Medical cost prediction using linear regression

# Overview

Insurance companies need to know the medical costs incurred by people based on some of the properties, so that they can better estimate their yearly premium. We have one such dataset of 1138 different individuals and their data for seven properties. We will be using this data to predict the annual medical costs for individuals given these properties.

## Goal

The goal of this analysis is to be able to predict the medical costs for individuals given data like their age, bmi, sex, smoking habits, etc so that insurance companies can make better decisions of the amount of premium to charge such that they do not have to spend more than what they charge customers.

## Data

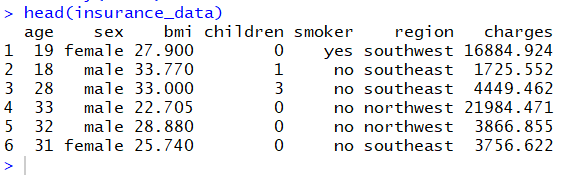
We have a dataset with 1138 observations and 7 variables. Below are the variables/columns in the given dataset and their interpretation.

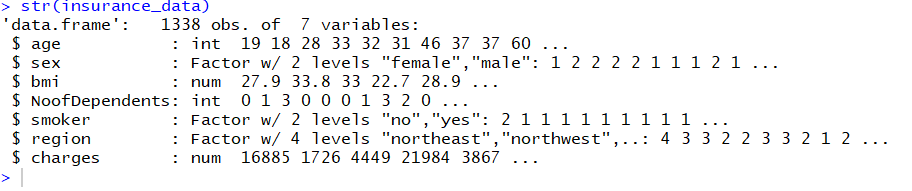
* **age**: An integer indicating the age of the primary beneficiary (excluding those above 64 years, since they are generally covered by the government).
* **sex**: The policy holder's gender, either male or female.
* **bmi**: The body mass index (BMI), which provides a sense of how over- or under-weight a person is relative to their height. BMI is equal to weight (in kilograms) divided by height (in meters) squared. An ideal BMI is within the range of 18.5 to 24.9.
* **children**: An integer indicating the number of children/dependents covered by the insurance plan.
* **smoker**: A yes or no categorical variable that indicates whether the insured regularly smokes tobacco.
* **region**: The beneficiary's place of residence in the US, divided into four geographic regions: northeast, southeast, southwest, or northwest.
* **charges**: Amount of medical costs incurred by individuals for one year. This will be the **response** **variable** of the model.

# Exploratory Data Analysis

Given data was copied to a data frame in R, named “insurance\_data”.

Take a look at the data:





“age” is integer number like 19,18, etc

“sex” is a categorical variable with two values “female” & “male”

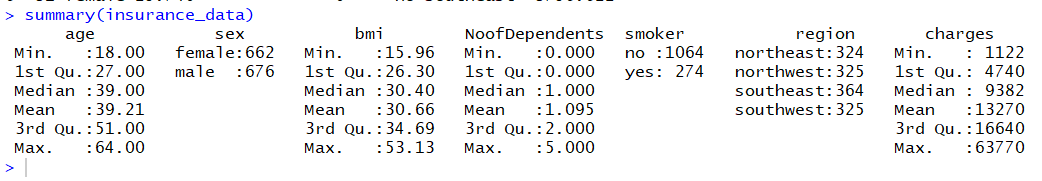
“bmi” is a continuous variable with decimal numbers

We renamed “children” variable to “Noofdependents” for better explanation of the data. It is a numeric variable with values like 0,1,2, etc

“smoker” is a categorical variable with two values “yes” and “no”

“region” is categorical variables and has four levels : northeast, northwest, southeast and southwest

## Summary of the data (mean, median, minimum and maximum values)



“age” has values between 18 and 64, which seems to make sense. So, no bad data in age column.

We have female:male ratio 662:676 which is pretty much balanced.

Minimum bmi is 16 and maximum is 53.13. Upon looking online, these seems to be the right values for bmi.

