Cepehr Alizadeh (ca3eh)
Seth Harrison (wdv7nu)
Said Mrad (sm2yk)
Max Ryoo (hr2ee)

STAT 6021 Final Project: Regression on Insurance Charges in America

Executive Summary

Medical insurance has become one of the most widely utilized forms of insurance, with 92% of Americans being either partly or fully covered in the year 2019. Because of this, it is important that people have an idea of the level of charges they will incur when utilizing their insurance. Our goal was to predict and see variables that affect the insurance charges for individuals.

Our dataset consists of seven variables, six of which will be used for regression analysis for determining whether a person will be charged below or above the average person for insurance and the amount they will be charged. These six variables include the given person's age, sex, BMI, number of children, smoker status, and location. Using these variables we developed multiple linear and logistic regression models for predicting insurance charge amounts and whether the individual would fall above or below the average, respectively. It is important to note that there may have been variables not included in the provided dataset that may have been impacting both models, which we made attempts to deal with and will discuss in a later section.

EDA

We conducted some preliminary exploratory data analysis (EDA) to see any trends or interesting observations about the data.

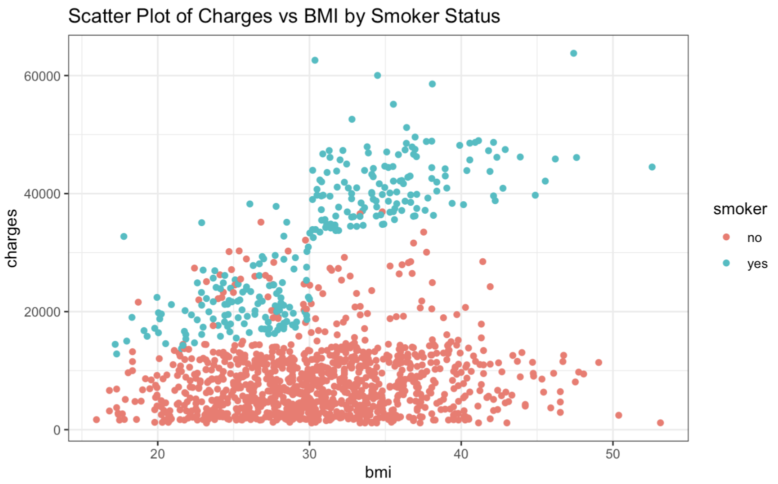


Figure 1

While looking at a scatterplot between charges and bmi, we noticed two different clusters, analyzing further we deduced that the clusters represented smoker status. When adding smokers as a color indication, the group observed that for smokers there was a strong positive linear relationship between BMI and charges, while the non smokers didn't seem to have the same strength as smokers.. This led the group to better understand the impact smoker status had on our response and the stark difference that will exist between the two groups. Therefore, the group chose to separate regressions for each level of smoker status.

When visualizing the relationship between ages and charges, it became clear that three separate relationships existed and while each had a positive linear association there were stark differences in the charges amongst them. To better understand the differences, the group included smoker status and BMI to see if they helped explain the associations seen.

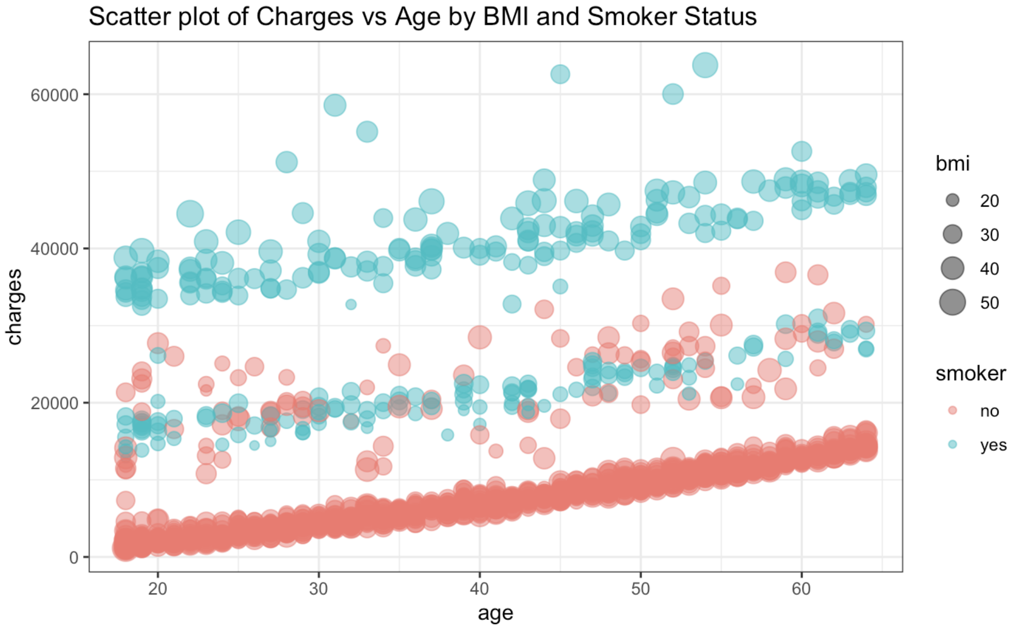


Figure 2

The grouping that had the lowest level of charges consisted of all non-smokers and mostly those with lower BMI. The group with the highest level of charges consisted of all smokers and mostly those with higher BMI. The grouping that fell somewhat in the middle had both smokers and non-smokers as well as individuals with varying BMIs.

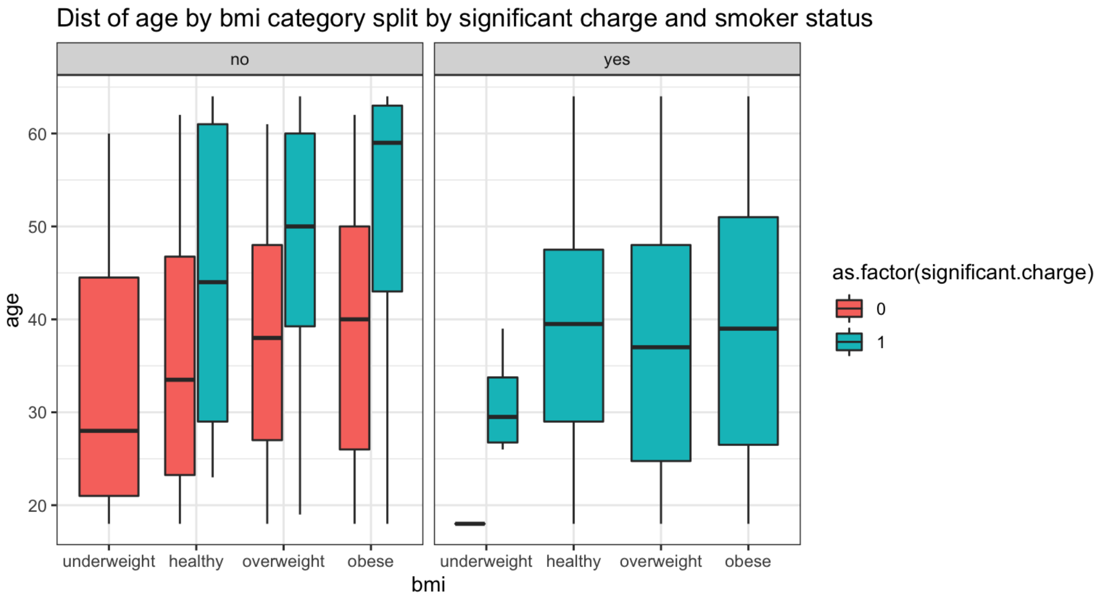


Figure 3

The group created a boxplot to showcase the distribution of age by BMI category split by significant charge and smoker status. As shown in Figure 3, for non-smokers it could be observed that significant charges occurred when the age was older, which can be observed from the increasing median age for each boxplot in the BMI category. Meanwhile, for the smokers regardless of age or bmi, the charges filled were significant, which gives us insight that smoking status plays a major role in determining a significant charge or non significant charge.

The exploratory plots that were generated helped us better understand the underlying relationships between our predictors as well as their impact on our response variable of interest. Moving forward, the group will use these visualizations to aid the framework of our regression analysis.

Multiple Linear Regression

From the exploratory data analysis, our first thought was to create a full model to generalize the trends in the data for the response variable (charges). The full model had the predictors age, bmi, children, smoker, and region. The following multilinear regression model was constructed from utilizing all the predictors.

$$\hat{y} =- 11938.5 + 256.9age - 131.3I_1 + 339.2bmi + 475.5children + 23848.5I_2 - 353.0I_3 - 1035.0I_4 - 960.0I_5$$

For this linear regression I_1 indicates whether the sex of the client is male. The value will be 0 for females. I_2 indicates whether a client smokes. The value will be 0 for non-smokers. I_3 indicates that the client is in the northwest region. I_4 indicates that the client is located in the southeast. I_5 indicates that the client is located in the southwest. If the client is in the northeast I_3, I_4, I_5 will be zero, since this is the reference class for our model.

The above model had a multiple R-squared of 0.7509, Adjusted R-squared of 0.7496, and a p-value of 2.2e-16, which means that the model performed well in measuring the costs given all the predictors in the model. However, for multiple linear regression we must check the assumptions first to validate our model. First we constructed a residual plot for the model as shown in Figure 4.

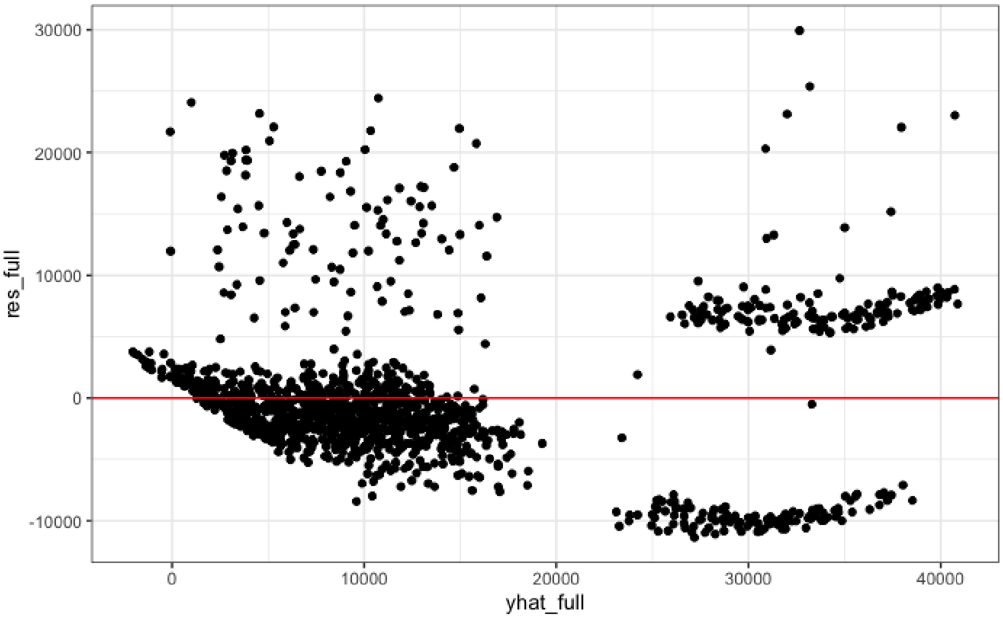


Figure 4

From the residual model we can clearly see that there are groupings happening as well as the variance not being held constant throughout the residuals thus validating the constant variance assumption of multiple linear regression. In order to combat this we decided to transform the response variable (charges) through a lambda value given from the boxcox plot of this model. The lambda value utilized was 0.15 as shown in Figure 5.

