Health Insurance Linear Regression

Aidan Mitchell

Abstract

This analysis examines the impact of **age, smoking status, and obesity** on health insurance charges using a dataset of policyholders. Through **exploratory data analysis**, I identified three distinct groups—**nonsmokers, non-obese smokers, and obese smokers**—each exhibiting unique trends in medical costs. A comprehensive **linear regression model with interaction terms** was constructed to assess how these factors influence charges.

Key findings indicate that **smokers, particularly those with obesity, experience the highest insurance costs**, with age playing a significant role in increasing charges. While transformations (Box-Cox, log, square root) were tested, they did not significantly improve model performance. The final model explained approximately **85% of the variance in insurance charges** (Adjusted $R^2 \approx 0.86$), confirming the strong influence of these risk factors.

These insights emphasize the need for **risk-adjusted insurance premiums and targeted health interventions** aimed at smoking cessation and obesity prevention. By integrating statistical modeling with policy implications, this study provides actionable insights for **insurers, healthcare providers, and policymakers** in managing healthcare costs effectively.

Introduction

Dataset Overview

The dataset details medical insurance costs billed to policyholders based on certain demographic characteristics and habits

Source

Kaggle: https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data

Variables

- Age
  - Data type: Integer
  - Meaning: The age (in years) of the primary beneficiary
- Sex
  - Data type:
  - Meaning:
- bmi
  - Data type: Float
  - Meaning: Body Mass Index, providing a simple numeric measure of a person's weight relative to height.
- Children
  - Data type: Integer
  - Meaning: Number of children covered by the insurance plan
- Smoker
  - Data type: String (commonly "yes" or "no")
  - Meaning: Indicates whether the individual smokes tobacco
- Region
  - Data type: String
  - Meaning: The residential area of the beneficiary (often "southwest," "southeast," "northwest," or "northeast")
- Charges
  - Data type: Float
  - Meaning: Individual medical costs billed by health insurance (the target variable) # Data Processing Describe the steps taken for data cleaning, handling missing values, and any necessary pre- processing. Mention any transformations or feature engineering applied to improve the model's performance.

Key Questions

1. How does the age of insured individuals affect the cost of health insurance?
2. Does being a smoker affect the cost of health insurance charges?

Hypotheses

1. H_0 : An increase in age is not associated with the cost of health insurance costs

   H_A : An increase in age is associated with the cost of health insurance costs

2. H_0 : Being a smoker is not associated with the cost of health insurance charges

   H_A : Being a smoker is associated with the cost of health insurance charges

Exploratory Data Analysis (EDA)

Data Cleaning

```
## [1] "Total missing values: 0"
```

```
## [1] "Total duplicate values: 0"
```

Descriptive Statistics

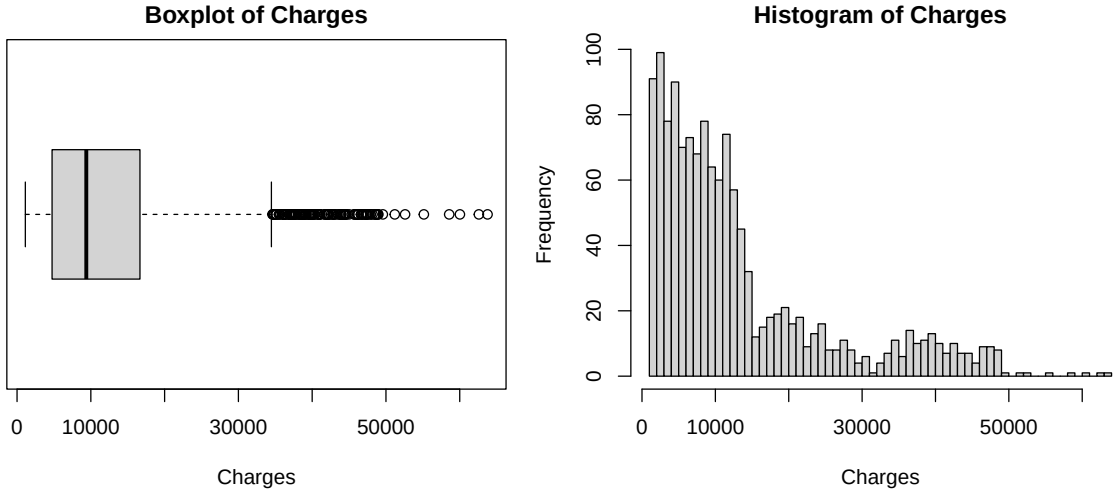
Table 1: Summary Statistics

age	sex	bmi	children	smoker	region	charges
Min. :18.00	Length:1337	Min. :15.96	Min. :0.000	Length:1337	Length:1337	Min. : 1122
1st Qu.:27.00	Class :character	1st Qu.:26.29	1st Qu.:0.000	Class :character	Class :character	1st Qu.: 4746
Median :39.00	Mode :character	Median :30.40	Median :1.000	Mode :character	Mode :character	Median : 9386
Mean :39.22	NA	Mean :30.66	Mean :1.096	NA	NA	Mean :13279
3rd Qu.:51.00	NA	3rd Qu.:34.70	3rd Qu.:2.000	NA	NA	3rd Qu.:16658
Max. :64.00	NA	Max. :53.13	Max. :5.000	NA	NA	Max. :63770