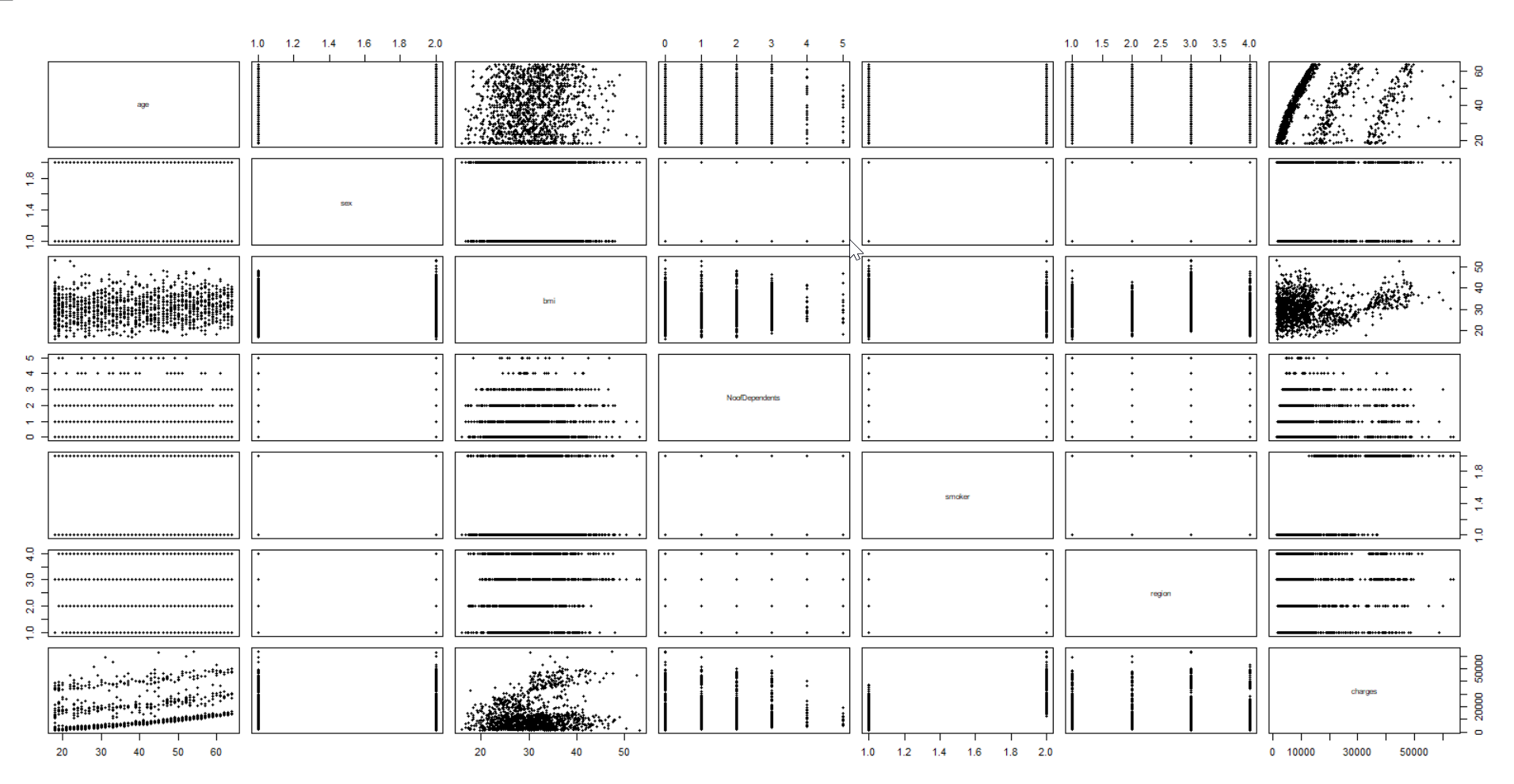
Smoker:non-smoker ratio is 274:1064

We see that all the four regions have almost equal number of observations.