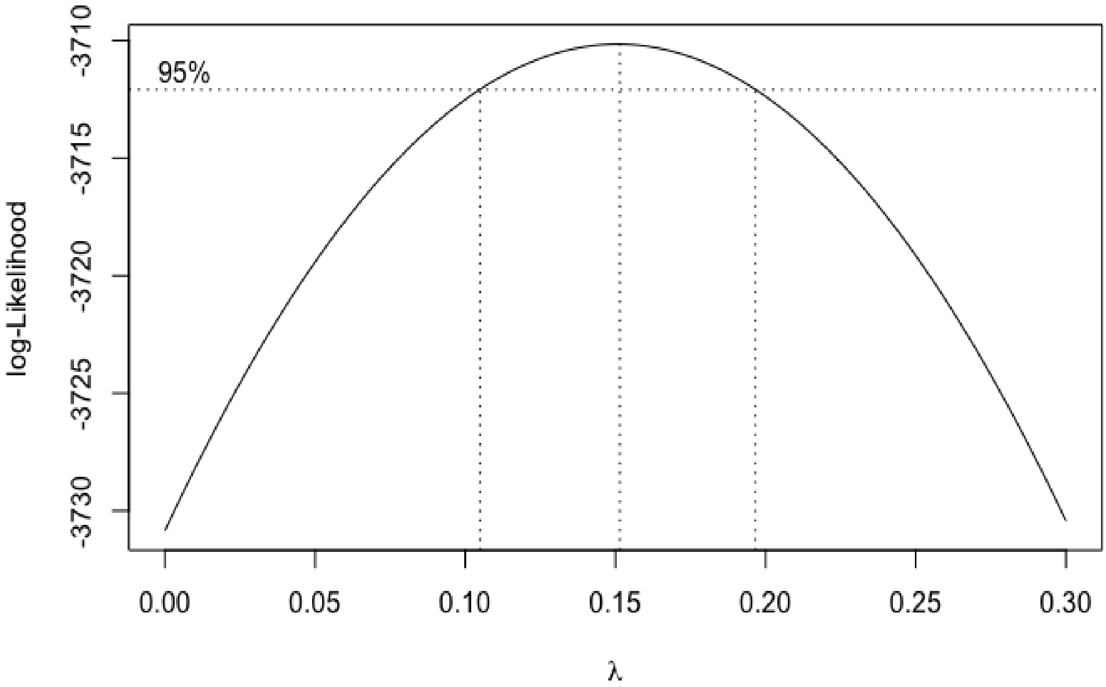


Figure 5

We reconstructed the residual plot with a new model that utilized the transformed dataset, which is shown below in Figure 6.

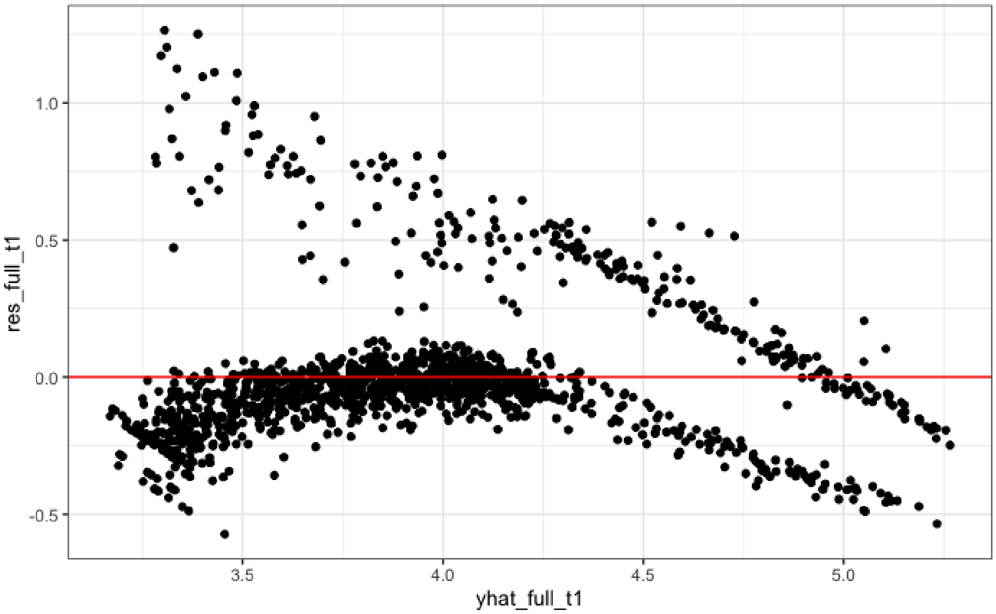


Figure 6

However, even with this transformation, we were still violating the constant variance assumption. With this new finding, we constructed the boxcox plot to check if we would need to do any more transformation, which yielded the figure 7 below.

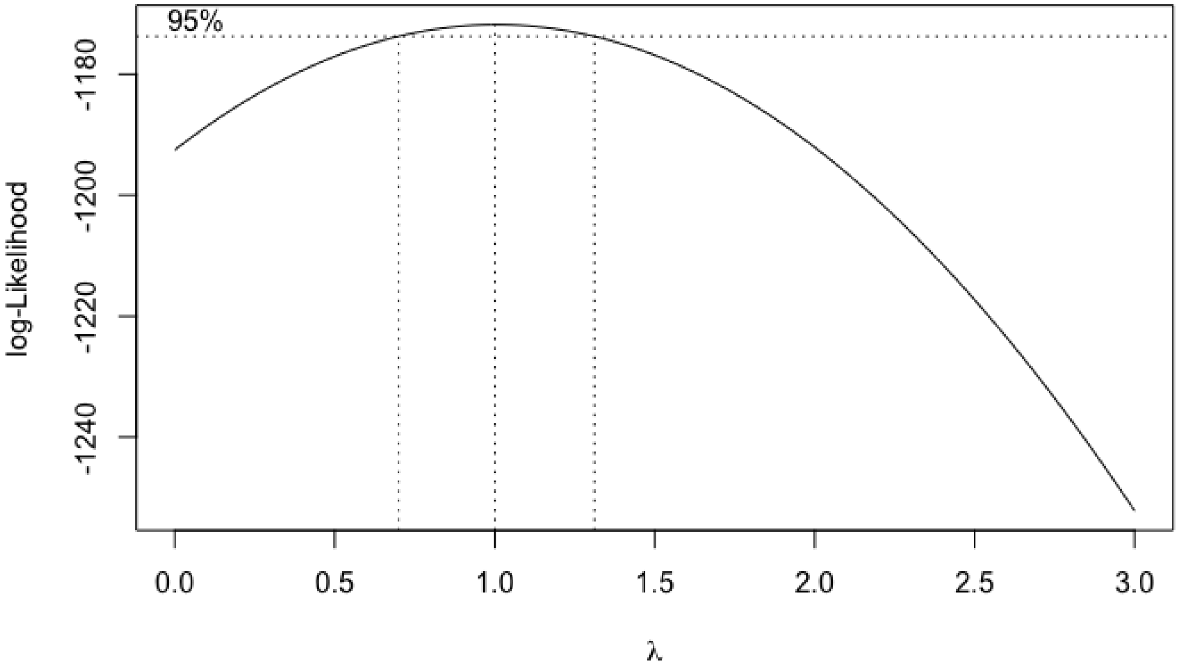


Figure 7

We clearly saw some patterns even in the transformed linear model. In order to find the indicator that contributed to these violations, we tried many combinations of transformations such as having region as an interaction with age + bmi, separating bmi into categorical groupings, and selection and dropping multiple predictors.

Since we were still violating the assumptions, we decided to look more in depth at the first residual plot for the model with no transformations and all the predictors. We saw that there is a slight divide in the plot for charges that are less than 20,000 and greater than 20,000. From this observation, we decided to do more in depth exploratory data analysis on the data by dividing the data into two subsets with the first subset including only the rows with lower than 20,000 charges.

When separating the data set into two subsets, we first looked at the relationship between charges and BMI by smoker status as shown in figure 8.

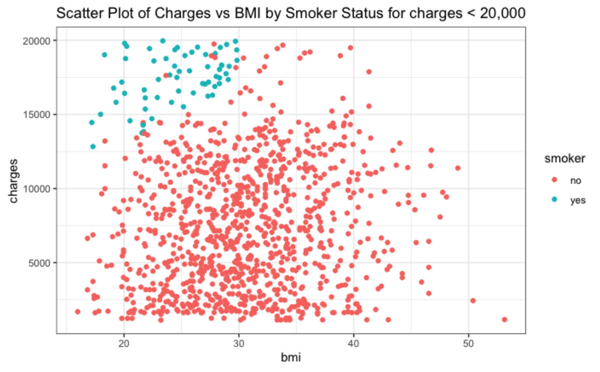
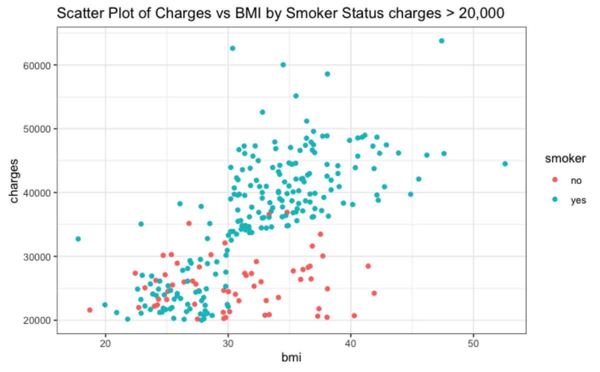
Charges < 20,000	Charges > 20,000
	

Figure 8

From the figure above, it is very clear that the for charges less than 20,000 the population was dominated by non smokers. Also for charges above 20,000 the population was dominated by mostly smokers. We also constructed a scatter plot for Charge by Age by smoker status for both the subsets of data shown in figure 9.

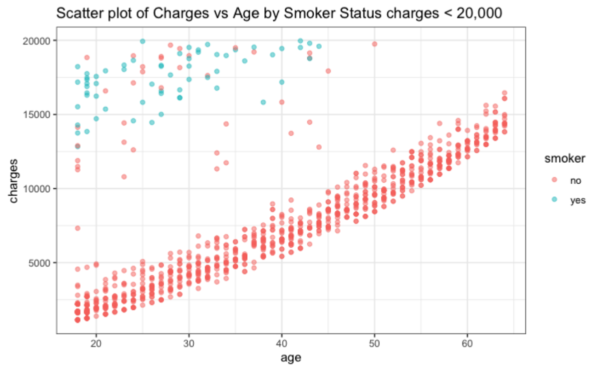
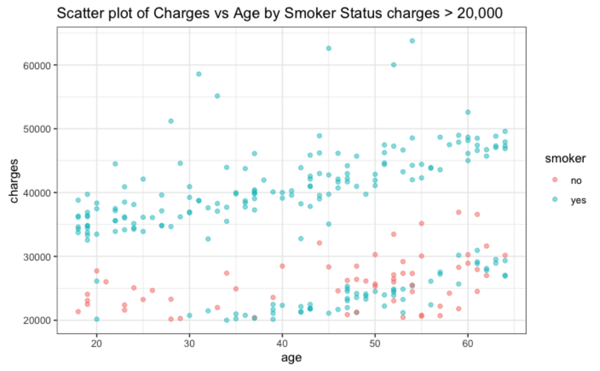
Charges < 20,000	Charges > 20,000
	

Figure 9

Similar to what was observed in charges vs bmi, we saw a clear distinction between the two subsets. Also, we were able to see a positive relationship between age and charges for charges < 20,000 for non smokers as well as a positive relationship between age and charges for charges >20,000 for smokers.

From this new finding, our decision was to look at the dataset as two datasets separated by smoker status. From the more in depth exploratory data analysis, we found that the trend for non smokers was different than the smokers. This could mean that when looking at clients, the insurance companies will first consider the client's smoking status and use different indicators to charge their client for a certain plan.

