STAT 6021: Project Two

Medical Insurnace Costs

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

1 Executive Summary :

The growing issue of higher medical costs per family has become a big concern to Americans. Increasing healthcare costs stop people from getting the needed care or fill prescriptions. Many families have difficulty in affording healthcare costs and this difficulty in paying bills has significant consequences for US families.

We selected a personal medical costs dataset. We want to explore what demographic characteristics affect the medical charges each family potentially pays in a year. So, we have considered Medical Cost Personal Dataset.

Dataset: datasets_13720_18513_insurance.csv

- The variables are 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 are –**
  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 a best fit model to predict the insurance **charges** based the demographic predictor variables and evaluate the validity and usefulness of this model.

  Additionally, we plan 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.

2 Exploratory Data Analysis :

We start our exploratory data analysis by taking a look at the dataset.

```
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
```

Our dataset looks clean and has no missing values.

At a glance, we have six predictors and a response variable **charges**. The dataset has 1338 rows, and non of the columns are 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 ...
```

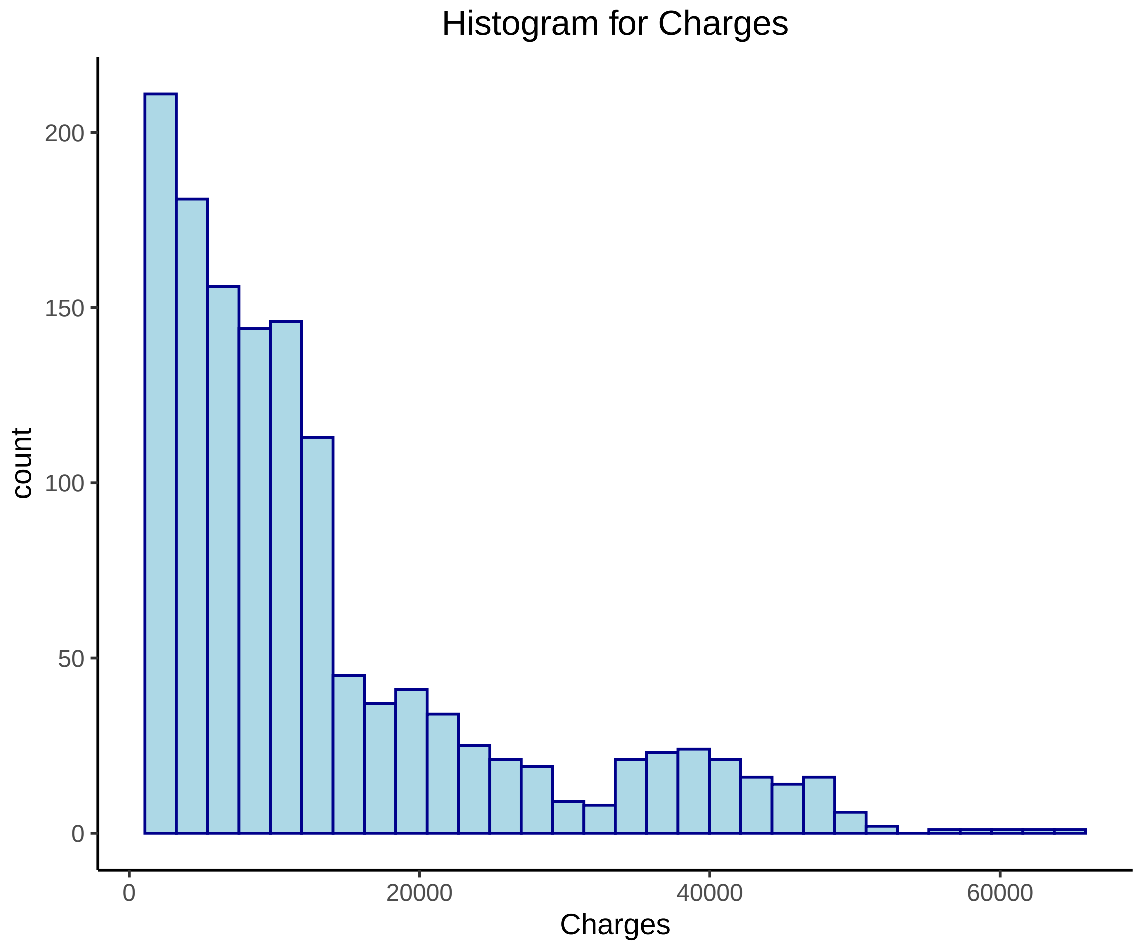
```
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