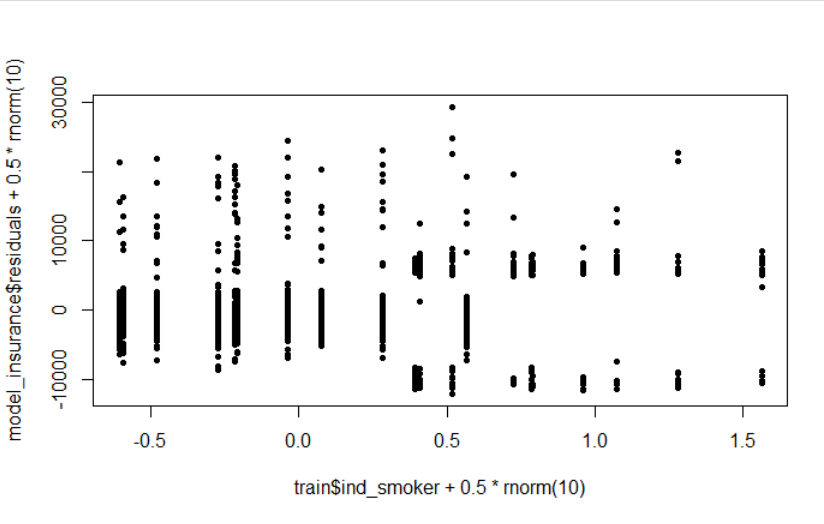
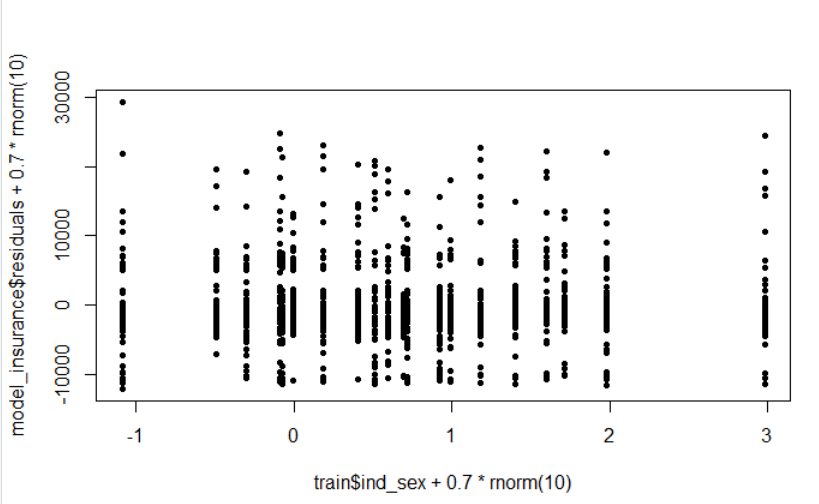
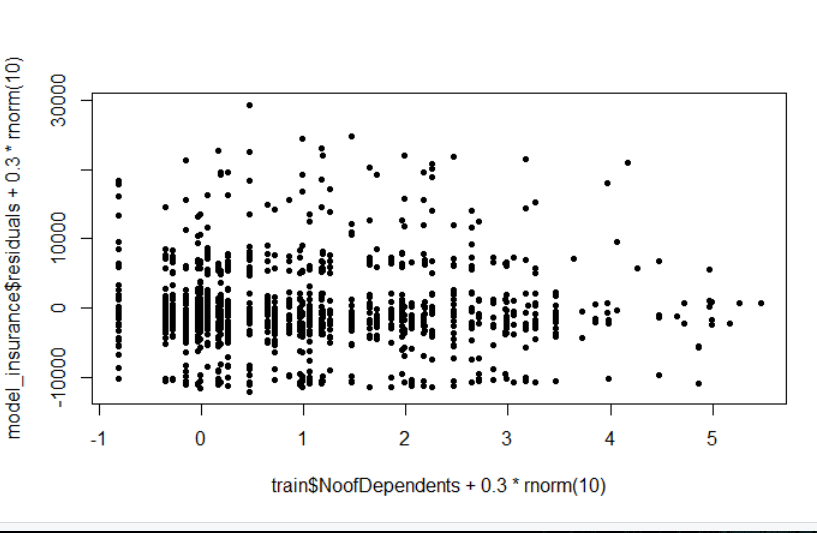
The range of “charges” is very wide with minimum value being $1122 and maximum value $63770.