Smokers

In order to find a model to represent these two datasets well, we decided to first use a selection method for the two datasets. We decided to use a forward selection method. For the smokers the predictors were only bmi and age with the AIC value of 4747.41. The two predictors were actually observed by the following EDA plots we created for smokers as shown in Figure 10.

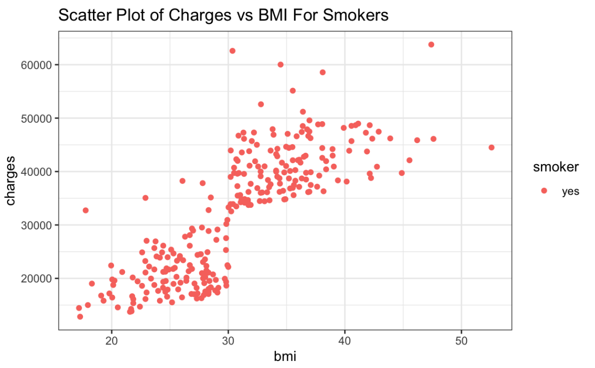
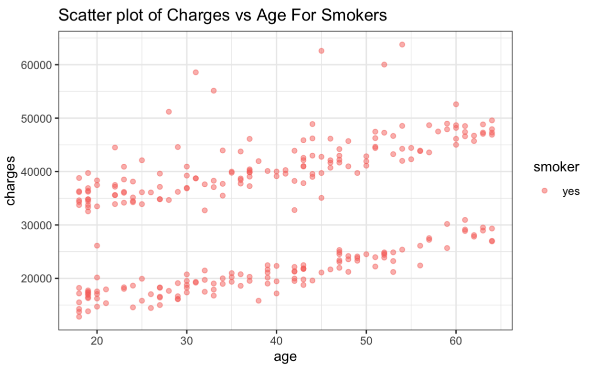
Charges vs BMI for Smokers	Charges vs Age for Smokers
	

Figure 10

From the figure above we can see that there are two trends for the Charges vs Age scatter plot. We initially suspected that this trend was occurring due to the two distinct groups of BMI for the smokers. It can be seen that for clients with BMI above 30, the charges spiked up significantly.

We wanted to first check the assumptions for linear regression for the model for smokers that only uses the predictors BMI and Age. Figure 11 shows the residual plot for the model.

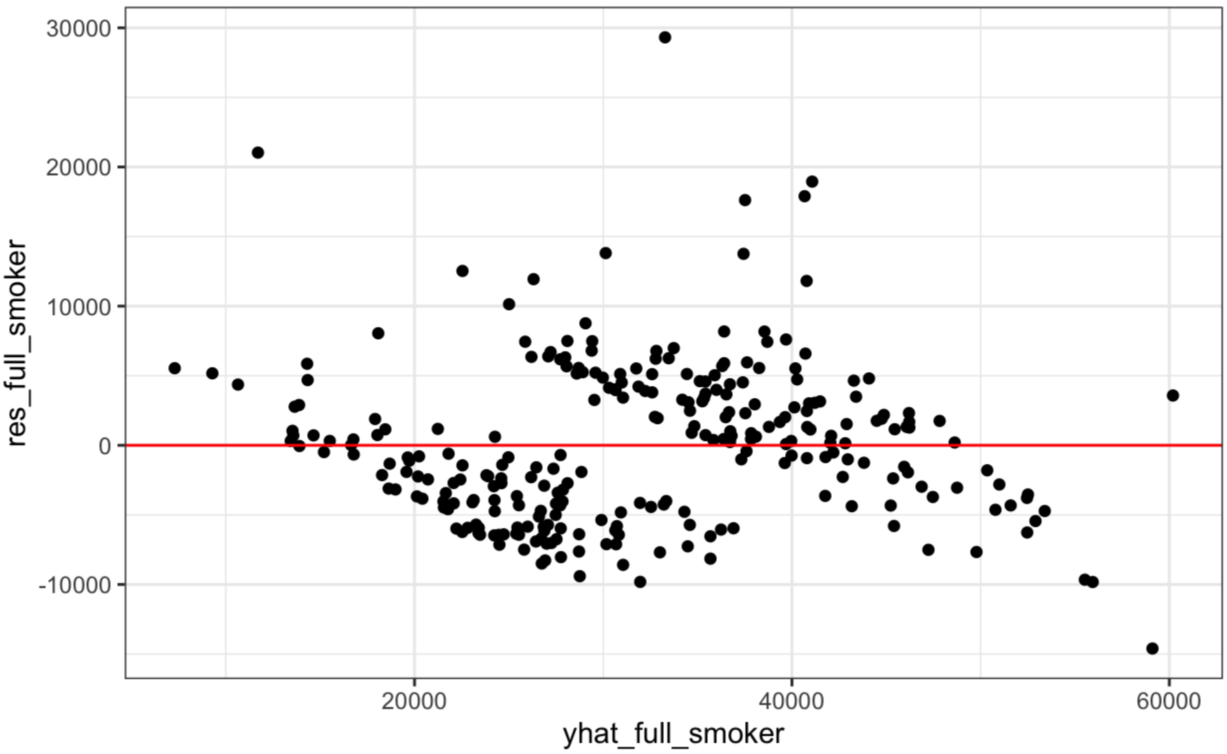


Figure 11

This residual plot seemed to have much better constant variance. We also constructed a boxcox plot to see if there is a need to change the response variable which is showcased in figure 12.

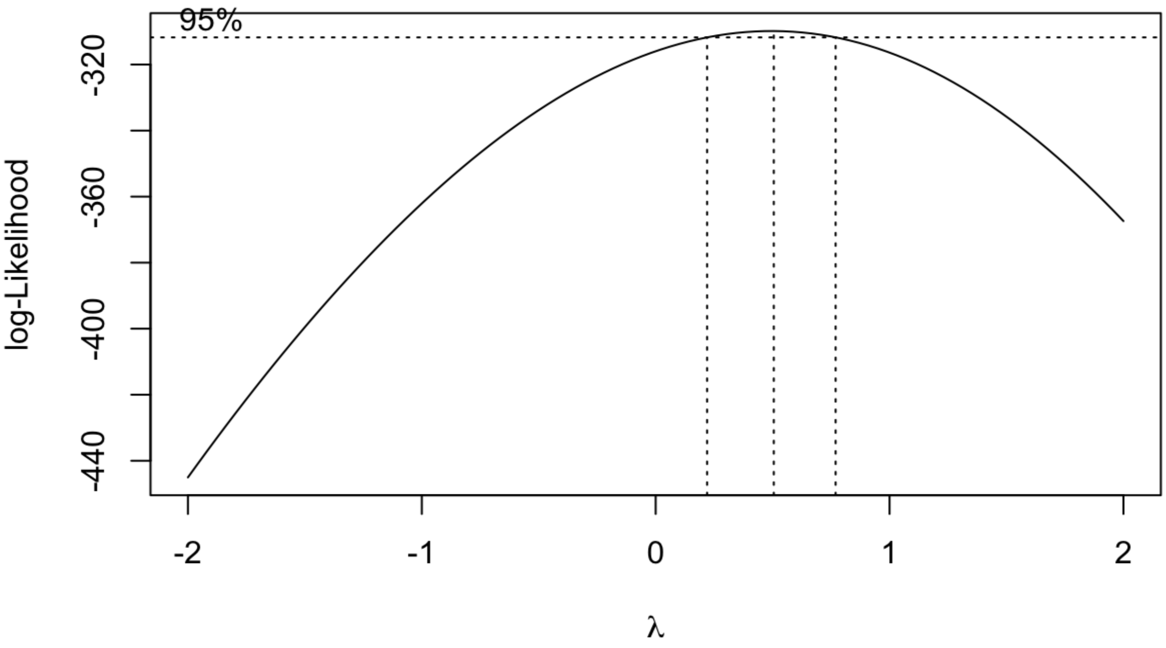


Figure 12

Because the value 1 was not in the 95% confidence interval for the lambda value, we decided to transform the response variable (charges) by the power of 0.5. The residual plot for the transformed response variable is shown in Figure 13 below.

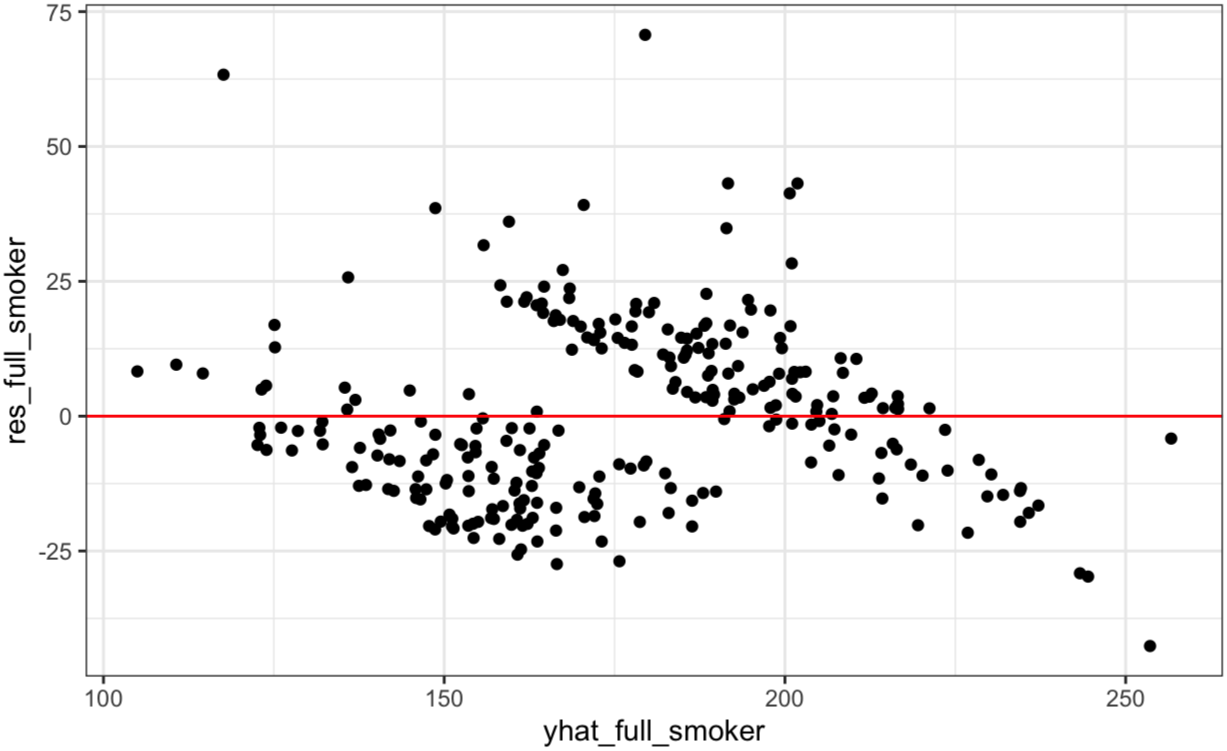


Figure 13

The residual plot seems to be a little better with the lower charges, however, the overall pattern is the same. Our group concluded that the residual plot does show constant variance. We wanted to double check the boxcox plot as a final procedure shown in Figure 14.

