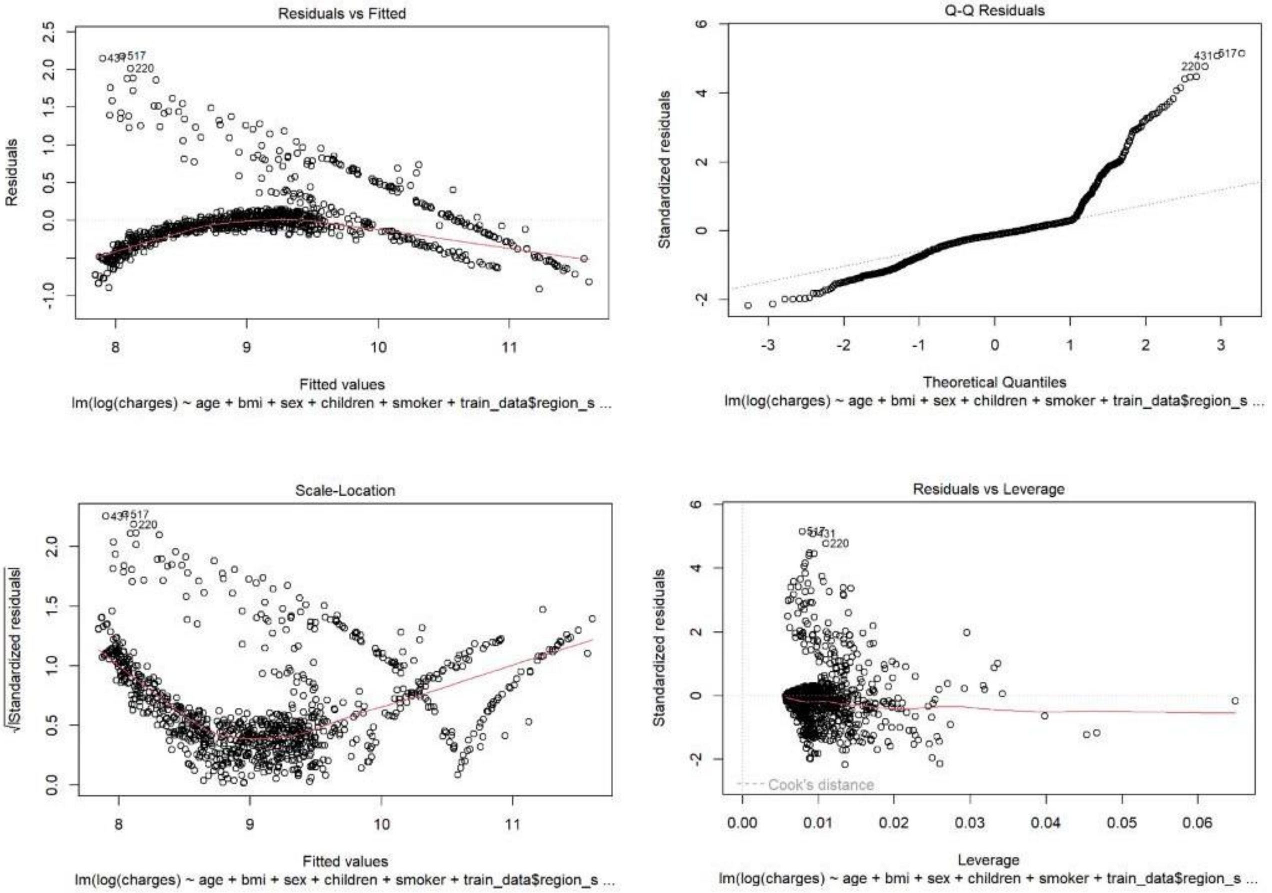


```
##
## Call:
## lm(formula = log(charges) ~ age + bmi + sex + children + smoker +
##     train_data$region_sw + train_data$region_se + train_data$region_ne +
##     train_data$interaction_term, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.91364 -0.18408 -0.05039  0.06842  2.17823
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  7.159983   0.092907  77.066  < 2e-16 ***
## age                          0.033859   0.001009  33.567  < 2e-16 ***
## bmi                          0.004148   0.002704   1.534  0.12533
## sex                          0.085343   0.028076   3.040  0.00243 **
## children                     0.107942   0.011354   9.507  < 2e-16 ***
## smoker                       0.121592   0.167659   0.725  0.46849
## train_data$region_sw        -0.040349   0.039903  -1.011  0.31219
## train_data$region_se        -0.039746   0.040219  -0.988  0.32329
## train_data$region_ne         0.107423   0.039851   2.696  0.00715 **
## train_data$interaction_term  0.046748   0.005346   8.745  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4245 on 926 degrees of freedom
## Multiple R-squared:  0.7859, Adjusted R-squared:  0.7838
## F-statistic: 377.7 on 9 and 926 DF,  p-value: < 2.2e-16
```