## Visualizing the data

Plotting each variable against others



Created below jitter plots for some covariates to better understand their relationship with response variable.

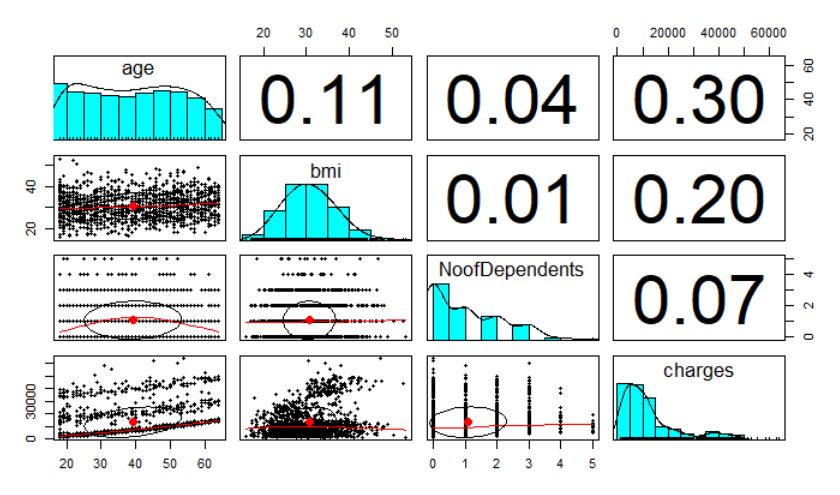


Linear relationship between age and charges can be clearly seen with three lines.

From bmi vs charges plot, we can see that bmi data is divided into two clusters.

Rest being categorical variables, we cannot really interpret anything.

Let us see histograms and correlation between numerical variables and response variable.



Here, we can see that our response variable “charges” is heavily right-skewed.

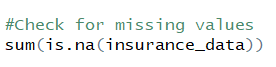
“age” has roughly two peaks which bmi is normally distributed.

“age” and “bmi” have not very weak correlation with “charges”

There is some weak correlation between age and bmi.

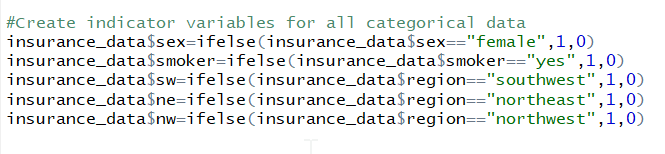
# Feature Engineering

1. Check for missing values in the data

We have no missing values in the dataset.

1. Since machine learning algorithms cannot operate on categorical variables, converting all the categorical variables to numeric variables using dummy variables.



Sex: female=1, male=0

Smoker: yes=1, no=0

**Region**: Created three indicator variables to depict four levels of region.

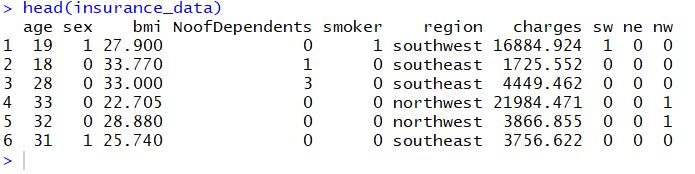
sw ne nw

Southwest 1 0 0

Southeast 0 0 0

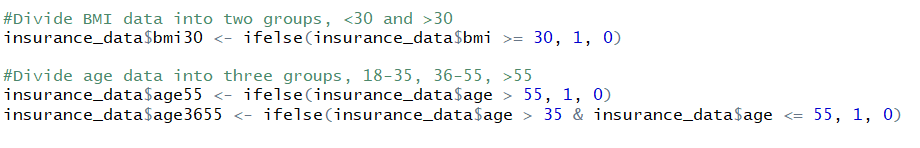
Northwest 0 0 1

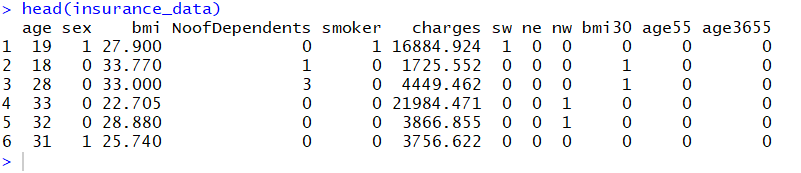
Northeast 0 1 0



We observed in plot bmi vs charges that it had two clusters, bmi < 30 and bmi > 30. So let us create a new variable “bmi30” with yes=1, no=0.