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

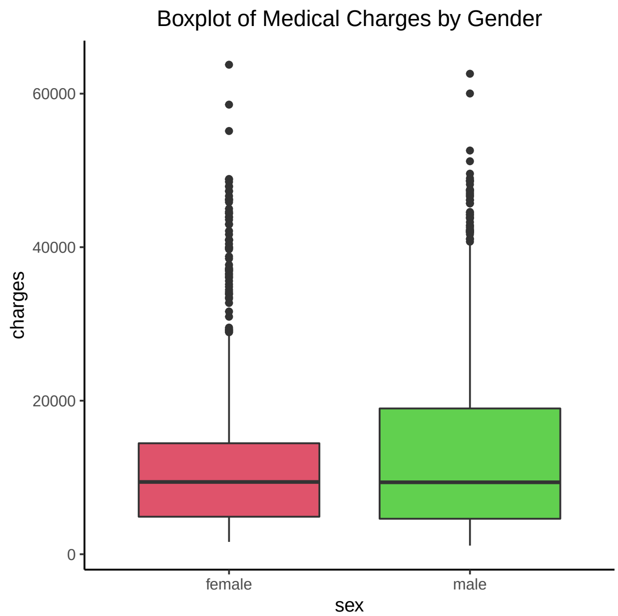
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1122    4740    9382   13270   16640   63770 
```

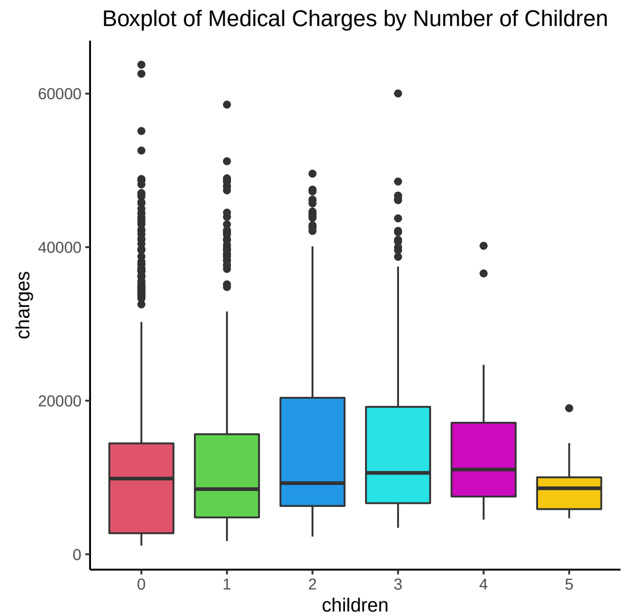
- From the summary we can make following observations :
  - The observations(records) are almost evenly distributed across region.
  - The age varies between low of 18 and a max of 64.
  - The observations are almost evenly distributed by sex.
  - The dataset has almost 4:1 non-smoker to smoker ratio or only 20.5% people smoke.
  - The bmi varies between a min of 15.96 and max of 53.13.
- The response variable mean is greater than median, this is an indication that data is right-skewed. This can be confirmed from the histogram we can confirm this from the histogram of **charges** shown below. The predictor age also seems to right skewed



Form the boxplot shown below for medical **charges** by **sex** the median value of the medical **charges** for both male and female is almost the same. the third quartile for male seems to greater than female, so the data may be skewed towards the men.

