ADSC2020

Regression for Applied Data

Thompson Rivers University

Winter-2024

Predicting Healthcare Expenditures: A Regression Analysis Approach

Submitted To:

Prof. Sean Hellingman

Submitted By:

Akansha Bhargavi-T00736533

Solomon Maccarthy- T00734513

Viswateja Adothi-T00736529

Table of Contents

1.Introduction

1.1 Background

The realm of medical insurance operates within a complex ecosystem of healthcare needs, risk assessments, and financial considerations. At the heart of this system lies the pivotal concern of medical insurance costs an aspect that profoundly affects consumers, healthcare providers, and insurers alike. The determination of these costs, commonly referred to as insurance premiums or charges, is influenced by an intricate interplay of demographic, behavioral, and health-related factors.

1.2 Data Overview

The dataset consists of comprehensive array of variables, spanning demographic, behavioral, and health-related dimensions. Demographic information includes age, sex, region, and family size, providing insights into the population being insured. Behavioral factors encompass lifestyle choices such as smoking habits, while health-related metrics such as BMI (Body Mass Index) shed light on the health status of individuals. Additionally, data pertaining to insurance charges serve as the cornerstone for our analysis, allowing us to discern patterns and relationships within the dataset.

Variable	Description
Age	Spans a wide range of ages, indicating impact on medical insurance costs
Sex	Data split between male and female participants, suggesting potential gender differences in medical costs
BMI	Body Mass Index (BMI) as a crucial health metric, indicating weight relative to height
Number of Children	Accounts for family size, offering insights into how dependents might affect insurance costs
Smoker Status	Indicates smoking habits, a key factor influencing health and insurance premiums
Region	Captures geographic variation, suggesting potential disparities in medical costs by location
Charges	Primary outcome variable, insurance charges, serving as basis for analyzing factors influencing medical expenses

1.3 Descriptive statistics

Descriptive statistics offer a comprehensive summary of the dataset's characteristics, providing insights into central tendencies, variability, and distributions of the variables. Tables containing measures such as mean, median, standard deviation, and quartiles for age, BMI, number of children, and insurance charges are presented to elucidate the central tendencies and dispersion of these variables. Additionally, frequency distributions for categorical variables such as sex, smoker status, and region are provided to illustrate the distribution of data across different categories. These descriptive statistics serve as foundational elements for understanding the dataset and formulating subsequent analyses and interpretations.

1.4 Overall Objective

The primary objective of this project is to analyze the interplay between various factors and their collective impact on medical insurance costs. By employing descriptive statistics and regression analysis techniques, we aim to gain deeper insights into how age, sex, BMI, number of children, smoker status, and region influence insurance charges. Furthermore, we seek to develop predictive models capable of accurately estimating insurance premiums based on these factors. Ultimately, our goal is to provide insurance companies, policymakers, and individuals with actionable insights that can inform decision-making processes, promote transparency, and contribute to the development of a more equitable healthcare system.

2. Research Questions and Hypothesis

2.1 Research Question

What are the key factors influencing medical insurance costs, and how do they interact to determine insurance charges?

2.2 Hypothesis

1. H0: There is no significant difference in medical insurance costs between smokers and non-smokers.

   H1: Smokers incur higher medical insurance costs compared to non-smokers.

2. H0: Age does not significantly affect medical insurance charges.
   H1: Increasing age is positively associated with higher medical insurance charges.

3. H0: BMI has no impact on medical insurance costs.
   H1: Higher BMI levels are associated with increased medical insurance charges.

4. H0: There is no significant interaction effect between sex and BMI on medical insurance charges.
   H1: The relationship between BMI and medical insurance charges differs between males and females, indicating a significant interaction effect.

5. H0: There is no regional variation in medical insurance charges.
   H1: Geographic location influences medical insurance costs, with certain regions experiencing higher charges than others.

3. Model Selection

Our analysis of medical insurance costs employed a comprehensive approach, focusing on factors like age, BMI, smoking status, number of children, sex, and region. We utilized Exploratory Data Analysis (EDA), variable selection, linear regression modeling, Generalized Linear Models (GLMs), and rigorous model diagnostics. These techniques allowed us to uncover insights into how these factors impact insurance charges, ensuring reliability and validity through assessments of model assumptions and prediction accuracy.

3.1 Data Analysis Techniques

3.1.1 Exploratory Data Analysis (EDA):

The EDA phase utilizes histograms, box plots, and bar plots to visually inspect the distribution of insurance charges and the prevalence of key variables like smoker status and number of children. This visual exploration aids in identifying trends, outliers, and potential relationships between the variables and the target outcome (insurance charges).