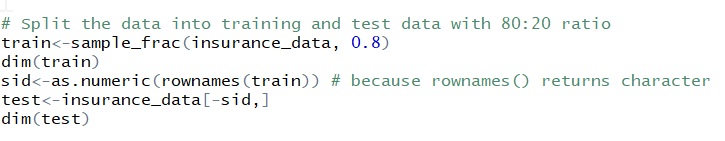
Also, assuming that medical costs for people between age 18-35 will be less, 36-55 medium and more for individuals >55 age, creating two new variables to group the age in given brackets.

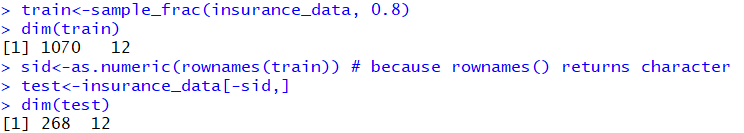




# Split training and test data

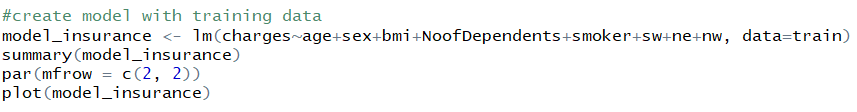
At this point, let us split our data into training (80%) and test (20%) data so as to not introduce any bias in the test data or the model.

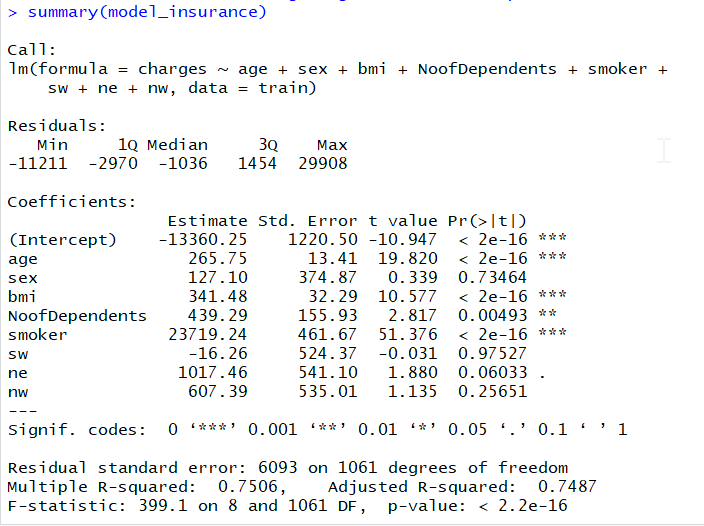




# Create Model

Let us create a basic linear regression model with all the initial covariates to predict the medical cost charges for individuals





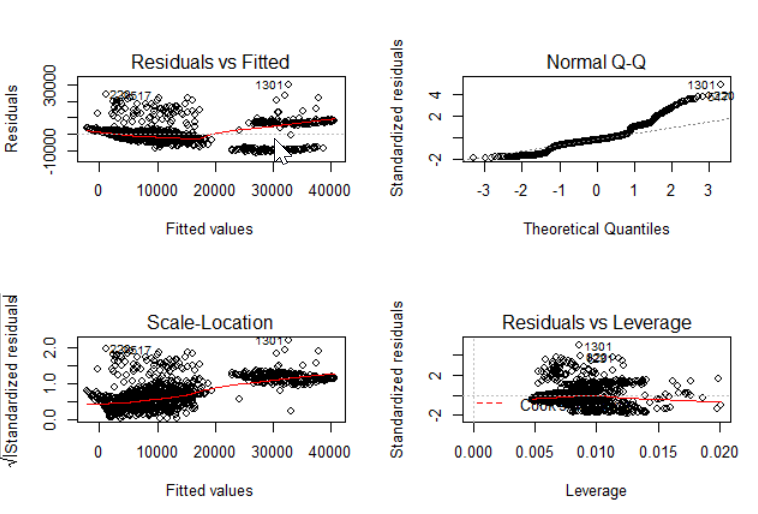
We can see from the summary of the model we created that,