The significant predictors are:

- **Age:** For every additional year of age, the log of charges increases by approximately 0.0339 holding all other factors constant.
- **Sex:** On an average the log of charges for a female is 0.085 more than that for a male candidate, holding all other factors constant
- **Number of children:** Each additional child increases the log of charges by 0.1079 holding all other factors constant.
- **Region (Northwest):** Residing in the Northeast is associated with a 0.0135 increase in the log of charges compared to reference region, northwest holding all other factors constant.

The Plots of the model after removing Heteroscedasticity:



- From the residual vs fitted plot of the corrected model we notice that the scale has changed, and residual spread has decreased there by reducing heteroscedasticity.

Check for Influential points:

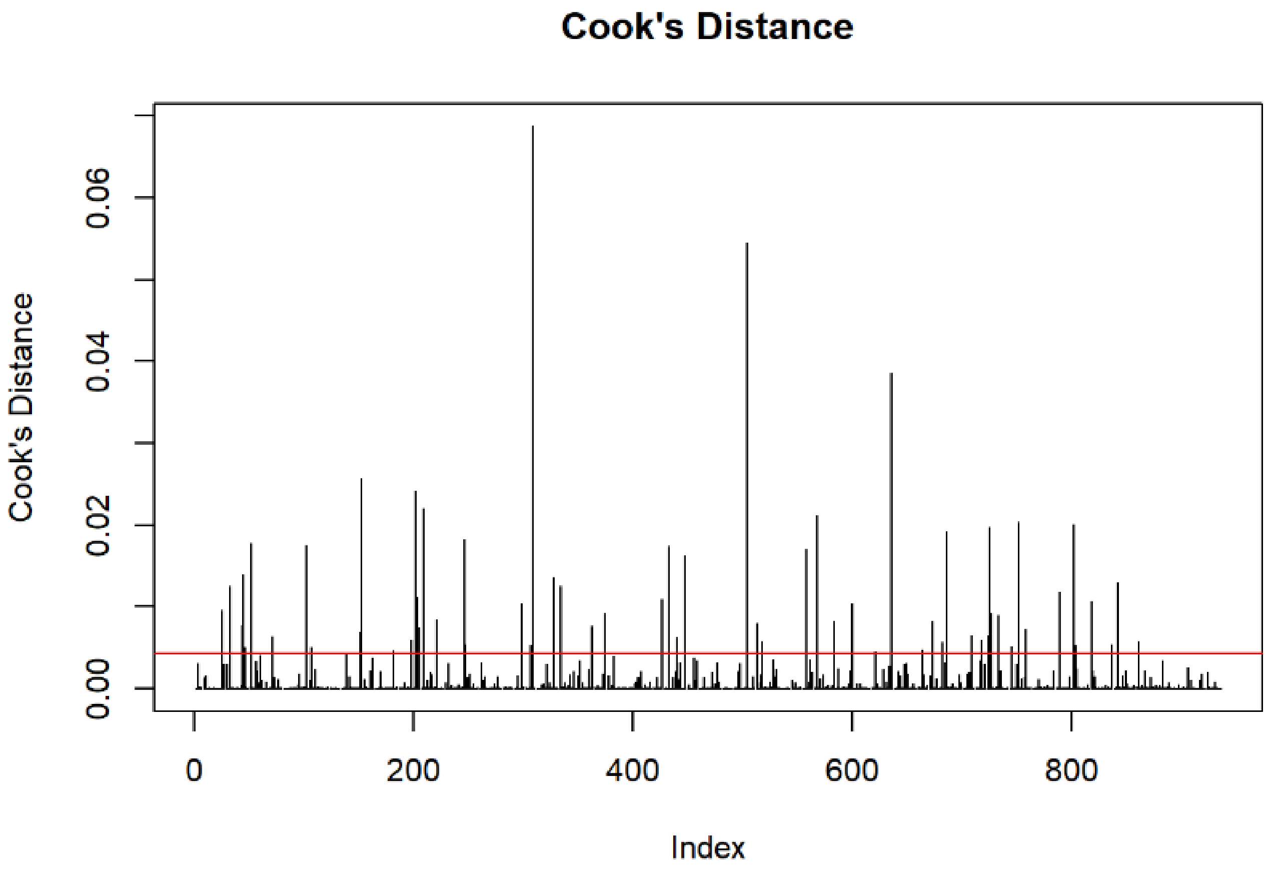
- Influential points are data points that significantly impact the model’s estimates. Cook’s Distance is a measure used to identify such influential points. From the basic plots we observed the presence of influential points or outliers and using cook’s distance we have removed these points from our training data set.

