ReGression Final Project

Tung Nguyen

5/5/2020

LInear Regression on Medical Insurance Costs

I. Data Description

This dataset comes from the book Machine Learning with R by Brett Lantz. The data contains medical information and costs billed by health insurance companies. There are 1338 observations and 7 variables in this dataset:

1. age: age of primary beneficiary
2. sex: insurance contractor gender, female, male
3. 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 ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
4. children: Number of children covered by health insurance / Number of dependents
5. smoker: Smoking
6. region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
7. charges: Individual medical costs billed by health insurance

There are 4 quantitative and 3 categorical variables that each describe a certain feature of the contractor. Below are the first five observations of the data.

```
# A tibble: 6 x 7
    age sex      bmi children smoker region    charges
  <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
1    19 female  27.9        0 yes    southwest  16885.
2    18 male    33.8        1 no     southeast   1726.
3    28 male    33          3 no     southeast   4449.
4    33 male    22.7        0 no     northwest  21984.
5    32 male    28.9        0 no     northwest   3867.
6    31 female  25.7        0 no     southeast   3757.
```

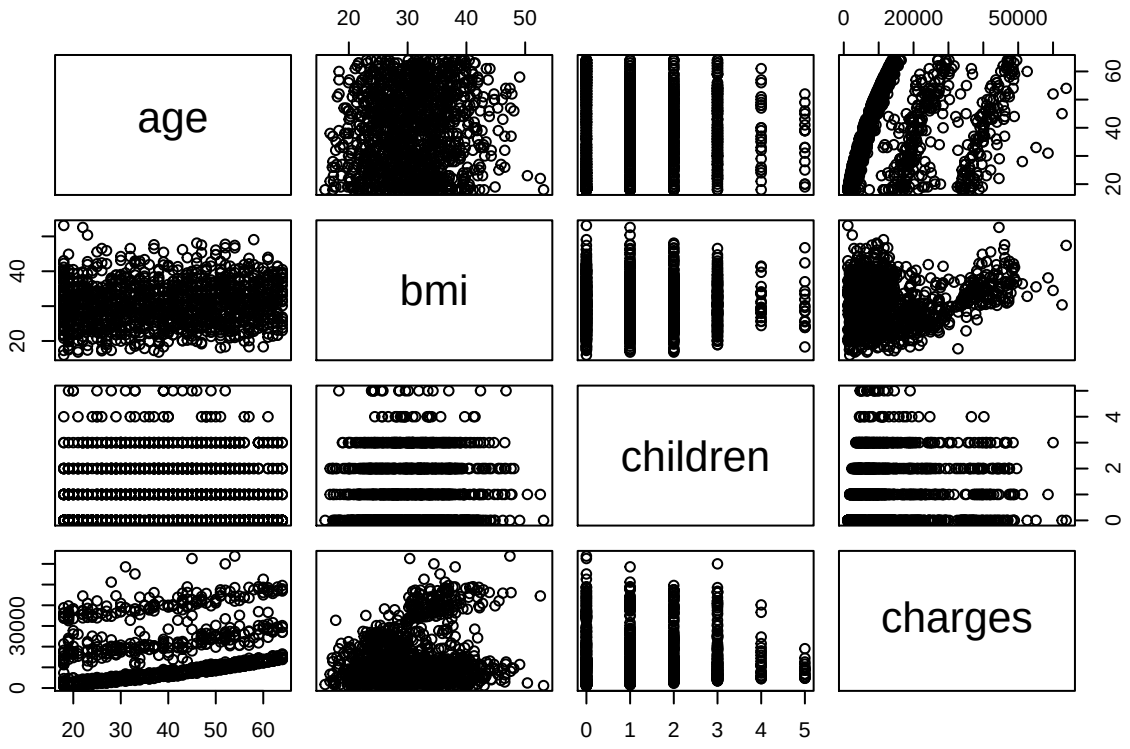
One question of interest is whether insurance costs can be determined from individual features. In our dataset, the insurance costs are the charges variable. It is a continuous variable in terms of dollars. age, bmi and children are numerical features while sex, smoker (whether this person smokes or not) and region are categorical features.

II. Data Exploration

Below is a summary of the basic statistics for our variables. Looking at the response variable, the minimum value is 1122 while the maximum value is 63770. Most points cluster between 4740 and 16640. This large variance in the response variable indicates that there are potential outliers. The other quantitative variables are reasonably varied.

```
      age            sex                 bmi           children
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
 Median :39.00   Mode  :character   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
    smoker             region             charges
 Length:1338        Length:1338        Min.   : 1122
 Class :character   Class :character   1st Qu.: 4740
 Mode  :character   Mode  :character   Median : 9382
                                       Mean   :13270
                                       3rd Qu.:16640
                                       Max.   :63770
```

Scatterplots of Quantitative Variables



Scatterplots of numerical variables with the response variable

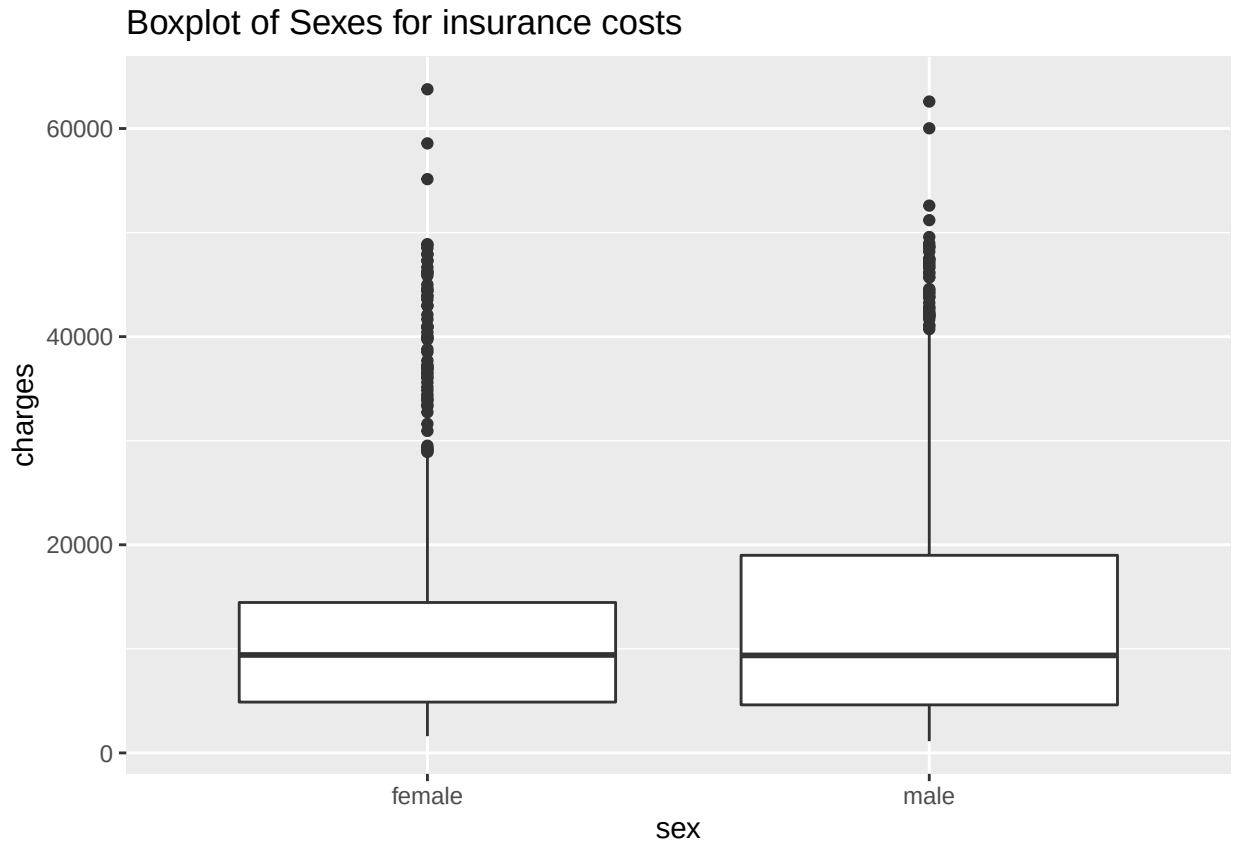
The Children variable seems to be a categorical variable because there are only a few values in the variable

(1-5). Below is the frequency distribution of children. It confirms that children should be treated as a categorical variable even though it is numerical.

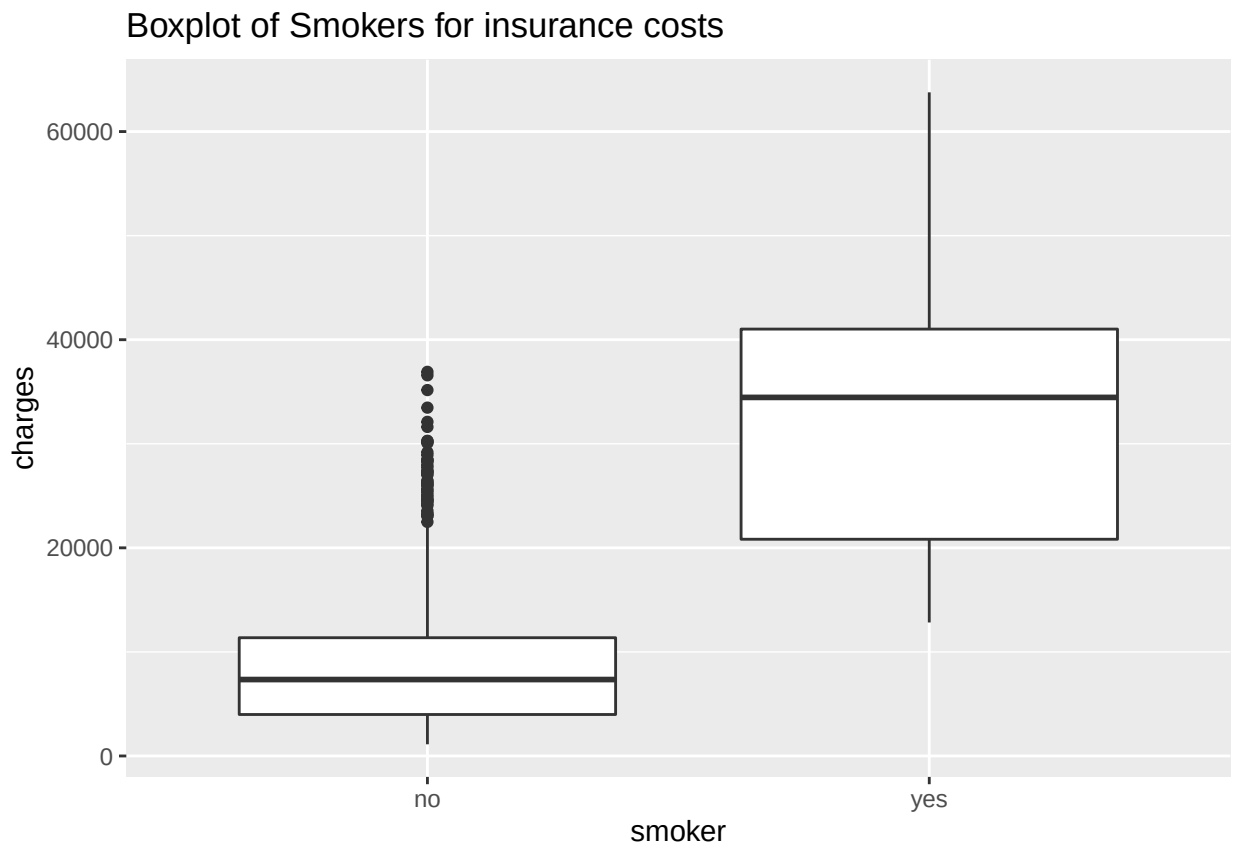
```
# A tibble: 6 x 2
  children count
     <dbl> <int>
1        0   574
2        1   324
3        2   240
4        3   157
5        4    25
6        5    18
```

Also from the plot above, age shows a linear relationship with our response, charges, albeit in clusters. bmi also shows a somewhat linear relationship with charges. In addition, bmi and age displays a certain level of correlation. The interaction of these two variables may play an important role in our model.

