# STAT 6021: Project Two

Medical Insurnace Costs

Niraja Bhidar(nd4dg), Derek Banks(dmb3ey), Jay Hombal (mh4ey), Ronak Rijhwani (rr7wq)

## 1 Executive Summary :

Insurance companies calculate premiums by using models based on certain demographics. We wanted to ascertain what variables influenced claims and how influential these predictors were. Through Kaggle, we were able to obtain a data set that contained the claims of people based on age, body mass index (BMI), sex, region, smoking, and children.

At first, we individually compared each variable against charges and saw that our data appeared to be right skewed. This implied that a transformation of the data would be necessary in order to normalize our data. Before we did that, we wanted to discern which variables were significant to keep in our model.

The linear regression output for our initial model, which contained all the variables, showed the northwest region was linearly related some other variable, therefore suggesting it should be dropped from the model. However, no matter how we manipulated the predictors, the data seemed to have non-constant variance which would give no value to any statistical analysis we did on these models. Due to the complexity of data manipulation, we used a logarithmic binomial model.

In this approach, we split the groups into charges that were above 20,000 USD dollars and charges that were below 20,000. The equation we found to be the best fit for our data was:

$$\log \frac{\pi_i}{1-\pi_i} = 0.04087 age + 0.10134 bmi + 4.75564 smoker$$

This shows that the odds of having charges over $20,000 gets multiplied by a factor of 116 if one is a smoker (while holding age and BMI constant), gets multiplied by a factor 1.13 for each BMI increase (while holding age and smoking constant), and gets multiplied by a factor of 1.04 for every addition year (while holding smoking and BMI constant).

## 2 Exploratory Data Analysis :

The data was contained in the file datasets_13720_18513_insurance.csv included with this project.

- The variables were as follows:
  - **Predictors**
    - **x1**: **age**: age of primary beneficiary.
    - **x2**: **sex**: insurance contractor gender, female, male.
    - **x3**: **bmi**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ˆ 2) using the ratio of

    height to weight, ideally 18.5 to 24.9.
    - **x4**: **children**: Number of children covered by health insurance / Number of dependents.
    - **x5**: **smoker**: Smoking
    - **x6**: **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
  - **Response Variable**
    - **Y**: **charges**: Individual medical costs billed by health insurance.
- **The main objectives for this project were:**
  1. Explore relationship between response variable **charges** & the six other predictor variables (x1–x6).
  2. Analyze the correlation and directionality of the dataset.
  3. Create a model that is the best fit model to predict the insurance **charges** based the demographic predictor variables and evaluate the validity and usefulness of this model.

Additionally, we planned to utilize model selection tools to give us a deeper understanding of how different potential models compare. We want to recommend a best fit model and end our section by exploring the pros and cons of our models under consideration.

Exploratory data analysis started with investigating the dataset.

```r
data <- read.csv("datasets_13720_18513_insurance.csv", header = TRUE, sep =",",
                 stringsAsFactors = TRUE)
head(data)
```

```
  age    sex    bmi children smoker    region   charges
1  19 female 27.900        0    yes southwest 16884.924
2  18   male 33.770        1     no southeast  1725.552
3  28   male 33.000        3     no southeast  4449.462
4  33   male 22.705        0     no northwest 21984.471
5  32   male 28.880        0     no northwest  3866.855
6  31 female 25.740        0     no southeast  3756.622
```

There were six predictors and a response variable **charges**. The dataset had 1338 rows, and the data appeared to need little cleaning and did not contain missing values.

```
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

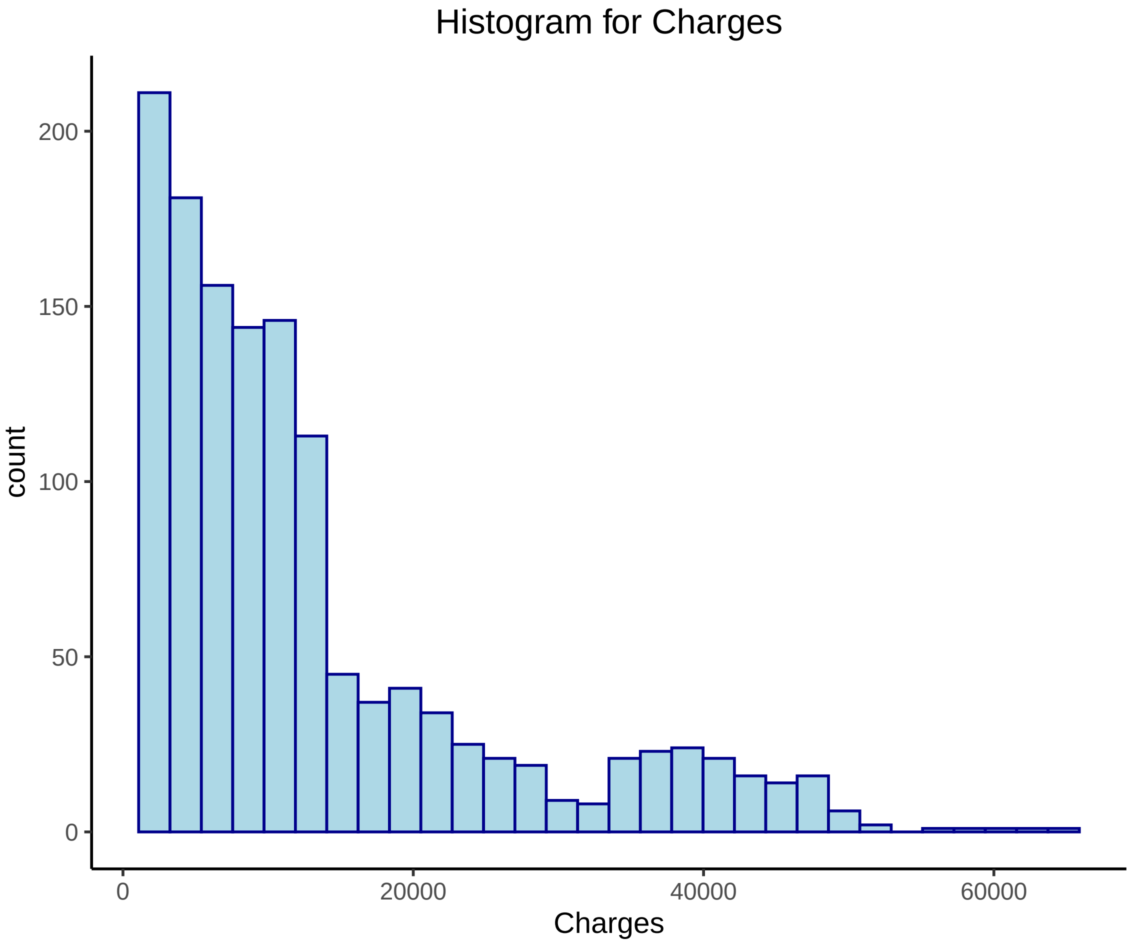
Inspecting the data types of variables, we see that the predictor variables sex, smoker, and region were categorical variables and automatically converted as factor by R when loading the dataset because because of the option **stringsAsFactors = TRUE** used while reading the csv file.