- Cook’s Distance quantifies the influence of each data point on the fitted regression model. It combines information about both the leverage and the residuals of each point.Points with a Cook’s Distance greater than the threshold (4/n, where n is the number of observations) are considered influential. These points could potentially have a undesirable effect on the regression coefficients.

Influential points noticed are:

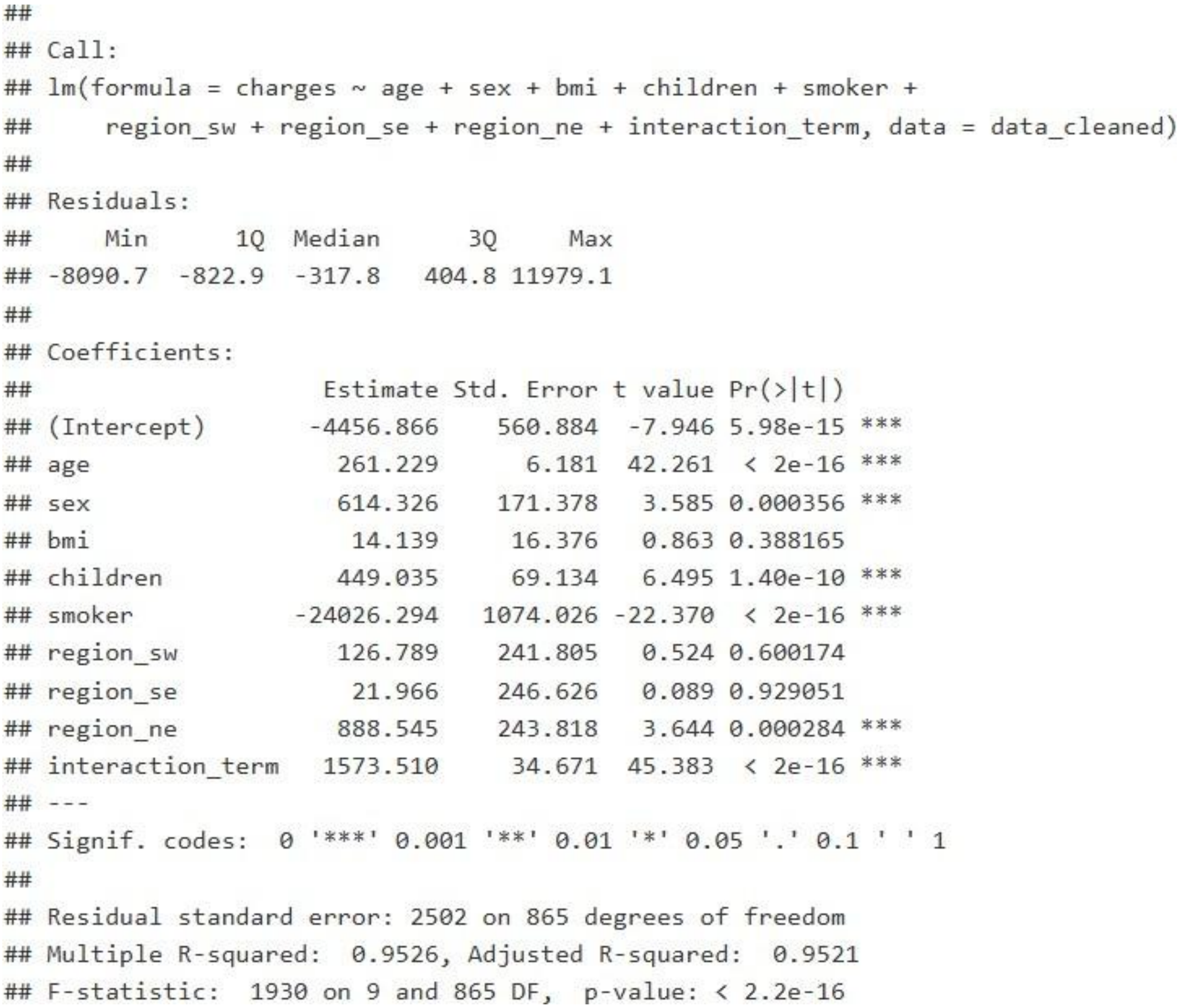
```
##   26  33  44  45  47  52  71 102 107 139 152 153 182 198 202 204 205 209 221 247
##   26  33  44  45  47  52  71 102 107 139 152 153 182 198 202 204 205 209 221 247
## 248 299 307 309 328 334 363 375 427 433 440 448 504 513 518 558 568 584 600 621
## 248 299 307 309 328 334 363 375 427 433 440 448 504 513 518 558 568 584 600 621
## 636 664 673 682 686 709 718 724 725 727 733 745 752 758 789 802 804 818 837 842
## 636 664 673 682 686 709 718 724 725 727 733 745 752 758 789 802 804 818 837 842
## 861
## 861
```