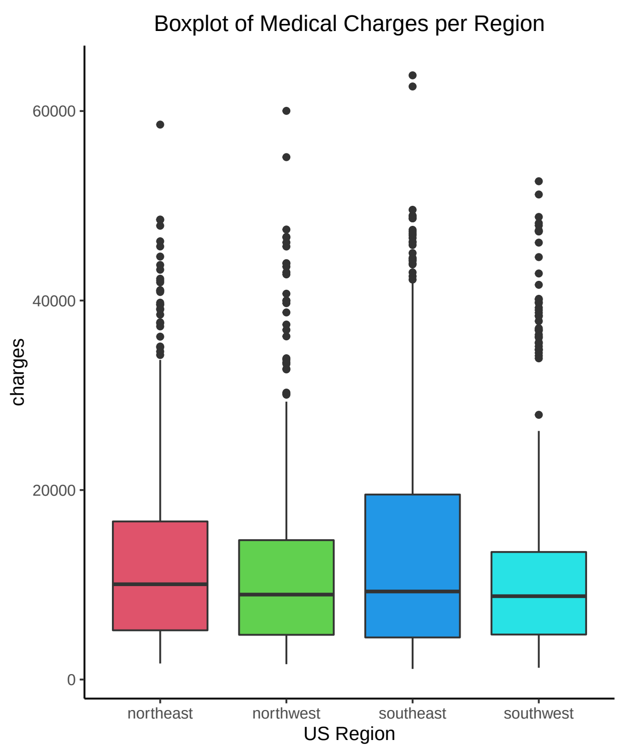
And, the boxplot of medical **charges** by **children**, we can make an interesting observation that the medical **charges** for people with 5 children are lower than people with one to four children and people with no children have the lowest medical charges

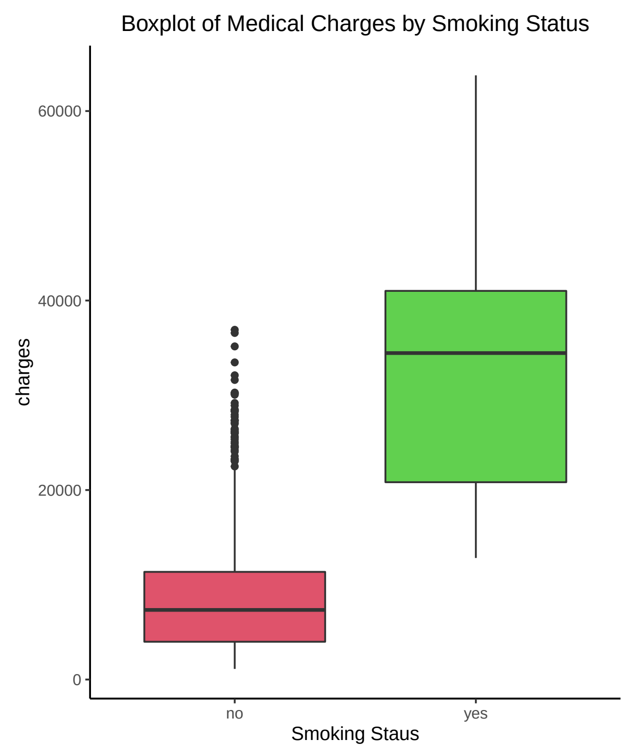


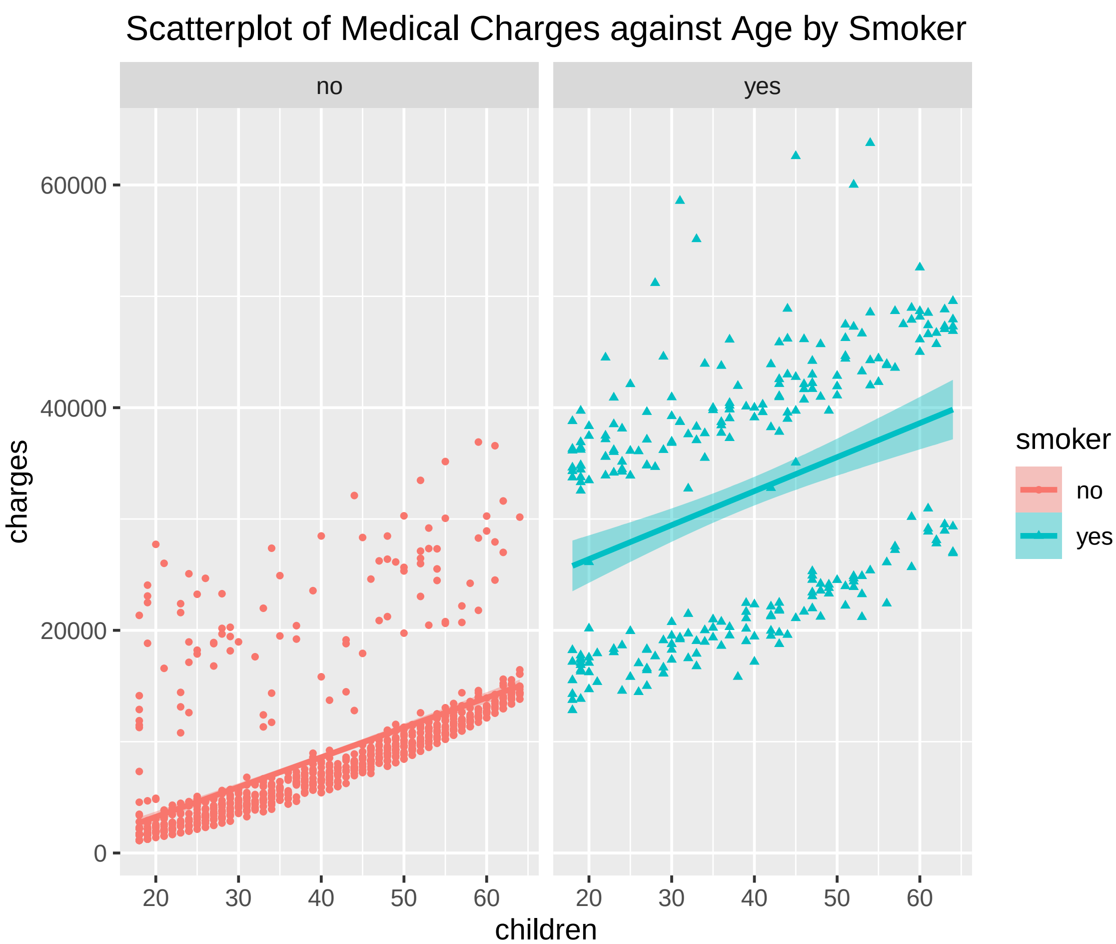


Form the boxplot of medical **charges** per **region** the median value of the medical **charges** across all four US regions is almost the same. The people in the southeast seem to have higher medical expenses then the people in the other areas.

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

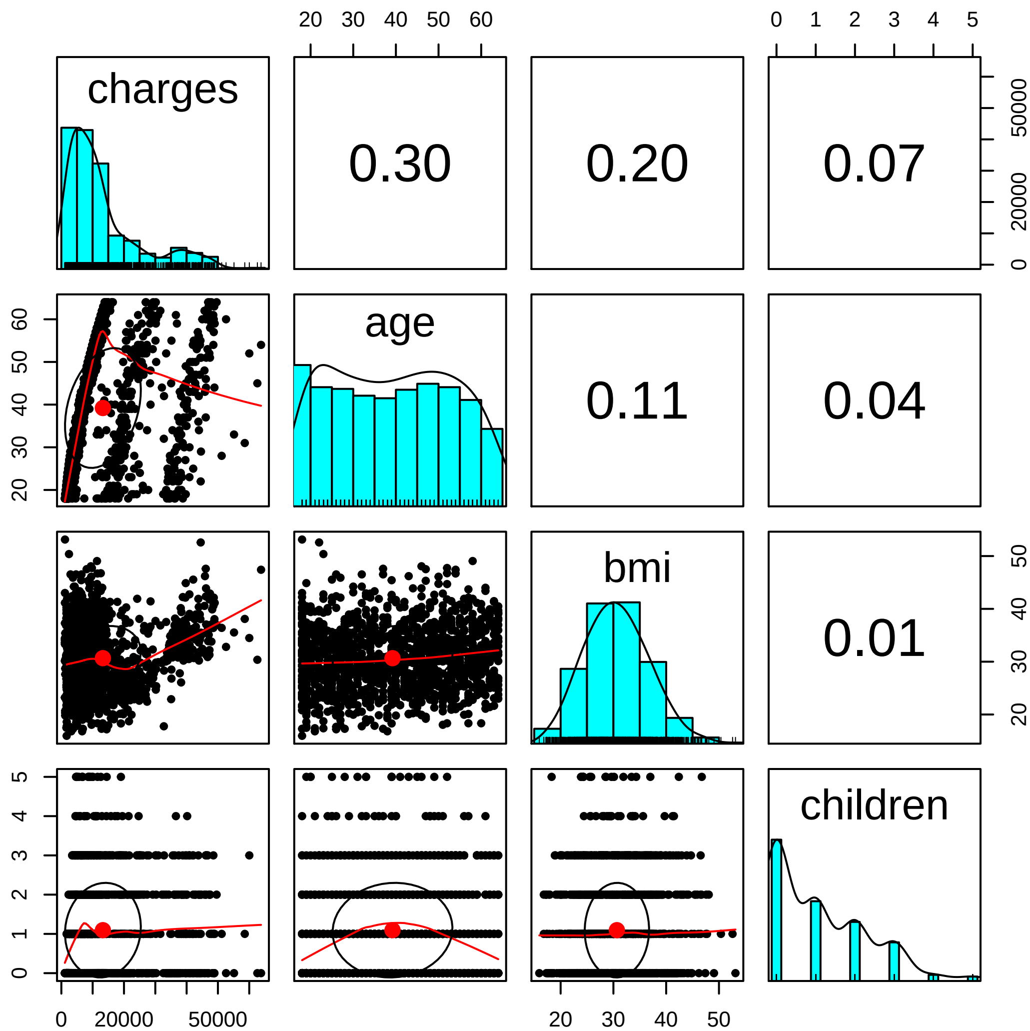






From the above scatter plot, we observe that medical charges increase with age for both smokers and non-smokers. However, those who smoke tend to have higher medical expenses than those who do not.

	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