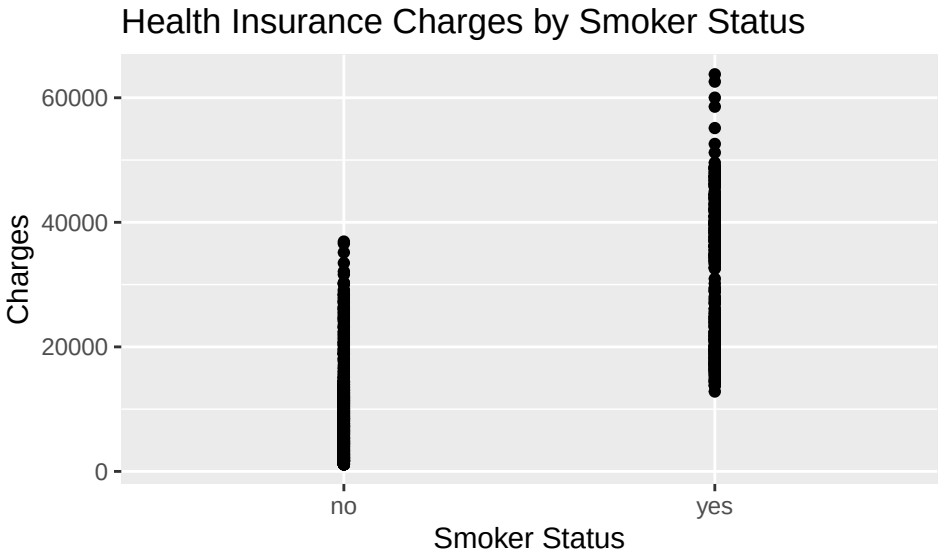
Table 2: Standard Deviations

	sd
age	14.044333
bmi	6.100468
children	1.205571
charges	12110.359656

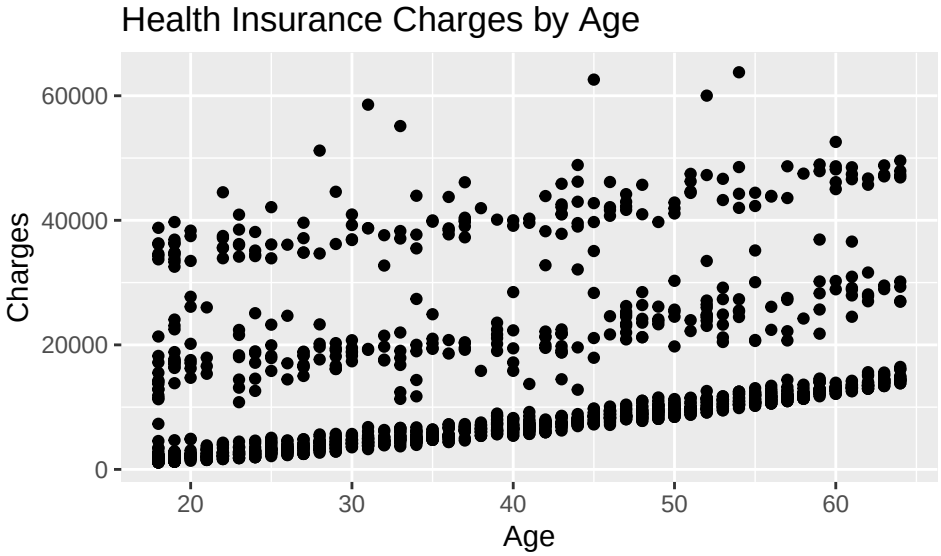
Visualizations



- From the Boxplot, the high number of outliers suggests an underlying factor influencing charges, that may require further exploration
- The histogram reveals three distinct peaks, suggesting there may be multiple distributions. This indicates that the reason for a high number of outliers may have been due to capturing multiple underlying groups as a single trend.