```
      age            sex           bmi           children      smoker
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
 Median :39.00                Median :30.40   Median :1.000
 Mean   :39.21                Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                Max.   :53.13   Max.   :5.000
       region       charges
 northeast:324   Min.   : 1122
 northwest:325   1st Qu.: 4740
 southeast:364   Median : 9382
 southwest:325   Mean   :13270
                 3rd Qu.:16640
                 Max.   :63770
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1122    4740    9382   13270   16640   63770
```

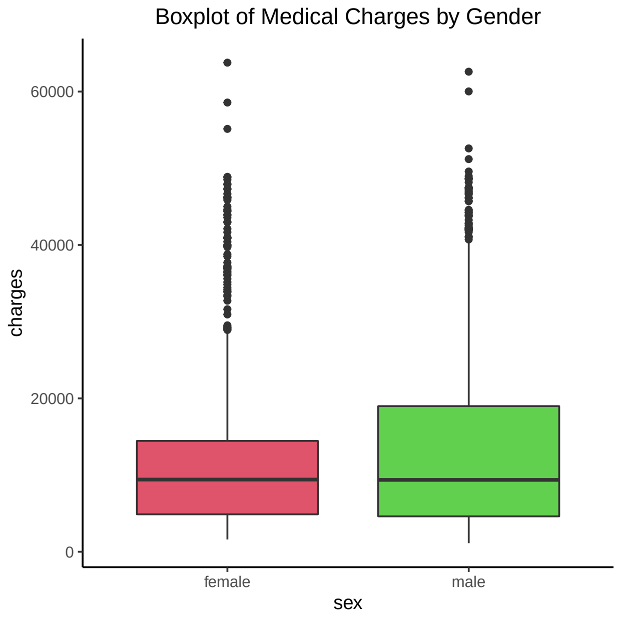
- From the summary we made the following observations :
  - The observations seemed to be evenly distributed across region.
  - The age varied between low of 18 and a max of 64.
  - The observations were almost evenly distributed by sex.
  - The dataset had almost 4:1 non-smoker to smoker ratio or only 20.5% people smoke.
  - The bmi varied between a min of 15.96 and max of 53.13.

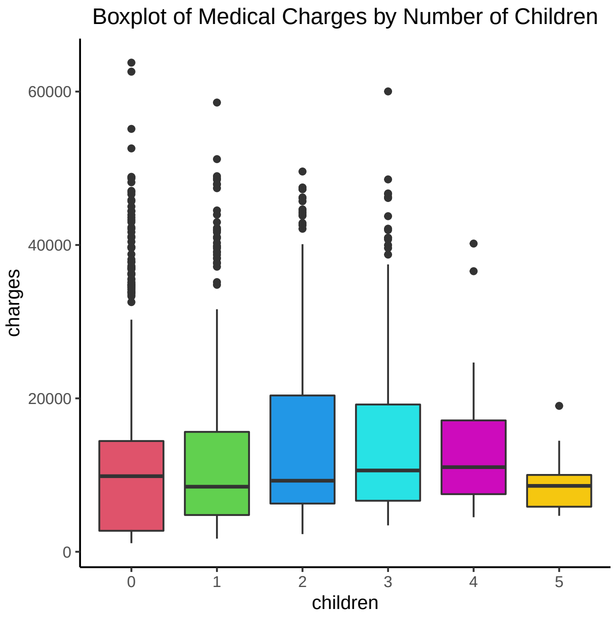
- The response variable mean was greater than median, this was an indication that data is right-skewed. This could be confirmed by the histogram of **charges** shown below.