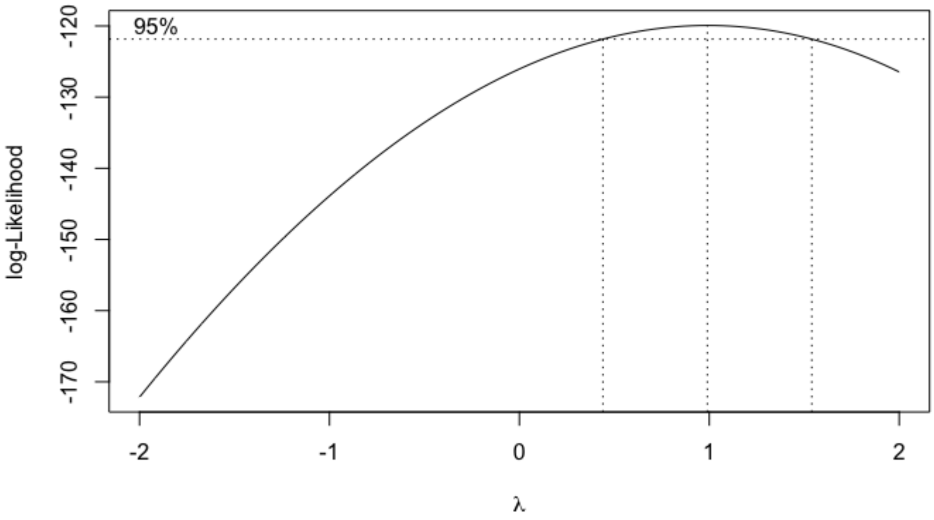


Figure 14

The boxcox 95% confidence interval shows that the lambda value includes the value 1, which means that no more transformation for the response variable is needed. The two other plots we generated were the QQ plot and the ACF plot as shown in Figure 15.

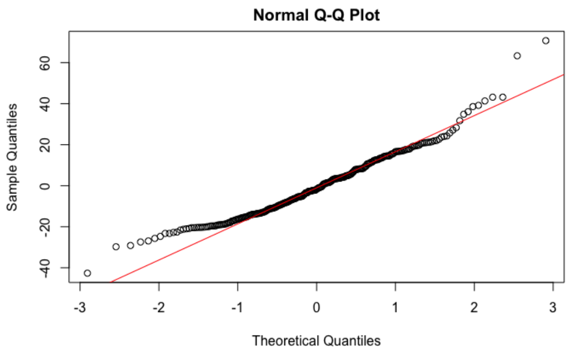
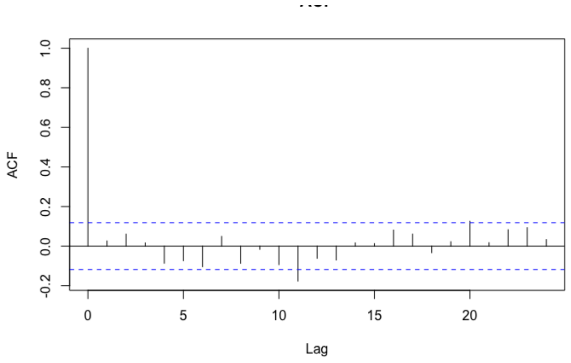
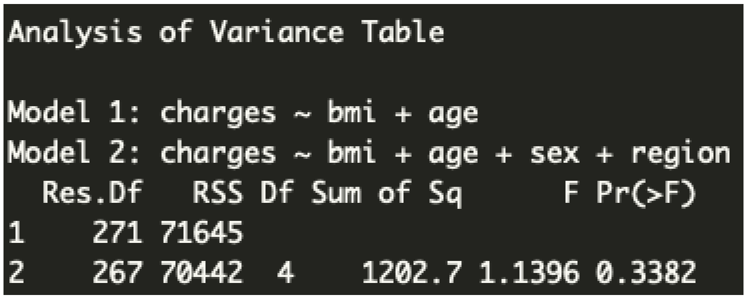
QQ Plot	ACF Plot
	

Figure 15

The qqplot follows the normality line very closely and the ACF plot seems to look good with only one or two lags. One final step on deciding that this model with age and bmi was the best model for our data, we decided to conduct a partial F tests for the two models. The analysis of the variance table is shown in Figure 16 below.

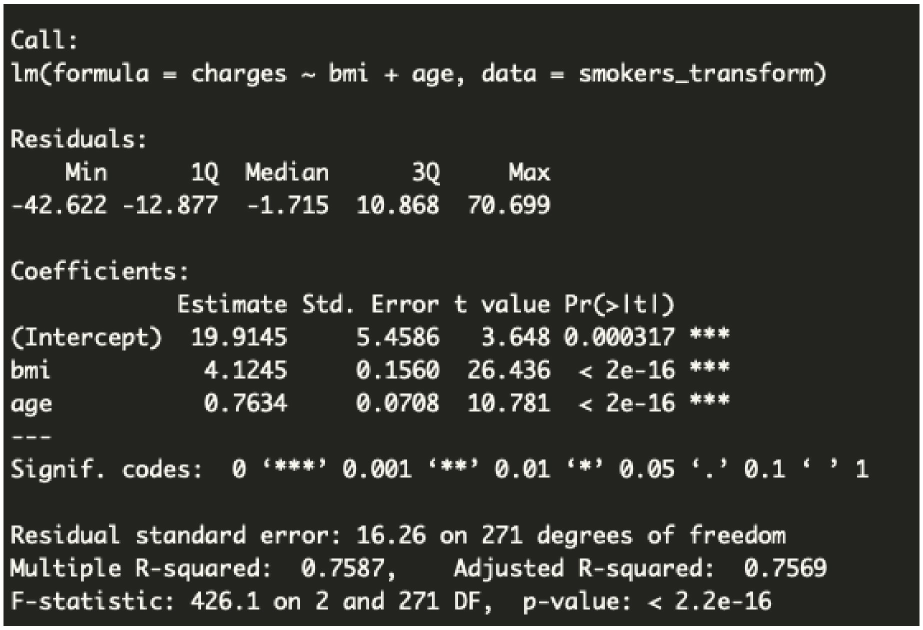


```
Analysis of Variance Table

Model 1: charges ~ bmi + age
Model 2: charges ~ bmi + age + sex + region
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    271 71645
2    267 70442  4    1202.7 1.1396 0.3382
```

Figure 16

From the Analysis of the Variance Table, we observed that the p-value was 0.3382. The Null Hypothesis of the partial F test can be stated as H_0: H_0: $\beta_3 = \beta_4 = 0$ and the alternative hypothesis can be stated as H_A: At least one of the coefficients in the null hypothesis is not zero. Since the p-value of this partial F test is greater than a threshold of 0.05, we fail to reject the null hypothesis. Therefore, we can drop these extra predictors and utilize the model with only bmi and age as the predictor. The R summary output of the model with the two main predictors is shown by Figure 17.



```
Call:
lm(formula = charges ~ bmi + age, data = smokers_transform)

Residuals:
    Min      1Q  Median      3Q     Max 
-42.622 -12.877  -1.715  10.868  70.699 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  19.9145     5.4586   3.648 0.000317 ***
bmi           4.1245     0.1560  26.436  < 2e-16 ***
age           0.7634     0.0708  10.781  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.26 on 271 degrees of freedom
Multiple R-squared:  0.7587,	Adjusted R-squared:  0.7569 
F-statistic: 426.1 on 2 and 271 DF,  p-value: < 2.2e-16
```

Figure 17

Using this summary the final multiple linear regression model for smokers is $\widehat{y\,*} = 19.9145 + 4.1245\,bmi + 0.7634\,age$, where $y\,* = y^{0.5}$. The slopes in context can be summarized by saying for every increase in bmi the charge variable increases by $(4.1245)^2 = 17.0115$ and for every increase in age the charge variable increases by $(0.7634)^2 = 0.5827796$. The intercept of the model can be interpreted as when the bmi and age values are zero, the cost of insurance is 396.5873.

One last component that we checked for this model for smokers was checking outliers or influential points. We conducted some analysis to find externally studentized residuals, leverage points, influential in terms of DFFITS and DFBETAS, and influential points in terms of Cook's Distance.

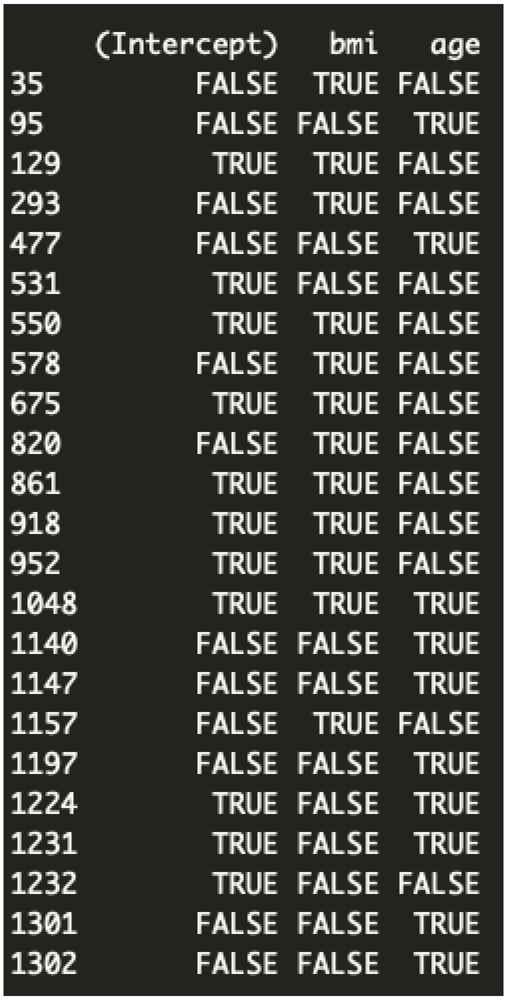
For externally studentized residuals we found that observations 129, 1301 with values 4.041989 and 4.510779 respectively were the outliers. With leverage we found 11 points to have high leverages as shown in the table below where the top row is the observation point and the bottom row is the measures.

251	293	413	544	550	665	804	861	1048	1125	1157
0.0268	0.0284	0.0226	0.0326	0.0258	0.0224	0.0251	0.0301	0.0547	0.0225	0.0307

In terms of influential observations in terms of DFFITS, we found 11 observations, which is showcased in the table below with the first row being observations and the second row being the measurements.

129	293	550	578	820	861	918	1048	1157	1231	1301
0.569	-0.231	-0.297	0.260	0.218	-0.329	0.246	-0.656	-0.225	0.244	0.302

The matrix with the influential points in terms of their predictors is shown below in the table for DFBETAS.



```
     (Intercept)   bmi   age
35         FALSE  TRUE FALSE
95         FALSE FALSE  TRUE
129         TRUE  TRUE FALSE
293        FALSE  TRUE FALSE
477        FALSE FALSE  TRUE
531         TRUE FALSE FALSE
550         TRUE  TRUE FALSE
578        FALSE  TRUE FALSE
675         TRUE  TRUE FALSE
820        FALSE  TRUE FALSE
861         TRUE  TRUE FALSE
918         TRUE  TRUE FALSE
952         TRUE  TRUE FALSE
1048        TRUE  TRUE  TRUE
1140       FALSE FALSE  TRUE
1147       FALSE FALSE  TRUE
1157       FALSE  TRUE FALSE
1197       FALSE FALSE  TRUE
1224        TRUE FALSE  TRUE
1231        TRUE FALSE  TRUE
1232        TRUE FALSE FALSE
1301       FALSE FALSE  TRUE
1302       FALSE FALSE  TRUE
```

Figure 18

For the model, there were no observations that were influential in terms of Cook's distance.

Non Smokers

Similar to the smoker dataset, we utilized a forward selection process to select a model, which produced a model with the predictors of age, children, region, and sex. This in relation to the dataset was a model that utilized all the predictor variables. The selection process model had an AIC of 17949.26. For the model with all the predictors we created a residual plot, which is showcased in figure 19.

