Project Report - 1

Name - Suraj Kumar Mishra
Student no - 8902749
Course - Multivariate Statistics
Date - 25/02/2024

Table of Contents

INTRODUCTION

Healthcare costs are a significant concern for individuals, families, and healthcare systems worldwide. Understanding the factors influencing medical expenses is crucial for effective financial planning and resource allocation in the healthcare sector. In this report, we employ predictive modelling techniques, specifically linear regression, to analyze a dataset of medical costs. By exploring demographic and lifestyle factors such as age, BMI, smoking status, region, and number of children, we aim to develop a model that accurately predicts medical charges billed by health insurance.

The selection of predictors in this analysis is guided by prior research highlighting their significant impact on medical costs. By leveraging past trends and findings from similar regression projects, we seek to build upon existing knowledge and contribute to understanding healthcare expenditure patterns.

Through a systematic approach, we examine potential non-linearities and interaction effects among variables, document coefficients obtained from the regression model, and assess model assumptions using diagnostic tests. Furthermore, we discuss the implications of our findings and propose potential model improvements to enhance predictive accuracy and reduce errors.

This report comprehensively explores predictive modelling with linear regression in the context of healthcare costs. By elucidating the relationships between predictors and medical charges, we aim to provide valuable insights for stakeholders in the healthcare industry and inform decision-making processes related to healthcare expenditure management.

1. Gathering Data

a. Clarify Data Source –

The dataset used for this analysis originates from a publicly available source provided by "Kaggle". The link for the dataset is given below.

https://www.kaggle.com/code/mariapushkareva/medical-insurance-cost-with-linear-regression/input

b. Metadata Details –

Dataset Name: Medical Cost Dataset

Source: https://www.kaggle.com/code/mariapushkareva/medical-insurance-cost-with-linear-regression/input

Description: Anonymized dataset containing medical cost data collected from healthcare insurance providers.

Format: CSV (Comma-separated values)

Variables: 7

Data Quality: Cleaned and anonymized

Data Source: Kaggle

c. Data Fields Description and Short Meanings –

Age: Age of the insured individual at the time of medical service.

Sex: Gender of the insured individual (Categorical: Male/Female)

BMI: Body Mass Index of the insured individual, calculated as weight (in kilograms) divided by height (in meters) squared.

Smoker: Smoking status of the insured individual, categorized as "Yes" for smokers and "No" for non-smokers.

Region: Geographic region of the insured individual's residence, categorized as Northeast, Southeast, Southwest, or Northwest.

Children: Number of children or dependents covered by the insurance plan.

Charges: Medical charges billed by the health insurance provider for the medical services received by the insured individual.

d. Data Count and Domain Context -

The dataset consists of 475 records, each representing an individual medical case. It provides a snapshot of healthcare expenses incurred by individuals within the context of the United States healthcare system. Understanding the domain context is crucial for interpreting the data accurately and deriving meaningful insights from the analysis.

e. Predictor and Response Variables –

Predictor Variables: Age, Sex, BMI, Smoking Status, Region, Number of Children.

Response Variable: Medical Charges.

2. Initial Modelling

a. Response Variable: Medical Charges -

Medical charges are the primary outcome of interest in this analysis as they directly reflect the financial impact of healthcare services on individuals and healthcare systems.

b. Predictor Variables – (Age, BMI, Sex, Smoker, Region, Children)

These predictors are chosen based on their known associations with healthcare costs, as evidenced by prior research and domain knowledge. Age, BMI, Sex, smoking status, and geographic region are commonly recognized as significant factors affecting medical charges.

c. Short Research Notes on Past Trends and Similar Regression Projects –

Past Trends -

Previous studies have consistently shown that factors such as age, BMI, and smoking status are strong predictors of medical charges, with older individuals, higher BMI, and smokers typically incurring higher healthcare costs.

Similar Regression Projects:

Regression analyses on medical cost datasets have demonstrated the importance of demographic and lifestyle factors in predicting healthcare expenses. Studies have also highlighted the need to consider interactions between variables to improve model accuracy.

d. Problem Statement for Linear Regression Analysis –

This project is being performed by the one of the healthcare industries to develop a predictive model that accurately estimates medical charges based on demographic and lifestyle factors. By understanding the relationships between predictors such as age, Sex, BMI, smoking status, region, and number of children, we aim to provide insights into healthcare expenditure patterns and inform decision-making processes related to resource allocation in the healthcare sector.

e. Describe Non-linearities and Interactions Amongst the Variables–

Non-linearities and interactions between variables will be explored during the regression analysis. This includes examining potential quadratic or higher-order relationships between predictors and the response variable and assessing interaction effects between variables (e.g., age and BMI) on medical charges.

f. Document Coefficients in Formatted Tables After Running Output in R –

(Intercept)	age	Sex male	BMI	children	Smoker yes	region northwest	region southeast	region southwest
-10177.0083	235.7135	-335.8178	311.0469	215.1597	23437.9035	110.4771	-947.0299	-420.3294

g. Include a Linear Model in the Form of y=f(x) –

$\text{Medical Charges}= \beta 0+\beta 1\times \text{age}+\beta 2\times \text{BMI}+\beta 3\times \text{Sex}+\beta 4\times \text{Smoker}+\beta 5\times \text{Region}+\beta 6\times \text{Children}+\epsilon$