and the plot.For example, we observe **somewhat(moderate)** correlated between **age** and **charges**, and **bmi** and **charges** and However, **bmi** and **age** and **children** and **charges**. have a weak correlation. We will further explore these relationships as we continue our analysis towards building a final module.

As we observed from the summary of the dataset, we can see that **smoker** status (class) has better correlation to **charges** the compared with **non-smoker** status (class). This could also mean that smokers have more medical expenses than non-smokers.

From our exploratory analysis more than one predictor can be considered for our initial model. We observed that **age** has somewhat better correlation with **charges** and **smokers** tend 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 We will explore candidate models applying model automatic predictor search procedures. We will use the R^2_{adj} and the BIC metrics to identify likely models since these both penalize for adding more terms.

- Based on our analysis -
  - The model with lowest BIC is: **chareges** $= B_0 + B_1(\text{age}) + B_2(\text{bmi}) + B_3(\text{children}) + B_4(\text{somkeyes})$
  - The model with highest adjusted R^2 is **chareges** $= B_0 + B_1(\text{age}) + B_2(\text{bmi}) + B_3(\text{children}) + B_4(\text{somkeyes}) + B_5(\text{regionsoutheast}) + B_6(\text{regionsouthwest})$

We also considered the models with the highest R^2, lowest Cp, and lowest MSE values. The best Cp and best MSE are both on the the same model as the best adjusted R^2.

The model with the best R^2 value has all predictors as adjusted R^2 inadditon to regionnorthwest

From the above, we see that our model with the lowest BIC (-1817.233) is the simple regression of **age, bmi30, children, smokeyes** against **charges** . The model with the highest adjusted R^2 is **age, bmi,children,smokeyes,regionsoutheast, and regionsouthwest** against medical **charges**.

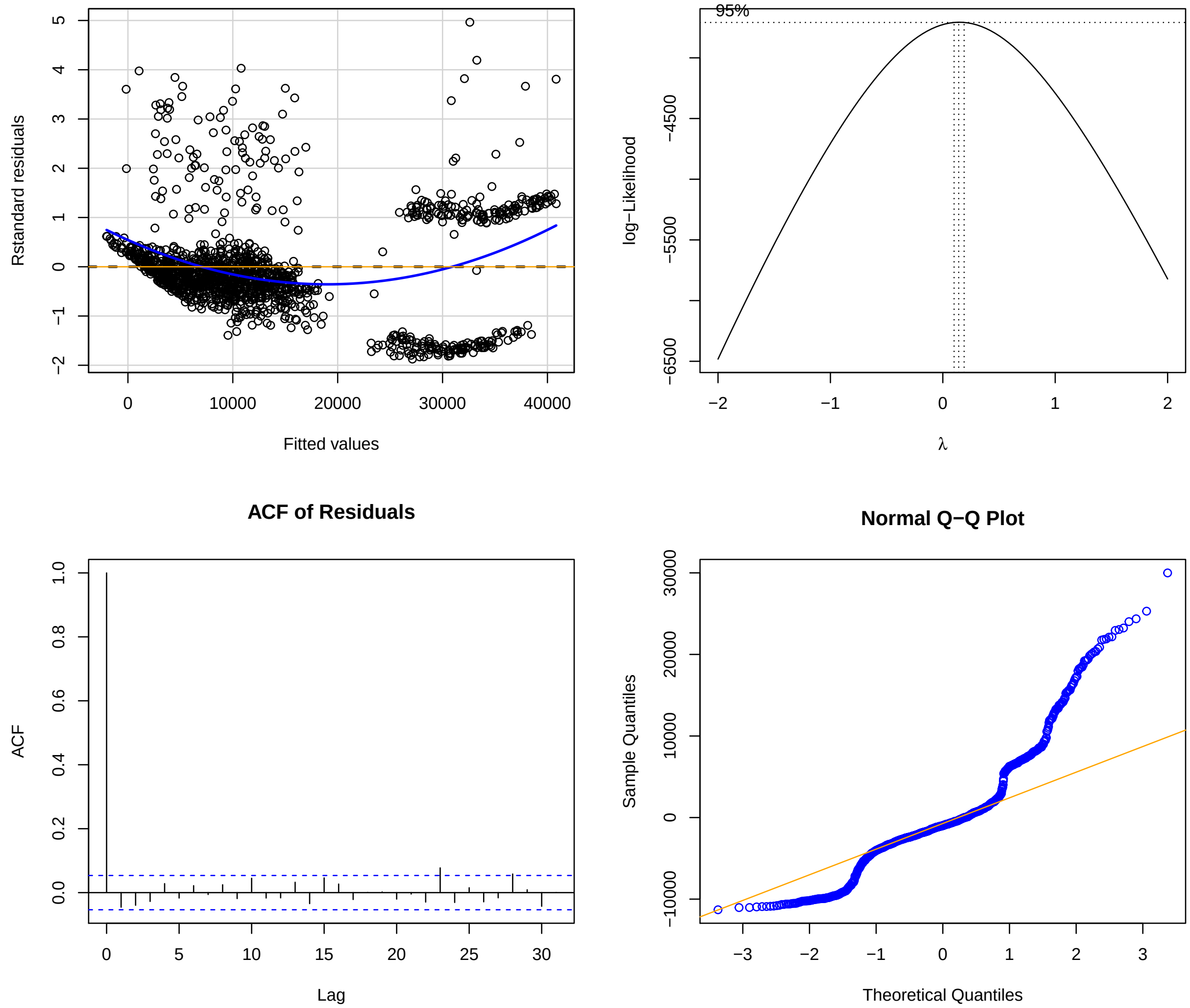
3. Initial Model Considered:

Based on results from the model search procedures, we ill choose a