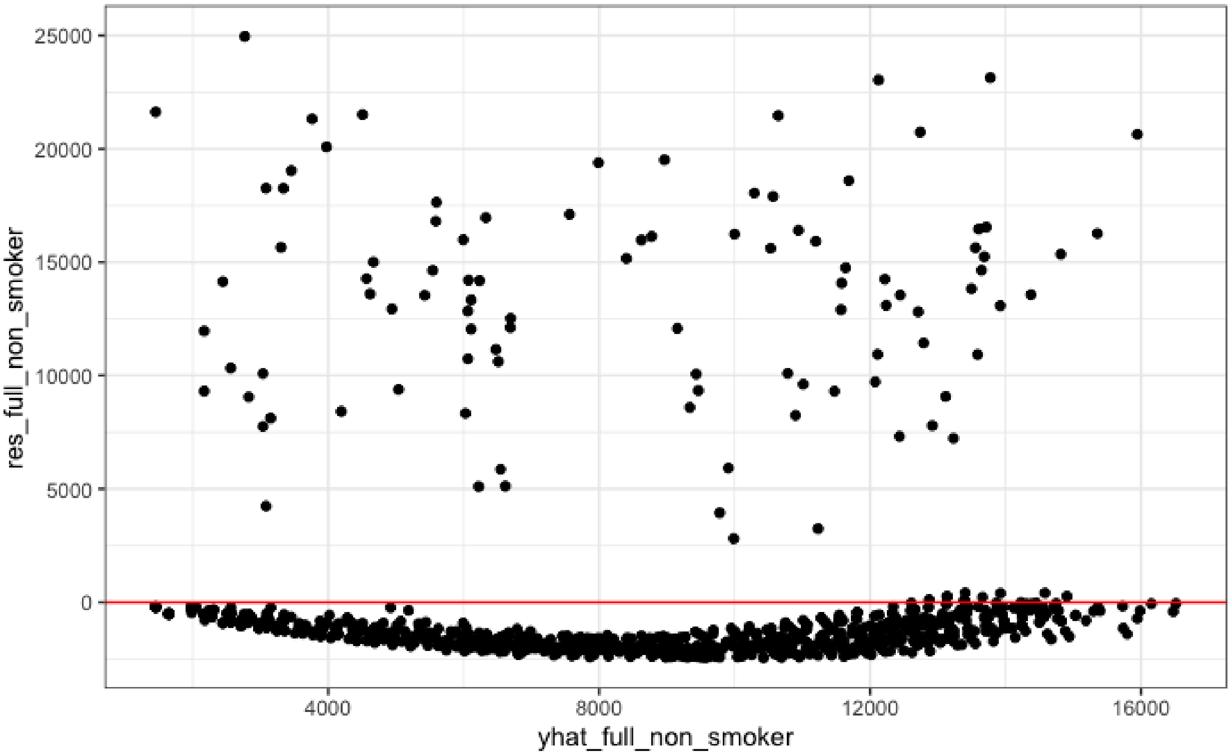


Figure 19

This residual plot obviously violates the constant variance assumption of multiple linear regression. Therefore, our first instinct was to create a boxcox to see if we could transform the response variable to meet the constant variance assumption, which is shown in figure 20.

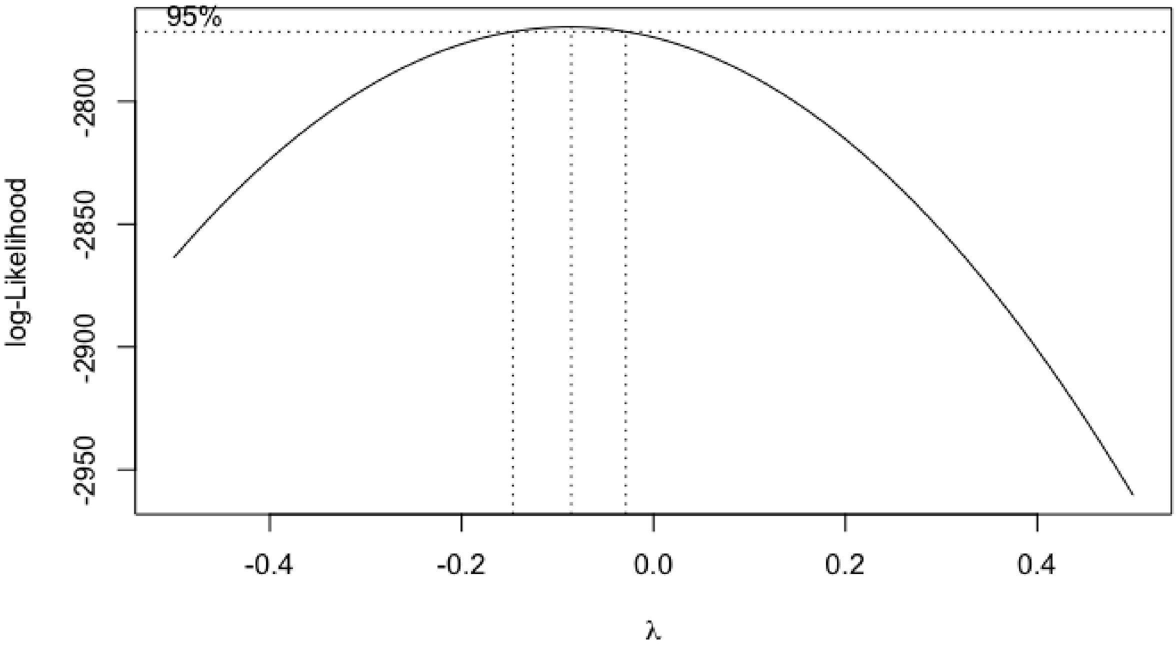


Figure 20

From this boxcox plot we can see that transforming the response variable to the power of -0.1 may help in the variance. However, even after transformation of the residual plot, figure 21, we could clearly show that it does not meet the constant variance assumption and there might be two separate groups in the data.

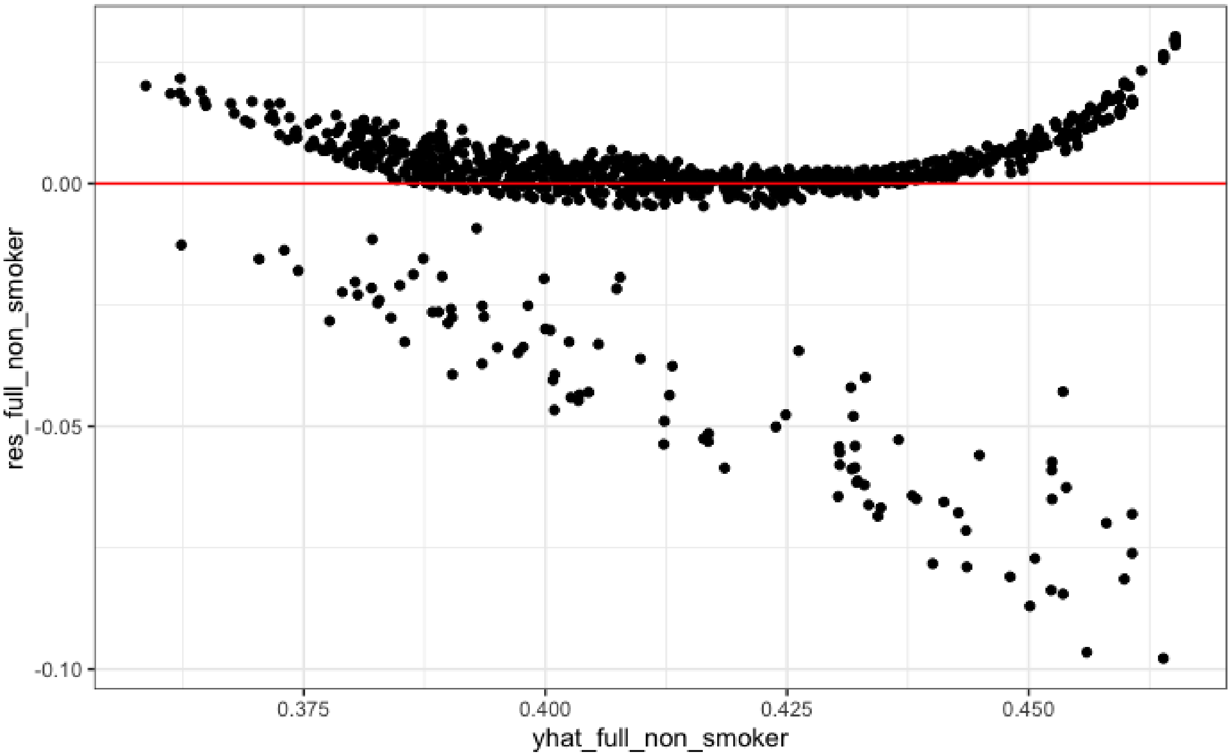


Figure 21

We also were able to observe that normality was not met with a generated QQ plot shown below.

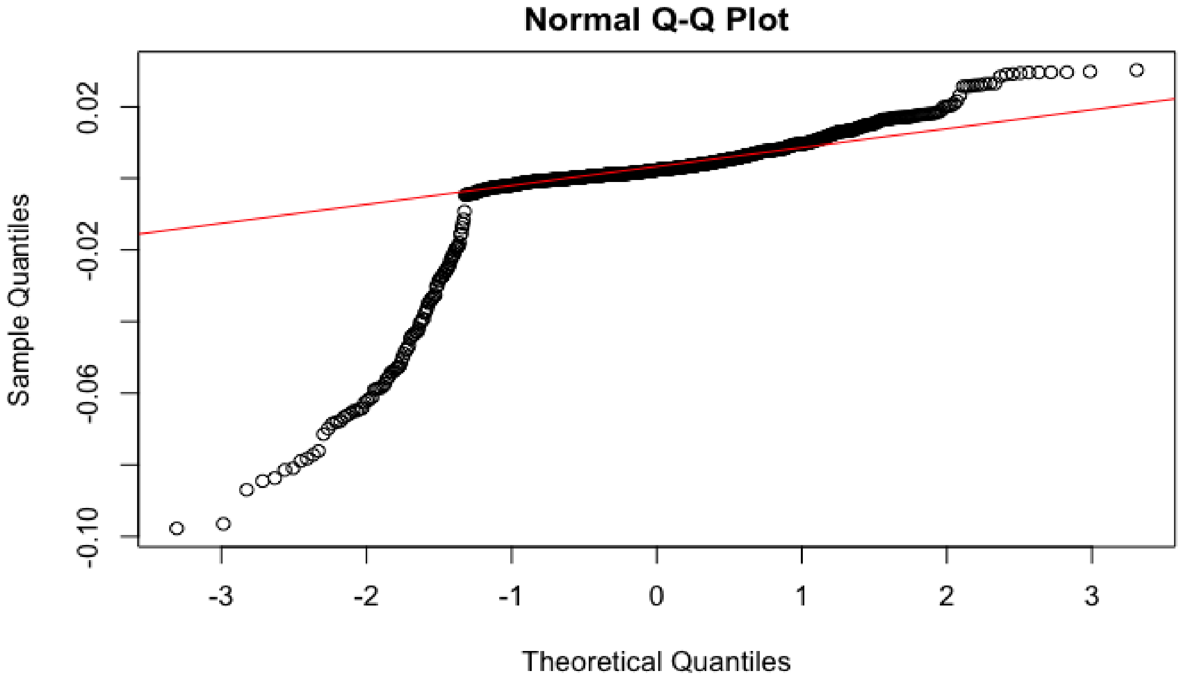


Figure 22

We decided to do some more exploratory data analysis on the data to see if we can split the smokers into two subsets by creating a scatter plot of Charges vs Age by Children by region shown in figure 23.