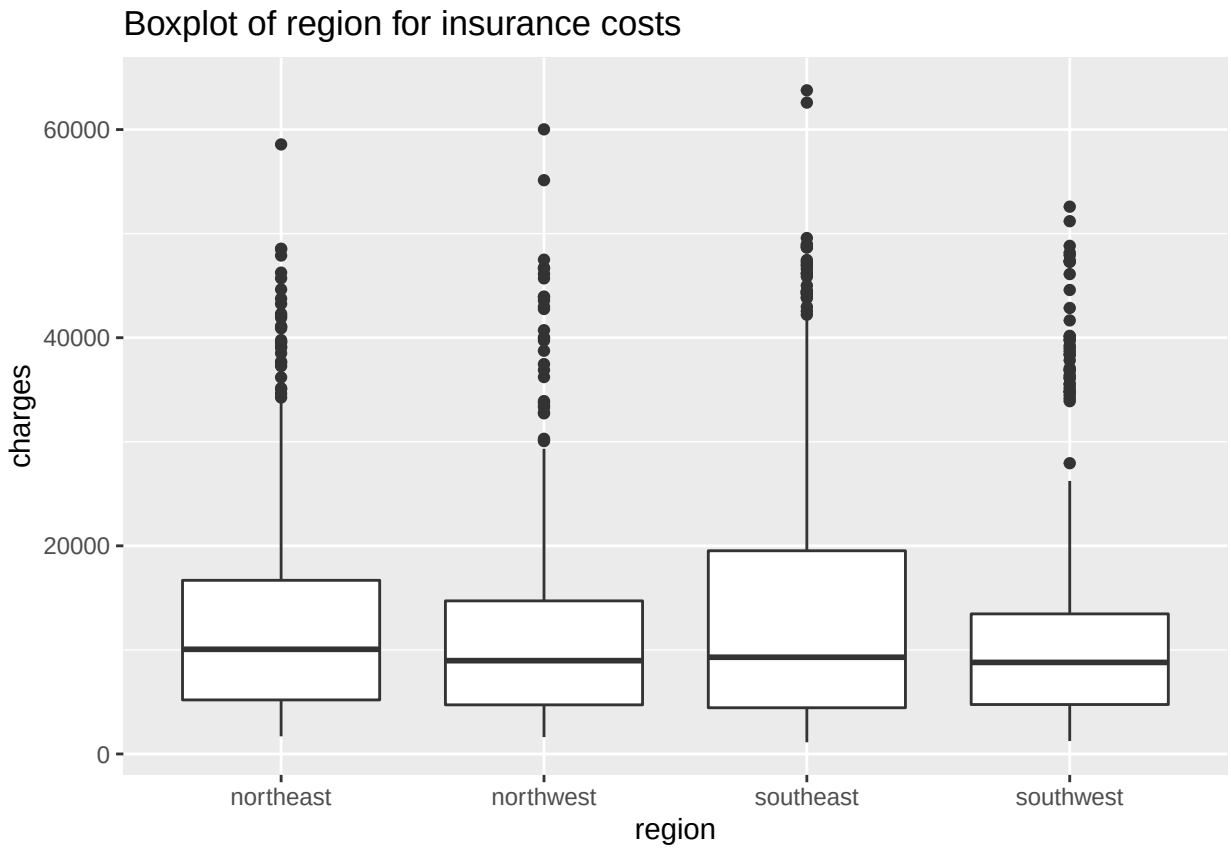
Boxplots of Categorical Variables



The plot above shows the boxplot of variable sex for insurance costs. The median costs for both sexes are pretty equal though there is more variance in insurnace costs for male.

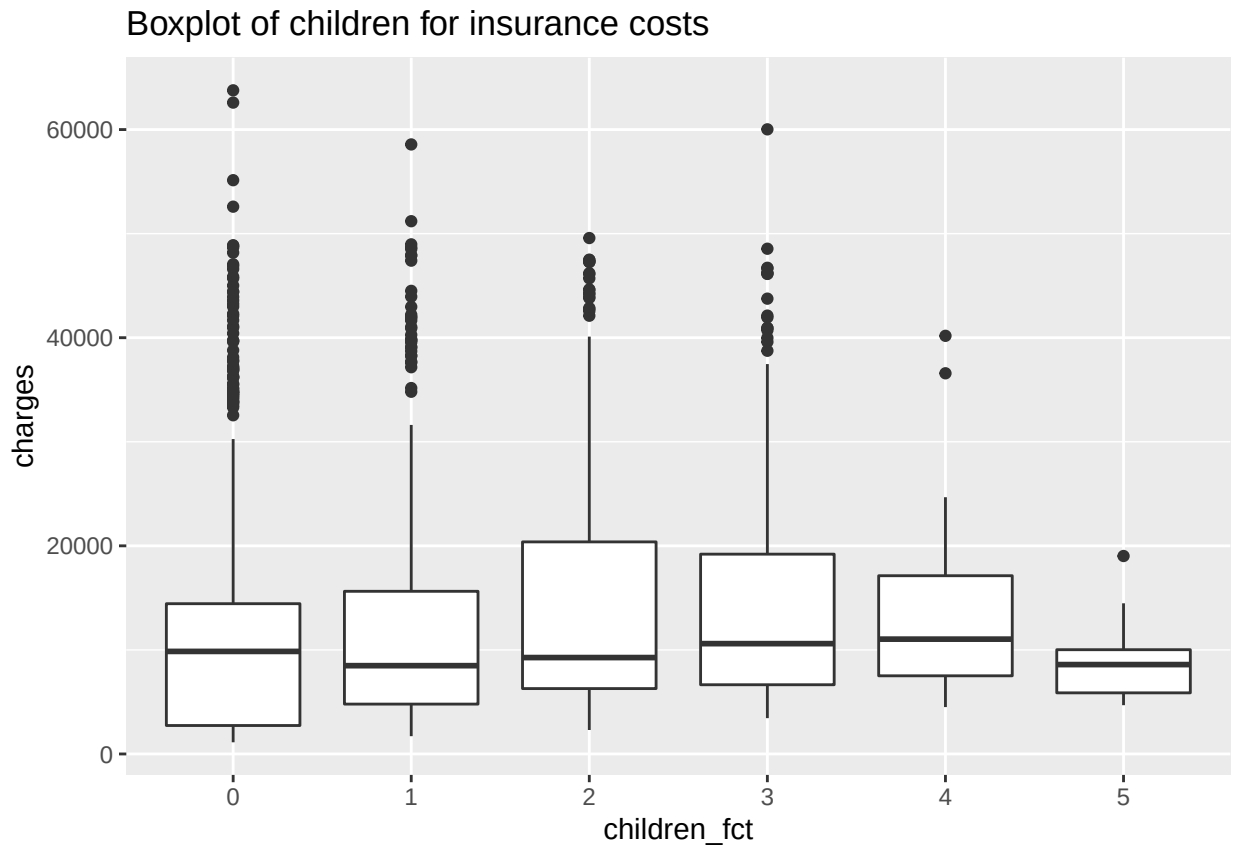


There's a clear trend here. Smokers have a much higher median insurance costs in comparison with non-



smokers.

There's not a clear trend for variable region in relation with insurance costs. The insurance costs decreases slightly from east to west, however.



The median insurance costs start high for contractors with zero children then goes down for 1 children contractors. The median costs keep increasing but then decreases when a contractor has 5 children. This could be due to the insurance companies policy to start with a high default cost. They give discount for contractors with children at a small rate then give really high discount for contractors with more than 5 children. One thing to note is that the boxplots show there are many outliers in our categorical variables. The outliers have the potential to influence the model so we'll come back to address this issue if necessary.

III. Model

Methodology

One interesting question is whether the insurance costs can be determined by individual features. I'll start with a full model that includes all the variables in the data. The full model equation is below.

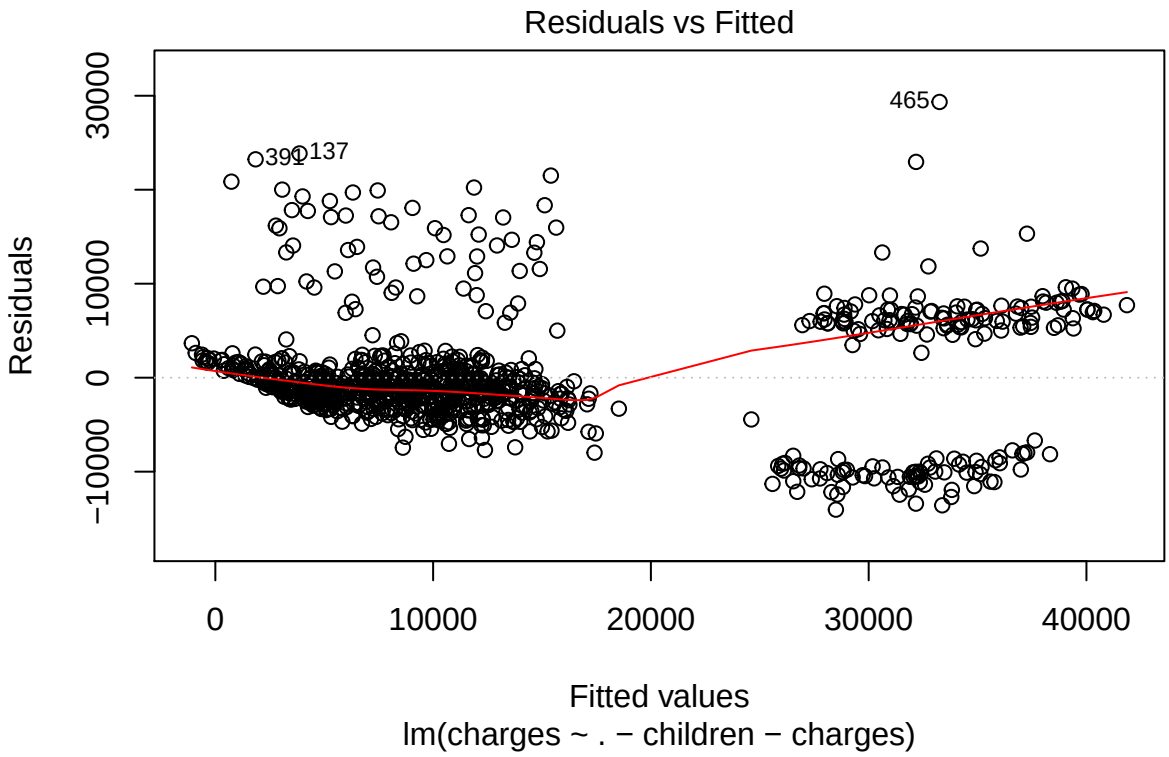
$charges = \beta_0 + \beta_1 age + \beta_2 bmi + \beta_3 sex + \beta_4 children + \beta_5 region + \beta_6 smoker$