In the box plot shown below for medical **charges** by **sex** the median value of the medical **charges** for both male and female appeared to be almost the same. The third quartile for male seemed to be greater than female, so the data may be skewed towards the men.

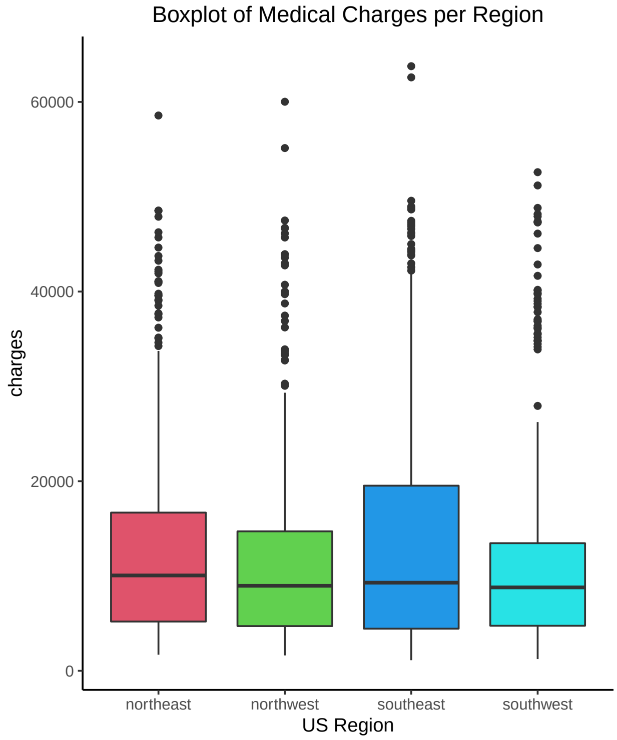
The box plot of medical **charges** by number of **children**, we could make an interesting observation that the medical **charges** for people with 5 children were lower than people with one to four children and people with no children had the lowest medical charges.

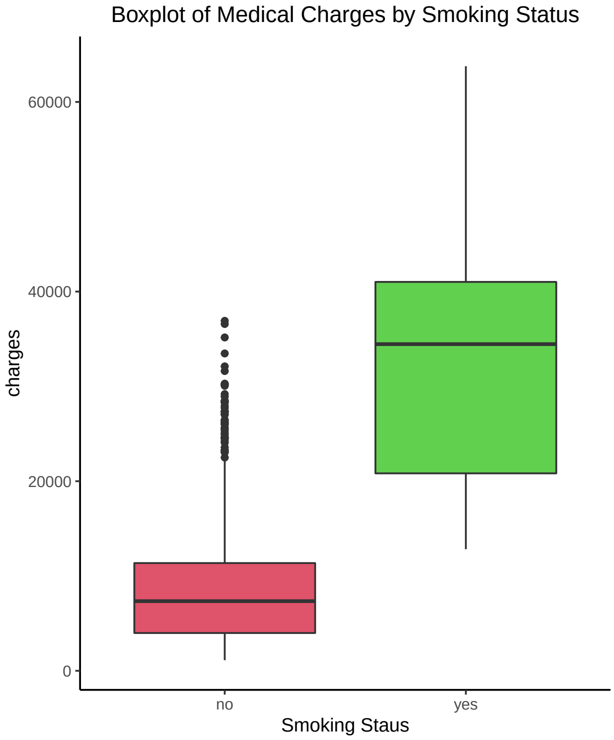

Boxplot of Medical Charges by Gender

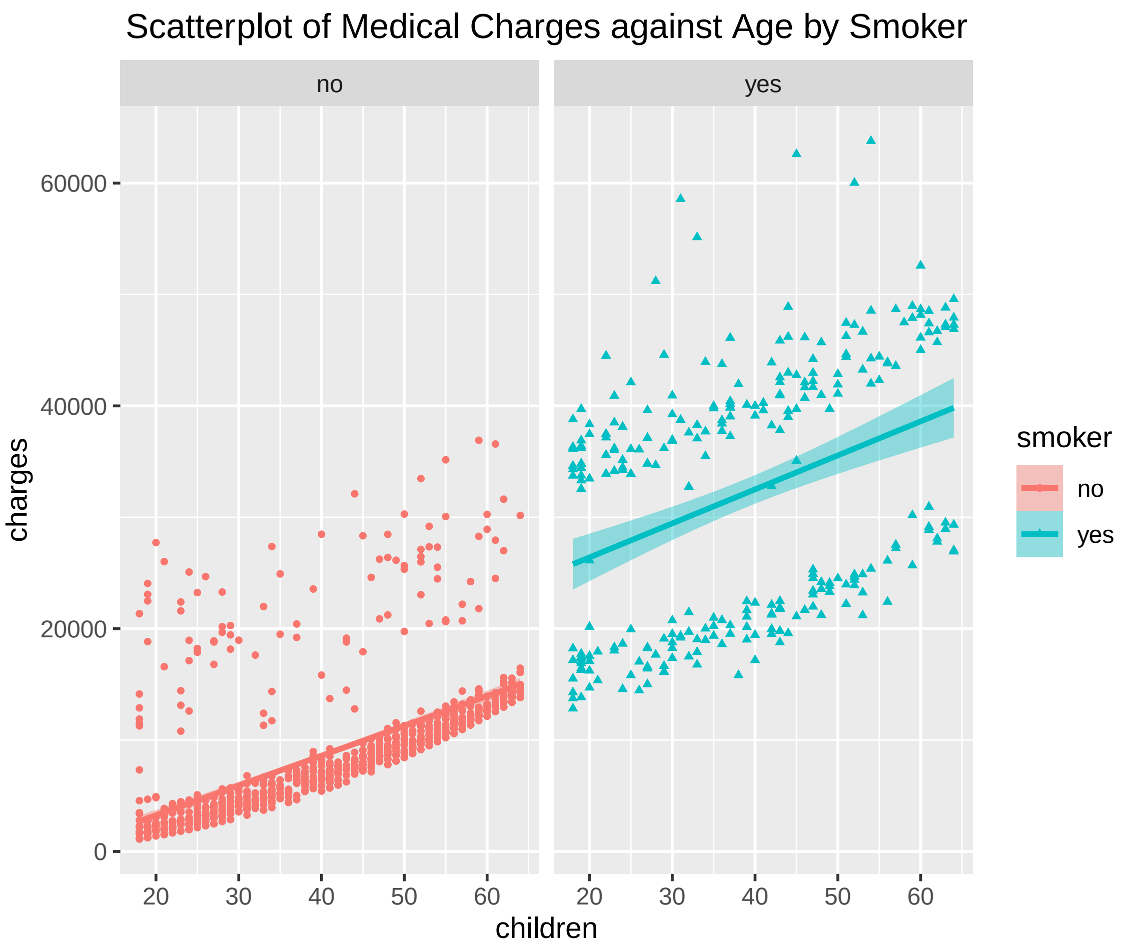

Boxplot of Medical Charges by Number of Children

In the box plot of medical **charges** per **region** the median value of the medical **charges** across all four US regions was almost the same. The people in the southeast seemed to have higher medical expenses then the people in the other areas.

However, exploring the box plot of medical **charges** by **smoking** status, we could see that the medical **charges** for those who smoke were much higher than those who do not smoke.

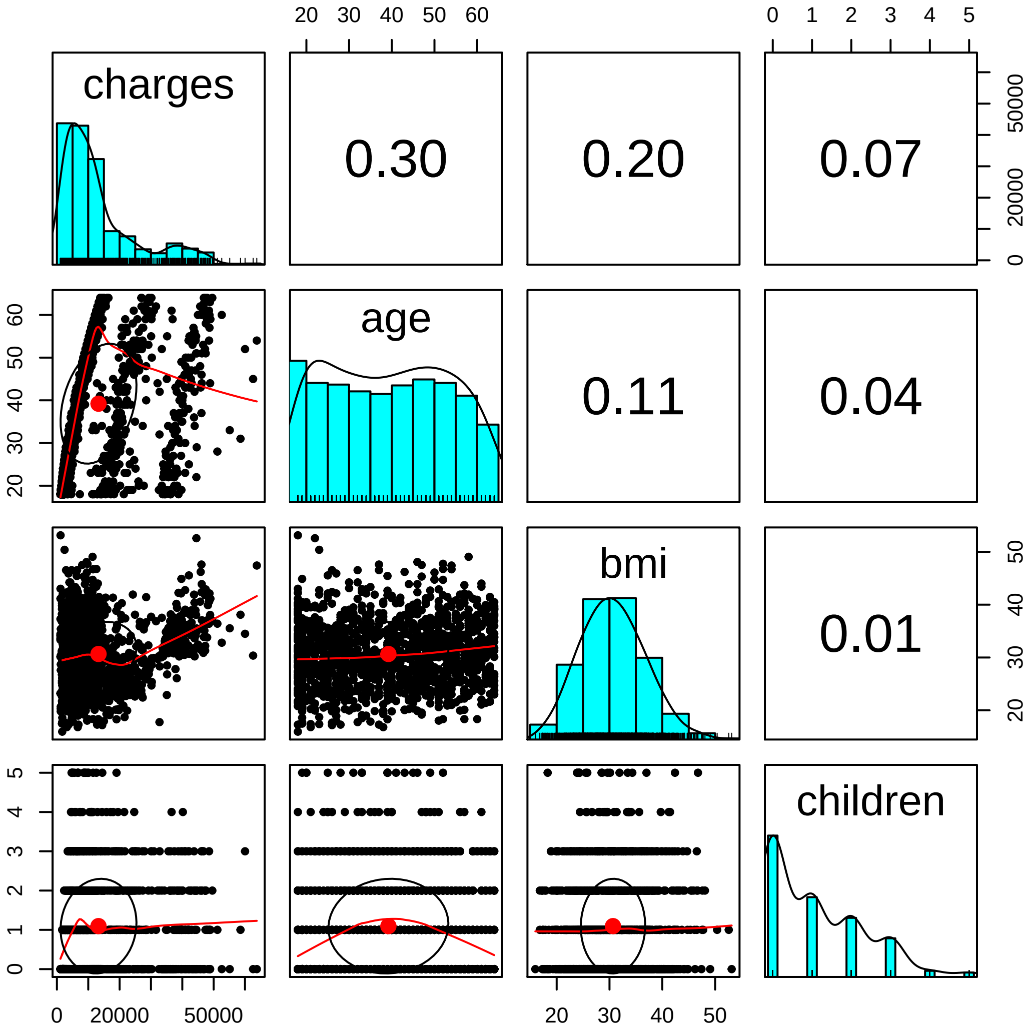

Boxplot of Medical Charges per Region


Boxplot of Medical Charges by Smoking Status



From the above scatter plot, we observed that medical charges increased with age for both smokers and non-smokers. However, those who smoked tended to have higher medical expenses than those who did not.