3.1.2 Variable selection:

Variable selection provides a foundation for exploring the relationships among predictor variables, identifying multicollinearity issues, and potentially selecting a subset of predictors for medical cost analysis

3.1.3 Linear Regression Analysis:

The core of the analysis involves fitting linear regression models to examine the relationship between predictors (age, BMI, etc.) and the response variable (insurance charges). Initial models assess the impact of individual predictors through simple linear regression, while multiple regression models explore the combined effects and interactions between variables.

3.1.4 Model Assumptions and Diagnostics:

Throughout the analysis, rigorous diagnostics are performed to assess the validity of model assumptions. This includes examining residuals for patterns indicating non-linearity or heteroscedasticity, conducting normality tests, and employing the Cook's distance measure to identify influential outliers. Adjustments and transformations, such as the Box-Cox and logarithmic transformations, are applied as necessary to address identified issues.

3.1.5 Generalized Linear Models (GLMs):

Given the presence of model assumption violations—such as non-normality of residuals and heteroscedasticity—alternative modeling approaches are considered. GLMs with an inverse Gaussian family are explored to accommodate the skewed nature of the insurance charges data, providing a flexible framework to model the expected mean of charges as a function of the predictors.

3.1.6 Prediction:

This method provides a robust assessment of model performance across different subsets of the data, ultimately aiding in model selection and refinement.

4. Results

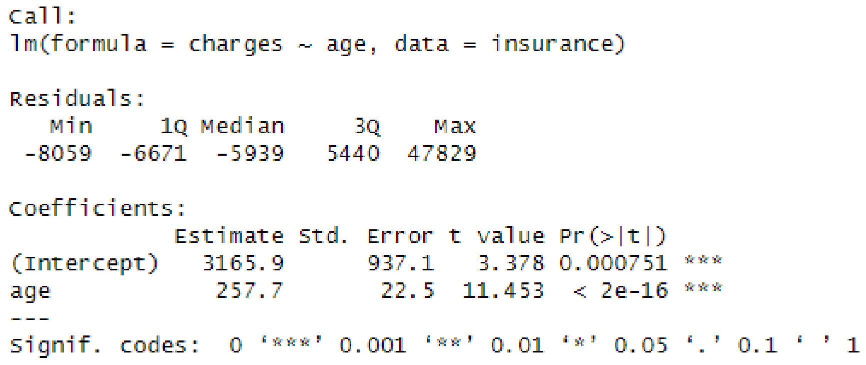
4.1 Variable selection

Correlation: Age and BMI are important to consider in models predicting insurance charges. Other factors also have significantly influence insurance charges

VIF: The multicollinearity diagnostics indicate that the variables selected for the model do not suffer from high multicollinearity

4.2 Simple Linear Regression

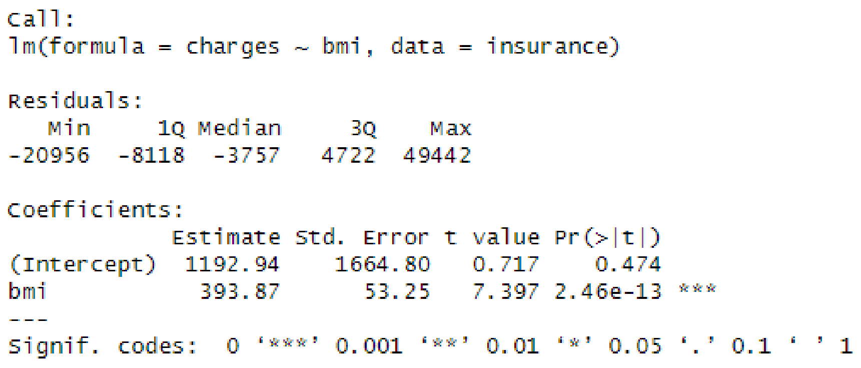
Model 1: Influence of Age on Insurance Charges



```
Call:
lm(formula = charges ~ age, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max 
 -8059  -6671  -5939   5440  47829 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3165.9      937.1   3.378 0.000751 ***
age            257.7       22.5  11.453  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This linear regression model examines the relationship between insurance charges and age. The output indicates a significant positive association between age and charges, with older individuals generally facing higher insurance costs. However, the model explains only a small portion of the variability in charges based on age.

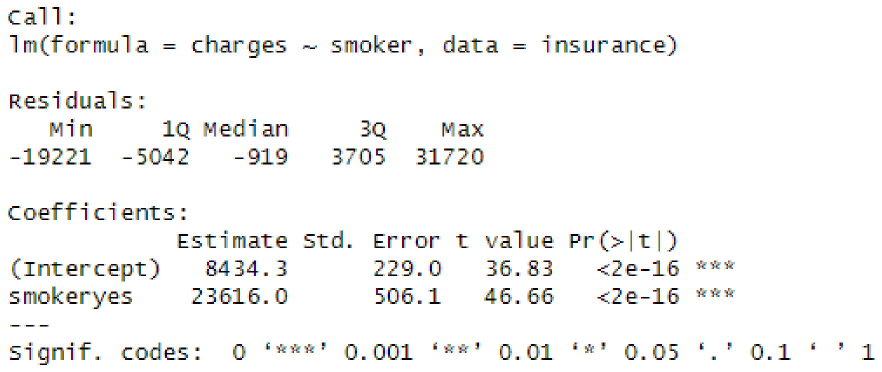
Model 2: Influence of BMI on Insurance Charges



```
Call:
lm(formula = charges ~ bmi, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max 
-20956  -8118  -3757   4722  49442 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1192.94    1664.80   0.717    0.474    
bmi           393.87      53.25   7.397 2.46e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The output suggests a significant positive association between BMI and charges, indicating that individuals with higher BMIs tend to have higher insurance costs.