Cooks Distance Plot:



- All the observations crossing the base red line in the cook's distance plot are considered influential.

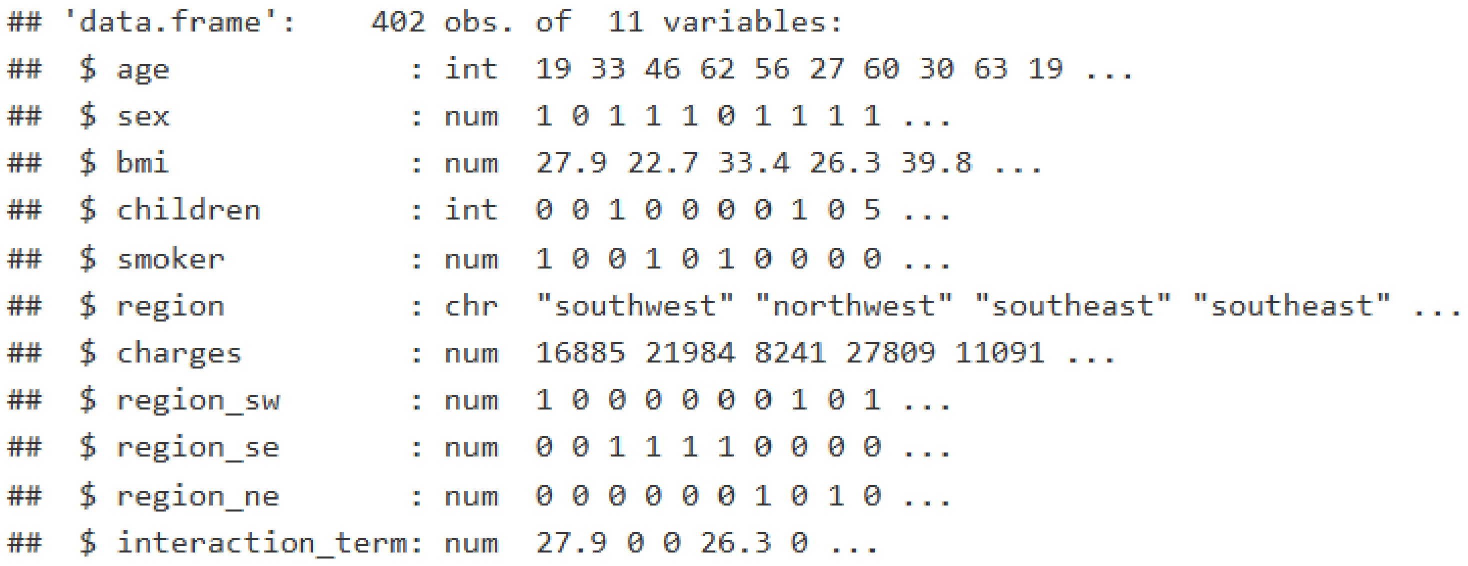
The Summary of the model is shown below



```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region_sw + region_se + region_ne + interaction_term, data = data_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8090.7  -822.9  -317.8   404.8 11979.1
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4456.866    560.884  -7.946 5.98e-15 ***
## age                 261.229      6.181  42.261  < 2e-16 ***
## sex                 614.326    171.378   3.585 0.000356 ***
## bmi                  14.139     16.376   0.863 0.388165
## children            449.035     69.134   6.495 1.40e-10 ***
## smoker           -24026.294   1074.026 -22.370  < 2e-16 ***
## region_sw           126.789    241.805   0.524 0.600174
## region_se            21.966    246.626   0.089 0.929051
## region_ne           888.545    243.818   3.644 0.000284 ***
## interaction_term   1573.510     34.671  45.383  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2502 on 865 degrees of freedom
## Multiple R-squared:  0.9526, Adjusted R-squared:  0.9521
## F-statistic:  1930 on 9 and 865 DF,  p-value: < 2.2e-16
```