```
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
```

Validating linear regression assumptions:



Looking at the above plots, we observe that 1. variance is not constant as seen in the box-cox plot and 2. non-linearity as seen in residual plot. We will fix address the constant variance the issue of non-constant variance.

In order to fix non-constant variance and non-linearity issues we will transform y first and the then predictors

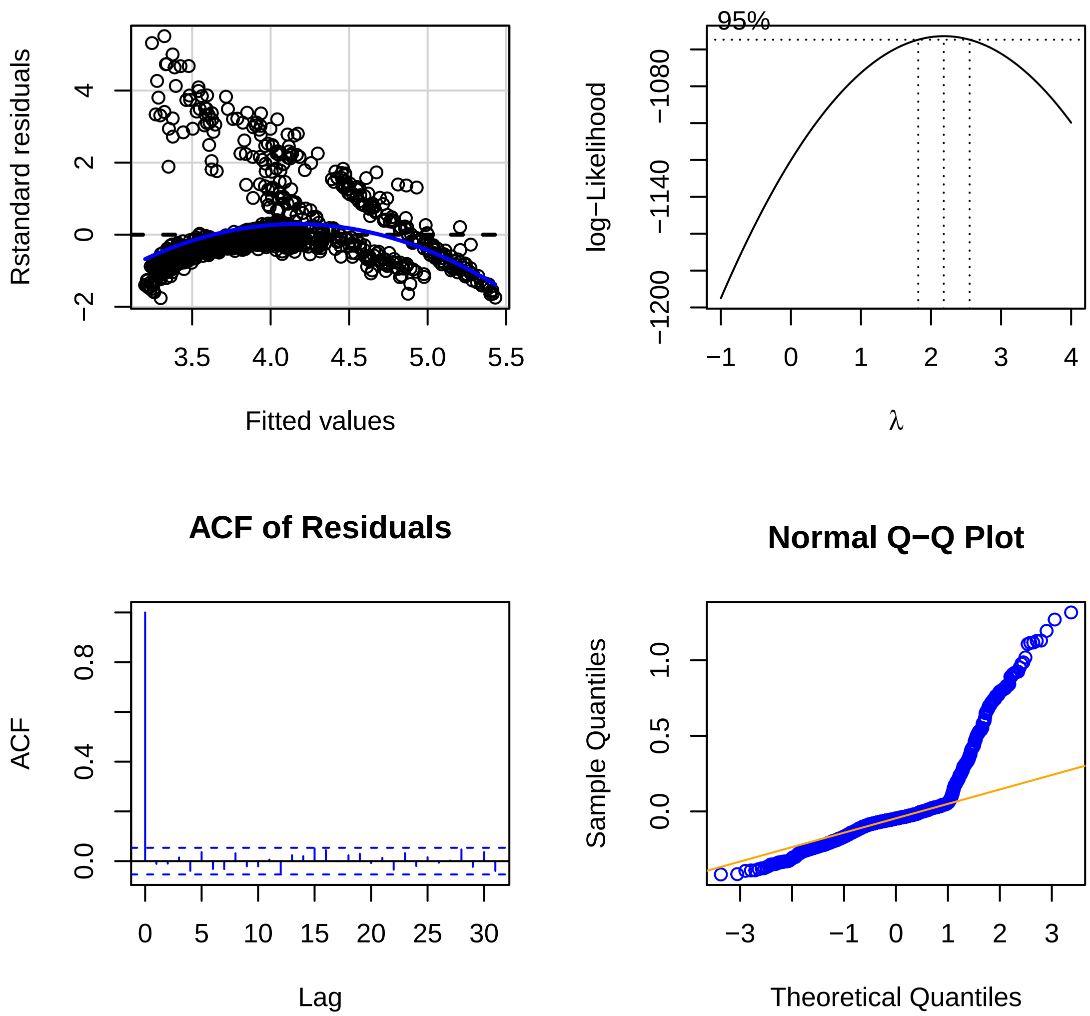
In our hypothesis, we said that, old age people, people who smoke and people with high bmi (bmi>30) may be at high risk and so their medical costs may be higher, based on that hypothesis and considering that our initial model suffers from non-linearity and non-constant variance issue. We will transform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

Following transformations will be applied 1. Transform **charges** (y) to fix non-constant variance 2. Transform age - by adding a non-linear term for age 3. Create a indicator variable for bmi (obesity indicator) 4. Specify and interaction between smokers and bmi indicator predictor

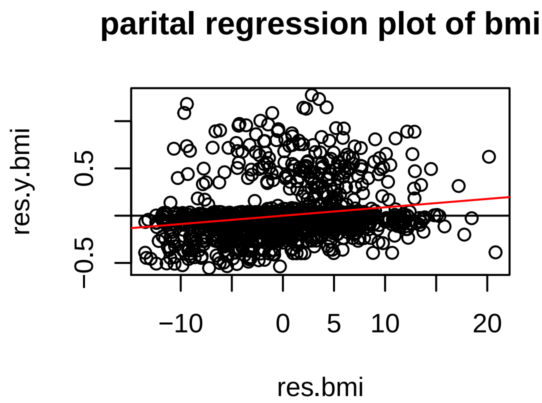
```
[1] TRUE