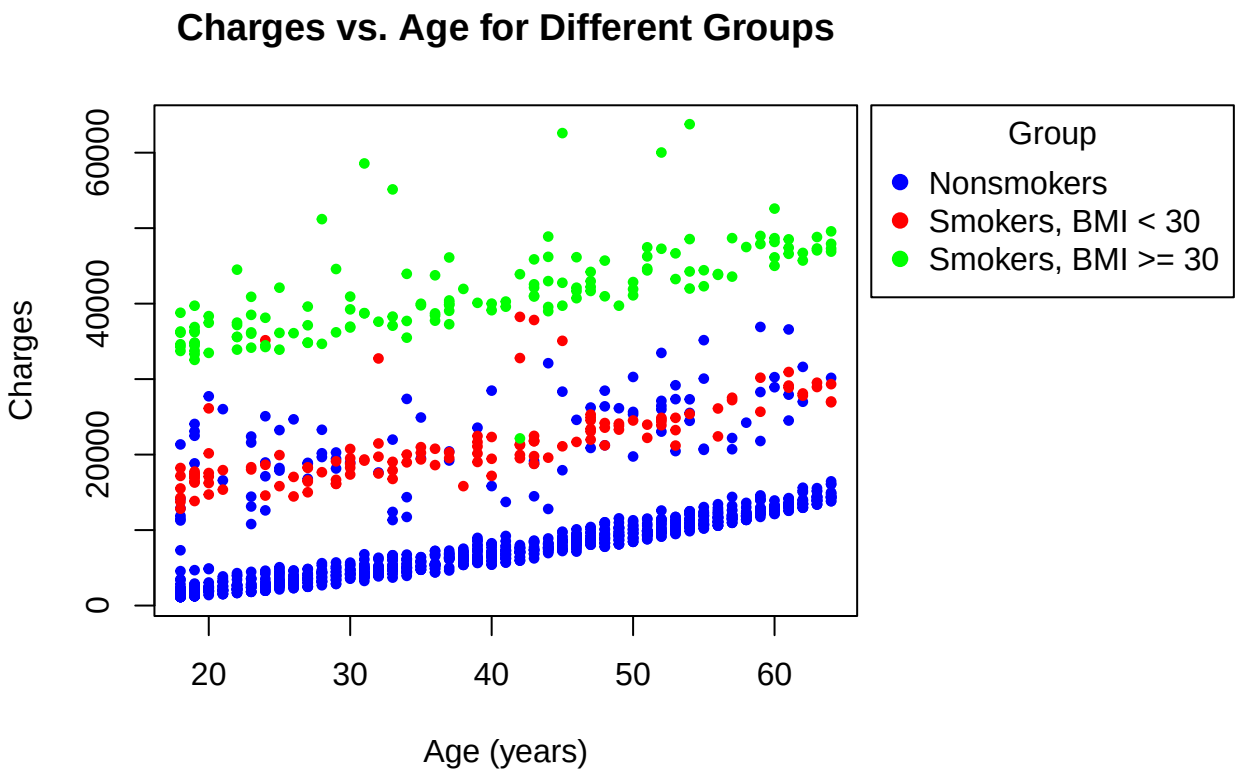


- After graphing `Charges` against `Smoker Status`, there seems to be a clear upwards trend corresponding to being a smoker. Those who smoke have on average higher charges, based on this initial plot.



- But when looking at the plot of `Charges` vs `Age`, we noticed that there are three clear, separate trends beginning from various Y intercepts. This led me to investigate into why this was the case, and what was separating the points into groups like this.

- I discovered that the trends that the data follows stems from three different groups
  - Nonsmokers
  - Smokers with Low BMI (BMI $<$ 30)
  - Smokers with High BMI (BMI $>=$ 30)
- This is imminent if you color code each point for the group that it falls into:



With this information, I decided to determine how age affects charges separately, for each of these groups:

Linear regression for Nonsmokers

```
##
## Call:
## lm(formula = charges ~ age, data = Nonsmoker)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3184.1 -1951.2 -1365.8  -664.5 24466.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2085.01     425.86  -4.896 1.13e-06 ***
## age           267.12      10.18  26.244  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4669 on 1061 degrees of freedom
## Multiple R-squared:  0.3936, Adjusted R-squared:  0.3931
## F-statistic: 688.8 on 1 and 1061 DF,  p-value: < 2.2e-16
```

- Model Summary: Charges = -2085.01 + 267.12 * Age
- Intercept (-2085.01): This indicates the estimated insurance charges when age is zero, which isn't practical but serves as a baseline.
- Slope (267.12): For every additional year of age, the insurance charges for nonsmokers increase by approximately $267.
- R^2 (0.3936): About 39.36% of the variability in charges among nonsmokers is explained by age. This is a moderate value, indicating that while age does impact charges, other factors are also significant.
- p-value (< 2.2e-16): The relationship between age and charges is statistically significant.

Linear regression for Smokers with High BMI

```
##
## Call:
## lm(formula = charges ~ age, data = Smoker_High_bmi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5587.0 -1960.8  -445.8   782.3 17388.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11503.36     962.92   11.95   <2e-16 ***
## age           260.64      23.99   10.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3662 on 127 degrees of freedom
## Multiple R-squared:  0.4818, Adjusted R-squared:  0.4777
## F-statistic: 118.1 on 1 and 127 DF,  p-value: < 2.2e-16
```

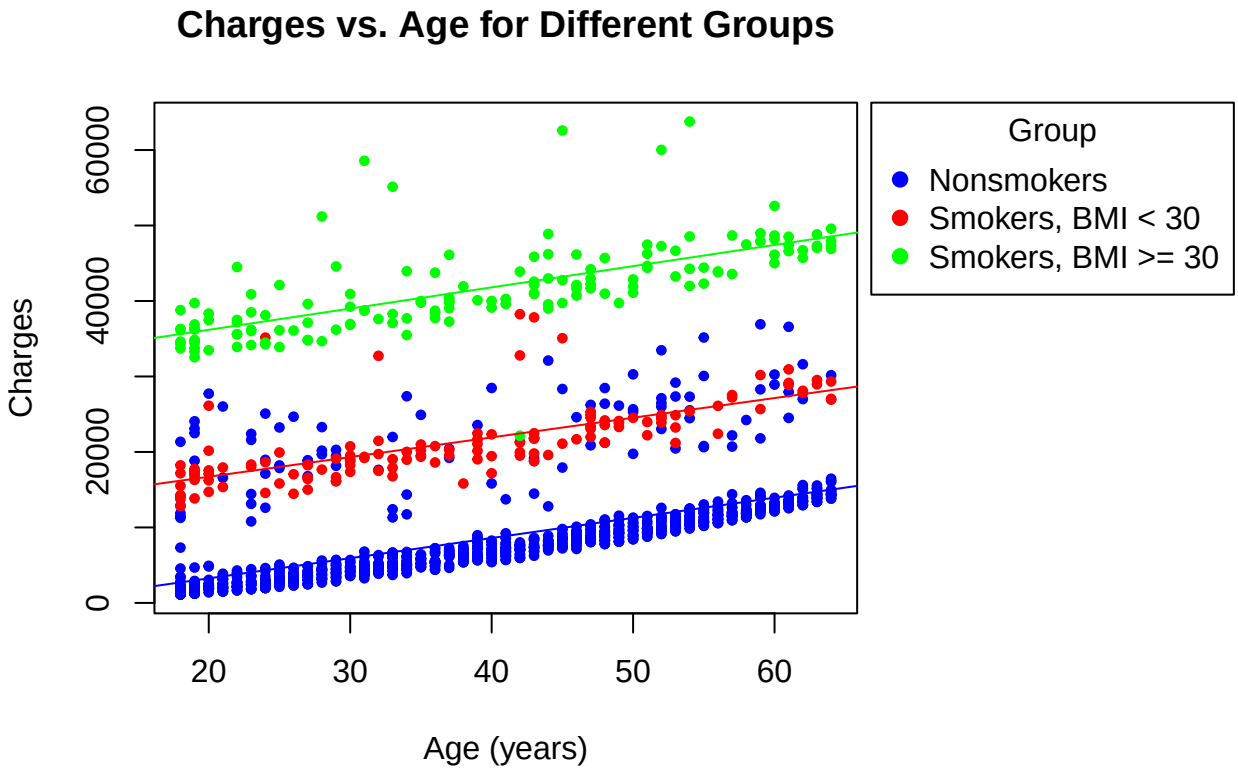
- **Model Summary**: `Charges = 11503.36 + 260.64 * Age`
- **Intercept (11503.36)**: This indicates the estimated insurance charges when age is zero for smokers with high BMI, reflecting significantly higher baseline charges possibly due to the combined risk factors of smoking and high BMI.
- **Slope (260.64)**: For every additional year of age, the insurance charges for smokers with high BMI increase by approximately $260.
- R^2 **(0.4818)**: About 48.18% of the variability in charges among smokers with high BMI is explained by age. This is a relatively higher value compared to nonsmokers, suggesting age is a more significant predictor in this group.
- **p-value (< 2.2e-16)**: The relationship between age and charges is statistically significant, confirming the importance of age in determining insurance costs for smokers with high BMI.