```
            charges       age       bmi   children
charges  1.00000000 0.2990082 0.1983410 0.06799823
age      0.29900819 1.0000000 0.1092719 0.04246900
bmi      0.19834097 0.1092719 1.0000000 0.01275890
children 0.06799823 0.0424690 0.0127589 1.00000000
```



We observed **somewhat(moderate)** correlated between **age** and **charges**, and **bmi** and **charges**. However, **bmi** and **age** and **children** and **charges** appeared to have a weak correlation. These relationships were futher explored in the computational exploration phase of the project while building to a final model.

We observed from the summary of the dataset that the **smoker** status (class) had a better correlation to **charges** compared with **non-smoker** status (class). This implied that smokers had more medical expenses than non-smokers.

From exploratory analysis we determined that more than one predictor should be considered for the initial model. We observed that **age** had somewhat better correlation with **charges** and **smokers** tended to have more medical expenses than **non-smokers**.

And in our computation analysis we saw that different classes sex and region categorical variables also indicated to the predictors that we could consider. So to start of we will consider all predictors.

**Computational Exploration** A model with all predictors was considered as an initial starting point. Additional candidate models were calculated by applying model automatic predictor search procedures.

The $\text{R}^2{}_{\text{adj}}$ value and the BIC metrics were used to identify likely models since these both approaches penalized for adding more terms.

- Based on the analysis -
  - The model with lowest BIC was:

$$charges = \beta_0 + \beta_1 age + \beta_2 bmi + \beta_3 children + \beta_4 smokeyes$$

  - The model with highest adjusted $\text{R}^2$ was:

$$charges = \beta_0 + \beta_1 age + \beta_2 bmi + \beta_3 children + \beta_4 smokeyes + \beta_5 regionsoutheast + \beta_6 region5southwest$$

We also considered the models with the highest $\text{R}^2$, lowest Cp, and lowest MSE values. The best Cp and best MSE were both on the the same model as the best adjusted $\text{R}^2$.

The model with the best $\text{R}^2$ and adjusted $\text{R}^2$ values and had all predictors in addition to regionnorthwest.

The model with the lowest BIC (-1817.233) was:

$$charges = \beta_0 + \beta_1 age + \beta_2 bmi30 + \beta_3 children + \beta_4 smokeyes$$

The model with the highest adjusted $\text{R}^2$ was:

$$charges = \beta_0 + \beta_1 age + \beta_2 bmi + \beta_3 children + \beta_4 smokeyes + \beta_5 regionsoutheast + \beta_6 region5southwest$$

## 3. Initial Model Considered:

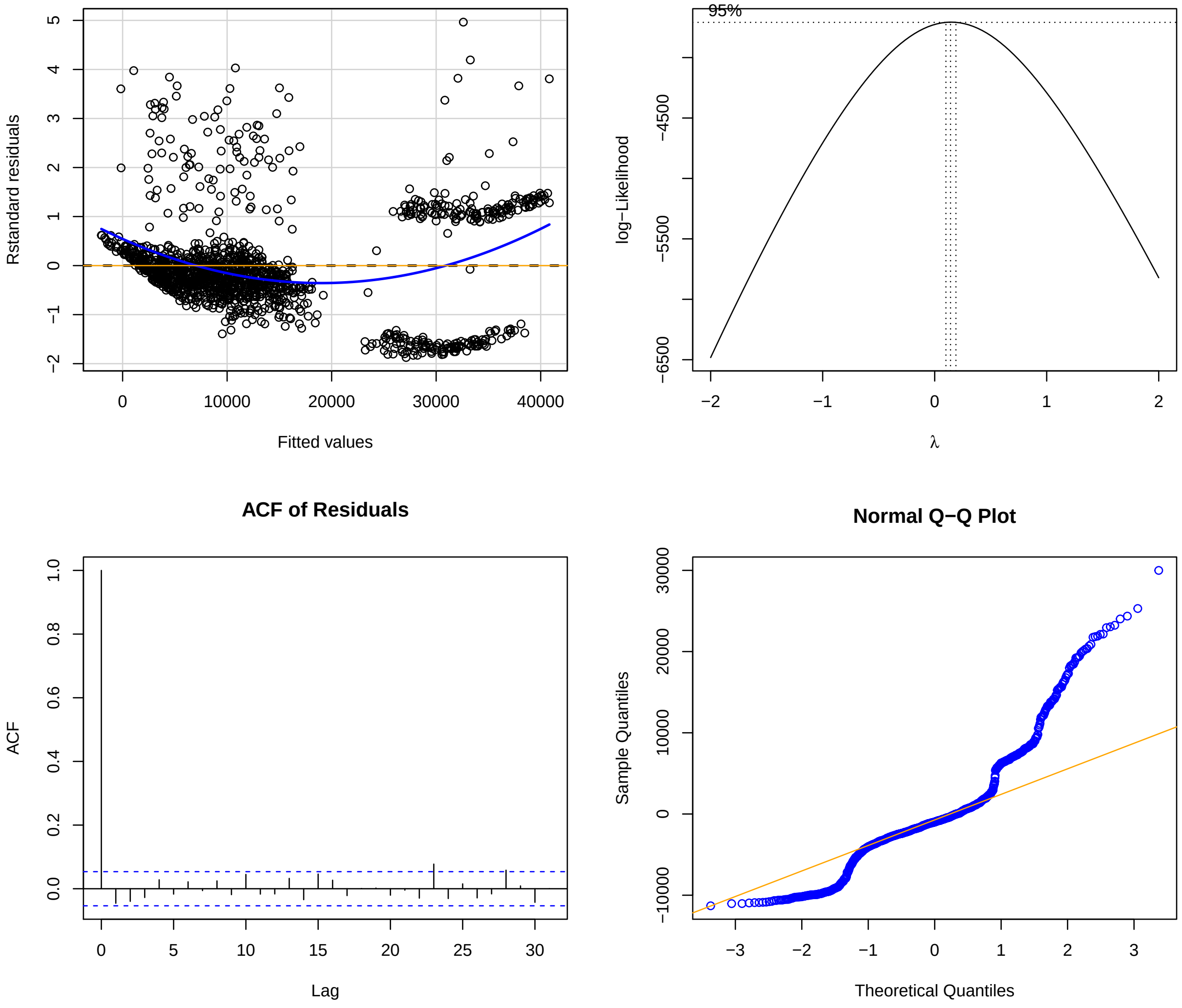
Based on the results from the model search procedures, the intial model considered was:

$$charges = \beta_0 + \beta_1 age + \beta_2 bmi + \beta_3 children + \beta_4 smokeyes + \beta_5 regionnorthwest + \beta_6 region5southeast + \beta_7 regionsouthwest + \beta_8 s$$