Medical Charges=

$-10177.0083+235.7135\times \text{age}+311.0469\times \text{BMI}-335.8178\times \text{Sex}+23437.9035\times \text{Smoker}-947.0299\times \text{Region}-420.3294\times \text{Region}+110.4771\times \text{Region}+215.1597\times \text{Children}+\epsilon$

h. Describe variables that are not of interest and not part of regression and the reasons for the same –

None.

i. Assumptions for the linear regression model –

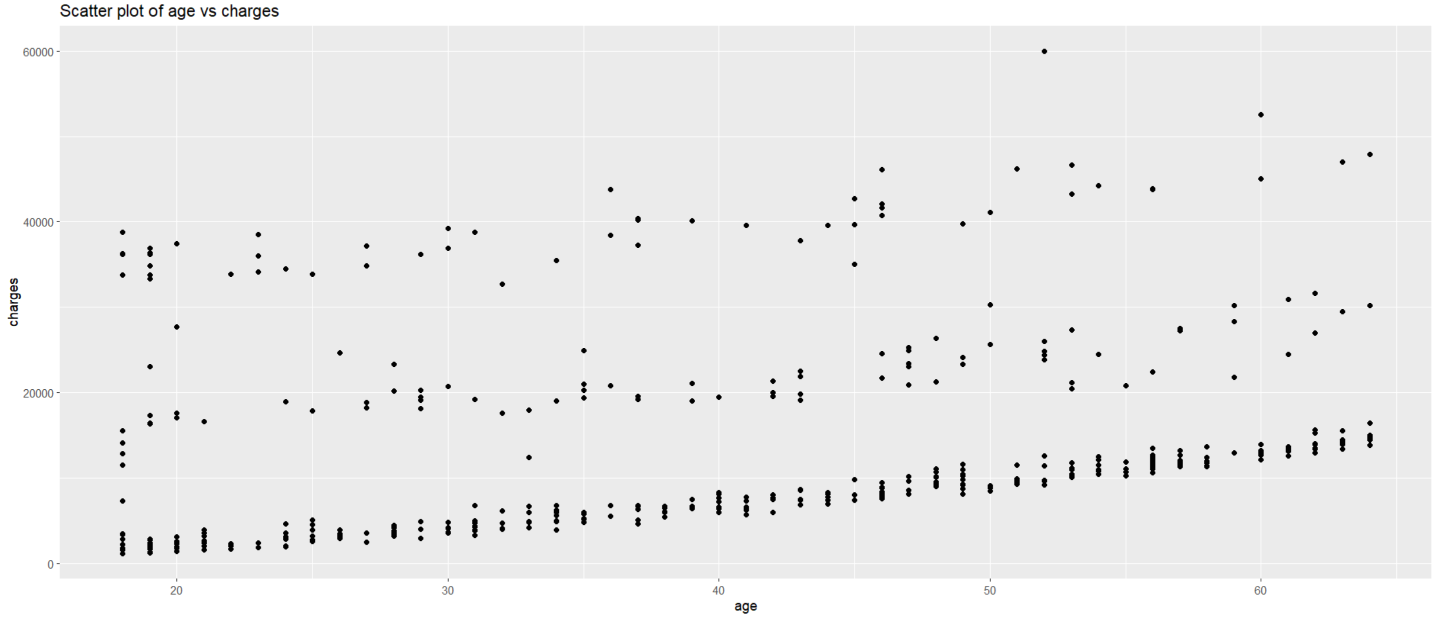
- **Linearity:** The relationship between predictors and the response variable is linear.
- **Independence of Errors:** Residuals are independent of each other.
- **Homoscedasticity:** Residuals have constant variance across predictor values.
- **Normality of Errors:** Residuals are normally distributed.
- **No Perfect Multicollinearity:** Predictor variables are not perfectly correlated.
- **No Autocorrelation:** Residuals are not correlated with each other.