p-value is equivalent to 0 which means that this linear regression model is significant in explaining the variation in “charges” based on the co-variates.

Three covariates – sex, sw and nw are not significant in this model.

R-square = 0.75 and Adjusted R-square = 0.749, which means this model can explain around 75% of variation in “charges”.

Residual standard error = 6093, which means that 95% of predictions will have error of around +-$12200. This is expected because the range of response variable “charges” is very wide.

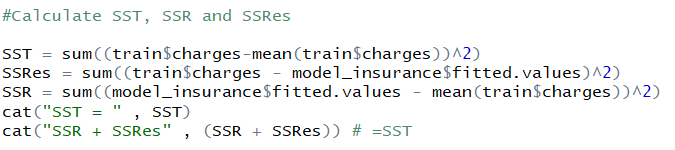


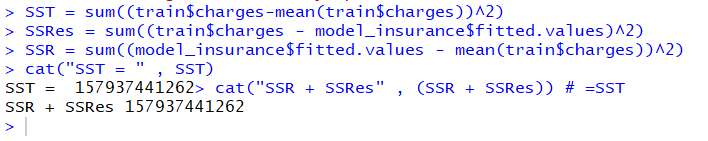
As seen from residual vs fitted value graph, variance is slightly increasing as fitted value increases. So, the model does not meet the “Equal Variance for Errors” assumption.

From the QQ-plot, we can derive that “Errors are not Normally Distributed”.

From the residuals vs leverage plot, we can say that we do not have any high residual, high leverage point which can be termed as outlier. However, we do have many high leverage points, but since they don’t have high residuals, they are not outliers.

## Calculating SST, SSR and SSRes and R-square



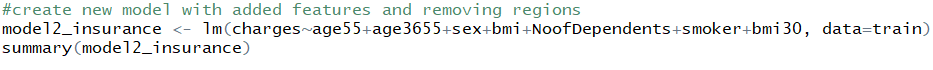


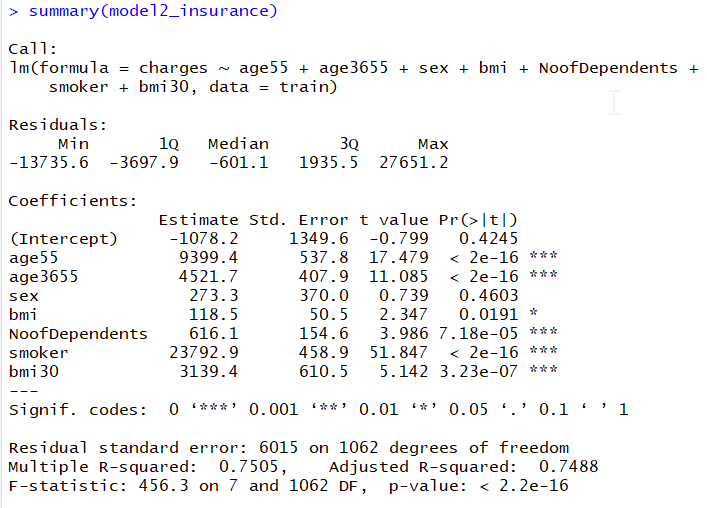




## Trying different models

Created new model with new age variables “age55” , “age3655” and “bmi30” and removing region covariates