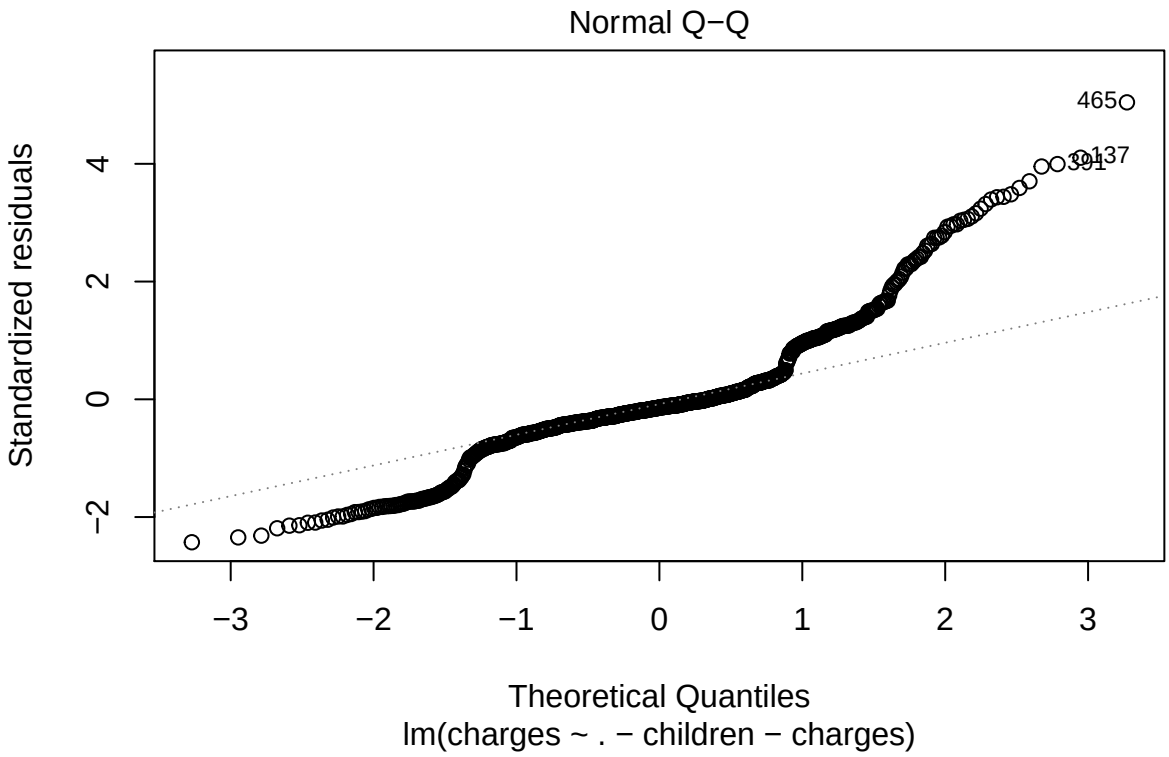
Train and test split Before modeling, we need to develop a rigorous approach for model selection.There are many methods for model selection such as AIC, BIC and cross validation, etc. From my own research AIC and cross validation strive to ahieve the same thing in different ways. Since the priority is the predictive power of the model, cross validation is chosen and the matrics are R-squared and RMSE. The greater the R-squared and the smaller the RMSE values, the better the predictive performance of our model is.

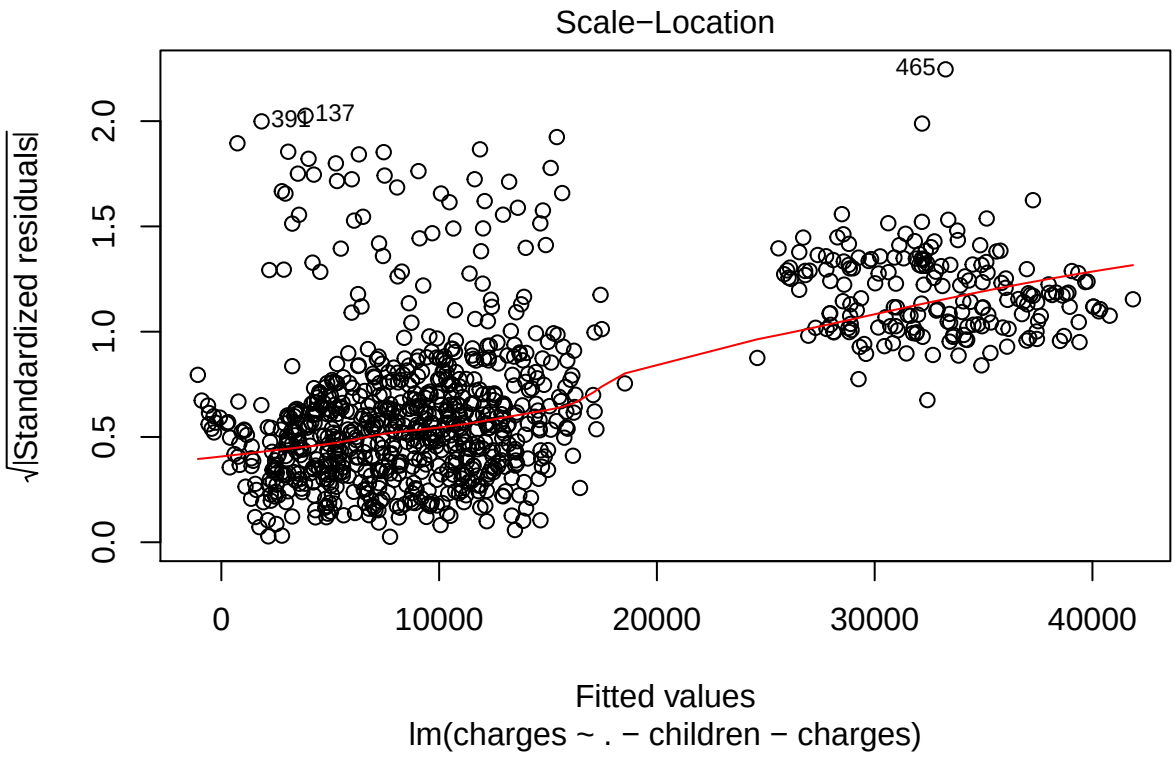
The data will be split into train and test sets using a 70 / 30 ratio. Cross validation is performed on the train set (70/30), then the test set will be used to assess the performance of the model.