3. Diagnostic Test

Plots for each predictor's vs response variables –

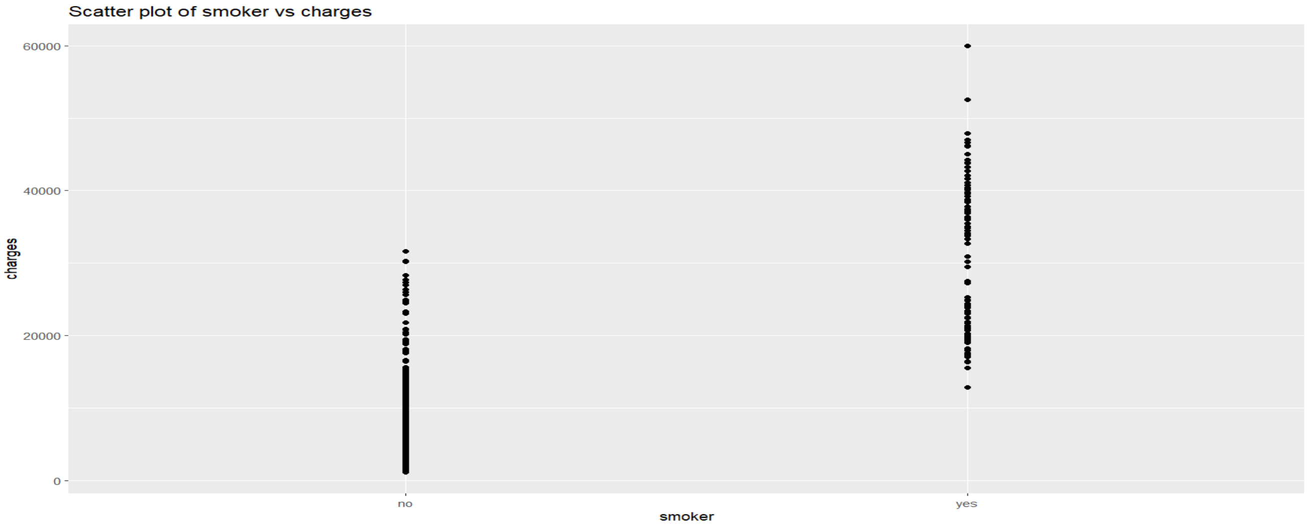
a. Age Vs Charges –

The scatter plot suggests that as age increases, medical charges also tend to increase.

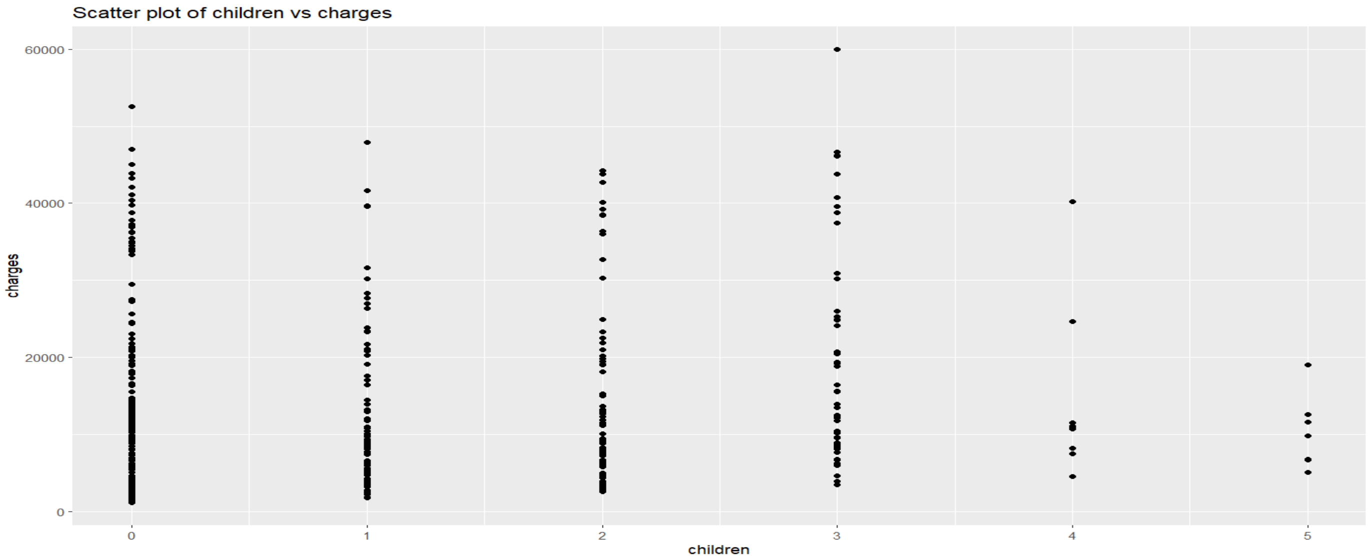


b. Smoker Vs Charges –

The scatter plot indicates that smokers tend to have higher medical charges compared to non-smokers. This observation aligns with the common understanding that smoking is associated with various health complications, which can lead to increased healthcare expenses.