Linear regression for Smokers with Low BMI

```
##
## Call:
## lm(formula = charges ~ age, data = Smoker_Low_bmi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20222.5  -2204.8  -1071.4    998.7  19382.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30558.13    1093.04   27.96   <2e-16 ***
## age           281.15      26.25   10.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4508 on 143 degrees of freedom
## Multiple R-squared:  0.4452, Adjusted R-squared:  0.4413
## F-statistic: 114.7 on 1 and 143 DF,  p-value: < 2.2e-16
```

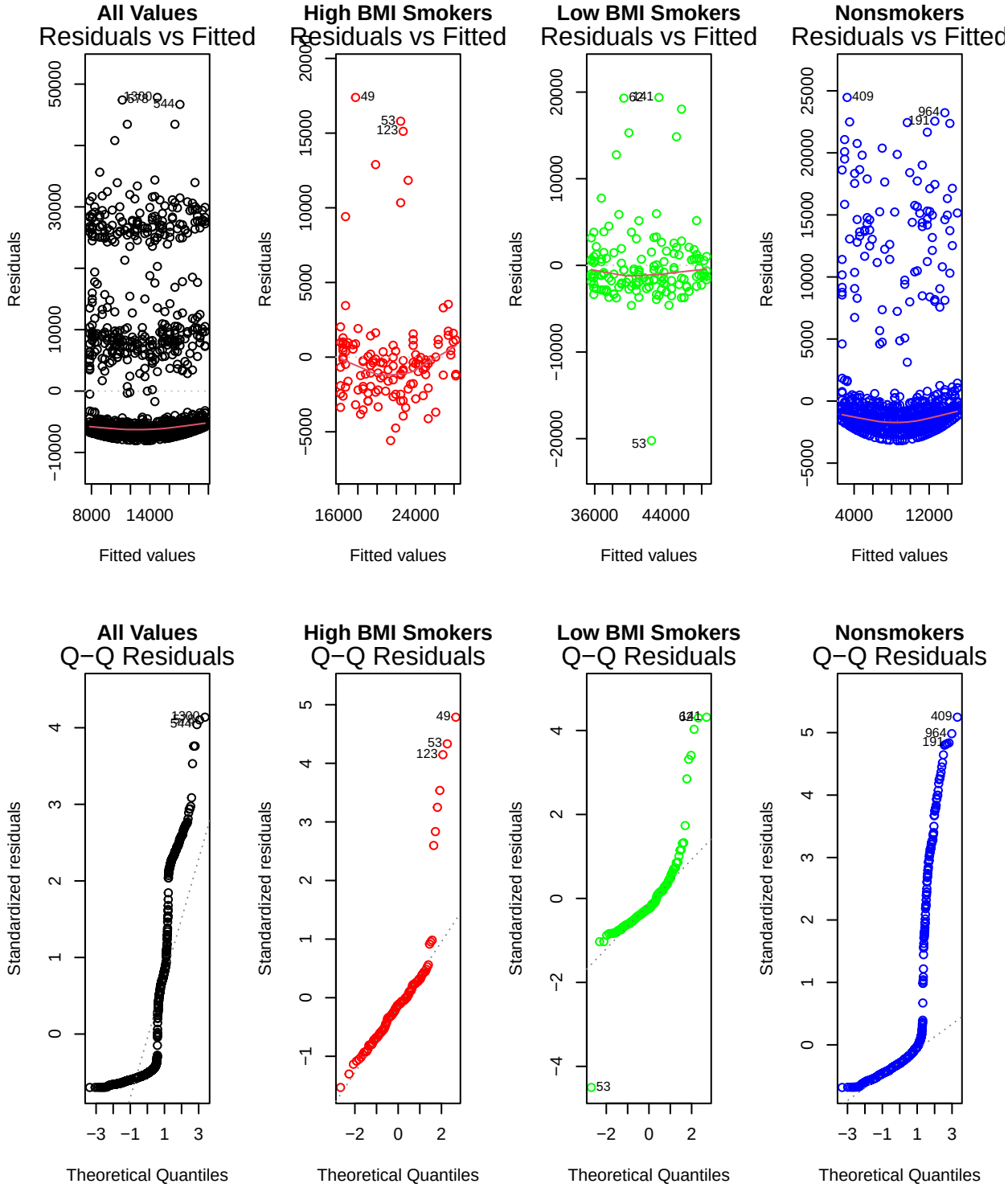
- **Model Summary**: `Charges = 30558.13 + 281.15 * Age`
- **Intercept (30558.13)**: The baseline charges for smokers with low BMI are significantly high at $30,558, indicating that smoking is a substantial risk factor for insurance costs, even when BMI is lower.
- **Slope (281.15)**: Each additional year of age increases charges by approximately $281, the highest among the three groups.
- R^2 **(0.4452)**: About 44.52% of the variability in charges can be explained by age for smokers with low BMI. This indicates a substantial but not exclusive influence of age on charges.
- **p-value (< 2.2e-16)**: The impact of age on insurance charges in this group is statistically significant.