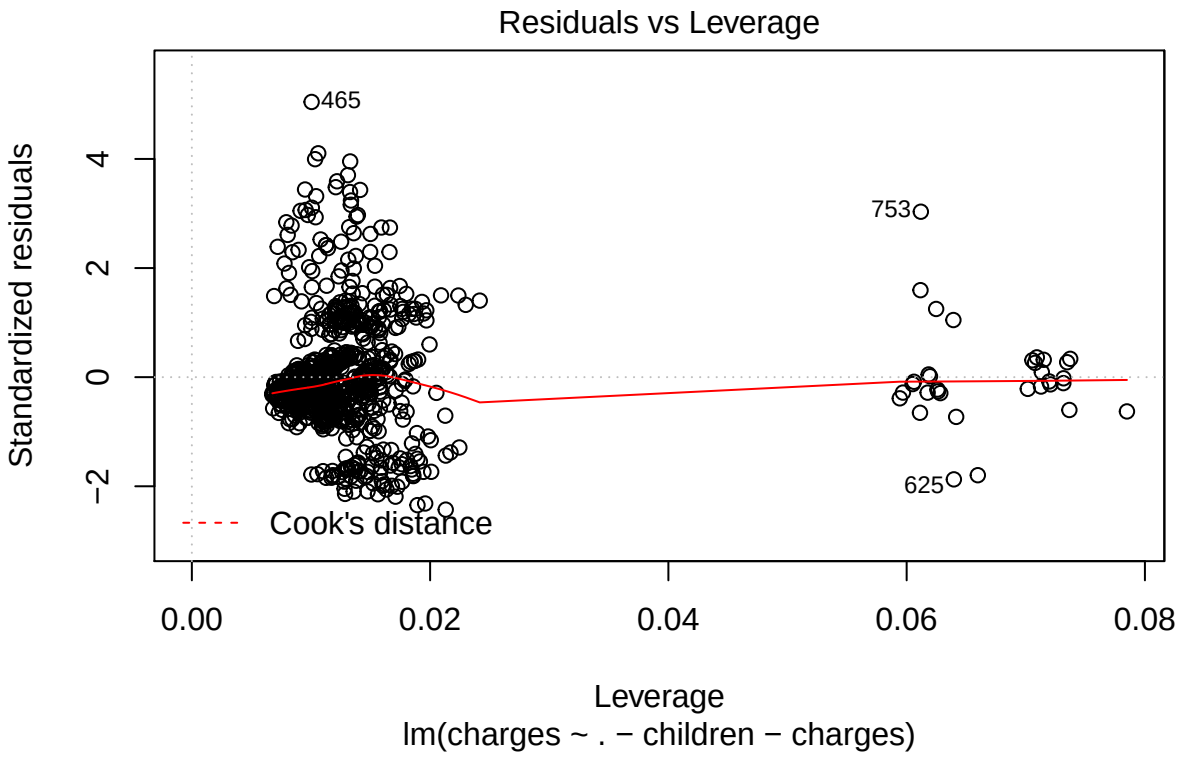
Model

```
Call:
lm(formula = charges ~ . - children - charges, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-14039.8  -2517.8   -822.7   1546.6  29337.1

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9101.54    1161.80  -7.834 1.30e-14 ***
age               241.16      13.83  17.436  < 2e-16 ***
sexmale           -13.17     386.42  -0.034  0.97282
smokeryes       24392.10     481.84  50.623  < 2e-16 ***
regionnorthwest  -838.04     560.75  -1.495  0.13538
regionsoutheast -1218.66     550.66  -2.213  0.02714 *
regionsouthwest -1493.82     555.91  -2.687  0.00734 **
children_fct1    -330.82     485.87  -0.681  0.49611
children_fct2    2208.49     549.32   4.020 6.28e-05 ***
children_fct3     622.69     635.66   0.980  0.32755
children_fct4    2181.29    1411.28   1.546  0.12254
children_fct5    1046.76    1547.26   0.677  0.49888
bmi               274.86      33.01   8.327 2.98e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5845 on 923 degrees of freedom
Multiple R-squared:  0.7696,    Adjusted R-squared:  0.7666
F-statistic:   257 on 12 and 923 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
30000
10000
0
−10000
391
137
465
0
10000
20000
30000
40000
Fitted values
lm(charges ~ . − children − charges)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ . − children − charges)
465
137
391
−2
0
2
4
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
10000
20000
30000
40000
465
391
137
Fitted values
lm(charges ~ . − children − charges)



The diagnostic plots show a lot of problems with our model. The residual vs fitted values plot show that our residuals are not linear as the residuals are clustered in groups. In addition, the QQ plot shows that our residuals are not approximately normally distributed since the values stray from the diagonal line as it goes further to the right. The scale-location plot also shows non-constant variance. The residuals get bigger as the fitted values get bigger. There are outliers (Some residuals are 4 standard deviations away from 0) in our model though no leverage or high influential points as demonstrated by no visible Cook's distance line.

The summary confirms our earlier analysis. Sex does not play an important role in determining the insurance costs of a contractor as shown by their large p-values. On the other hand, age, bmi and smoker are important factors. Region and children, however, show statistical significance on some levels.

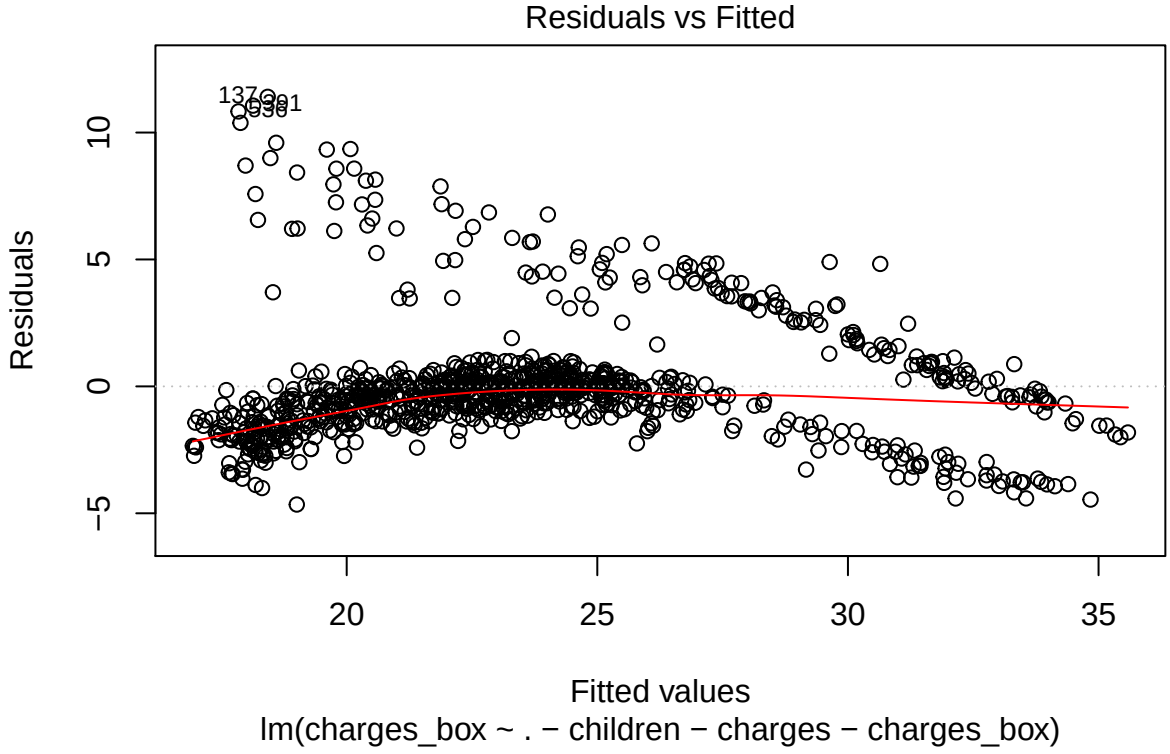
Addressing Violation From the initial diagnosis, the linear regression model built above violates the linearity, normal distribution and constant variance assumptions. A few ways to address these problems are presented below. ##### Box-Cox transformation

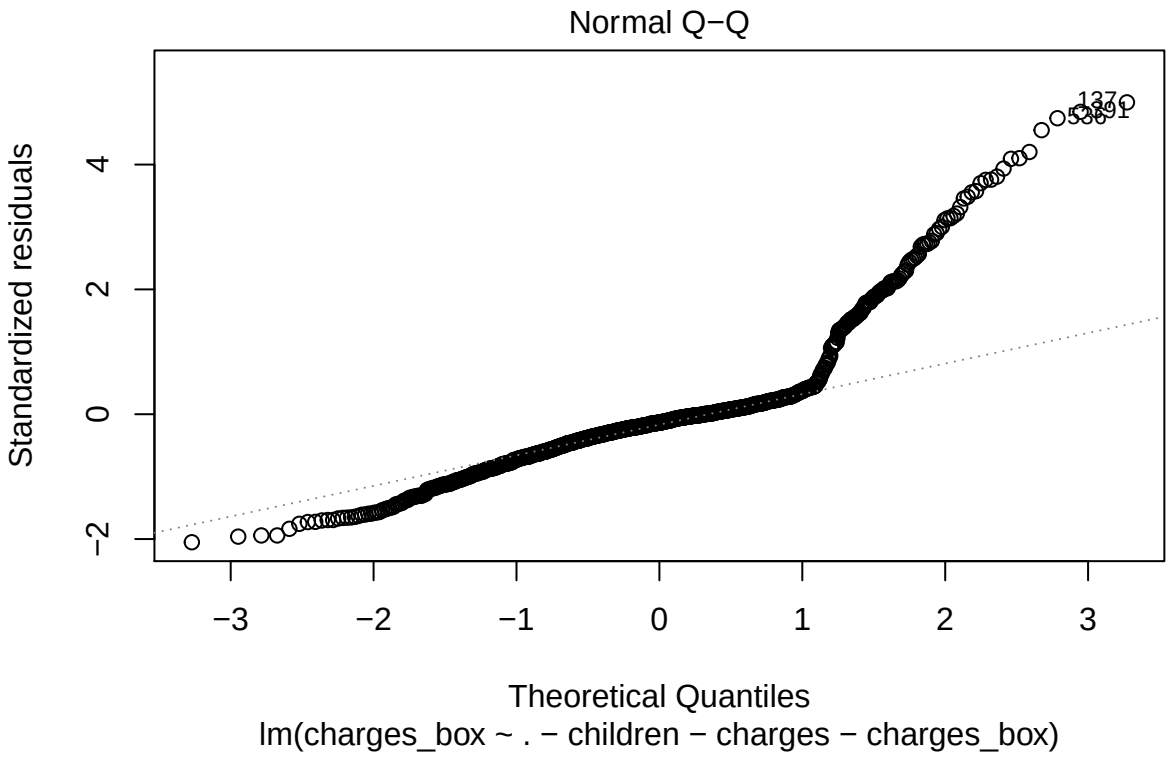
Box-Cox transformation is one way to address non linearity nad non-normality. The transformation is applied to both the response and predictors where appropriate. The idea is to find a lambda value that maximizes the log-likelihood of the data. The transformation formula is below: $y_i^\lambda = \frac{y_i^\lambda - 1}{\lambda} if\ \lambda \neq 0\ else\ log(y_i)$

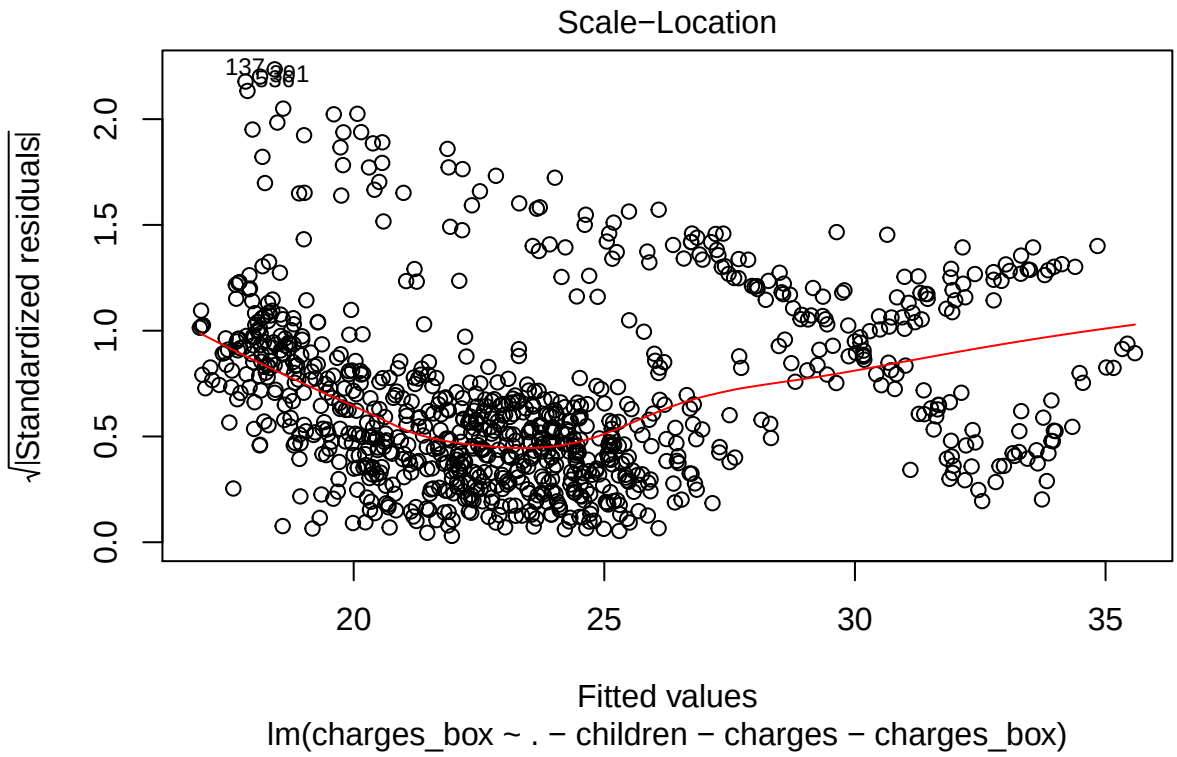
Transformation of the Response (charges)

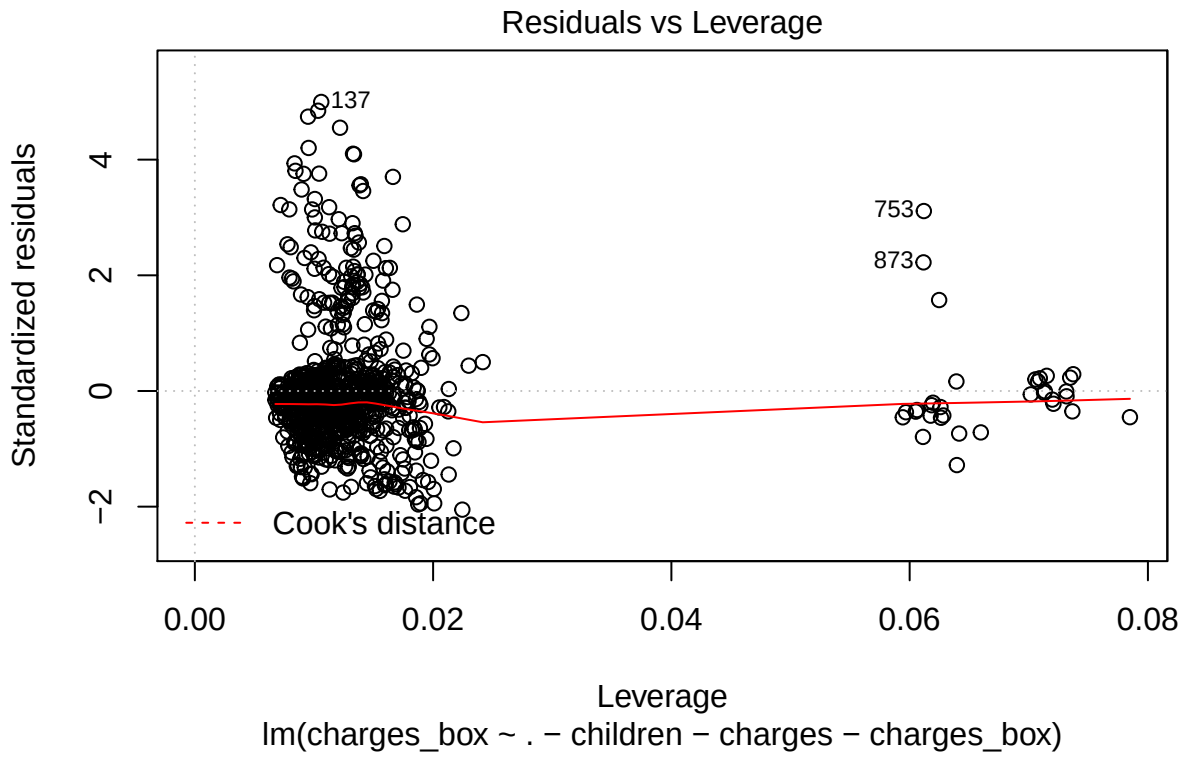
```
Call:
lm(formula = charges_box ~ . - children - charges - charges_box,
    data = train_data1)