```
initalmodel <- lm(charges ~ age + bmi + children + smoker + region +sex, data=data)
summary(initalmodel)
```

```
Call:
lm(formula = charges ~ age + bmi + children + smoker + region +
    sex, data = data)

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 ***
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 *
sexmale           -131.3      332.9  -0.394 0.693348
---
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
```

Next the linear regression assumptions were validated:



In the plots displayed above, we observed that the variance was not constant as seen in the box-cox plot and that the residual plot did not have constant variance.

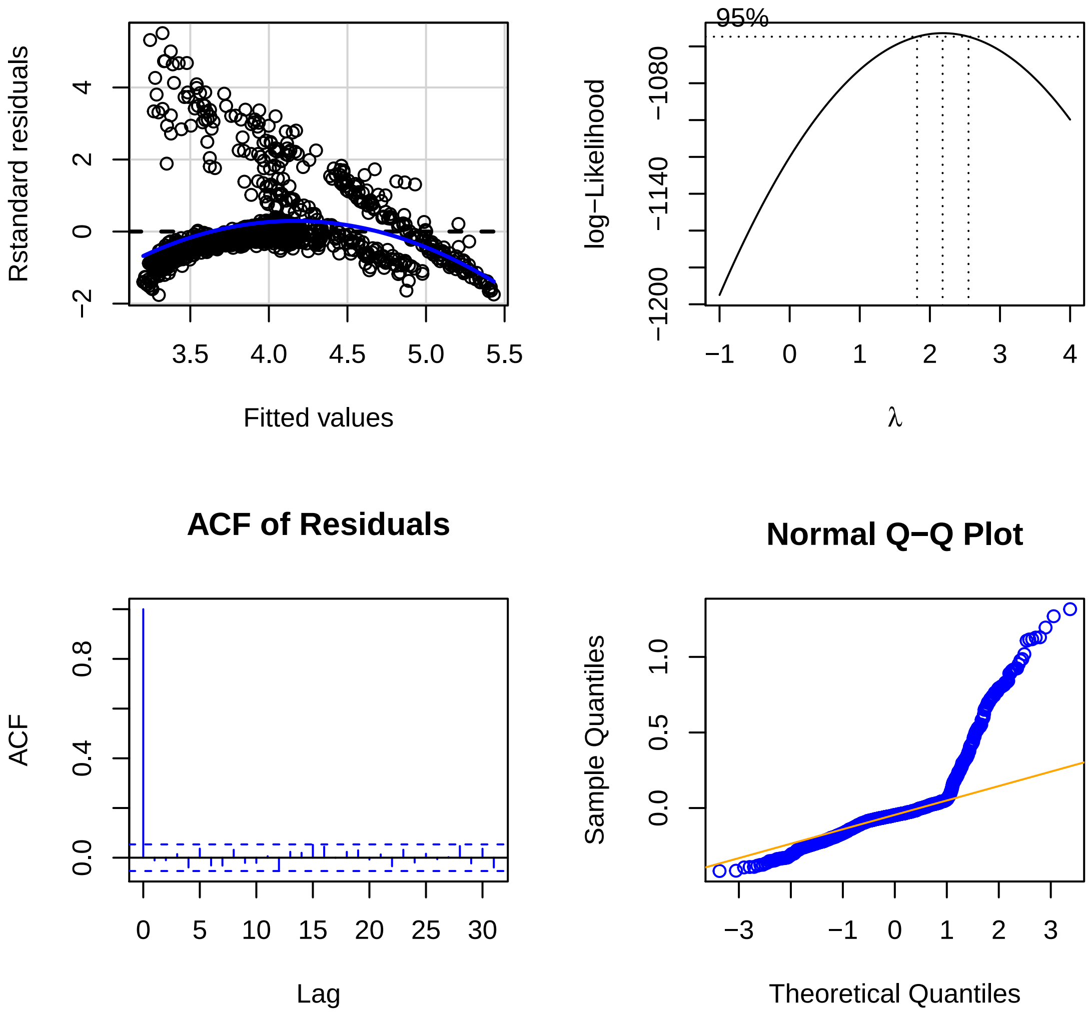
In our hypothesis, we said that, age, people who smoke and people with high bmi (bmi>30) may be at high risk and so their medical costs may be higher. Based on that hypothesis and considering that our initial model suffered from non-linearity and non-constant variance issues. We will transformed both response variable and predictors.

The following transformations were applied: 1. Transformed **charges** (y) to fix non-constant variance 2. Transformed age by adding a non-linear term 3. Created a indicator variable for bmi (obesity indicator) 4. Specified an interaction between smokers and bmi indicator predictor

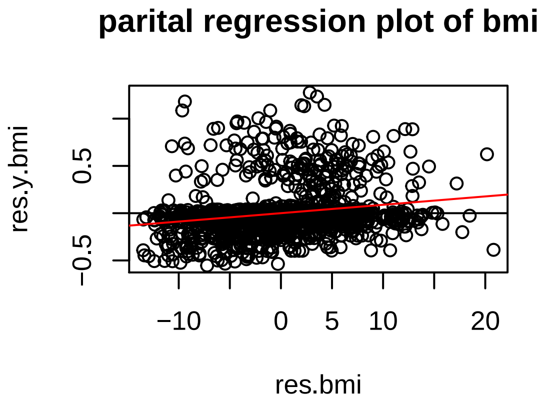
```
[1] TRUE
```

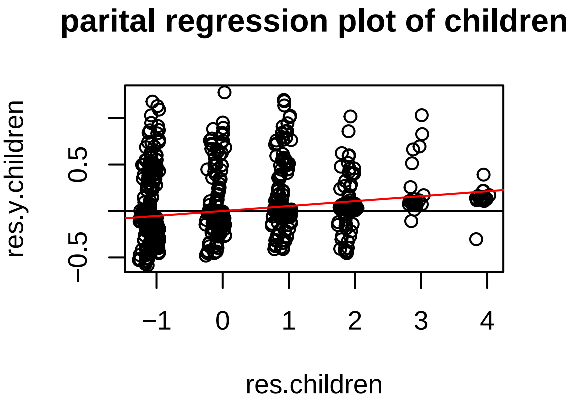
```
Call:
lm(formula = charges^0.15 ~ age + age2 + children + bmi + sex +
    bmi30 * smoker + region, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.4167 -0.1094 -0.0469  0.0192  1.3158 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.816e+00  7.348e-02  38.323  < 2e-16 ***
age               2.428e-02  3.226e-03   7.527 9.53e-14 ***
age2             -6.299e-05  4.024e-05  -1.565 0.117753    