Call:
lm(formula = charges^0.15 ~ age + age2 + children + bmi + sex +
```

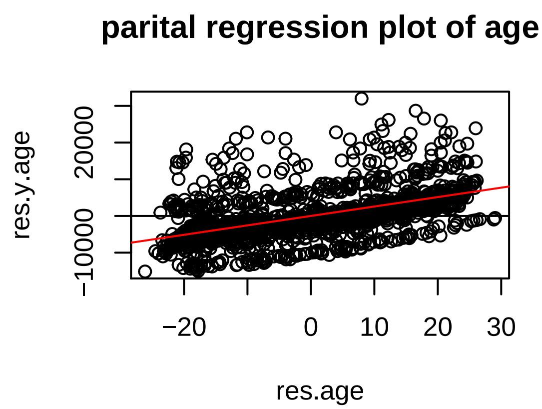
```
    bmi30 * smoker + region, data = data)

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 measure how well our model explains the response variable. The transformed model has improved Multiple R^2 and Adjusted R^2 compared to initial model.



While bob-cox plot now shows that non-constant variance issue is addressed, but from the residual plot it is not clear have solved the non-constant and non-linearity issue, we can further explore which predictors can be removed by creating partial regression plots

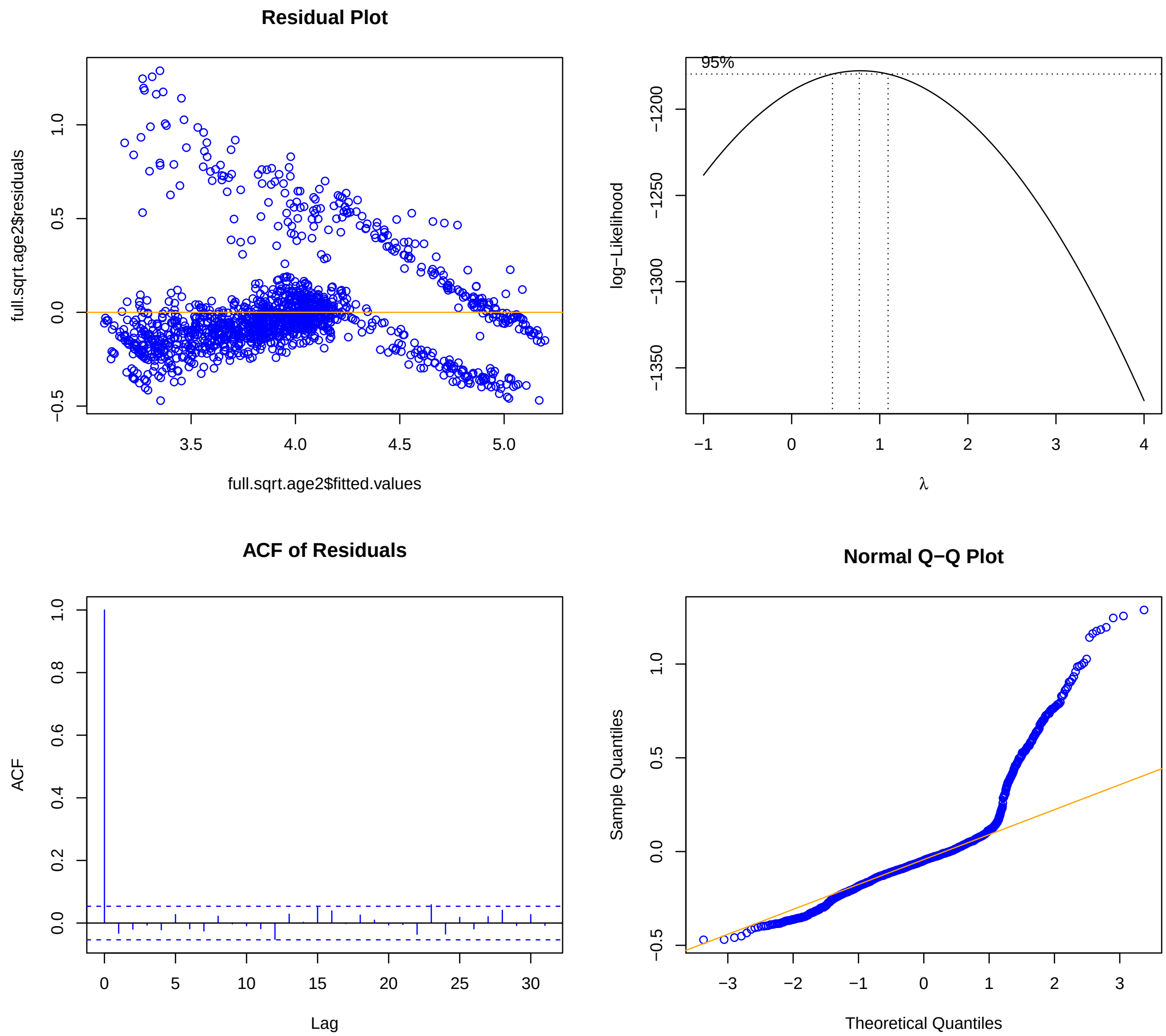




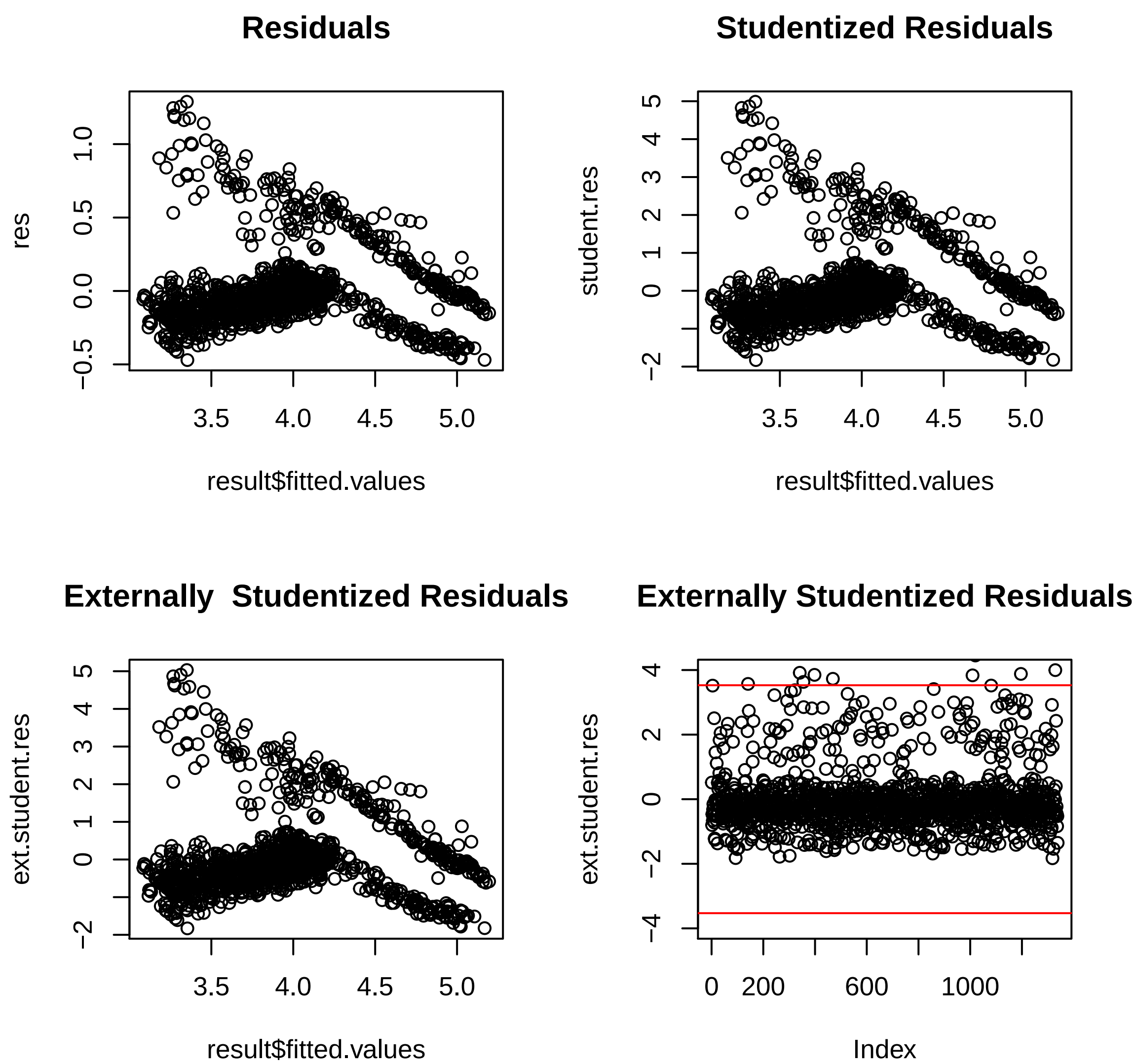
From the above partial regression plot, we see a leaner pattern for all three quantiative variables, this means the linear terms for the predictors **bmi**, **age** and **children** is appropriate.

```
'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
```




```
[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 not satisfying linear regression assumptions. We still see the presence of non-linearity and non-constant variance. It may be due to outliers in the data. This model is good enough to explore the relationship between the predictors and the response variable. However, the predicted values may be unrealistic.

4. Alternate Model Considered:

In the EDA section, we observed that our response variable is 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.

So we will consider a logistic regression by converting the response variable into a categorical variable.

Our goal now is to answer the question -

3. find the best fit model that can predict if the medical charges are greater than or less than/equal $20000?

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

we will first use the training dataset to fit the model.

```
lrmodel1<-glm(lrcharges ~ age + bmi + smoker + region + sex + children , family="binomial" , data = lrda
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. Our data supports the claim that our logistic regression model is useful in estimating the log odds of whether medical **charges** are greater or less than $20000