Residuals:
```

```
    Min      1Q  Median      3Q     Max
-4.6542 -1.1396 -0.3081  0.3638 11.4065

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     13.57832    0.45611  29.770  < 2e-16 ***
age              0.16404    0.00543  30.211  < 2e-16 ***
sexmale         -0.19211    0.15171  -1.266 0.205720
smokeryes        8.77037    0.18917  46.363  < 2e-16 ***
regionnorthwest -0.56005    0.22014  -2.544 0.011121 *
regionsoutheast -0.97345    0.21618  -4.503 7.56e-06 ***
regionsouthwest -0.92839    0.21824  -4.254 2.31e-05 ***
children_fct1    0.31220    0.19075   1.637 0.102026
children_fct2    1.37302    0.21566   6.367 3.04e-10 ***
children_fct3    0.96183    0.24955   3.854 0.000124 ***
children_fct4    2.29867    0.55405   4.149 3.65e-05 ***
children_fct5    1.76913    0.60744   2.912 0.003673 **
bmi              0.06850    0.01296   5.286 1.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.295 on 923 degrees of freedom
Multiple R-squared:  0.7796,    Adjusted R-squared:  0.7767
F-statistic:   272 on 12 and 923 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges_box ~ . − children − charges − charges_box)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
20
25
30
35
Fitted values
lm(charges_box ~ . − children − charges − charges_box)



The diagnostic plots again show improvements in our model. Residuals vs Fitted plot shows a curved pattern though the red line is much aligned with the 0 line. The residuals are also scattered around the 0 line more symmetrically. The QQ plot still shows non-normality while the scale-location plot displays improvement but the variance still increases as fitted values get bigger..

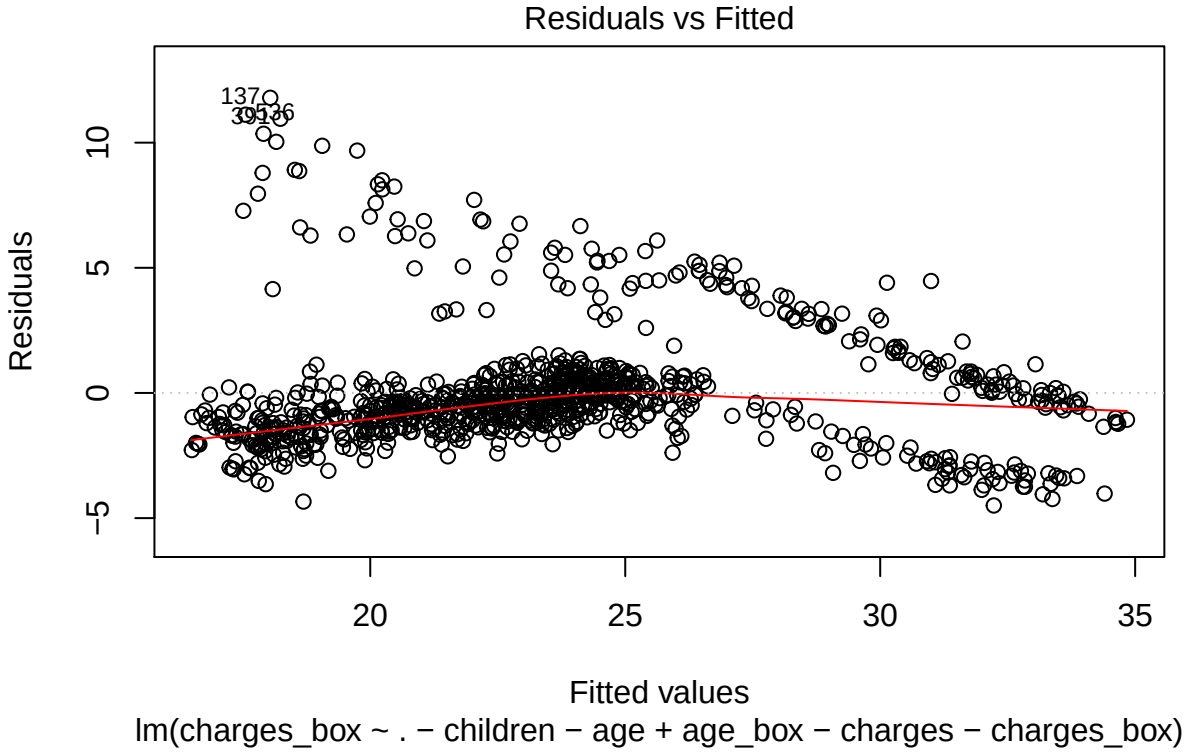
The summary shows an improvement from the non-transformed model as demonstrated by a jump of R-squared adjusted values from 0.7653 to 0.7868. Aside from age, smoker, bmi, most other variables have also become statistically significant (p-values < 0.05).

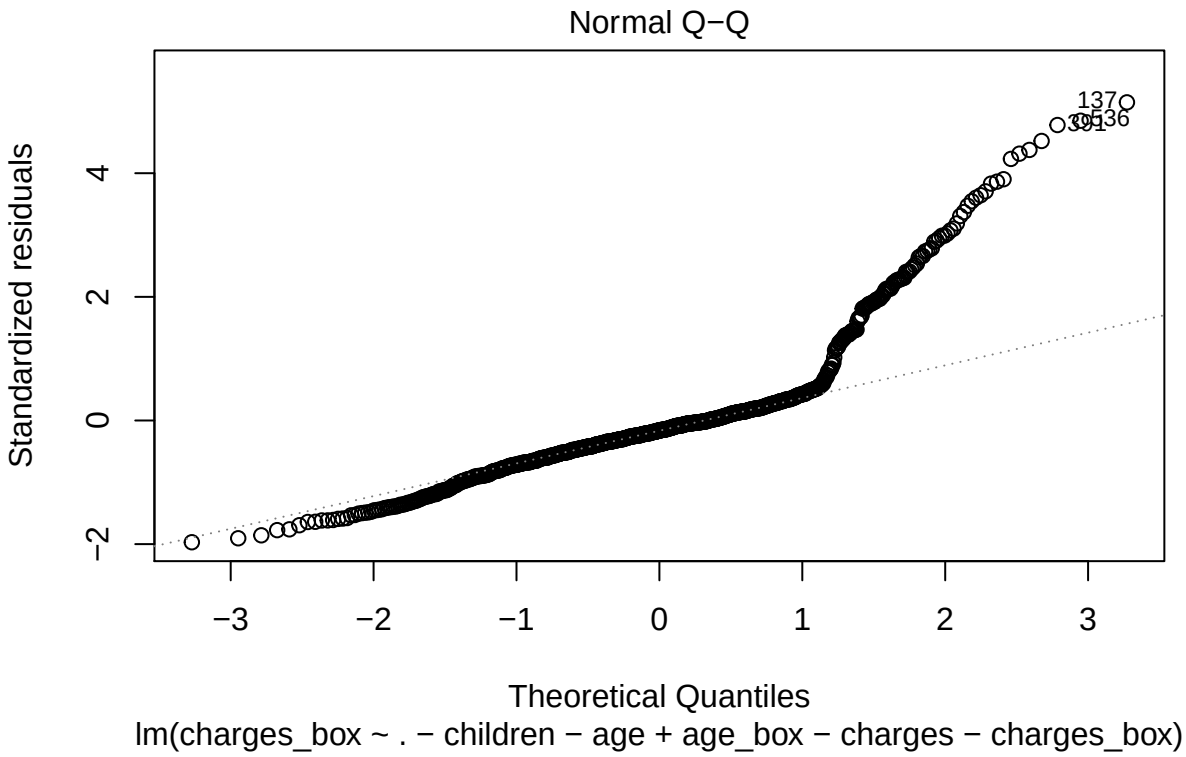
Transformation of Age

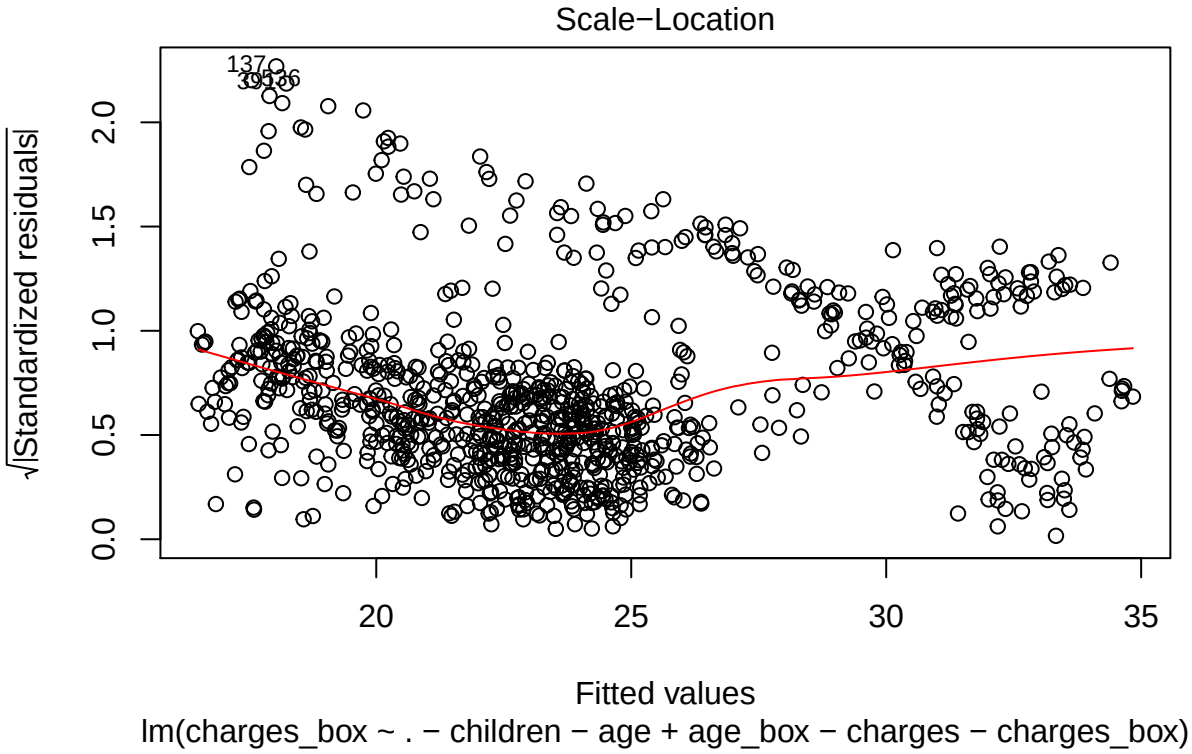
In the previous analysis, the scatter plots of age vs charges show non-linearity pattern as age is clustered in 3 groups. It is reasonable to apply a transformation to age.