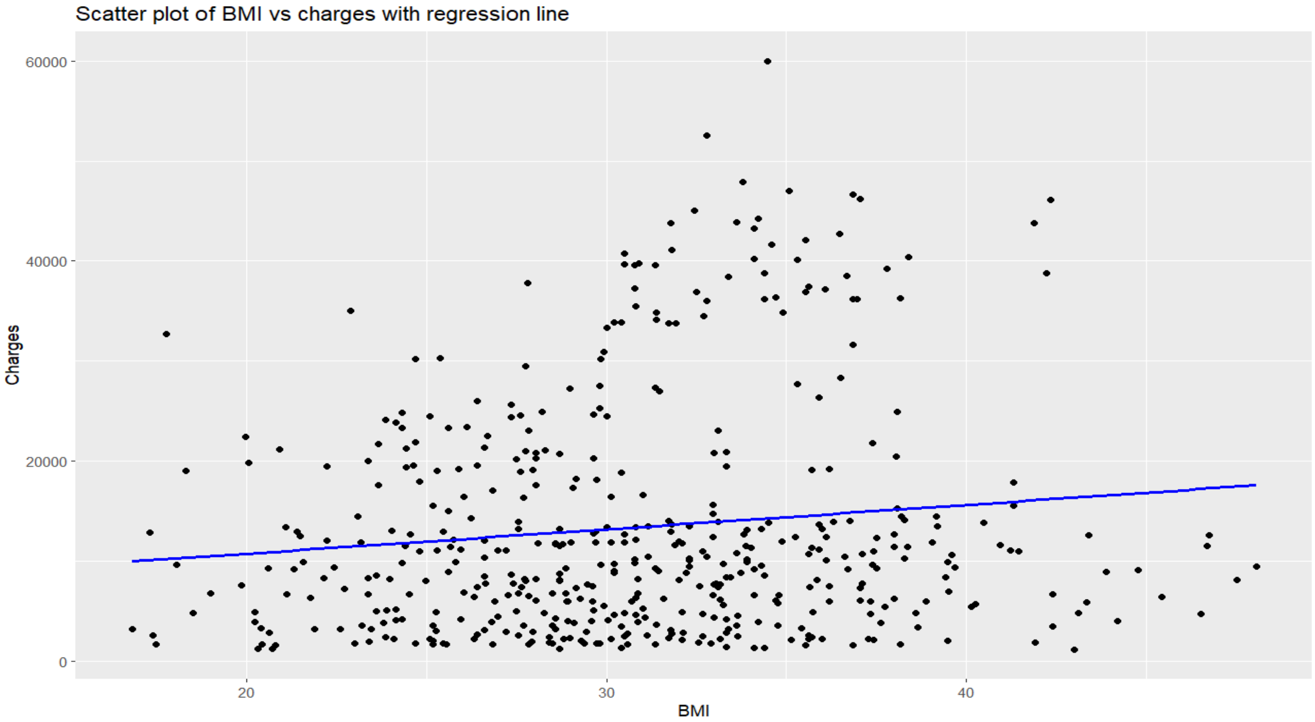


c. Children Vs Charges –



d. BMI Vs Charges –

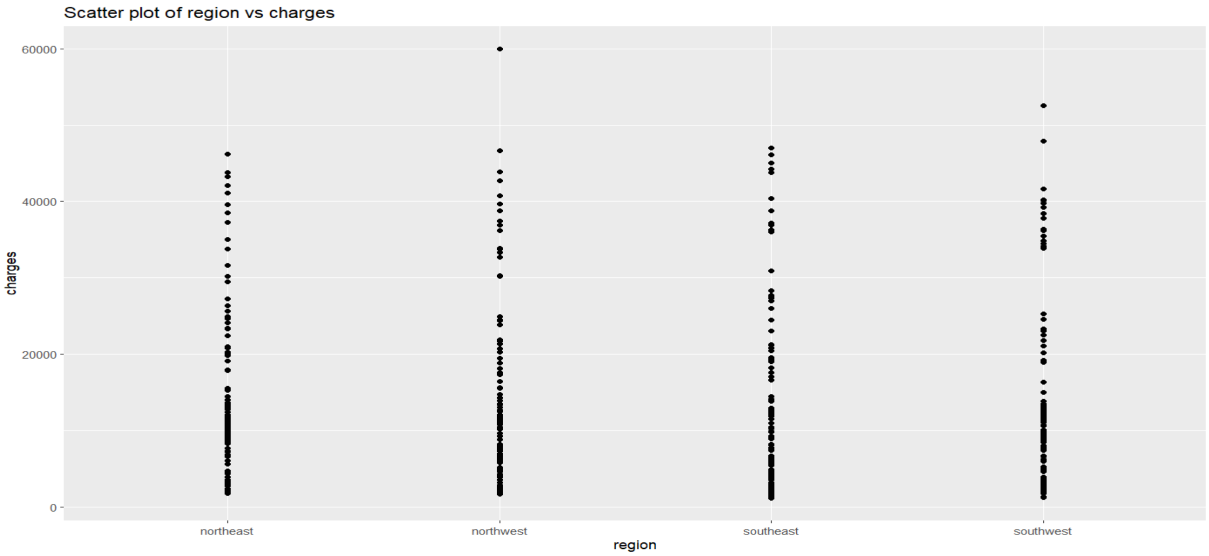
The regression line in the scatter plot of BMI versus medical charges is upward-sloping, suggesting a positive correlation between BMI and charges. This means that as BMI increases, medical charges also tend to increase.