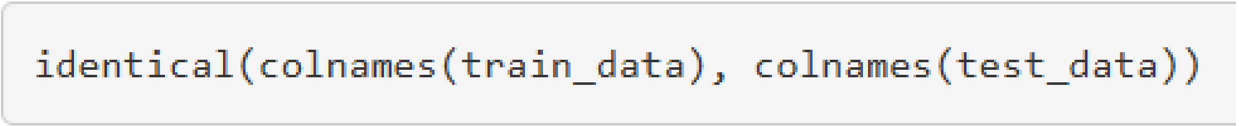
- We can observe that the R-squared of the model is 0.9526, showing that 95.26% of the variability in charges is explained by the model. The Adjusted R – Squared =0.9521 suggesting a high goodness of fit.

Examining the test data:

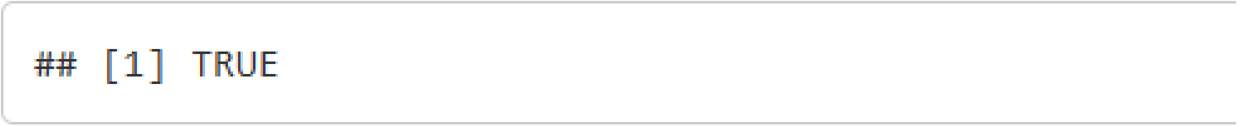


```
## 'data.frame':    402 obs. of  11 variables:
##  $ age             : int  19 33 46 62 56 27 60 30 63 19 ...
##  $ sex             : num  1 0 1 1 1 0 1 1 1 1 ...
##  $ bmi             : num  27.9 22.7 33.4 26.3 39.8 ...
##  $ children        : int  0 0 1 0 0 0 0 1 0 5 ...
##  $ smoker          : num  1 0 0 1 0 1 0 0 0 0 ...
##  $ region          : chr  "southwest" "northwest" "southeast" "southeast" ...
##  $ charges         : num  16885 21984 8241 27809 11091 ...
##  $ region_sw       : num  1 0 0 0 0 0 0 1 0 1 ...
##  $ region_se       : num  0 0 1 1 1 1 0 0 0 0 ...
##  $ region_ne       : num  0 0 0 0 0 0 1 0 1 0 ...
##  $ interaction_term: num  27.9 0 0 26.3 0 ...
```

Checking if column names of test and train data are the same:

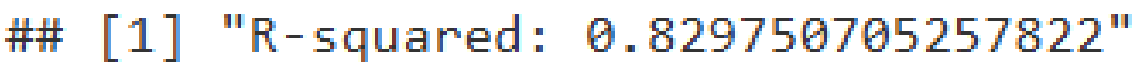


```
identical(colnames(train_data), colnames(test_data))
```



```
## [1] TRUE
```