Model 3: Influence of smoker on Insurance Charges

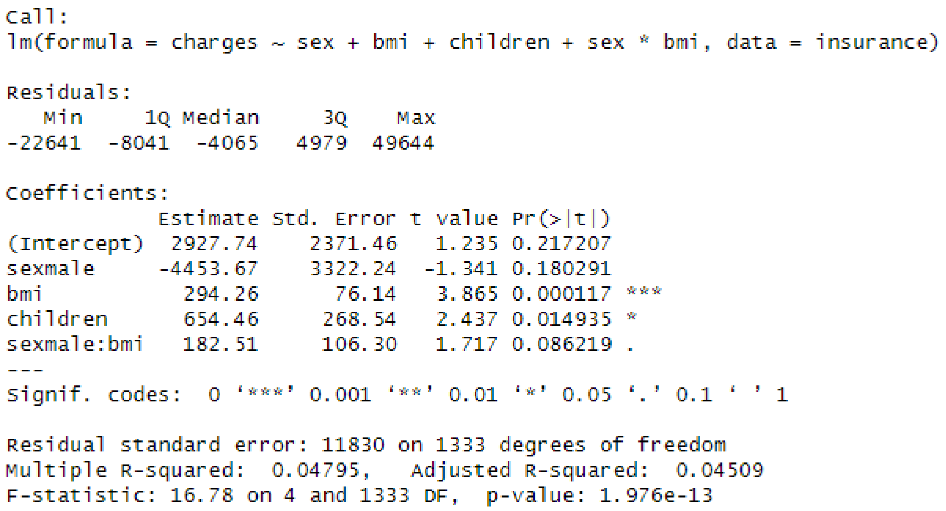


```
Call:
lm(formula = charges ~ smoker, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max 
-19221  -5042   -919   3705  31720 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8434.3      229.0   36.83   <2e-16 ***
smokeryes    23616.0      506.1   46.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The output indicates a significant difference in charges between smokers and non-smokers, with smokers having substantially higher charges. The model explains a considerable proportion of the variability in charges based on smoking status.

4.3 Multiple Linear Regression

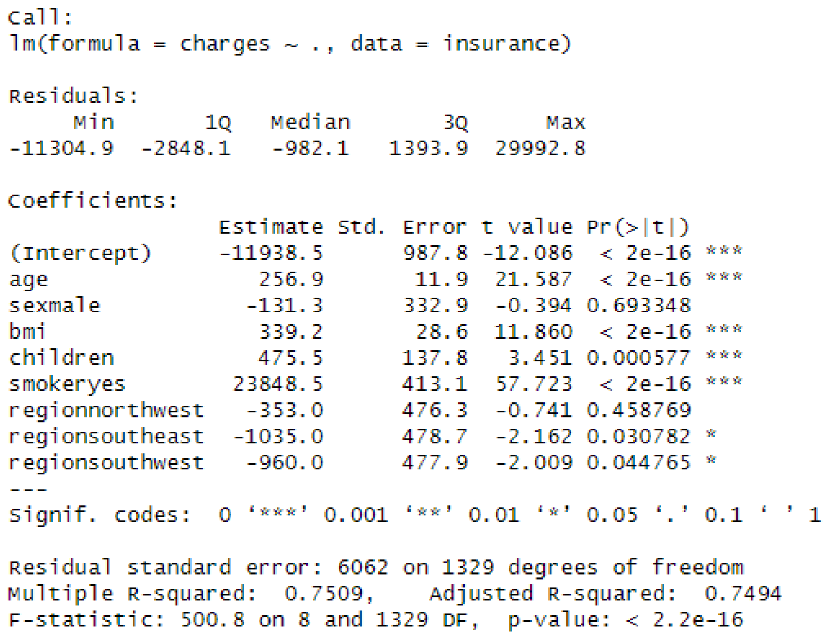
Model 1: Including interaction term (sex*bmi)



```
Call:
lm(formula = charges ~ sex + bmi + children + sex * bmi, data = insurance)