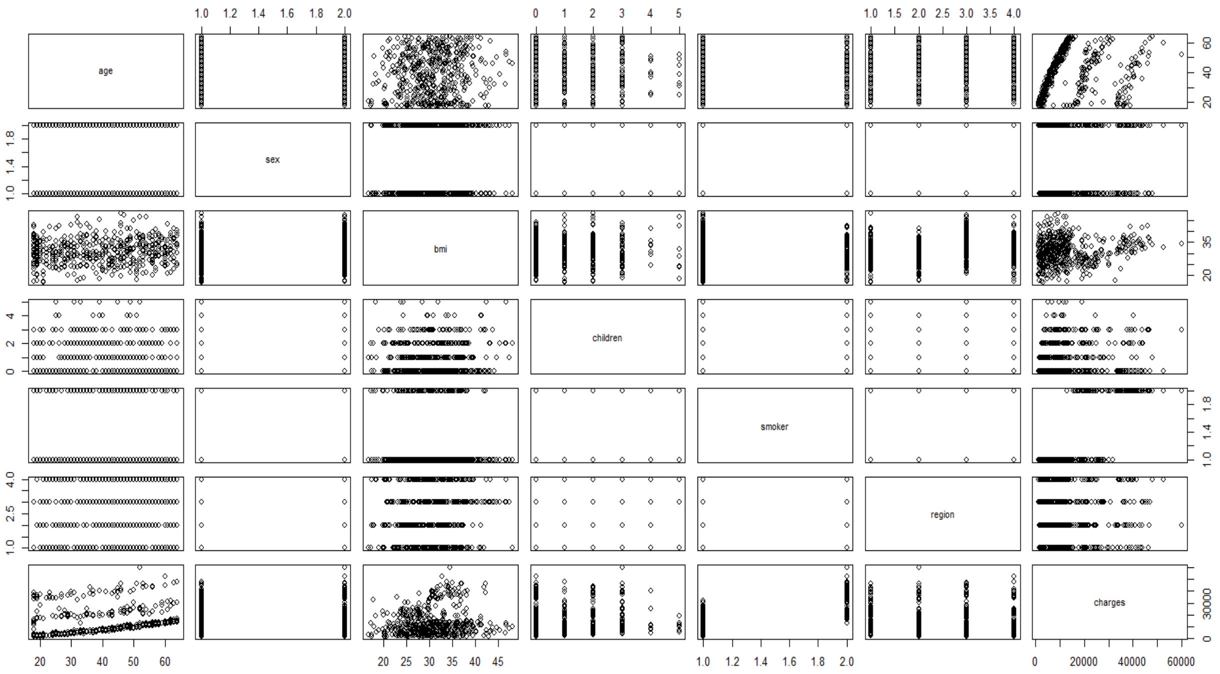
e. Region Vs Charges –

In this plot, the medical charges vary, differing region-wise, as the southwest region has the highest medical charges.

Also, the trend is upward-sloping, which indicates a positive relationship.