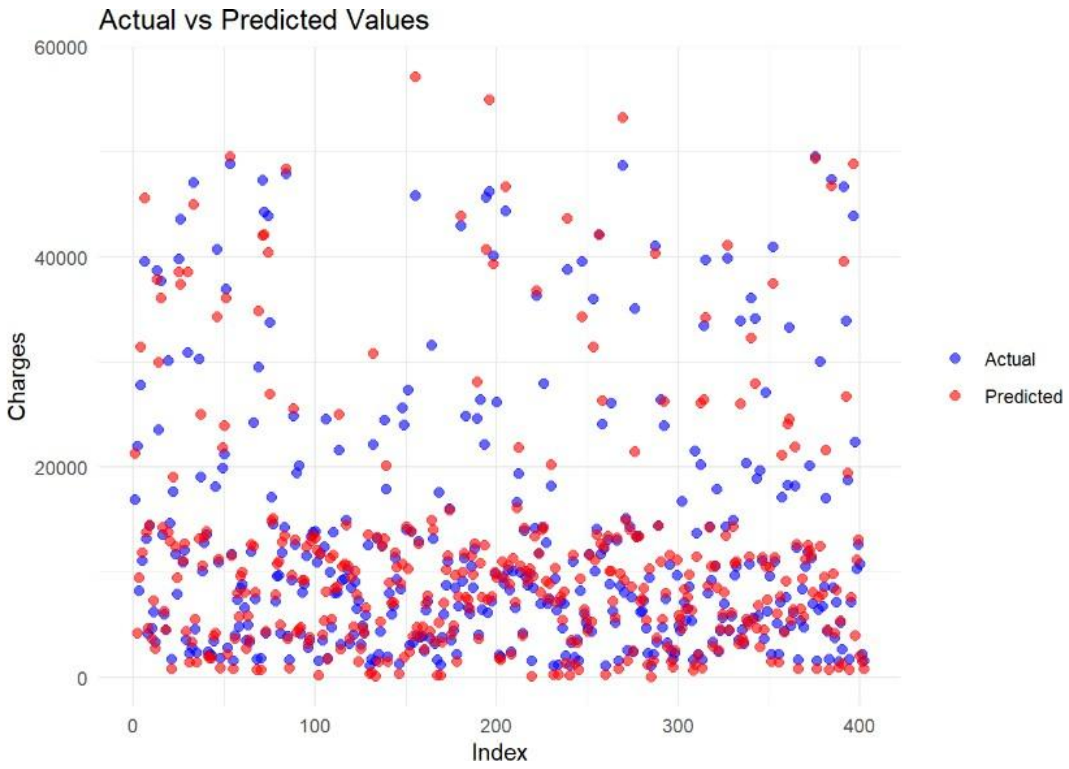
Evaluating the model on Predicted values:



```
## [1] "R-squared: 0.829750705257822"
```

The Test R squared of the model is 0.8298 indicating that it captured 82.98 percent of the variance in new or unseen data.

Comparison of Actual vs Predicted Values:



The plot provides a comparison between actual and predicted insurance charges for each observation in the dataset. Each point on the x-axis represents an individual observation, while the y-axis shows the charge amount. Actual values are displayed as blue dots, and predicted values are shown as red dots. This visual allows for an assessment of the model's accuracy.

CONCLUSION

This project looked at the factors that affect health insurance premiums using regression analysis. After looking at the data, we found the key predictors of health insurance premiums are age, number of children and smoking status. Sex and region were found to be less significant in predicting premiums. The initial analysis showed non-linearity in the data as seen in the residual vs fitted plot. We added an interaction term to capture the non-linearity between the predictors where we observed the relationship between BMI and smoker predictor where the charges were higher for a smoker compared to a non-smoking candidate with similar BMI values. This improved the R2 of the model and the residuals vs fitted plot looked much better, indicating a better model fit.

To check the robustness of the model:

1. Multicollinearity: VIF was calculated and there was no multicollinearity between the predictors.
2. Autocorrelation: DW test showed no autocorrelation, so the residuals were independent.
3. Heteroscedasticity: Residuals vs fitted plot showed heteroscedasticity. We applied logarithmic transformation to the response variable, and it reduced the spread of the residuals and the problem of non-constant variance.
4. Influential Points: Cook's Distance was used to identify influential points, and they were removed from the training data. This improved the statistical performance of the model and made it more robust.

After doing all the diagnostic checks and transformations, the final model was validated using a separate test set. The model forecasted health insurance premiums with reasonable accuracy and predicted unseen data.

The insights from this study are valuable for both consumers and insurers. For consumers, this analysis gives them a better understanding of how their personal characteristics (age, BMI, smoking status, number of children) affect their health insurance premiums. By knowing these relationships, consumers can make better decisions when shopping for health insurance. For insurers, the model can be used to improve pricing strategy and make premium calculations more transparent and fairer.

Overall, this project helps to understand the factors that affect health insurance pricing. It provides a data driven model to make health insurance more affordable and fairer. This study also opens opportunities for future research on health insurance pricing models and fair pricing.

REFERENCES

1. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Applications in R*. Springer.
2. R Documentation - Libraries: car, dplyr, ggplot2 and lmtest (https://www.r-project.org).

ACKNOWLEDGEMENTS

- We extend our gratitude to **Professor [Kiah Wah Ong]**, for their guidance and support
- Thanks to **Illinois Institute of Technology** for providing access to the data and additional resources.