Residuals:
   Min     1Q Median     3Q    Max 
-22641  -8041  -4065   4979  49644 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2927.74    2371.46   1.235 0.217207    
sexmale     -4453.67    3322.24  -1.341 0.180291    
bmi           294.26      76.14   3.865 0.000117 ***
children      654.46     268.54   2.437 0.014935 *  
sexmale:bmi   182.51     106.30   1.717 0.086219 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11830 on 1333 degrees of freedom
Multiple R-squared:  0.04795,	Adjusted R-squared:  0.04509 
F-statistic: 16.78 on 4 and 1333 DF,  p-value: 1.976e-13
```

The model's explanatory power is relatively low (Adjusted R-squared = 0.04509), suggesting that the included variables explain only a small proportion of the variability in insurance charges. Additionally, the p-value associated with the F-statistic indicates that the overall regression model is statistically significant (p-value = 1.976e-13), meaning that at least one of the predictor variables has a significant effect on insurance charges.

Model 2: Including all the independent factors

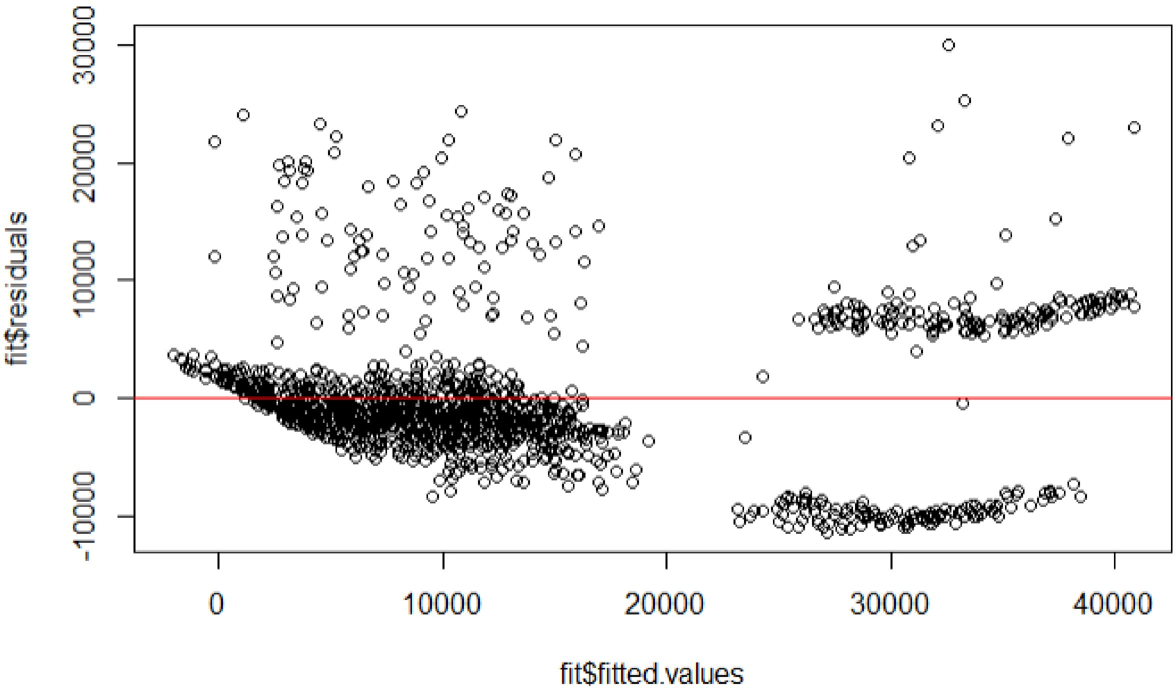


```
Call:
lm(formula = charges ~ ., data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-11304.9  -2848.1   -982.1   1393.9  29992.8 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
age                256.9       11.9  21.587  < 2e-16 ***
sexmale           -131.3      332.9  -0.394 0.693348    
bmi                339.2       28.6  11.860  < 2e-16 ***
children           475.5      137.8   3.451 0.000577 ***
smokeryes        23848.5      413.1  57.723  < 2e-16 ***
regionnorthwest   -353.0      476.3  -0.741 0.458769    
regionsoutheast  -1035.0      478.7  -2.162 0.030782 *  
regionsouthwest   -960.0      477.9  -2.009 0.044765 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,	Adjusted R-squared:  0.7494 
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

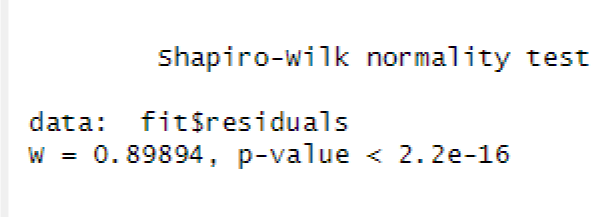
The model has a relatively high explanatory power (Adjusted R-squared = 0.7494), suggesting that the included variables collectively explain a substantial proportion of the variability in insurance charges. The p-value associated with the F-statistic is very low, indicating that the overall regression model is statistically significant (p-value < 2.2e-16).