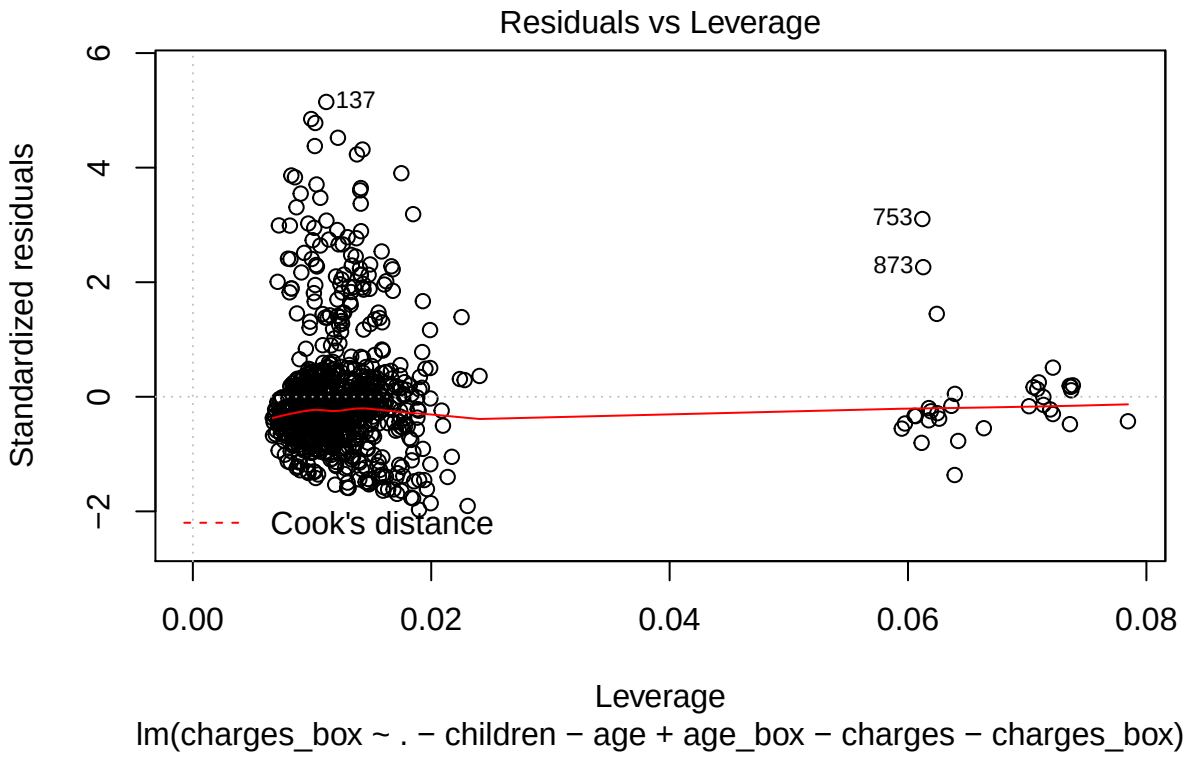
```
Call:
lm(formula = charges_box ~ . - children - age + age_box - charges -
    charges_box, data = train_data2)

Residuals:
    Min      1Q  Median      3Q     Max
-4.4951 -1.1967 -0.3689  0.4387 11.7920

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.20892    0.64299   6.546 9.80e-11 ***
sexmale        -0.19321    0.15233  -1.268 0.204980
smokeryes       8.76425    0.18994  46.142  < 2e-16 ***
```

```
regionnorthwest -0.53226    0.22104  -2.408 0.016237 *
regionsoutheast -0.95742    0.21709  -4.410 1.15e-05 ***
regionsouthwest -0.90991    0.21915  -4.152 3.60e-05 ***
children_fct1    0.09672    0.19194   0.504 0.614456
children_fct2    1.11429    0.21693   5.137 3.41e-07 ***
children_fct3    0.77221    0.25124   3.074 0.002177 **
children_fct4    2.05264    0.55623   3.690 0.000237 ***
children_fct5    1.52497    0.60972   2.501 0.012553 *
bmi              0.07318    0.01300   5.631 2.38e-08 ***
age_box          3.08870    0.10309  29.962  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.304 on 923 degrees of freedom
Multiple R-squared:  0.7778,    Adjusted R-squared:  0.7749
F-statistic: 269.2 on 12 and 923 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges_box ~ . − children − age + age_box − charges − charges_box)
137
536
−3
−2
−1
0
1
2
3
−2
0
2
4


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
20
25
30
35
Fitted values
lm(charges_box ~ . − children − age + age_box − charges − charges_box)



The summary and diagnostic plots show similar results as compared to the model with only the response variable transformed. In order to ensure the models do not overfit. Cross validation will be run on 3 models to assess their performances.

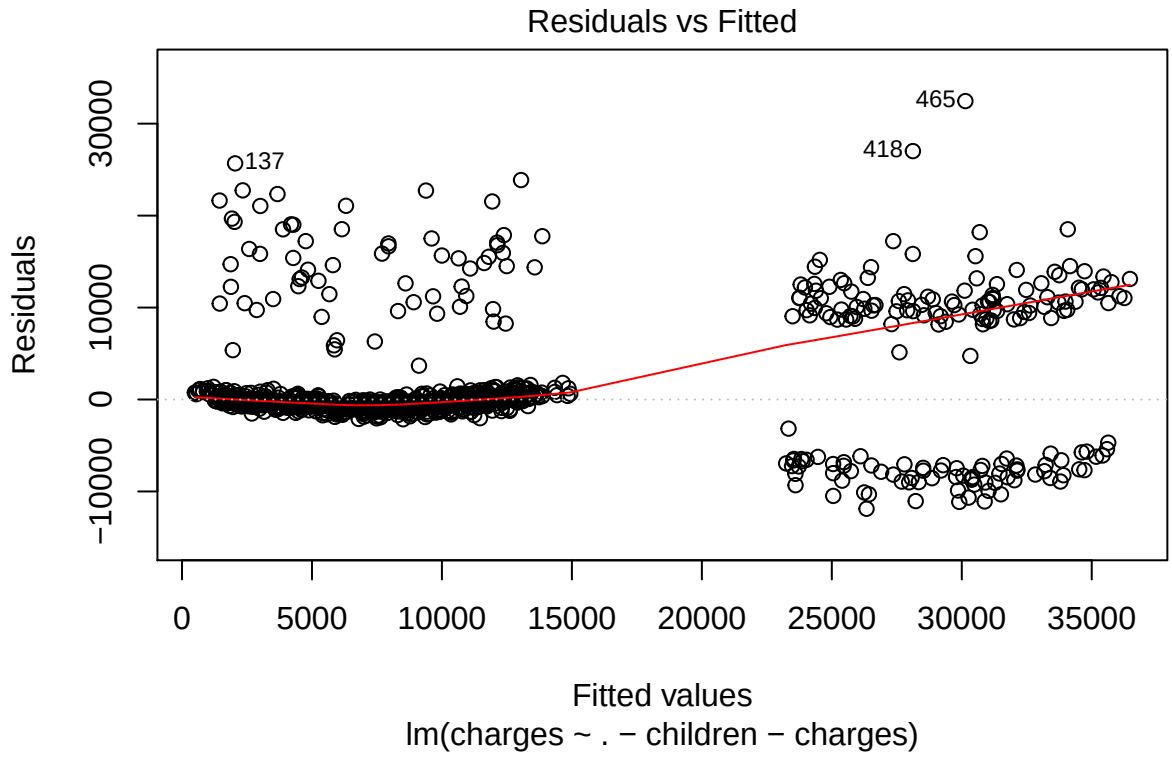
5 Fold Cross Validation The table below shows the results of three models using cross validation. Out of three models, the untransformed model perferms the best with the biggest R2 and smallest RMSE . Clearly, our transformation overfits the data.

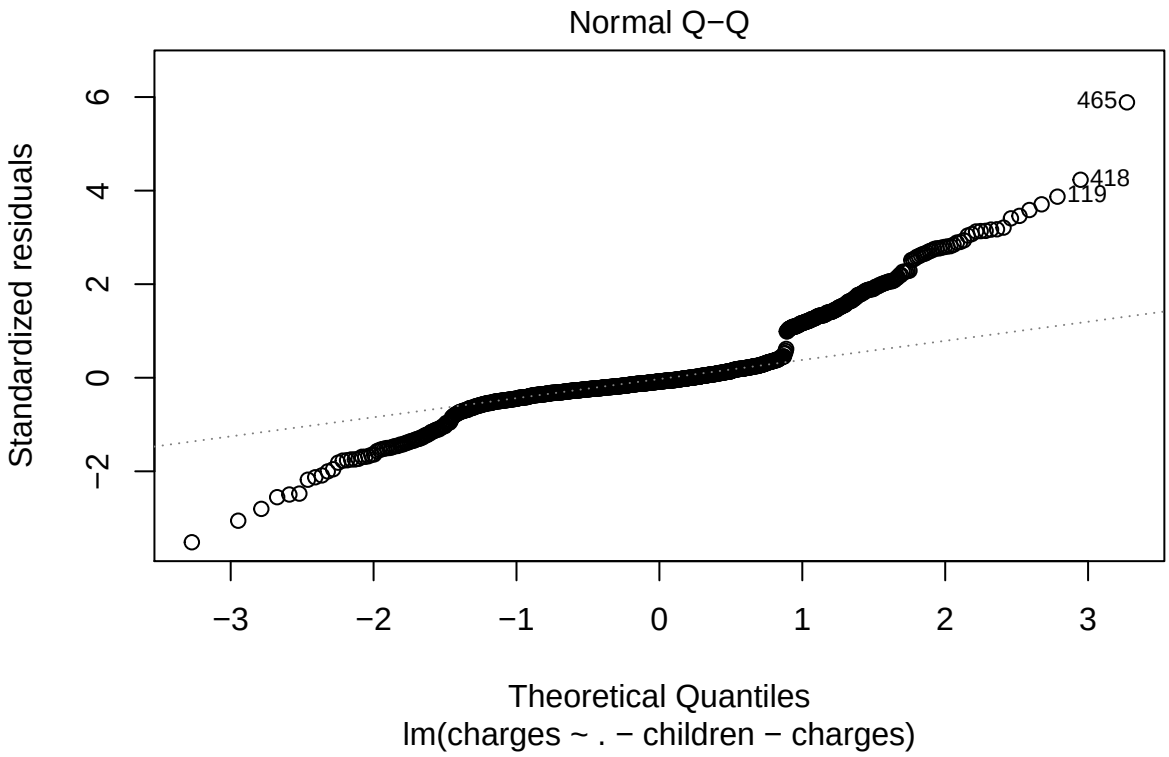
```
             full      fullbox  fullboxwage
R2      0.7007255    0.6829274    0.6837907