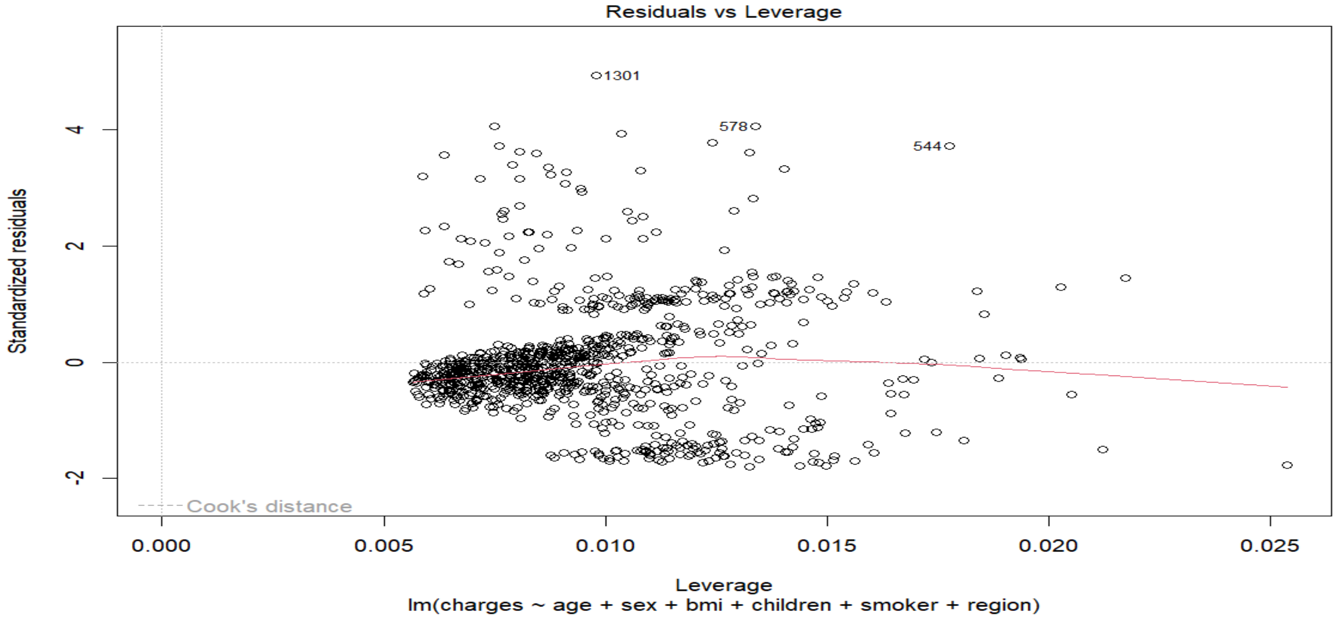


Dataset Plot -



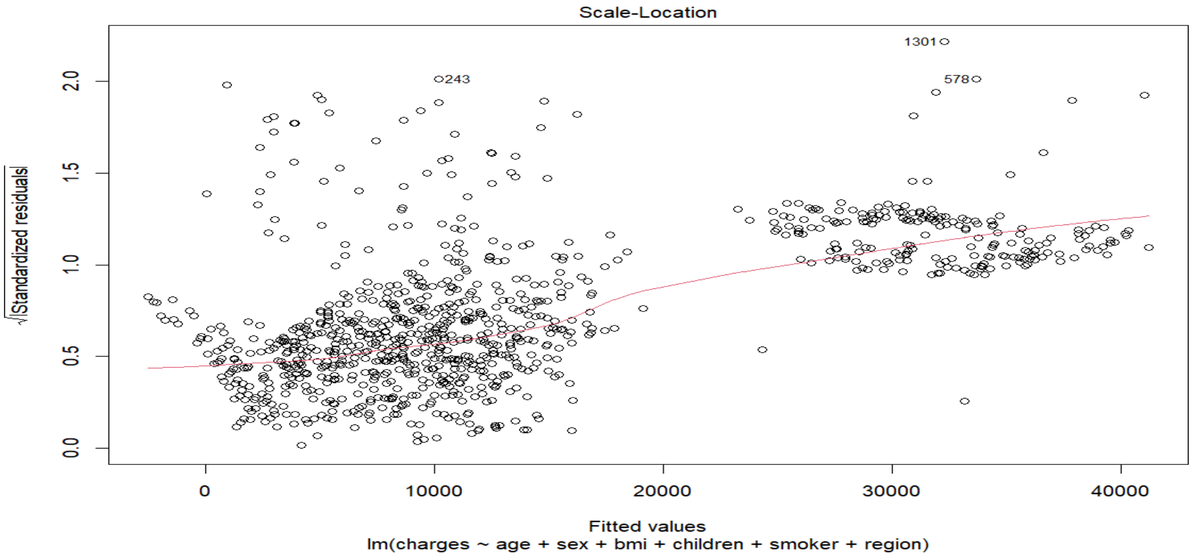
Plot for the initial model -

1. **Residuals vs Leverage plot -** This plot examines the influence of each observation on the regression coefficients by displaying Cook's distance against leveraged values. Cook's distance measures how much the predicted values of the response variable would change if a particular observation were removed from the analysis. In this plot, influential observations exceed a certain threshold, indicating that they disproportionately impact the regression model's coefficients. These influential observations can significantly affect the model's results and interpretations. Therefore, it's important to identify and understand these influential observations to ensure the validity and reliability of the regression analysis.