4.4 Model Diagnostics:

Linearity: This implies that the relationship between the explanatory variables and the response variable charges might not be strictly linear, which could impact the model's accuracy.

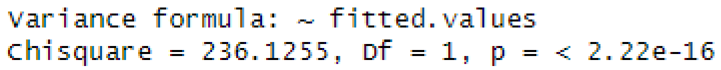


Normality: The Shapiro-Wilk normality test on the residuals yielded a p-value of < 2.2e-16, indicating a significant deviation from normality.



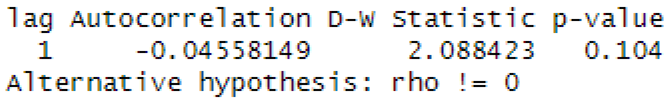
```
	Shapiro-Wilk normality test

data:  fit$residuals
W = 0.89894, p-value < 2.2e-16
```

Homoscadicity: The non-constant variance score test showed a significant violation of the assumption of constant variance (homoscedasticity) with a p-value of < 2.22e-16.



```
Variance formula: ~ fitted.values
Chisquare = 236.1255, Df = 1, p = < 2.22e-16
```

Independence: The lag Autocorrelation Durbin-Watson statistic yielded a value of 2.088423 with a corresponding p-value of 0.104, indicating no significant autocorrelation at lag 1. Therefore, we fail to reject the null hypothesis that there is no autocorrelation in the residuals at lag 1.



```
lag Autocorrelation D-W Statistic p-value
   1     -0.04558149      2.088423   0.104
 Alternative hypothesis: rho != 0
```

4.5 Transformations:

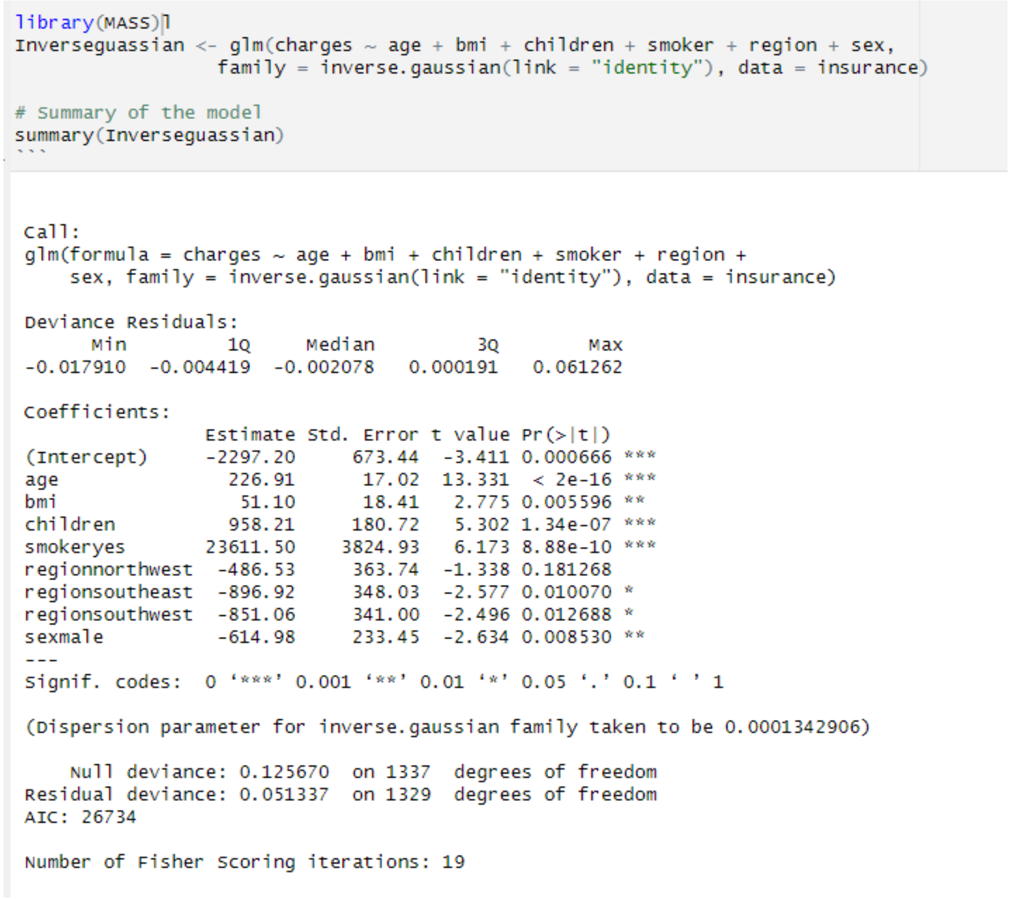
As per our diagnostics linearity,normality,homoscadacity are violated hence we are performing

- Box-Cox transformation to find an optimal lambda and transform the response variable
- Created a function to identify and remove outliers based on Cook's distance

4.6 Generalized Linear Models:

Due to persistent violations of assumptions even after transformation attempts, we have shifted our modeling approach to Generalized Linear Models (GLMs). Specifically, we are utilizing the Gamma distribution, which is well-suited for with non-normally distributed data as indicated by histograms.