- **We can plot these models on the same graph as before to show the three trends within the points:**



We can now check our model assumptions for each of these groups:

Assumptions of Linearity/Normality



- Based on the residuals and Q-Q plots, linearity and normality cannot be assumed. Along with attempts of log and square root transformations, the models show that the variability of the data is inconsistent.

After initially examining the impact of age on health insurance charges within three distinct groups: non-smokers, smokers with low BMI, and smokers with high BMI, it became evident that each group exhibited unique trends and influences on charges.

Model Selection Approach: Forward Selection

To determine the most appropriate predictors for our regression model, I employed **stepwise selection techniques**. These methods help streamline the model by including only variables that significantly contribute to explaining variations in `charges`, improving both **interpretability and predictive accuracy**.

Stepwise Selection Methods Considered

1. **Forward Selection**: Starts with an **empty model** (only an intercept) and **adds predictors one by one** based on their statistical significance.
2. **Backward Elimination**: Begins with the **full model** (all predictors included) and **removes insignificant predictors** sequentially.
3. **Both Directions**: Combines forward and backward approaches, **adding or removing** variables based on model fit criteria.

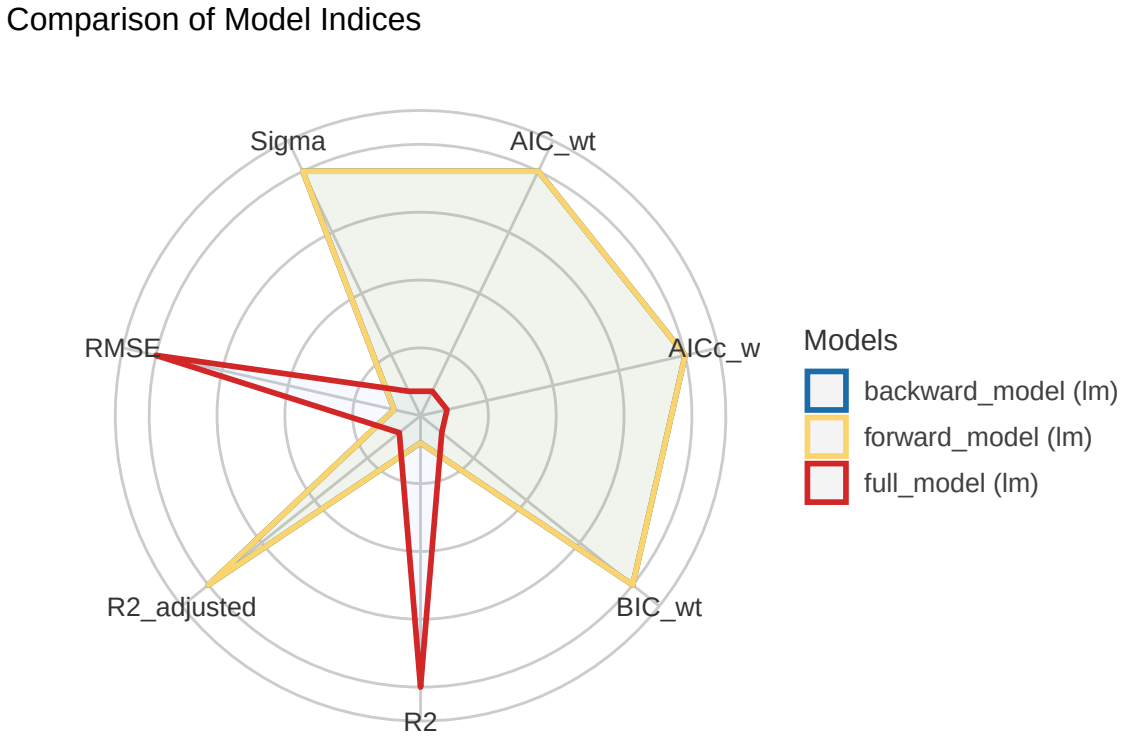
Why I Chose Forward Selection

After comparing different selection methods using the `step()` function, we opted for **forward selection** for the following reasons:

- **Avoids Overfitting**: Since forward selection starts with **no predictors**, it only adds variables that improve model performance, making it **less prone to overfitting** than backward selection.

- **Better Interpretability**: By **sequentially adding** significant variables, forward selection ensures that only the **most relevant predictors** are included, resulting in a **simpler and more interpretable model**.