children          5.109e-02  5.710e-03   8.948  < 2e-16 ***
bmi               4.007e-03  1.848e-03   2.169 0.030288 *  
sexmale          -4.518e-02  1.318e-02  -3.429 0.000625 ***
bmi301           -2.074e-02  2.280e-02  -0.910 0.363243    
smokeryes         7.202e-01  2.372e-02  30.360  < 2e-16 ***
regionnorthwest  -3.253e-02  1.883e-02  -1.727 0.084396 .  
regionsoutheast  -8.001e-02  1.896e-02  -4.220 2.61e-05 ***
regionsouthwest  -7.718e-02  1.890e-02  -4.083 4.71e-05 ***
bmi301:smokeryes  4.531e-01  3.261e-02  13.896  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2397 on 1326 degrees of freedom
Multiple R-squared:  0.8063,    Adjusted R-squared:  0.8047 
F-statistic: 501.9 on 11 and 1326 DF,  p-value: < 2.2e-16
```

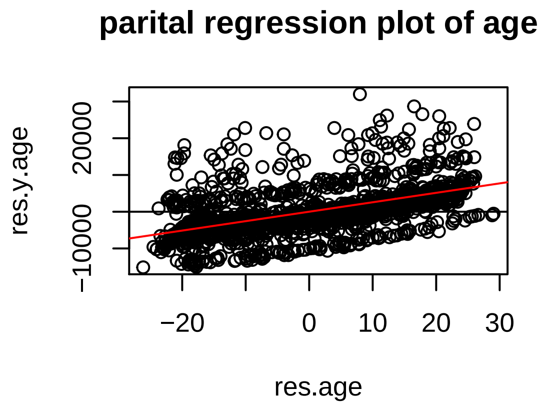
Multiple $R^2$ and Adjusted $R^2$ measured how well our model explained the response variable. The transformed model had improved Multiple $R^2$ and Adjusted $R^2$ compared to initial model.



While box cox plot now showed that non-constant variance issue was addressed, the residual plot still appeaed to have non-constant variance. It was not clear that the transform solved the non-constant and non-linearity issue, we further explored which predictors could be removed by creating partial regression plots.



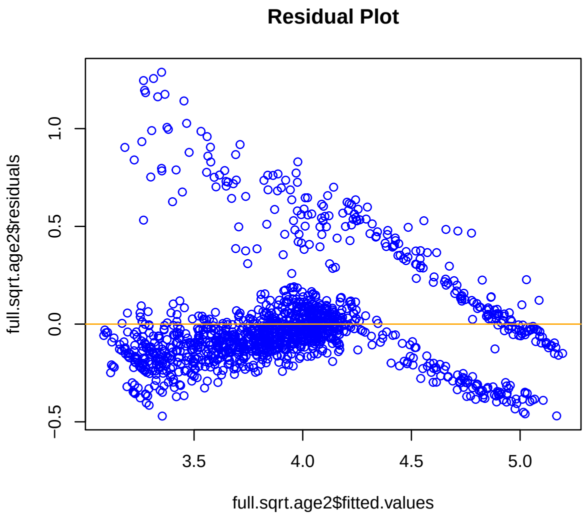


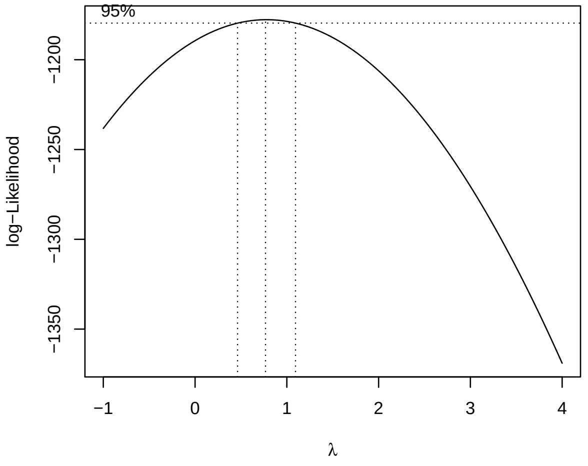


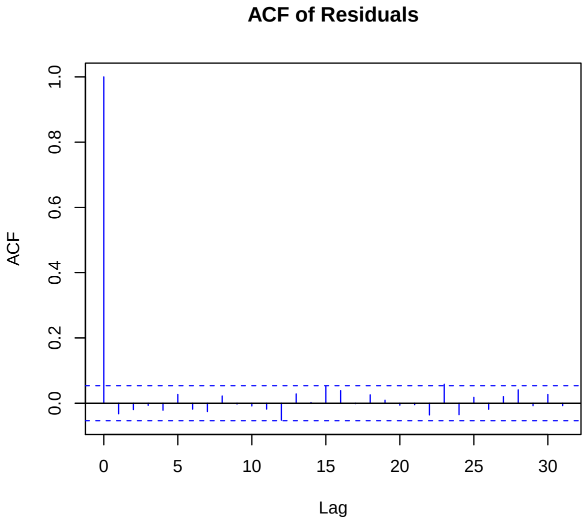
From the partial regression plot, we observed a leaner pattern for all three quantiative variables, this means the linear terms for the predictors **bmi**, **age** and **children** seemed appropriate.

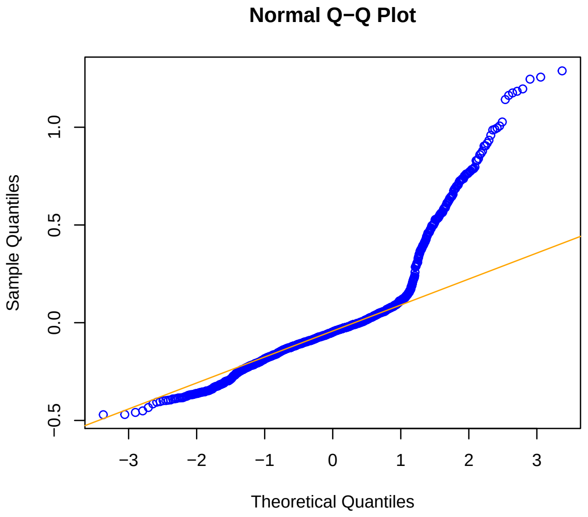
```
'data.frame':   1338 obs. of  9 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
 $ age2    : num  361 324 784 1089 1024 ...
 $ bmi30   : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 1 1 1 ...
```