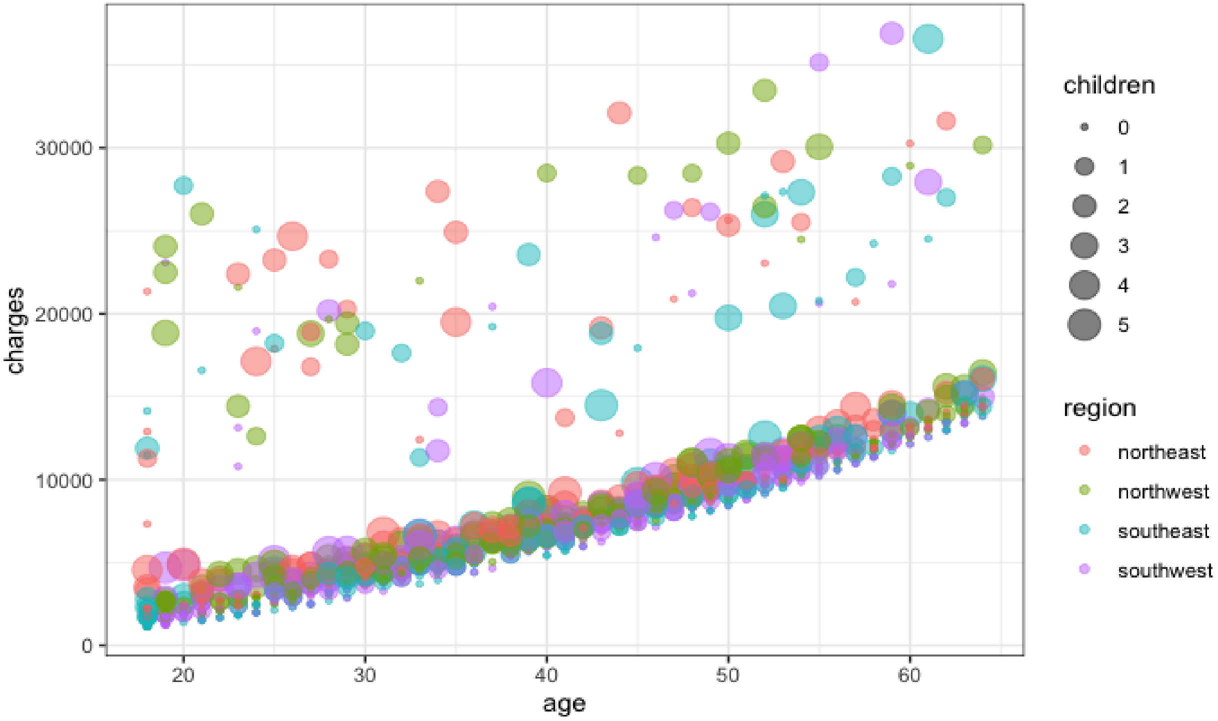


Figure 23

We couldn't find a variable that clearly separates the two groups. From the above scatter plot it seems like most clients follow a general trend (clustered points at the bottom) while some other clients have higher charges. However, no variable dominates the higher charged clients.

One idea we had was that for the higher charged clients there is an underlying variable causing these high charges. We decided to remove outliers based on this linear model by using DFFITS. Using this removal process 89 observations were removed, which was 8.36% of the dataset. With this outlier removed dataset, we decided to run another linear regression model with the same variables, which lead to the following residual plot.

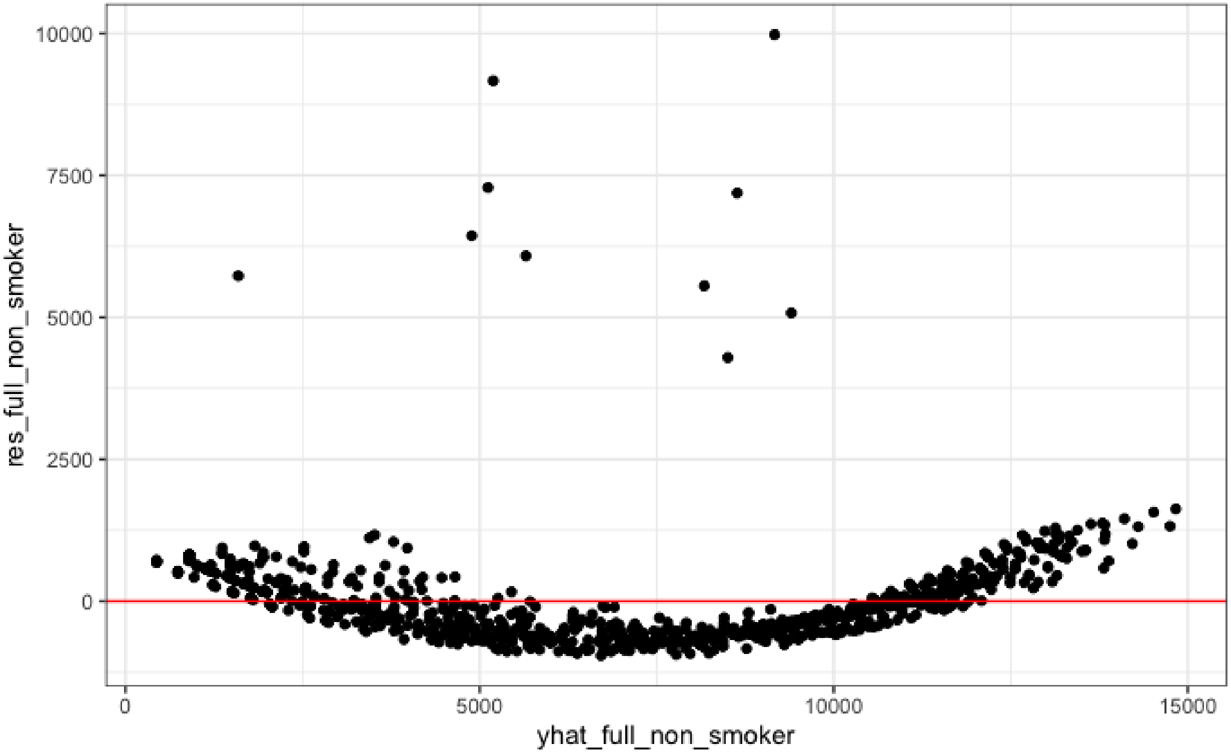


Figure 24

Although still the assumption of constant variance was not met, we at least got one group instead of two groupings. From the removed dataset we wanted to check what our data looks like through a scatter plot shown below.

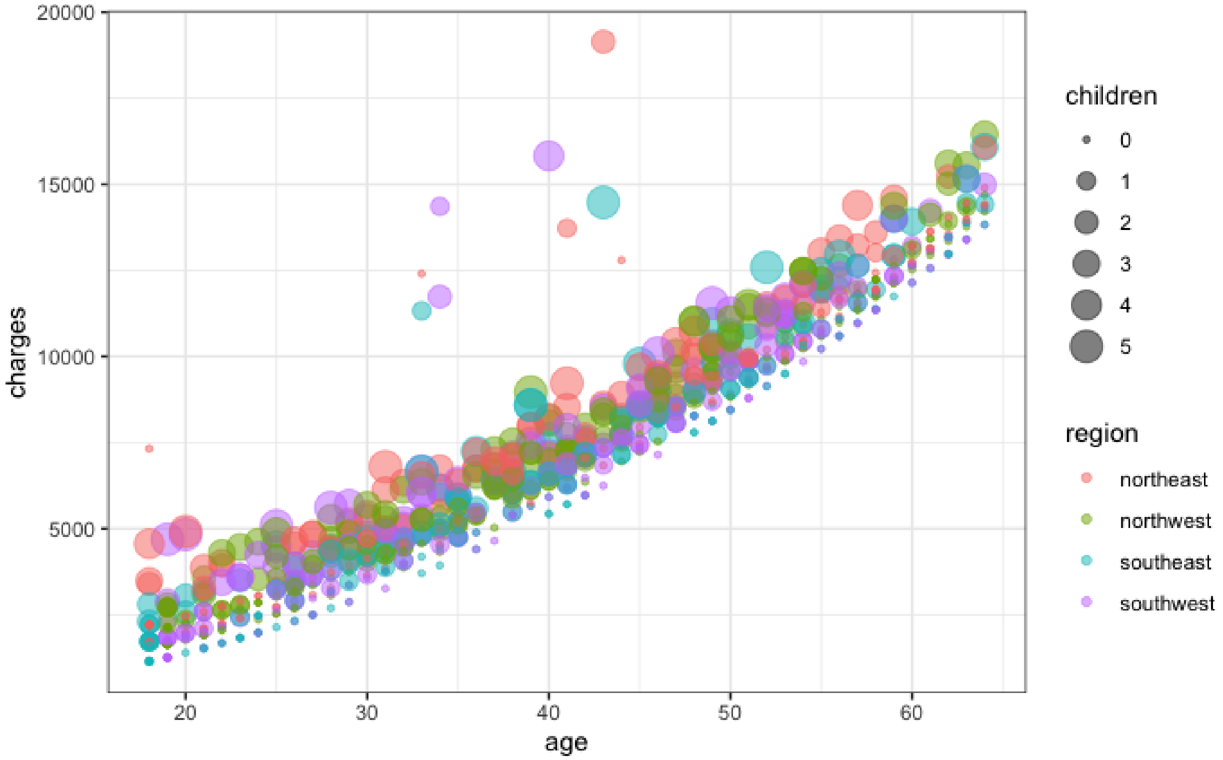


Figure 25

We then utilized the boxcox plot to find which lambda value to use to transform the response variable. From the boxcox shown below, we were able to transform our response variable with a lambda value of 0.5.

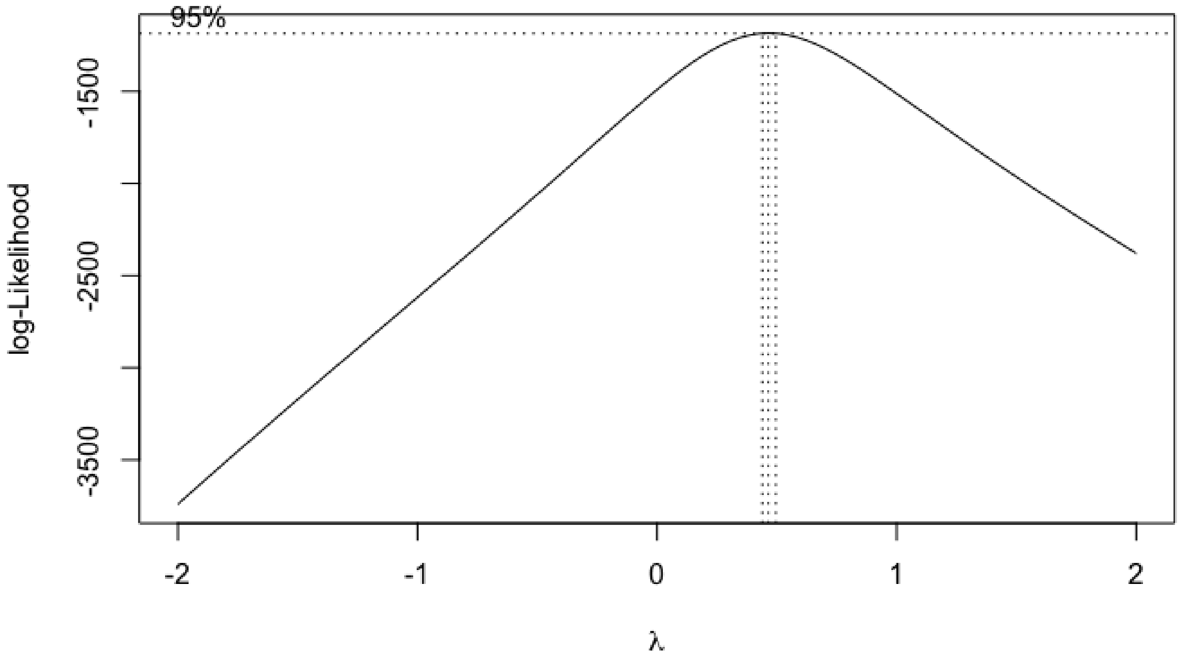


Figure 26

After the transformation was conducted on the response variable another residual plot was generated,which showed more constant variance with a fan in and fan out behavior as showcased in the plot below.