RMSE 6651.9781793 7085.1301643 7035.0548614
```

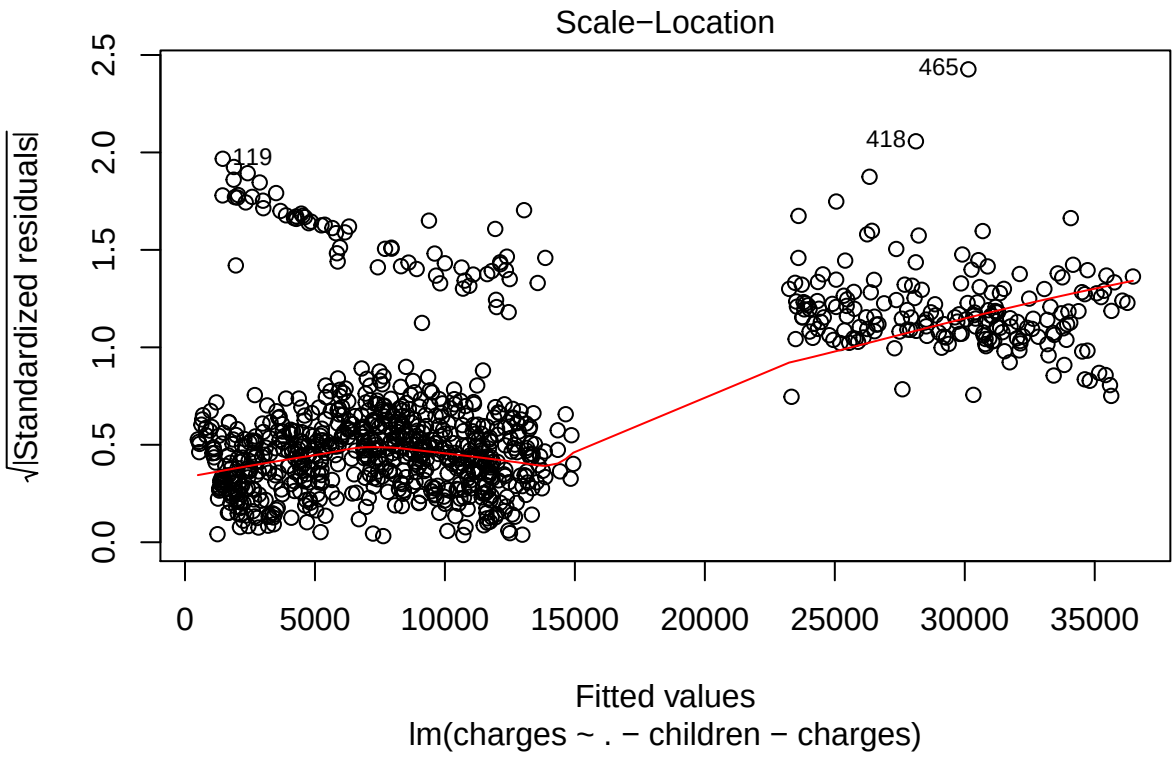
Weighted Least Square (WLS) As discussed above, the non constant variance assumption was violated. One way to address this issue is to put more weights on the non-outliers. Below is fitted model with weighted least square. From now on, when referring to the base model, the model at hand is the first model (non transformed one)

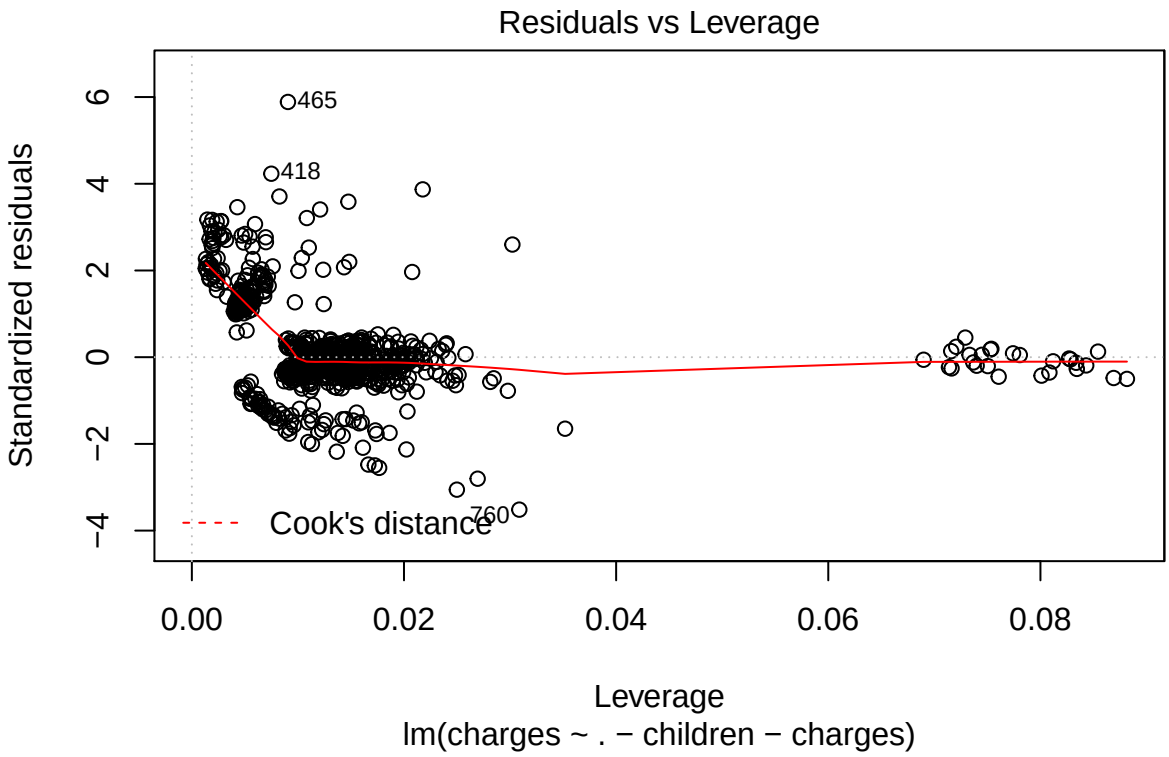
```
Call:
lm(formula = charges ~ . - children - charges, data = train_data,
    weights = 1/sighat)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-262.27  -22.83   -6.03   18.63  444.02
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4726.487    598.979  -7.891 8.47e-15 ***
age               255.355      7.492  34.083  < 2e-16 ***
sexmale          -256.407    201.707  -1.271 0.203983
smokeryes       22185.953    492.113  45.083  < 2e-16 ***
regionnorthwest  -306.554    294.723  -1.040 0.298545
regionsoutheast  -725.664    296.706  -2.446 0.014642 *
regionsouthwest  -790.717    290.835  -2.719 0.006675 **
children_fct1     120.244    250.114   0.481 0.630804
children_fct2    1007.446    296.403   3.399 0.000706 ***
children_fct3    1184.519    341.607   3.467 0.000550 ***
children_fct4    2119.571    750.886   2.823 0.004864 **
children_fct5    2140.908    714.089   2.998 0.002790 **
bmi                71.499     17.100   4.181 3.17e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 75.77 on 923 degrees of freedom
Multiple R-squared:  0.7765,    Adjusted R-squared:  0.7736
F-statistic: 267.2 on 12 and 923 DF,  p-value: < 2.2e-16
```




Normal Q−Q
Standardized residuals
6
4
2
0
−2
465
418
119
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(charges ~ . − children − charges)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
119
418
465
0
5000
10000
15000
20000
25000
30000
35000
Fitted values
lm(charges ~ . − children − charges)



The model diagnostic plots and summary are pretty similar to the base. The adjusted R-squared is a little bit higher than the non-WLS one. Below is the cv results of the WLS model.