The model summary shows, 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

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

```
[1] 0
```

small p-value we reject the null hypothesis that at least one of these coefficients is 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 already know from the Wald test that the region sex and children predictors are insignificant, so we will conduct the delta G^2 test to see if these predictors can be removed from the model.

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

```
[1] 0.2701389
```

p-value is 0.0941 greater than 0.05, so we cannot reject the null so that we will choose the simpler model with just the three predictors **age**, **bmi** and **smoker**.

Logistic Regression model validation

Next, we will go over how well-chosen logistic regression model does in predicting an outcome that medical **charges** are 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's actually no, how often does it predict yes = 0.0625

False Negative Rate: When it's actually Yes, how often does it predict yes = 0.2198582

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

Specificity determines the proportion of actual negatives that are correctly identified = 0.9375

The AUC value for our model is 0.8999704. The AUC value is higher than 0.5, which means the model does better than random guessing the classifying observations.

5. Conclusion:

1. Even after applying transformations, the model fit is still not satisfying linear regression assumptions.
2. We still see non-linearity, and non-constant variance issues are still not addressed in the model.
3. It could be due to skewed data or outliers in the dataset.
4. So we conclude that our initial transformed model is useful for exploring the relationship between predictor and response variables. However, the predicted values will be unreliable.
5. The alternate logistic regression model has the better predictability

Our team recommendation:

The logistic regression model has better predictability.