- **Better Performance Compared to Other Models**: I evaluated the **performance of forward, backward, and full models** using the `compare_performance()` function. The **forward selection model outperformed the others**, as demonstrated in the performance plot:

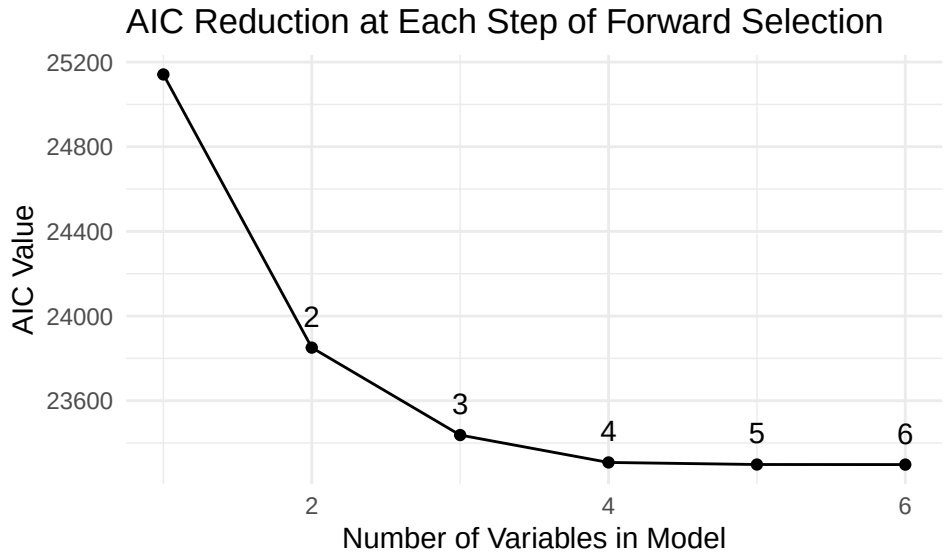


We can then look at each step of the forward selection process to see how AIC is affected:

```
## Start:  AIC=25141.46
## charges ~ 1
##
##             Df  Sum of Sq        RSS   AIC
## + smoker     1 1.2143e+11 7.4508e+10 23851
## + age        1 1.7436e+10 1.7850e+11 25019
## + bmi        1 7.7127e+09 1.8823e+11 25090
## + children   1 8.8982e+08 1.9505e+11 25137
## + region     3 1.2819e+09 1.9466e+11 25139
## + sex        1 6.6015e+08 1.9528e+11 25139
## <none>                    1.9594e+11 25142
##
## Step:  AIC=23850.72
## charges ~ smoker
##
##             Df  Sum of Sq        RSS   AIC
## + age        1 1.9883e+10 5.4625e+10 23438
## + bmi        1 7.4852e+09 6.7023e+10 23711
## + children   1 7.4398e+08 7.3764e+10 23839
## <none>                    7.4508e+10 23851
## + sex        1 1.0017e+06 7.4507e+10 23853
## + region     3 1.0465e+08 7.4403e+10 23855
##
## Step:  AIC=23437.68
```

```
## charges ~ smoker + age
##
##            Df  Sum of Sq        RSS   AIC
## + bmi       1 5113680989 4.9511e+10 23308
## + children  1  458348288 5.4166e+10 23428
## <none>                   5.4625e+10 23438
## + sex       1    2326134 5.4622e+10 23440
## + region    3  137958782 5.4487e+10 23440
##
## Step:  AIC=23308.27
## charges ~ smoker + age + bmi
##
##            Df Sum of Sq        RSS   AIC
## + children  1 433569708 4.9077e+10 23298
## + region    3 232954323 4.9278e+10 23308
## <none>                  4.9511e+10 23308
## + sex       1   3786242 4.9507e+10 23310
##
## Step:  AIC=23298.51
## charges ~ smoker + age + bmi + children
##
##          Df Sum of Sq        RSS   AIC
## + region  3 233790258 4.8844e+10 23298
## <none>                4.9077e+10 23298
## + sex     1   5360623 4.9072e+10 23300
##
## Step:  AIC=23298.13
## charges ~ smoker + age + bmi + children + region
##
##        Df Sum of Sq        RSS   AIC
## <none>              4.8844e+10 23298
## + sex   1   5553651 4.8838e+10 23300
```

Continuing on with the forward selection past the addition of smoking status, age, and BMI has virtually no benefit to the model, and only adds risk of overfitting, as shown below:



To explore the combined effects of age, smoking status, and BMI more holistically, a comprehensive model was developed. This model incorporated all groups into a single analysis framework, allowing for the examination of both the individual and interactive effects of these factors on insurance charges. By transitioning to this model, the analysis could leverage interaction terms (age:group) to precisely capture how each group's age-related increase in charges differs, providing a deeper understanding of the underlying patterns observed in the preliminary group-specific analyses.

Creation and Usage of the `group` Variable for Interaction Terms

In the comprehensive regression model, the `group` variable was created to classify individuals into distinct categories based on their smoking status and body mass index (BMI).

Steps to Create the `group` Variable:

1. **BMI Classification**: Individuals were first categorized as `obese` or `not obese`, with obesity defined as having a BMI of 30 or higher.
2. **Smoking Status**: Each individual's smoking status was already recorded as 'yes' or 'no', and was re-coded into a numeric format where `1` represents smokers and `0` represents non-smokers.
3. **Group Definition**: Using the `case_when` function from the `dplyr` package, individuals were segmented into three groups:
   - `not smoker`: Individuals who do not smoke.
   - `not obese smoker`: Smokers who are not obese.
   - `obese smoker`: Smokers who are obese.