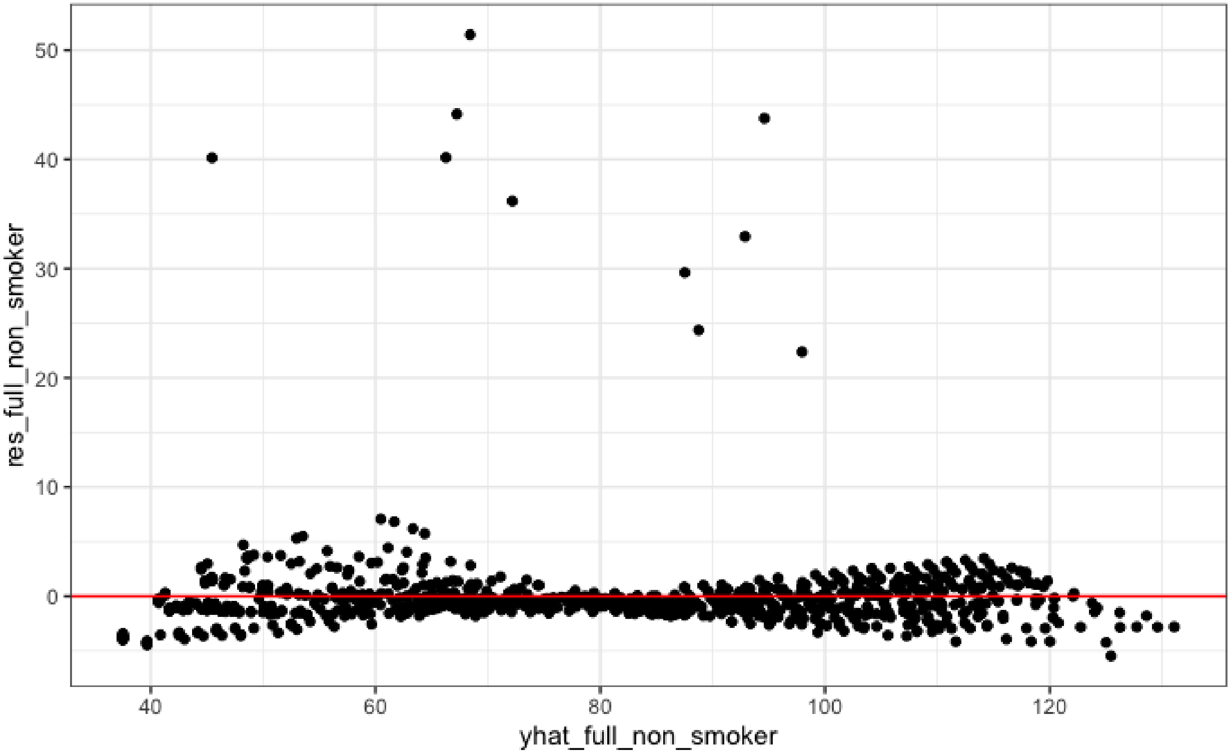


Figure 27

Although most of the higher charged clients were removed from DFFITS, from the scatter plot of the removed dataset, we still saw a group of highly charged clients. In hopes of removing this group we again used the DFFITS values to remove the outliers once again. The percent of removed values was 1.54% from the removed dataset and in total 9.77% was dropped where the total dataframe for smokers had 1064 rows. The data with the outliers removed a total of two times showed the following scatter plot, which showed a positive linear relationship.

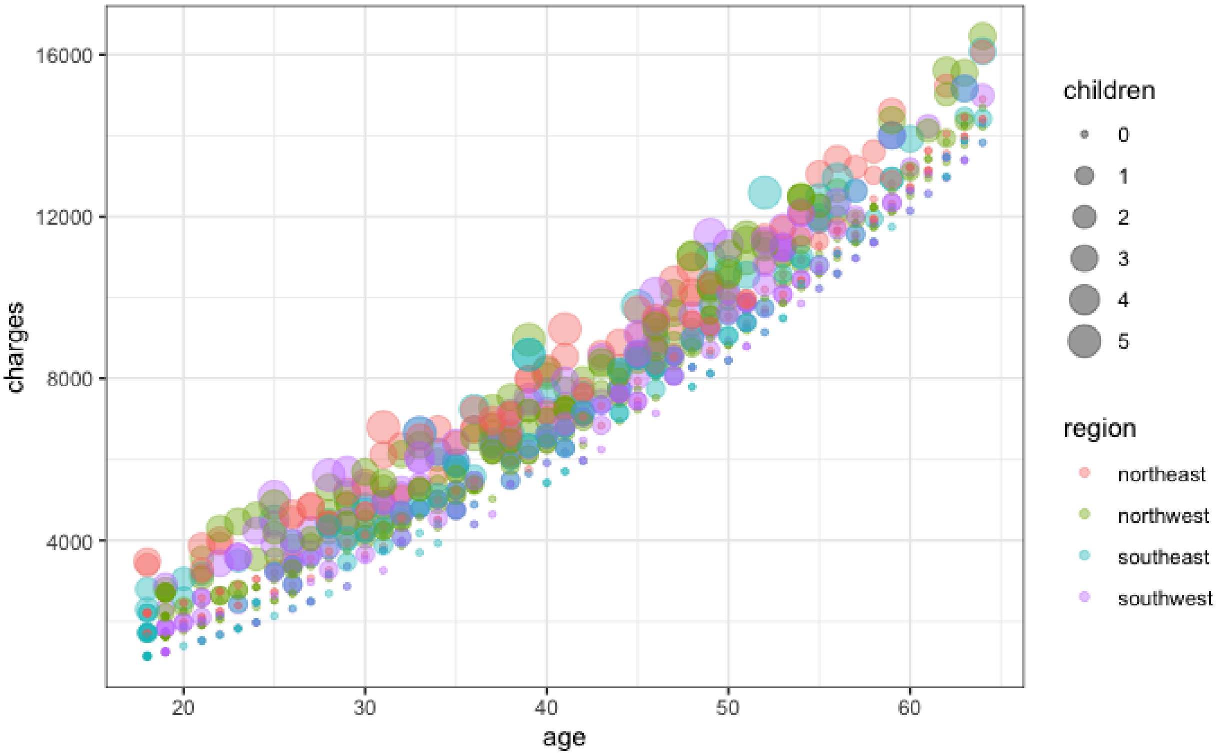


Figure 28

With this dataset, our group again ran a linear regression plot with all the predictors, which yielded the following residual plot.

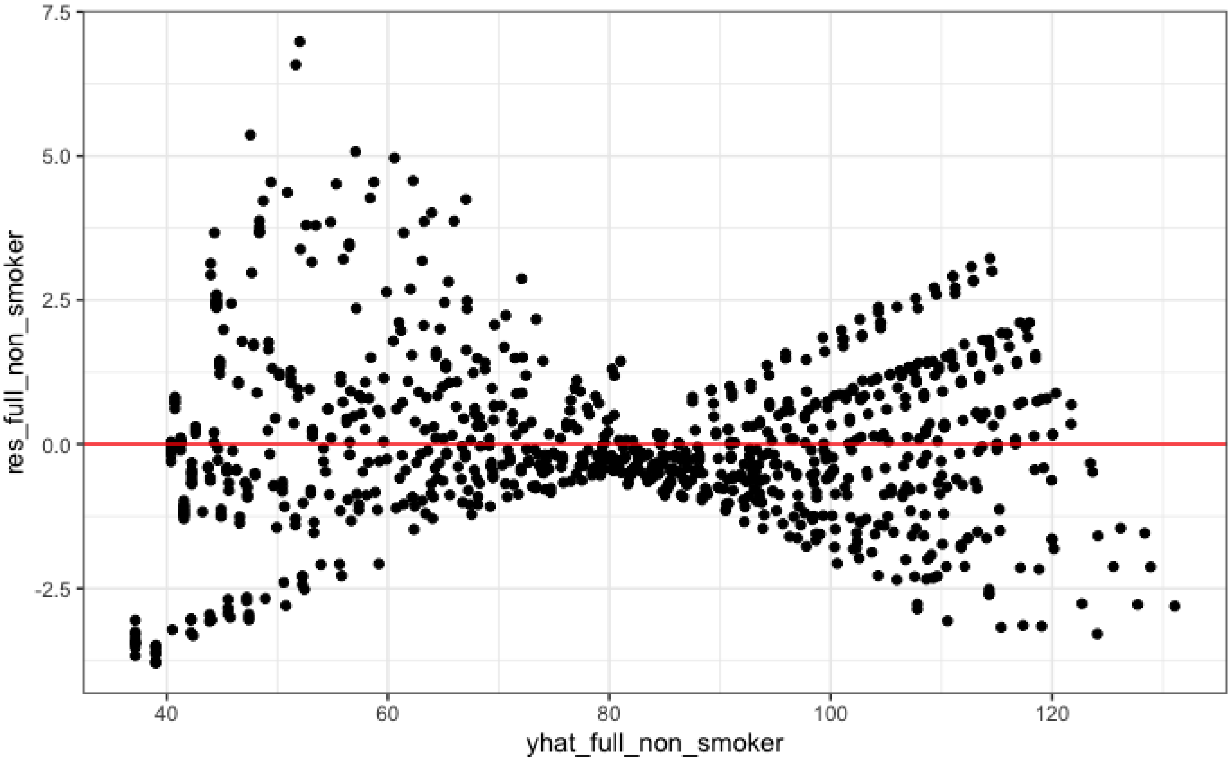


Figure 29

From the residual plot we clearly saw there was one group, but there was a fan in and fan out behavior, which means the constant variance assumption is not met. However, with many transformations and many trials and errors, this was the closest we were able to clean the data for this project. Our generated model can be summarized as the following.

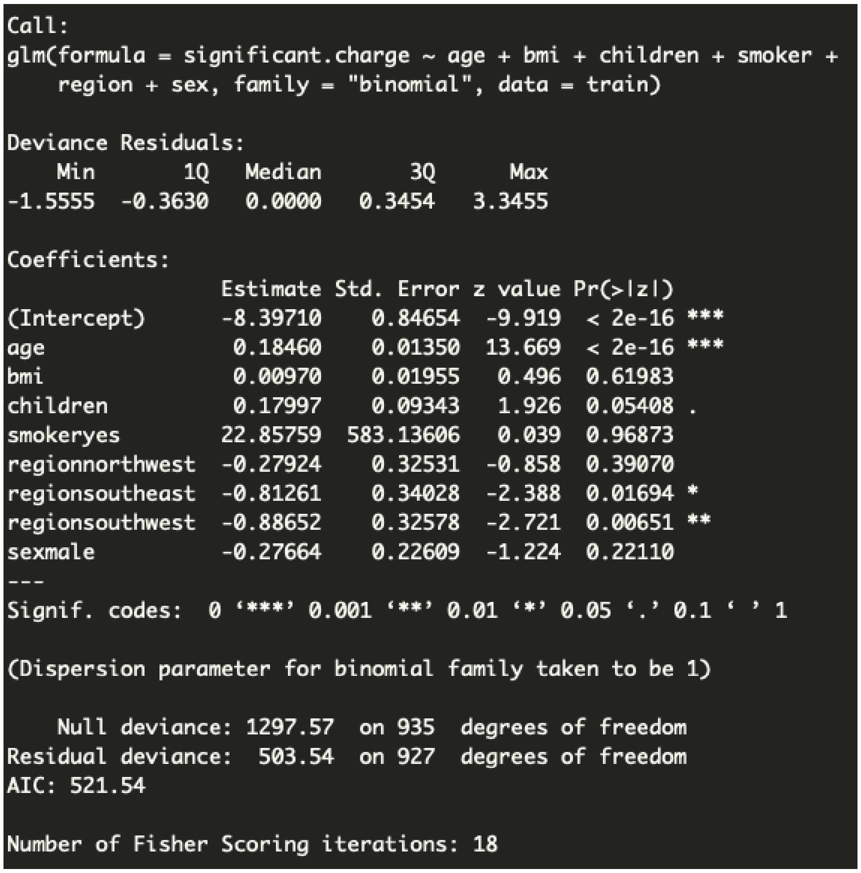
$$\widehat{y\,*} = 14.284 - 1.388I_1 - 4.108I_2 - 3.919I_3 + 1.679age + 3.585children - 3.230sex$$

From this multiple linear regression model, the indicator variables were I_1 = northwest, I_2 = southeast, and I_3 = southwest and sex was an indicator for whether a client was male. The response variable $\widehat{y\,*} = y^{0.5}$. The model's multiple R-squared and Adjusted R-squared was 0.9962 for both values with a p-value for the model being < 2.2e-16.