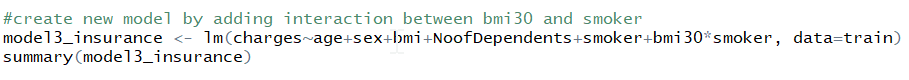


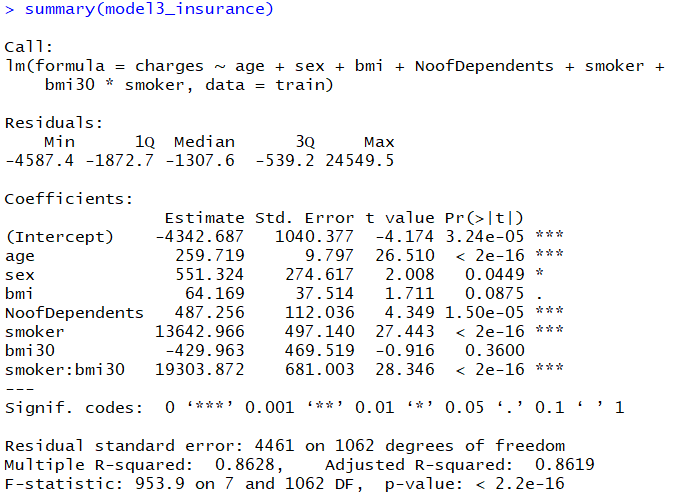


This adjusted-Rsquare for both the models is almost equal, but residual standard error decreased from 6093 to 6015 in the new model.

## Considering Interaction between covariates

In order to further improve the model, we considered interactions between the terms and tried adding different interaction terms and checking the performance of the model. **The significant improvement in the model was found upon adding interaction term between covariates “bmi30” and “smoker”.**





Adjusted R-square increased from 0.75 to 0.862 in this new model with interaction between covariates “bmi30” and “smoker”.

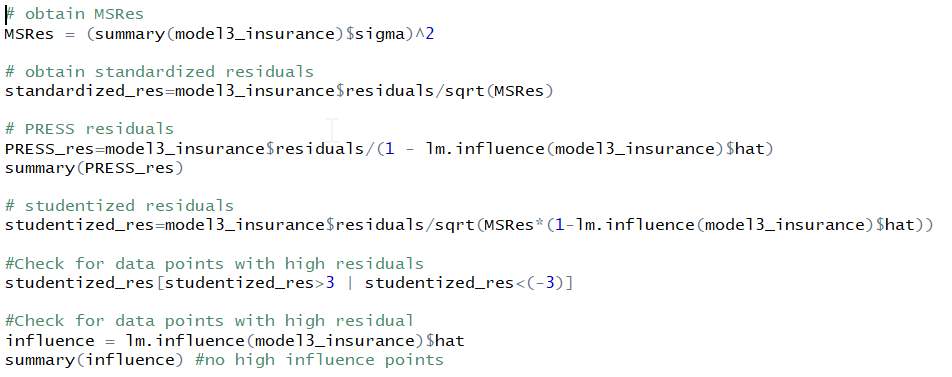
Residual standard error also decreased significantly from 6015 to 4461.

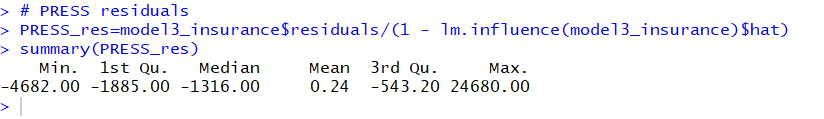
So, this model is quite better than the earlier models.

## Checking the model adequacy

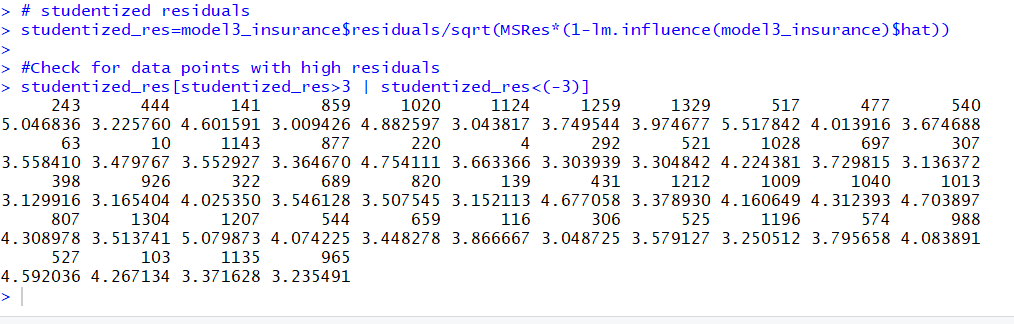
### Residual Analysis

To check the model adequacy, let us calculate it’s PRESS residual, studentized residual.