```
             full      wlm4_cv
R2      0.7007255    0.6869969
RMSE 6651.9781793 6928.6263306
```

The base model has higher R2 score and lower RMSE scores which signals the base one is still the best model so far.

Variable Selection

Adding non-linear term From the scatterplots, age displays behaviour that are not strictly linear. It follows that improvements could be made by adding non-linear terms to the model.

```
             full      age2_cv
R2      0.7007255    0.7045399
RMSE 6651.9781793 6610.7094775
```

This time the model with non-linear term in age performs slightly better than the base model. Next, variables that are not statistically significant in the base model will be removed to see if improvements could be made.

```
             full    no_sex_cv   no_chld_cv    no_reg_cv       red_cv
R2      0.7007255    0.7007007    0.7069823    0.7029727    0.7084115
RMSE 6651.9781793 6652.2254304 6576.5634766 6624.1697360 6558.9003454
```

From left to right, respectively, the models are base one, model without sex variable, model without children variable, model without region variable and model without sex, children and region variables. The best model is the one without region variable as it has the biggest R2 and smallest RMSE. Although the model without the region variable only performs the best, the model without sex, children and region performs the second best. Their results are only slightly different. Since parsimonious model is the goal of building a model. The red_cv model should be chosen as the next base model. We know from previous analysis that adding non-linear terms led to better performance. Below shows the cv results when removing region variable from the non-linear model.

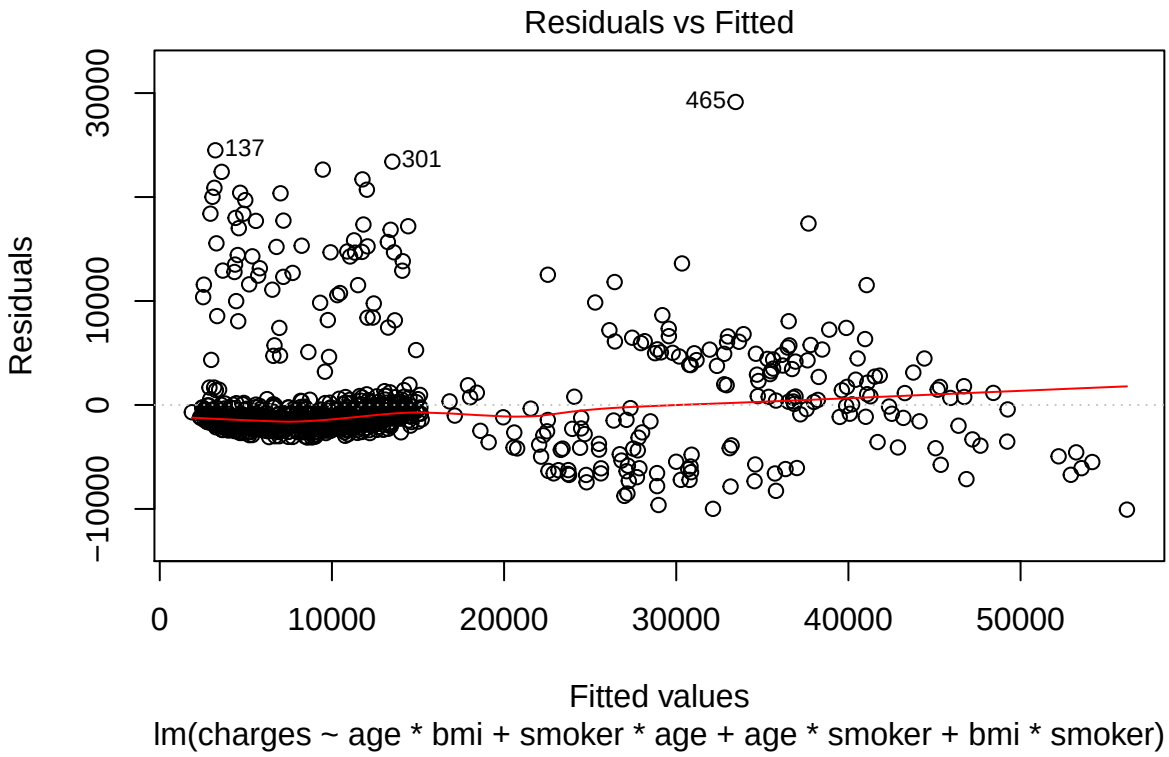
```
             full       red_cv age2_noreg_cv
R2      0.7007255    0.7084115     0.7102093
RMSE 6651.9781793 6558.9003454  6538.0669398
```

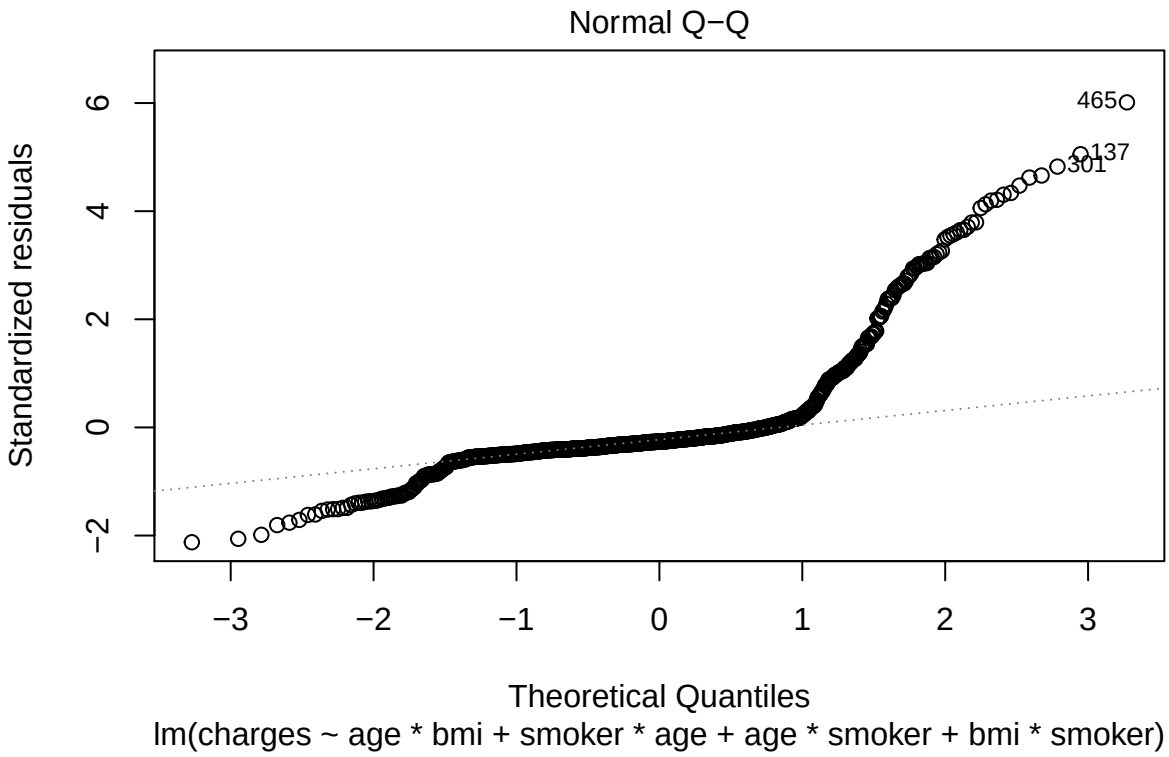
Although the performance is lightly better for the model with non-linear term, the base model is still preferred as it is most parsimonious and performs almost as well as the one with non-linear term. Next is the results of a model fitted with interactive terms.

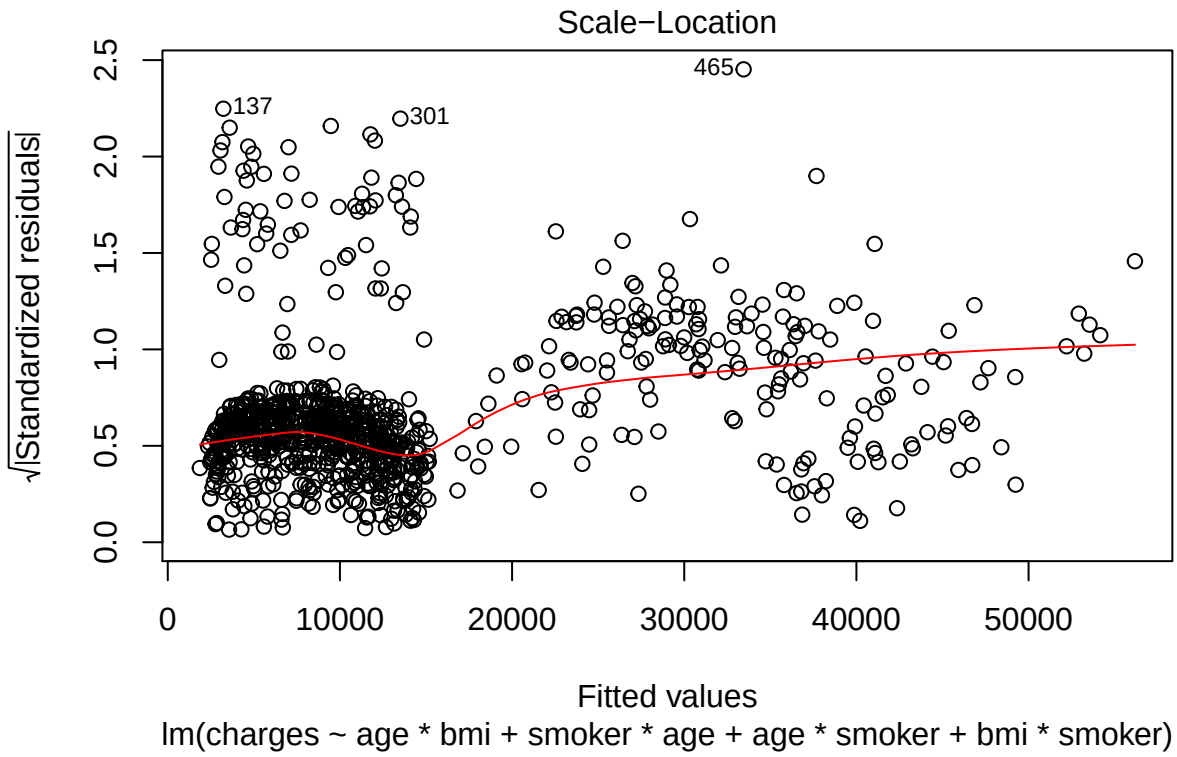
```
             full age2_noreg_cv age2_inter_cv
R2      0.7007255     0.7102093     0.8285419
RMSE 6651.9781793  6538.0669398  5028.3938277
```

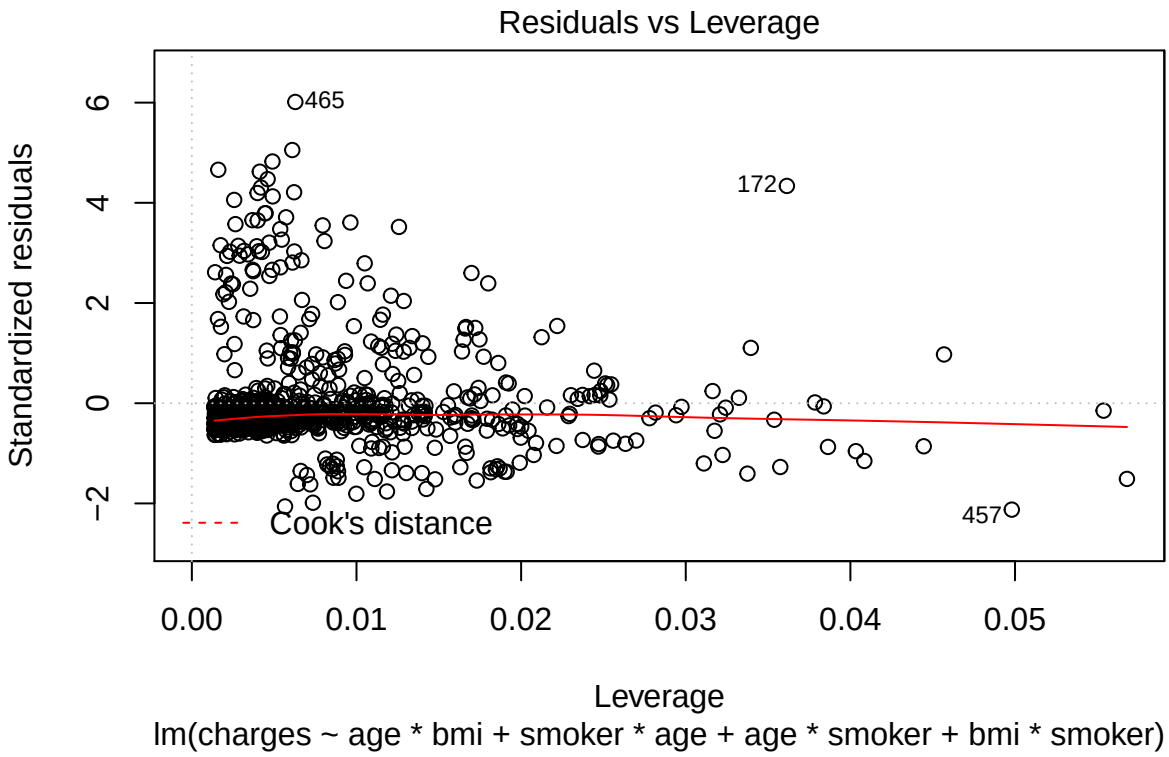
The model performance improves significantly in comparison with the base models. R-square value jumps from ~0.71 to ~0.803 while RMSE decreases from ~6267 to ~5121. Below is the model summary.

```
Call:
lm(formula = charges ~ age * bmi + smoker * age + age * smoker +
    bmi * smoker, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10065.9  -1977.7  -1284.7   -212.9  29144.5

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1196.868   2442.144   0.490  0.62419
age              175.578     60.177   2.918  0.00361 **
bmi              -92.211     77.769  -1.186  0.23605
smokeryes     -20535.842   2412.306  -8.513  < 2e-16 ***
age:bmi            2.515      1.882   1.337  0.18170
age:smokeryes     11.788     28.430   0.415  0.67851
bmi:smokeryes   1440.046     68.183  21.120  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4862 on 929 degrees of freedom
Multiple R-squared:  0.8396,    Adjusted R-squared:  0.8385
F-statistic: 810.3 on 6 and 929 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
30000
10000
0
−10000
137
301
465
0
10000
20000
30000
40000
50000
Fitted values
lm(charges ~ age * bmi + smoker * age + age * smoker + bmi * smoker)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(charges ~ age * bmi + smoker * age + age * smoker + bmi * smoker)
465
137
301
−2
0
2
4
6
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
137
301
465
0
10000
20000
30000
40000
50000
Fitted values
lm(charges ~ age * bmi + smoker * age + age * smoker + bmi * smoker)



The diagnostic plots show improvements accross assumptions for our model, though these improvements still violate the assumptions.

Once the interactive terms were added to the model, most predictors become statistically insignificant. Smoking and age^2 are the only two statistically significant predictors. Since keeping age^2 while discarding age would make no sense in terms of interpretability, we'll eliminate age^2 from the model. Below is the model with only smoker as the predictor.

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

Residuals:
   Min     1Q Median     3Q    Max
-18490  -4769   -798   3615  29820

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8329.9      258.9   32.17   <2e-16 ***
smokeryes    24443.1      576.2   42.42   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7076 on 934 degrees of freedom
Multiple R-squared:  0.6583,    Adjusted R-squared:  0.658
F-statistic:  1800 on 1 and 934 DF,  p-value: < 2.2e-16

          R2         RMSE
```

```
0.5364007 8336.4479265
```

Modeling with only smoker as the predictor sees a significant drop from around 0.80 for R2 values and a huge increase in RMSE. These results confirm that our best model is the base model with interactive terms.

Conclusion

In this analysis, a linear model with all the variables was fitted on the insurance data. The model shows a lot of violations of linear regression assumptions. Non-linearity, non-norammlly distributed

A few approaches were taken to address the issue. Box-Cox transformations were performed on both the response and age variables. The diagnostic plots show no apparent fix of the non-linear assumption. Weighted-Least square method was used to fit on the data to address non-constant variance. The diagnostic plots show an improvement over the base one. However, the performance of the model decreases as showcased by smaller R-squared and RMSE values for both methods.

Once statistically insignificant predictors and non-linear terms were added to the model, there was a big increase in the model's performance. Nonetheless, the increase in performance comes with the cost of less interpretability. Most of the predictors become non-statistically significant. Only smoking and agˆ2 are significant. The model confirms that smoking severely affects the insurance costs of an individual.