These categories were then converted into a factor variable with levels explicitly ordered to ensure that the model's intercept corresponds to the `not smoker` group, serving as the baseline category against which the other groups are compared.

```
##
## Call:
## lm(formula = charges ~ group + age:group, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20222.5  -1960.8  -1297.5   -393.5  24466.9
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2085.008    416.392  -5.007 6.26e-07 ***
## groupnot obese smoker      13588.364   1270.649  10.694  < 2e-16 ***
## groupobese smoker          32643.135   1182.529  27.605  < 2e-16 ***
## groupnot smoker:age          267.118      9.952  26.841  < 2e-16 ***
## groupnot obese smoker:age    260.640     29.903   8.716  < 2e-16 ***
## groupobese smoker:age        281.153     26.578  10.578  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4565 on 1331 degrees of freedom
## Multiple R-squared:  0.8584, Adjusted R-squared:  0.8579
## F-statistic:  1614 on 5 and 1331 DF,  p-value: < 2.2e-16
```

Interpret Results

Model Overview

- **Formula**: `charges ~ group + age:group`
- The model addresses both the absolute differences in charges among the groups and the interaction effects between age and group categories, to discern how age-related changes in charges differ across these groups.

Results

Coefficients

- **Intercept**: The baseline charge for the reference group (non-smokers) when age is zero is estimated at -2085.008. This negative value, while not practical, sets a baseline for the model.
- **Non-obese smoker**: This group is associated with an increase in charges of approximately $13,588 over non-smokers, adjusting for age.
- **Obese smoker**: Smokers who are obese incur about $32,643 more in charges than non-smokers, indicating a significant impact of combined smoking and obesity on health costs.
- **Age effects**:
  - **Non-smokers**: Each additional year of age increases charges by approximately $267.
  - **Non-obese smokers**: Each year of age adds about $261 to charges, slightly less than non-smokers.
  - **Obese smokers**: Each year of age results in an approximate $281 increase in charges, the highest among the groups.

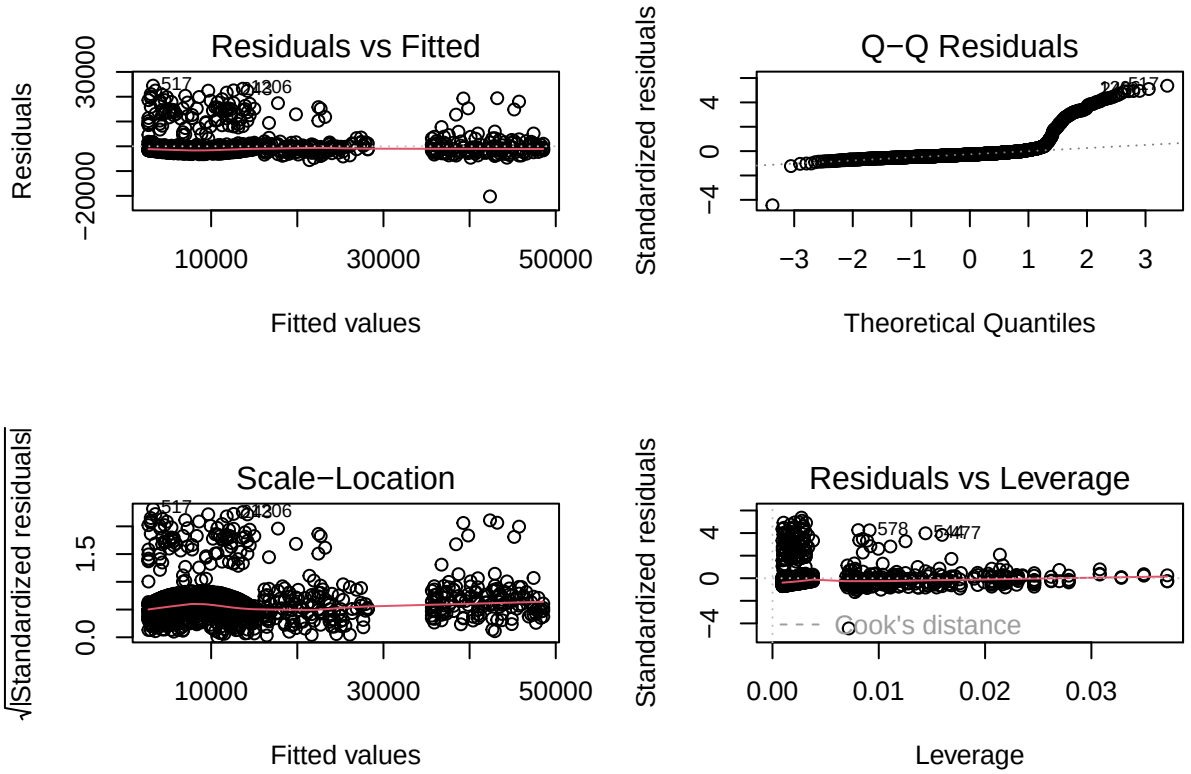
Statistical Significance

- All predictors have p-values $< 2.2e\text{-}16$, affirming the robustness of the findings.

Model Fit

- **Residual standard error**: 4565 on 1331 degrees of freedom
- **Multiple R-squared**: 0.8584; **Adjusted R-squared**: 0.8579
  - The model explains approximately 85.84% of the variance in insurance charges, indicating excellent model fit.
- **F-statistic**: 1614 on 5 and 1331 DF, p-value $< 2.2e\text{-}16$
  - The overall model is statistically significant, demonstrating strong explanatory power.

Assumptions Checking



1. Residuals vs Fitted: This plot checks for non-linearity, homoscedasticity (equal variance), and the presence of outliers.

- The residuals do not display any clear patterns, indicating no obvious non-linearity.
- The dispersion of residuals around the horizontal zero line suggests that the variance of residuals is relatively consistent across the range of fitted values (homoscedasticity).
- Several outliers are evident, as some residuals are significantly far from the zero line.

2. Q-Q Plot: Assesses whether the residuals are normally distributed.

- The points mostly follow the theoretical line, indicating that the residuals are decently normal.
- Deviations from the line at both tails suggest the presence of outliers with potentially heavy tails.