Logistic Regression

Although our multiple linear regression model can accurately predict how much a client will have to pay for a certain plan, we wanted to also relay information on whether or not a client will be charged significantly for providing their information to the company. Our group engineered the charges response variable to a classifier of significant charges where yes would be significant and no would be insignificant. The threshold to engineer this new column in the dataset was by dividing the data into two groups where group 1 (yes / significant) was below the median charges of the population while group 2 (no / significant) was above the median charges of the population.

For this model, we decided to use all the predictors given in the dataset (age, bmi, smoking status, region, sex, and number of children). In order to make this model, first we had to split the dataset into training and testing sets, where the split was a ratio of 7:3 respectively.



```
Call:
glm(formula = significant.charge ~ age + bmi + children + smoker +
    region + sex, family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5555  -0.3630   0.0000   0.3454   3.3455

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -8.39710    0.84654  -9.919  < 2e-16 ***
age               0.18460    0.01350  13.669  < 2e-16 ***
bmi               0.00970    0.01955   0.496  0.61983
children          0.17997    0.09343   1.926  0.05408 .
smokeryes        22.85759  583.13606   0.039  0.96873
regionnorthwest  -0.27924    0.32531  -0.858  0.39070
regionsoutheast  -0.81261    0.34028  -2.388  0.01694 *
regionsouthwest  -0.88652    0.32578  -2.721  0.00651 **
sexmale          -0.27664    0.22609  -1.224  0.22110
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1297.57  on 935  degrees of freedom
Residual deviance:  503.54  on 927  degrees of freedom
AIC: 521.54

Number of Fisher Scoring iterations: 18
```

Figure 30

From this model, we were able to also create a ROC curve to see the performance of the model when classifying significant charges as shown in Figure 31.

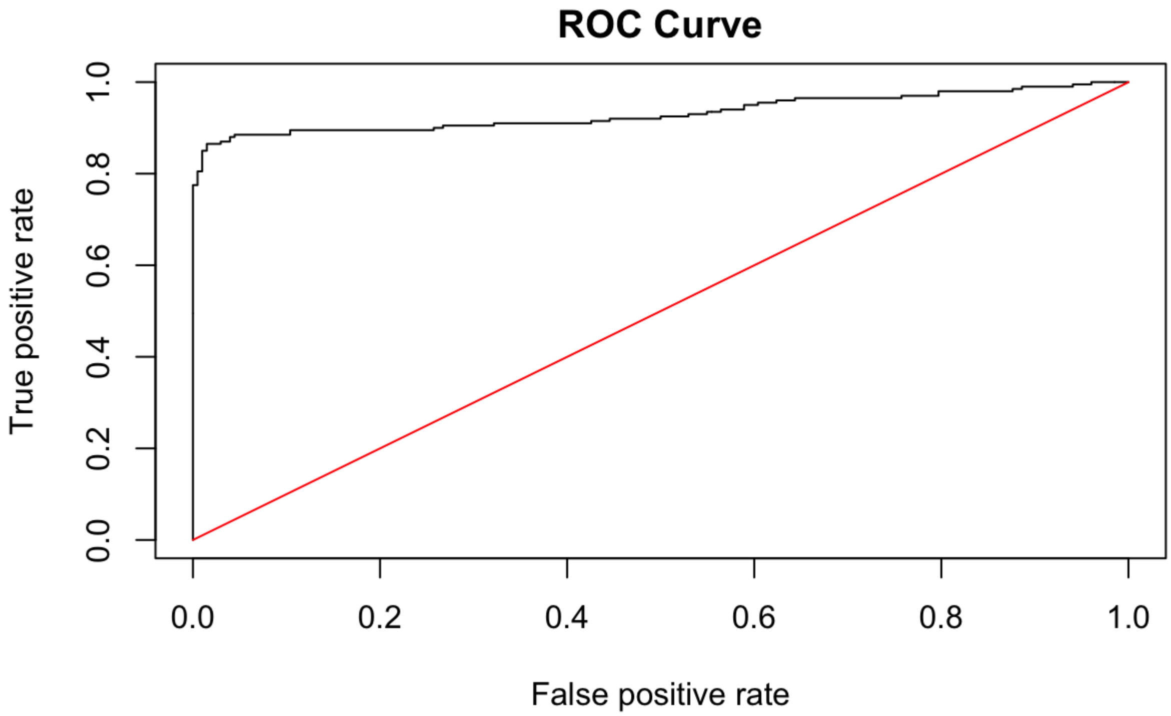


Figure 31

Also, for this model the AUC value was 0.9327475. From these two observations of the ROC Curve and AUC value, we can say that our model performs well when compared to random guessing. However, just looking at the ROC Curve and AUC value is not enough to safely say that this model is performing well. Below is the confusion matrix when the threshold was set to 0.5 for predictions.

	FALSE	TRUE
FALSE	181	21
TRUE	21	179

From the confusion matrix, we can compute the False Positive rate to be = 21 / (21+181) = 0.1039604 , False Negative Rate to be = 21 / (21 + 179) = 0.105 , and error rate to be = 1 - (181 + 179) / (179 + 181+ 21 + 31) = 0.1262136. Although these statistics look great, one aspect that we need to consider is the context of the problem. Threshold selection may need more subject matter expertise, but for the purpose of this project we assumed that people will want to be safer than sorry. It will be better for some clients to be thinking that they will be charged significantly and not be charged significantly rather than thinking that they will not be charged significantly and being charged significantly.

In order to see this scenario, we decided to lower the threshold to 0.25, which yielded the confusion matrix displayed below.

	FALSE	TRUE
FALSE	147	55
TRUE	20	180

From this confusion matrix we can observe that the people who would not be significantly charged will receive from the model that they will receive a pleasant surprise from the insurance company, which is showcased by the shift of (55 - 21) insignificant charged people being predicted to be charged significantly. However, in the grand context of the problem, from the group's perspective it will not play a huge role because of the density plot of predicted probabilities as shown in Figure 32 below.

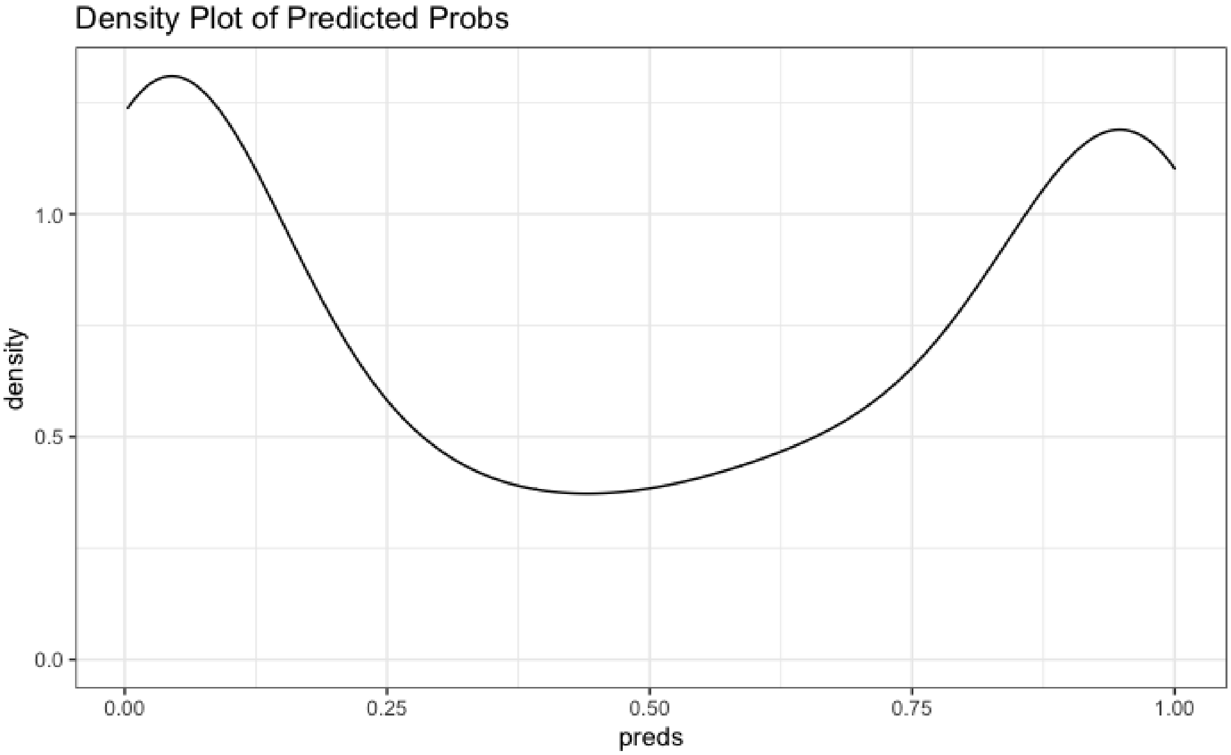


Figure 32

As one can observe the predictions have two main groupings where the middle predictions are less populated. However, as a group we cannot make a final say on the threshold value since we are not subject matter experts.

Conclusion

The findings from the EDA and initial attempt at MLR using the entirety of the dataset led to the necessary decision to split the dataset by smoker status to most accurately predict the charge an insurance company will levy. The impact smoker status had on charge made it very clear that smokers and non-smokers had distinctive relationships. Creating a singular model that did not account for this distinction would not have been accurate and would have led to predictions that were unreliable.

Although the split datasets were utilized for the MLR modeling, the original dataset was used in the logistic regression setting. The analysis of the logistic regression model proved that its performance was better suited than random guessing based on its ROC curve and AUC value. Of equal and if not more importance was the model's ability to reduce false negatives which we deemed to be crucial. It's also noted that this may change after receiving the advice of subject matter experts on what type of error the industry is most concerned with.

The application of the models ties back to the greater motivation of analyzing this dataset which is to help individuals and families navigate the ever increasing costs associated with medical insurance. The ambiguity behind the true cost of obtaining and maintaining insurance can be anxiety inducing for persons on tight budgets. This is compounded for individuals who are just beginning to obtain medical insurance and may not fully realize the associated costs. Bringing more transparency to the costs will aid consumers in signing up for the plans that are most beneficial to them but also less costly.

Limits of dataset will be one of the inspirations behind possible future research. Specifically, the missing information or descriptors of the individuals sampled from a socioeconomic standpoint. Factors such as race, income, education can all have major impacts on analyses for the given scenario. Moving forward, the models can be fine-tuned and made better by the inclusion of these sorts of variables. The analysis will not be blind to what may be assumed to have a major impact on the insurance costs individuals have. These variables may also better explain some of the relationships seen in our existing dataset like smoker status or region.

Another desired aspect of future research is the pursuit of similar data from different sources to ensure that the analysis is well balanced. Pulling cost data from different companies or different regions of the country will give the analysis full understanding of the medical insurance vertical. The findings from this dataset may not be identical to the findings from a similar dataset and if the insights are similar then it would further confirm what has been discovered to most impact insurance charges.