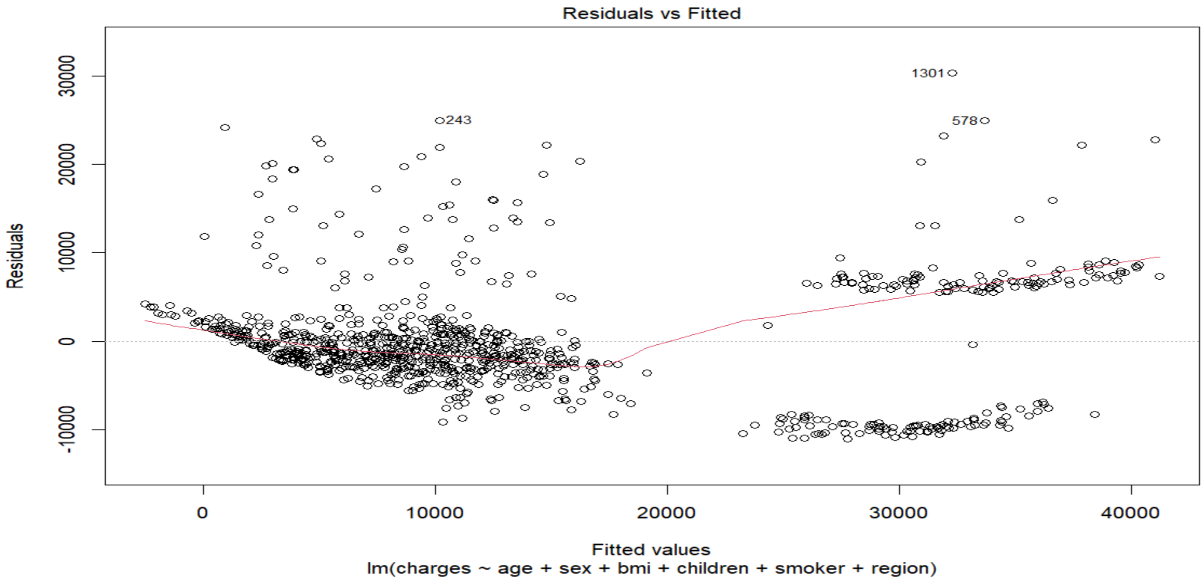


2. Scale -Location plot -

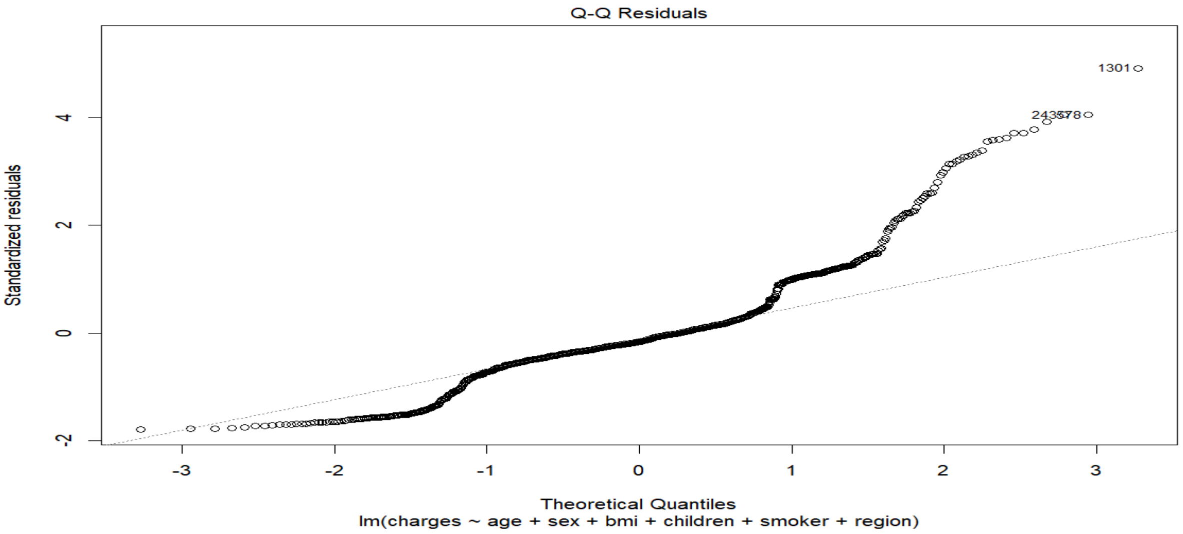
This plot assesses homoscedasticity by displaying the square root of standardized residuals against fitted values. The points form a horizontal line with no clear pattern, suggesting that the homoscedasticity assumption is met. However, suppose the points would fan out or form a funnel shape. In that case, it indicates heteroscedasticity, suggesting that the variance of the residuals is not constant across all levels of the predictor variable.



3. Residuals vs Fitted - This plot examines linearity and identifies outliers or influential observations by plotting residuals against fitted values. Most residuals are randomly scattered around zero, indicating that the linear relationship between predictors and the response variable is appropriate. However, there are some outliers where residuals deviate significantly from zero, suggesting the presence of influential observations. Additionally, there are instances where residuals tend to cluster towards the bottom and top of the zero line. These concentrations of residuals may indicate specific patterns or behaviours within the data, and they warrant further investigation to understand their potential impact on the model's assumptions and results.



4. Q-Q Plot - This plot assesses the normality of residuals by comparing their distribution to a theoretical normal distribution. Most points closely follow the diagonal line, suggesting that the residuals are approximately normally distributed. However, there are some deviations from the diagonal line, particularly towards the tails. These deviations indicate departures from normality, with some residuals exhibiting positive or negative skewness. Further examination of these deviations is necessary to understand their potential implications for the model's assumptions and results



4. Model Selection –

I employed the forward selection method to systematically add predictors to my model based on their individual contribution to improving model fit and predictive accuracy.

a. Detection of Non-linearities:

In the previous step, we examined diagnostic plots such as residuals vs. fitted values, QQ plots, and scale-location plots to detect non-linear relationships between predictors and the response variable. The non-linear patterns were observed in the plots, so just to avoid them, we could consider incorporating polynomial terms or transformations to address them.

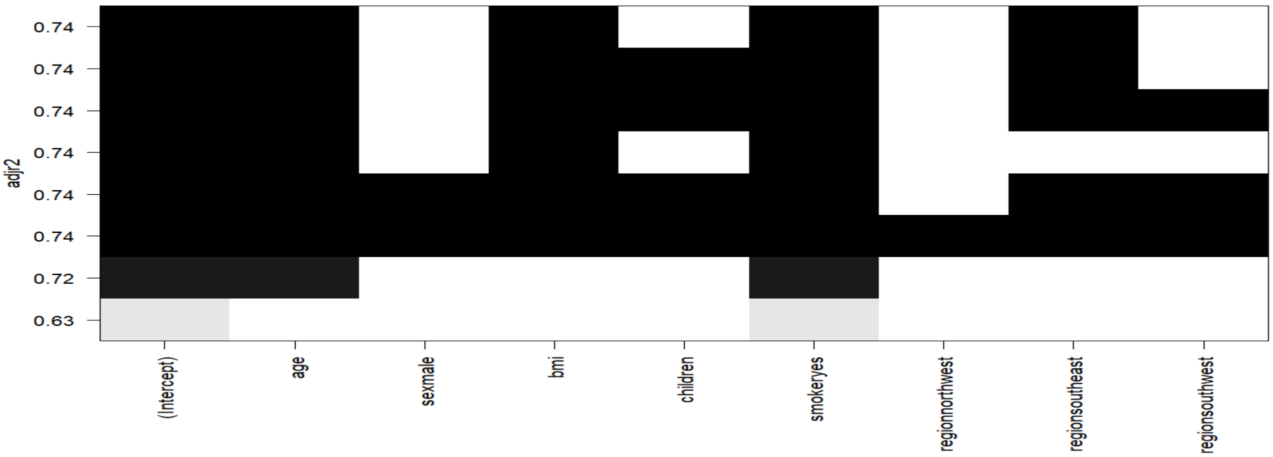
b. Variable Selection:

Variable selection is crucial for identifying the subset of predictors that contribute the most to explaining the variation in the response variable while minimizing overfitting. In our case, we used the best subset selection method to systematically evaluate all possible combinations of predictors and identify the subset that yields the best model fit.

The regsubsets function considers all possible combinations of predictors (age, sex, BMI, children, smoker, region) to find the subset that best fits the best model. This exhaustive search allows for identifying the optimal subset of predictors that maximizes the adjusted R-squared value, indicating the best model fit.