Model: this GLM formulation aims to capture the relationship between the predictor variables and the response variable charges, we are specifying the family distribution to be inverse Gaussian (family = inverse.gaussian) and utilizing the identity link function (link = "identity").



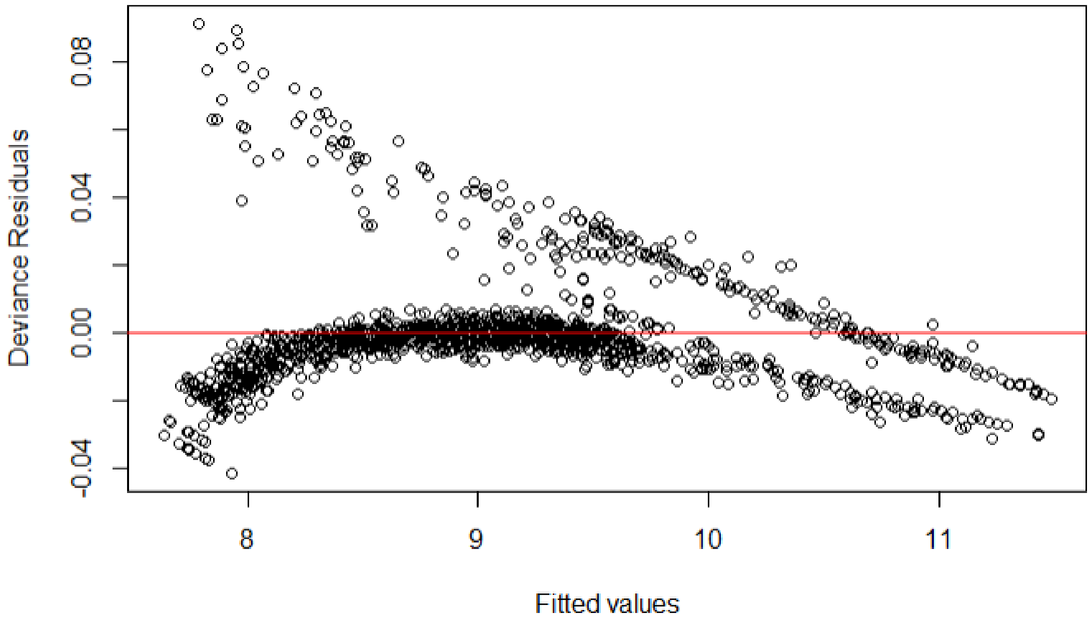
```
library(MASS)
Inverseguassian <- glm(charges ~ age + bmi + children + smoker + region + sex,
                  family = inverse.gaussian(link = "identity"), data = insurance)

# Summary of the model
summary(Inverseguassian)
```

```
Call:
glm(formula = charges ~ age + bmi + children + smoker + region + 
    sex, family = inverse.gaussian(link = "identity"), data = insurance)

Deviance Residuals: 
      Min         1Q     Median         3Q        Max  
-0.017910  -0.004419  -0.002078   0.000191   0.061262  

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -2297.20     673.44  -3.411 0.000666 ***
age               226.91      17.02  13.331  < 2e-16 ***
bmi                51.10      18.41   2.775 0.005596 ** 
children          958.21     180.72   5.302 1.34e-07 ***
smokeryes       23611.50    3824.93   6.173 8.88e-10 ***
regionnorthwest  -486.53     363.74  -1.338 0.181268    
regionsoutheast  -896.92     348.03  -2.577 0.010070 *  
regionsouthwest  -851.06     341.00  -2.496 0.012688 *  
sexmale          -614.98     233.45  -2.634 0.008530 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for inverse.gaussian family taken to be 0.0001342906)

    Null deviance: 0.125670  on 1337  degrees of freedom
Residual deviance: 0.051337  on 1329  degrees of freedom
AIC: 26734

Number of Fisher Scoring iterations: 19
```

This results provide insights into the relationship between predictor variables and insurance charges, accounting for the non-normal distribution of the response variable and potentially improving the model's performance compared to linear regression models. The overdispersion is 0.3601945 which states that the model is not overdispersed.