3. Scale-Location Plot: Checks if residuals are spread equally along the ranges of predictors (homoscedasticity).

- The plot shows a relatively constant spread across the range of fitted values, supporting the assumption of equal variance.
- Like the Residuals vs Fitted plot, the presence of outliers is noticeable.

4. Residuals vs Leverage: Identifies influential observations that might disproportionately influence the model's estimates.

- Most data points exhibit low leverage, indicating they do not unduly influence the model.
- A few points have high leverage and exceed the Cook's distance threshold, suggesting they are influential and could be distorting the regression results.
- Specific points labeled (e.g., 0578, 05447) appear particularly influential, warranting further investigation.

Evaluating Transformations for Model Improvement

To assess whether transformations could improve model fit and better meet linear regression assumptions, I applied several transformations to the dependent variable (`charges`), including:

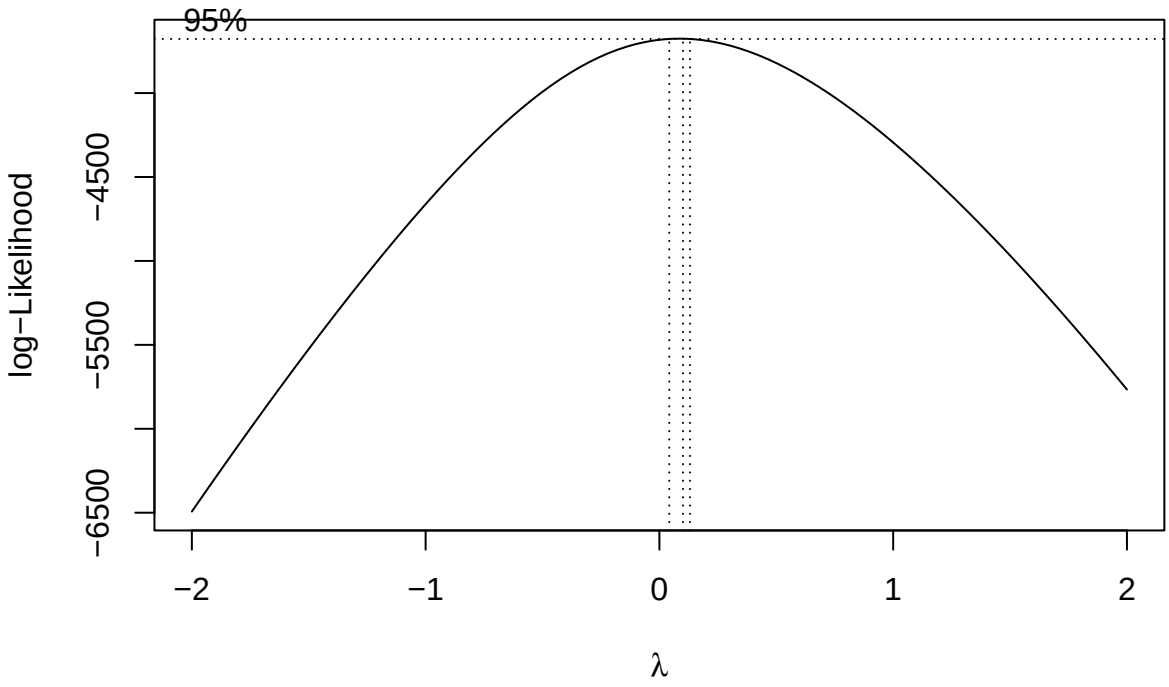
1. **Inverse Transformation (`1/charges`)**
2. **Square Root Transformation (`sqrt(charges)`)**
3. **Box-Cox Transformation (to identify an optimal power transformation)**

Applying Transformations

The following transformations were applied, and models were refitted to evaluate changes in performance:

```
## [1] "Inverse R-squared: 0.641225700783616"
```

```
## [1] "Square Root R-squared: 0.776877511927726"
```




```
## [1] "BoxCox R-squared: 0.78360438970551"
```

Thus, the attempted transformations do not have beneficial impacts on the model

Conclusion

Key Findings

This analysis has demonstrated that both smoking and obesity are associated with significant increases in health insurance charges, and these effects are further influenced by the age of the insured individuals. Key findings from the analysis include:

- **Smokers with obesity face the highest charges**, with costs significantly higher than those for non-smokers and smokers without obesity.
- **Age-related increases in charges** are slightly more pronounced in smokers with high BMI, indicating that aging in these risk groups is associated with higher health costs.

Implications

The findings underscore the need for health insurance policies and health interventions that are tailored to address the specific risk profiles of individuals:

- **Policy Adjustments**: Insurance companies might consider adjusting premiums and coverage terms to more accurately reflect the increased risks associated with smoking and obesity. This could include developing tiered premium systems or offering incentives for lifestyle changes.
- **Targeted Interventions**: Healthcare providers and policymakers could develop targeted health interventions aimed at smoking cessation and weight management, especially as these factors together significantly drive up healthcare costs.
- **Preventive Measures**: Early intervention in younger populations who are at risk of becoming smokers or developing obesity could reduce long-term costs and improve health outcomes.

Limitations

While the analysis provides compelling insights, there are several limitations:

- **Causal Inference**: The study design is observational, which limits the ability to draw causal conclusions. The associations observed may be influenced by unmeasured confounding factors such as genetics, socio-economic status, or other lifestyle habits.
- **Data Scope**: The analysis is based on existing datasets that may not capture all relevant variables, such as diet, physical activity, or detailed medical history, which could influence insurance charges.
- **Generalizability**: The results are dependent on the specific demographic and geographic characteristics of the dataset. Findings may differ in other populations with different health, economic, or cultural backgrounds.