```
Call:
lm(formula = charges^0.15 ~ log(age) + children + log(bmi) +
    smoker + region, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47069 -0.13211 -0.04842  0.04722  1.28840

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.318640   0.133283   2.391 0.016955 *
log(age)         0.685523   0.018340  37.379  < 2e-16 ***
children         0.041804   0.005906   7.078 2.36e-12 ***
log(bmi)         0.286829   0.036637   7.829 9.98e-15 ***
smokeryes        0.955017   0.017604  54.249  < 2e-16 ***
regionnorthwest -0.035919   0.020353  -1.765 0.077820 .
regionsoutheast -0.085747   0.020425  -4.198 2.87e-05 ***
regionsouthwest -0.073450   0.020434  -3.595 0.000337 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.259 on 1330 degrees of freedom
Multiple R-squared:  0.7731,    Adjusted R-squared:  0.7719
F-statistic: 647.3 on 7 and 1330 DF,  p-value: < 2.2e-16
```

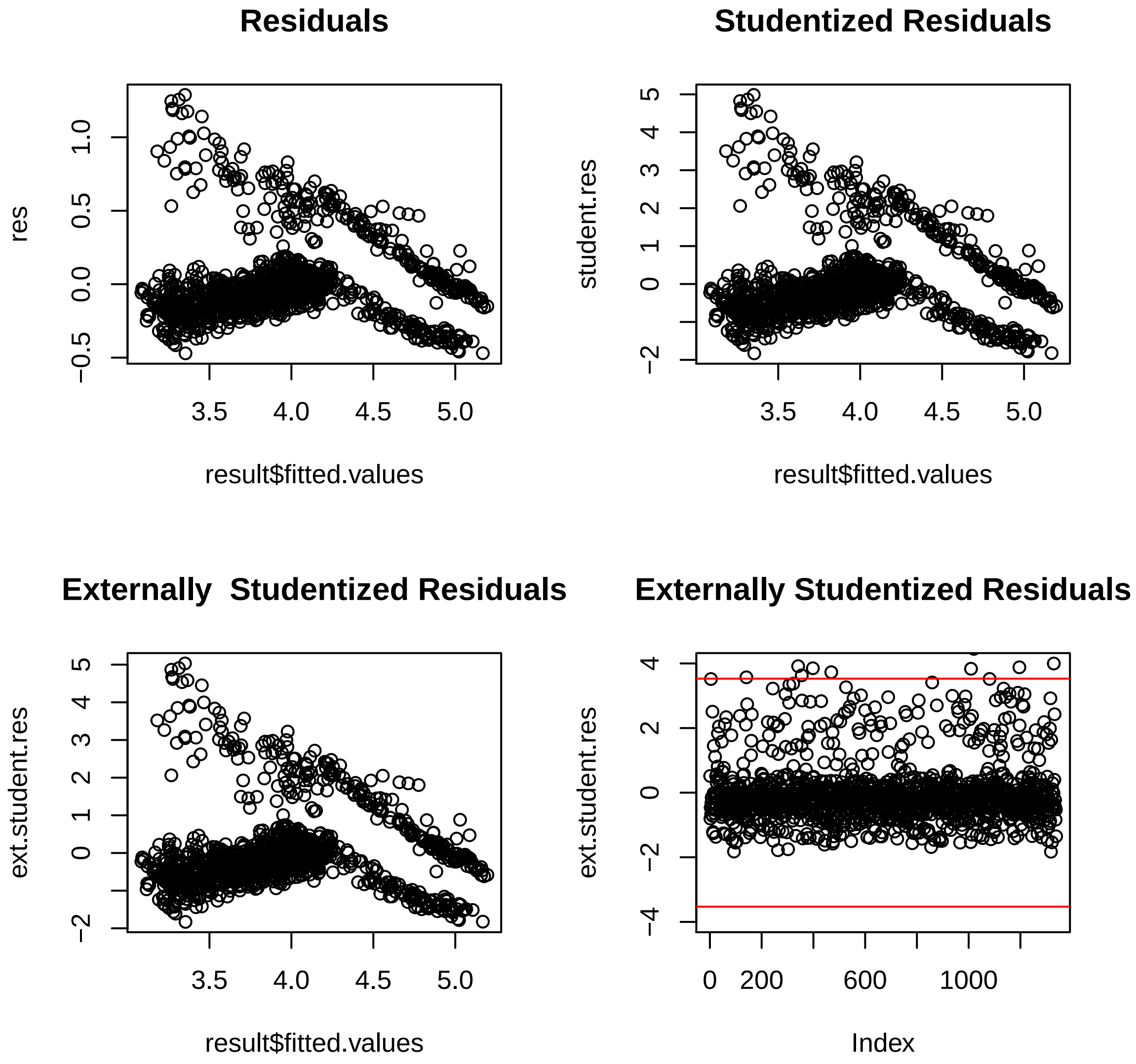

Residual Plot
full.sqrt.age2$residuals
−0.5
0.0
0.5
1.0
3.5
4.0
4.5
5.0
full.sqrt.age2$fitted.values


95%
log−Likelihood
−1350
−1300
−1250
−1200
−1
0
1
2
3
4
λ


ACF of Residuals
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag


Normal Q−Q Plot
Sample Quantiles
−0.5
0.0
0.5
1.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles