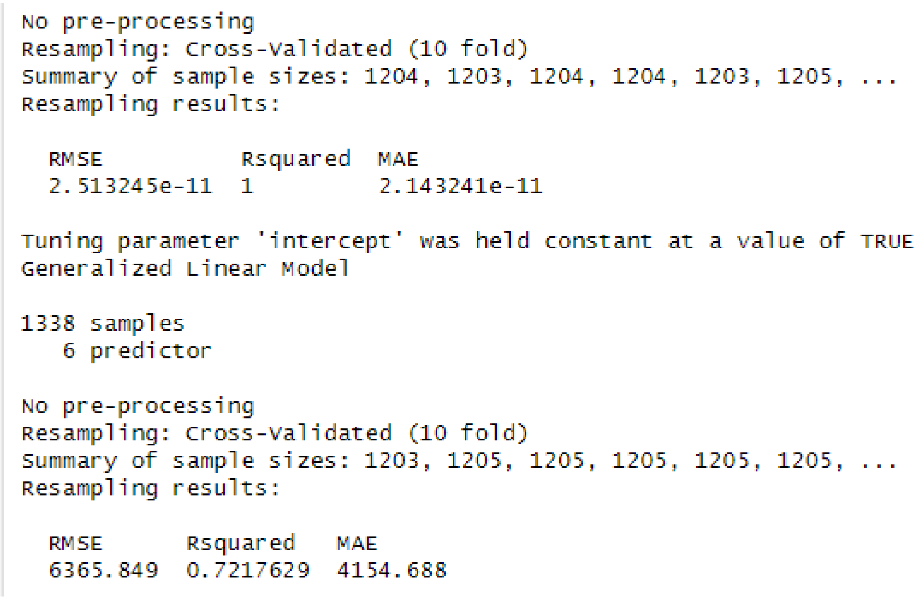
4.6.1 Model Diagnostics:

Linearity: This implies that the relationship between the explanatory variables and the response variable charges is approximately linear



4.7 Prediction:

Due to the high precision and variability of our response variable, we observe elevated RMSE and MAE values. These findings highlight the model's predictive performance, with the generalized linear model indicating higher RMSE and MAE values compared to the linear regression model.



```
No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1204, 1203, 1204, 1204, 1203, 1205, ...
Resampling results:

  RMSE          Rsquared  MAE
  2.513245e-11  1         2.143241e-11

Tuning parameter 'intercept' was held constant at a value of TRUE
Generalized Linear Model

1338 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1203, 1205, 1205, 1205, 1205, 1205, ...
Resampling results:

  RMSE      Rsquared   MAE
  6365.849  0.7217629  4154.688
```

5. Conclusion

This analysis explored insurance charges using a range of regression models and diagnostics. Initial simple linear and multiple regression models revealed significant associations but encountered challenges with assumptions. Generalized Linear Models, specifically employing the inverse Gaussian distribution, effectively tackled assumption violations. Despite elevated RMSE and MAE values attributed to high variability, GLMs demonstrated potential in capturing intricate relationships. Further refinement holds promise for improving predictive capabilities regarding insurance charge determinants.