Once the best subset of predictors is identified, the adjusted R-squared values for different numbers of predictors are visualized using the plot function. This visualization helps select the subset of predictors that yields the highest adjusted R-squared value, indicating the best model fit.

As the plot below shows, age and smoke predictors have a higher impact on medical charges than others, so we can consider taking the polynomial steps. We could get more linearity in the model and increase its accuracy.



c. Model Evaluation and Selection:

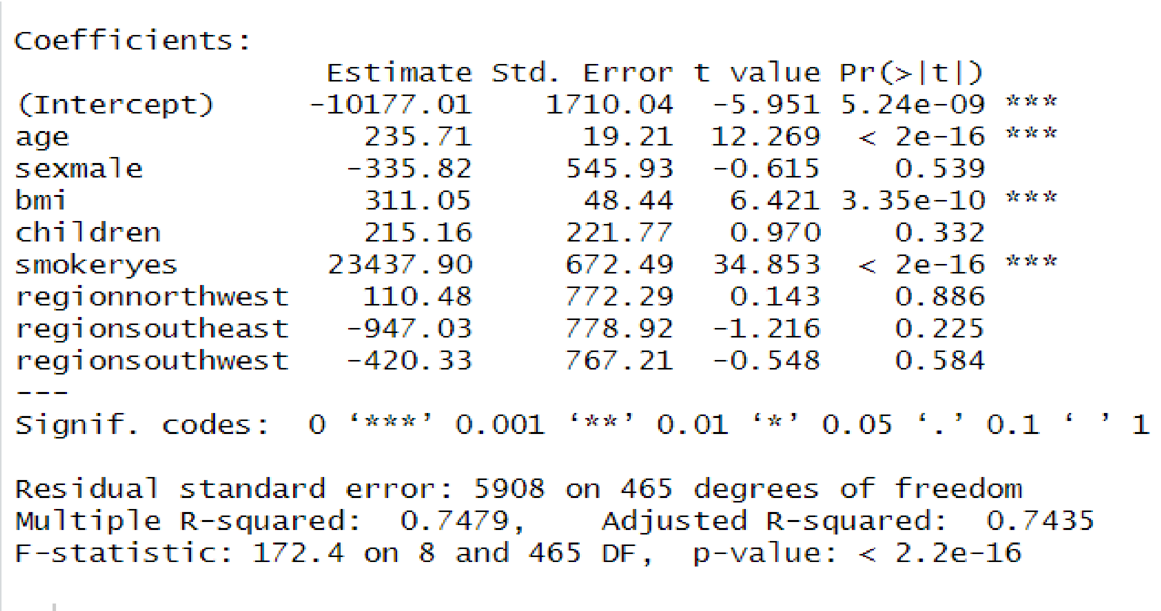
After selecting the subset of predictors, a linear regression model is trained using the selected predictors. This is done using the train function from the caret package.

The train function trains the linear regression model using the selected subset of predictors (age, sex, BMI, children, smoker, region) and the response variable (medical charges). Cross-validation with 6 folds is employed to evaluate the model's performance.

The resulting trained model, referred to as Train_Model2, captures the relationship between the predictors and the response variable. It can then be evaluated for its performance using various metrics such as R-squared, mean squared error, or root mean squared error. Additionally, the trained model can be used to make predictions on new data, providing insights into the factors influencing medical charges.

d. Screenshot of Summary Statistics:

Below is a screenshot of the summary statistics for the selected model, showcasing key metrics such as coefficients, standard errors, t-values, and p-values:



```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -10177.01    1710.04  -5.951 5.24e-09 ***
age                235.71      19.21  12.269  < 2e-16 ***
sexmale           -335.82     545.93  -0.615    0.539    
bmi                311.05      48.44   6.421 3.35e-10 ***
children           215.16     221.77   0.970    0.332    
smokeryes        23437.90     672.49  34.853  < 2e-16 ***
regionnorthwest    110.48     772.29   0.143    0.886    
regionsoutheast   -947.03     778.92  -1.216    0.225    
regionsouthwest   -420.33     767.21  -0.548    0.584    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5908 on 465 degrees of freedom
Multiple R-squared:  0.7479,    Adjusted R-squared:  0.7435 
F-statistic: 172.4 on 8 and 465 DF,  p-value: < 2.2e-16
```

5. Prediction and Summary –

Using our final model, we performed a prediction for an individual with the following characteristics: age 18, female, BMI 26.315, no children, non-smoker, and residing in the northeast region. The predicted medical charges for this individual were $2251.034. Upon comparison with the actual charges of $2198.18985, we observed a slight difference between the predicted and actual values.

Conclusion –

While our model provides reasonably accurate predictions, there is potential for improvement by considering the incorporation of polynomial terms for certain predictors. By increasing the polynomial degree for relevant predictors such as age or BMI, we can capture more complex relationships that may exist between these variables and medical charges. This approach could lead to a more flexible and adaptive model capable of better capturing the underlying patterns in the data. Additionally, further fine-tuning of the model parameters and feature engineering may contribute to enhancing its performance. Overall, by iteratively refining our model and exploring different modelling techniques, we aim to develop a robust predictive tool that can accurately estimate medical charges for individuals based on their demographic and lifestyle factors.