```
[1] 3.529468
```



```
     103      141      220      341      355      398      431      469
4.619736 3.570630 4.907768 3.915308 3.631488 3.850610 4.865511 3.728462
     517      527     1009     1020     1028     1040     1196     1329
5.031708 4.585884 3.835164 4.450827 4.672036 4.535482 3.880408 3.997432
```

Even after applying transformations, the model fit is still did not seem to satisfy linear regression assumptions. We still observed the presence of non-linearity and non-constant variance potentially due to outliers in the data. This model may be adequate to explore the relationship between the predictors and the response variable. However, the predicted values may be unrealistic.

## 4. Alternate Models Considered:

In the EDA section, we observed that our response variable was right-skewed. From the validation of the initial-model assumptions, we acknowledged that our initial model could be used to explore the relationship between response and predictor variables. However, predictions may not be accurate.

Next we considered a logistic regression by converting the response variable into a categorical variable and changed the goal to be to determine the likelyhood of charges above or below a threshold value.

### 4a. Find the best fit model that can predict if the medical charges are greater than or less than/equal $20000?

We convereted the response variable to categorical variable and splitting the data into training & testing dataset

First we use a training dataset to fit the model.

```
lrmodel1<-glm(lrcharges ~ age + bmi + smoker + region + sex + children , family="binomial" , data = lrd
summary(lrmodel1)
```

```
Call:
glm(formula = lrcharges ~ age + bmi + smoker + region + sex +
    children, family = "binomial", data = lrdata_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8184  -0.3834  -0.2428  -0.1373   3.1248

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.91973    1.07038  -7.399 1.37e-13 ***
age              0.03953    0.01153   3.428 0.000607 ***
bmi              0.11077    0.02582   4.289 1.79e-05 ***
smokeryes        4.85604    0.37410  12.981  < 2e-16 ***
regionnorthwest  0.22102    0.44125   0.501 0.616451
regionsoutheast -0.47459    0.41783  -1.136 0.256019
```

```
regionsouthwest -0.81043    0.45047  -1.799 0.072008 .
sexmale          0.05906    0.30390   0.194 0.845916
children         0.03180    0.12272   0.259 0.795552
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 664.52  on 668  degrees of freedom
Residual deviance: 321.06  on 660  degrees of freedom
AIC: 339.06

Number of Fisher Scoring iterations: 6
```

The higher the difference between null deviance and residual deviance, the better the model's predictability would be. Our data supported the claim that our logistic regression model was useful in estimating the log odds of whether medical **charges** are greater or less than $20000

The model summary showed, based Z-value (Wald test) age, bmi, and smoker are significant predictors with p-value of less than 0.05. Furthermore, the region sex and children predictors seem insignificant, hence removing the model.

Hypothesis-testing: $H_0$: coefficients for all predictors is = 0 and
$H_1$: at least one coefficient is not zero

```r
1-pchisq(lrmodel1$null.deviance - lrmodel1$deviance,8)
```

```
[1] 0
```

Based on the small p-value we rejected the null hypothesis that at least one of these coefficients was not zero.

```r
lrmodel2 <-glm(lrcharges ~ age + bmi  + smoker, family="binomial" , data = lrdata_train)
summary(lrmodel2)
```

```

Call:
glm(formula = lrcharges ~ age + bmi + smoker, family = "binomial",
    data = lrdata_train)
```

```
Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.7030  -0.3739  -0.2481  -0.1506   3.1720  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -7.86104    1.03377  -7.604 2.87e-14 ***
age          0.04087    0.01150   3.554 0.000379 ***
bmi          0.10134    0.02465   4.111 3.94e-05 ***
smokeryes    4.75564    0.35844  13.268  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 664.52  on 668  degrees of freedom
Residual deviance: 327.44  on 665  degrees of freedom
AIC: 335.44

Number of Fisher Scoring iterations: 6
```

We knew from the Wald test that the region sex and children predictors were insignificant, so we will conducted the delta $G^2$ test to see if these predictors can be removed from the model.

```
#test if additional predictors have coefficients equal to 0
1-pchisq(lrmodel2$deviance - lrmodel1$deviance,5)
```

```
[1] 0.2701389
```

The p-value was 0.0941 greater than 0.05, so we could not reject the null. We then chose a simpler model with just the three predictors **age**, **bmi** and **smoker**.

### 4b. Logistic Regression model validation

We then tested how well this logistic regression model performed in predicting an outcome that medical **charges** were greater than or less than $20000 given the values of other predictors, using the probability of the observations in the test data of being in each class, we will choose a threshold of 0.5 for the confusion matrix.

False Positive Rate: When it was actually no, how often would it predict yes = 0.0625

False Negative Rate: When it was actually yes, how often would it predict yes = 0.2198582

Sensitivity out of all the positive classes, how much was predicted correctly = 0.7801418

Specificity determined the proportion of actual negatives that were correctly identified = 0.9375

The AUC value for our model was 0.8999704. The AUC value was higher than 0.5, which meant the model did better than random guessing the classifying observations.

## 5. Conclusion:

1. Even after applying transformations, the model fit is still did not satisfy linear regression assumptions.
2. We still observed non-linearity, and non-constant variance issues were still not addressed in the model.
3. The regression assumption issues could be due to skewed data or outliers in the dataset.
4. We then conclude that our initial transformed model is useful for exploring the relationship between predictor and response variables. However, the predicted values would be unreliable.
5. The alternate logistic regression model would be a better predictor of the likelihood of charges above 20,000 USD when other variables were held constant.

Our team recommendation for a prediction model:

**The logistic regression model has better predictability. This model would be:**

$$\log \frac{\pi_i}{1-\pi_i} = 0.04087age + 0.10134bmi + 4.75564smoker$$