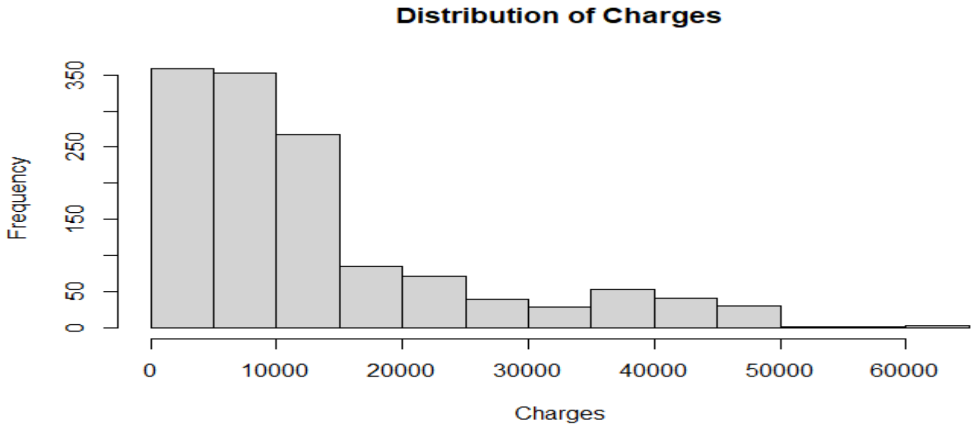
6. Appendices

6.1 Data Visualization

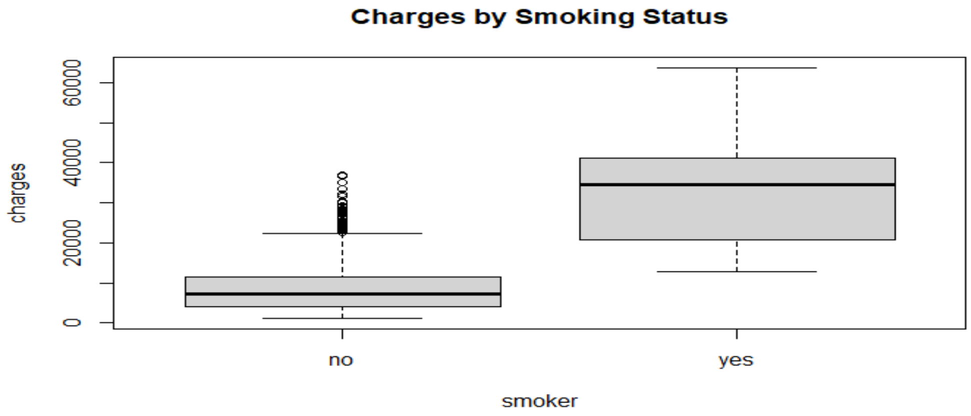
6.1.1 Distribution of charges

Visualizations, including Bar chart, histogram, boxplot were employed to discover trends and associations within the data.

A pictorial view of the distribution of Charges (Histogram)



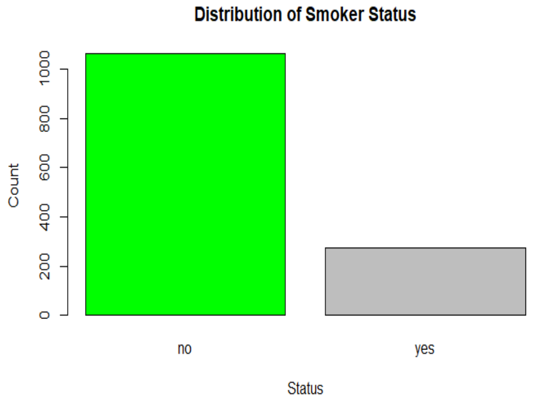
The histogram shows that our distribution is Right positively skewed and also clearly states the distribution is not normal.

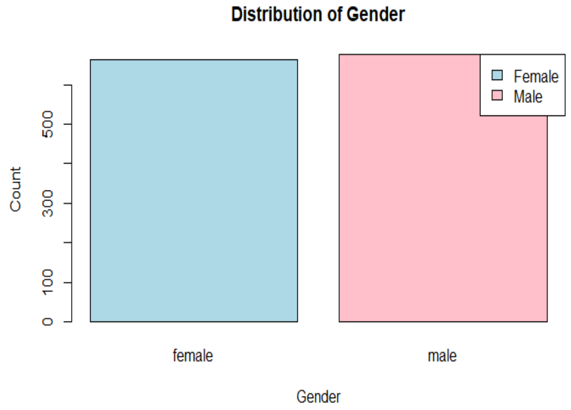


Based on the visualization we can say that individuals with “Yes” Smoking status have a high median and with the “No” status have a high variability.

6.1.2 Distribution of smoker status and Gender:

A pictorial view of distributions of smokers’ status,Sex (Bar Chart)



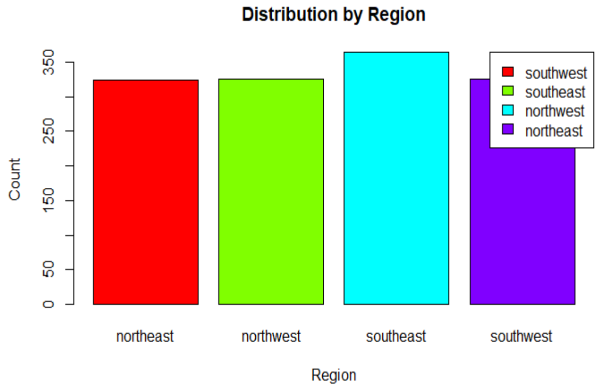


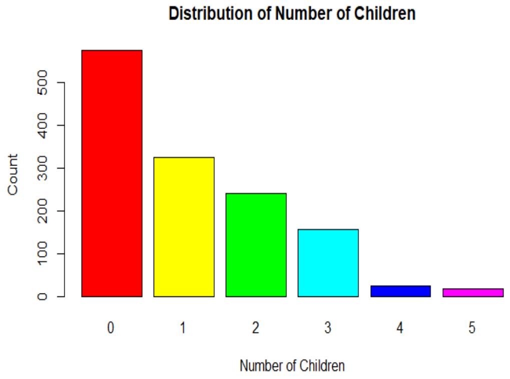
Based on distribution of smoker status we can say that we have more non smokers than smokers.

Based on gender distribution we can say that both female and male are approximately equally distributed.

6.1.3 Distribution if Region and Children:

A bar graph of our categorical variables verifying trends of how they perform.





Based on distribution of region we can say that most of them are from south east and remaining all are almost same.

Based on children distribution we can say that most of them does not have children or have only one child.

6.2 Code

draft_project.Rmd

6.3 References

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

Team Work:

Our team worked cohesively to analyze and interpret the dataset findings. Each member contributed equally to modeling, interpretation, and documentation tasks, ensuring a comprehensive and balanced approach to the analysis.