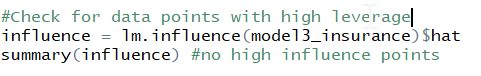


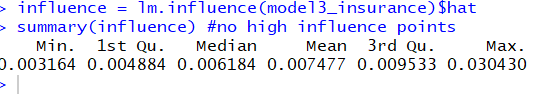


Let us see how many points can be identified as having high PRESS residuals by calculating studentized residual. We are looking for points which have studentized residual >3 and <-3. We see that there are quite some points which have high PRESS residual.



However, in order to be qualified as outlier, data point should have high PRESS residual and high leverage. So, let us check for the high leverage points.

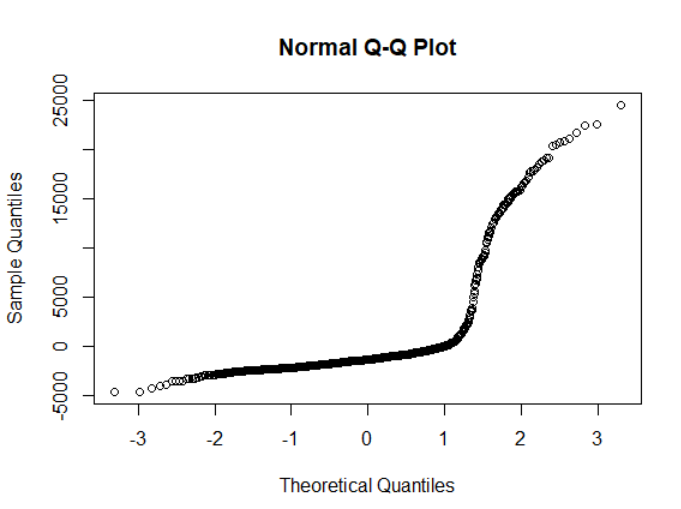




We can see that maximum value of influence is 0.03, which means that **none of the points have high leverage. So, no outliers in the dataset.**

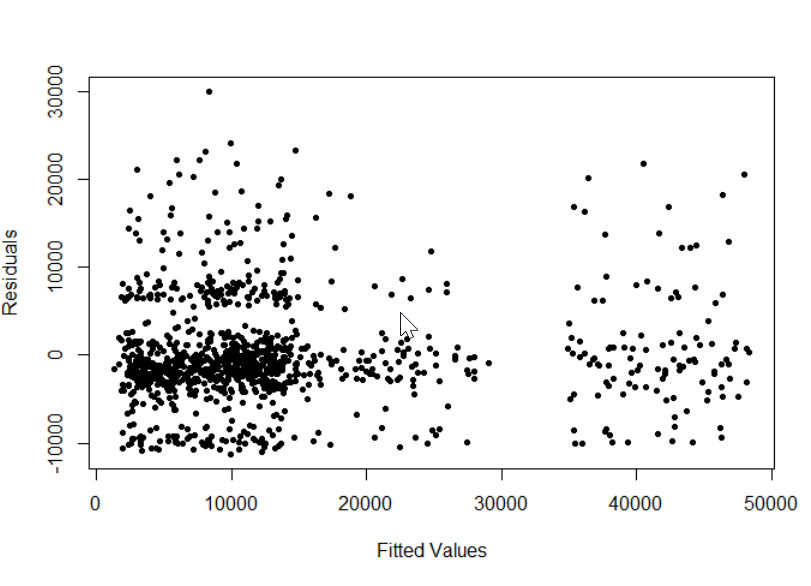
### Residual Plots

Let’s plot the QQ-plot to check for the **normality** of errors

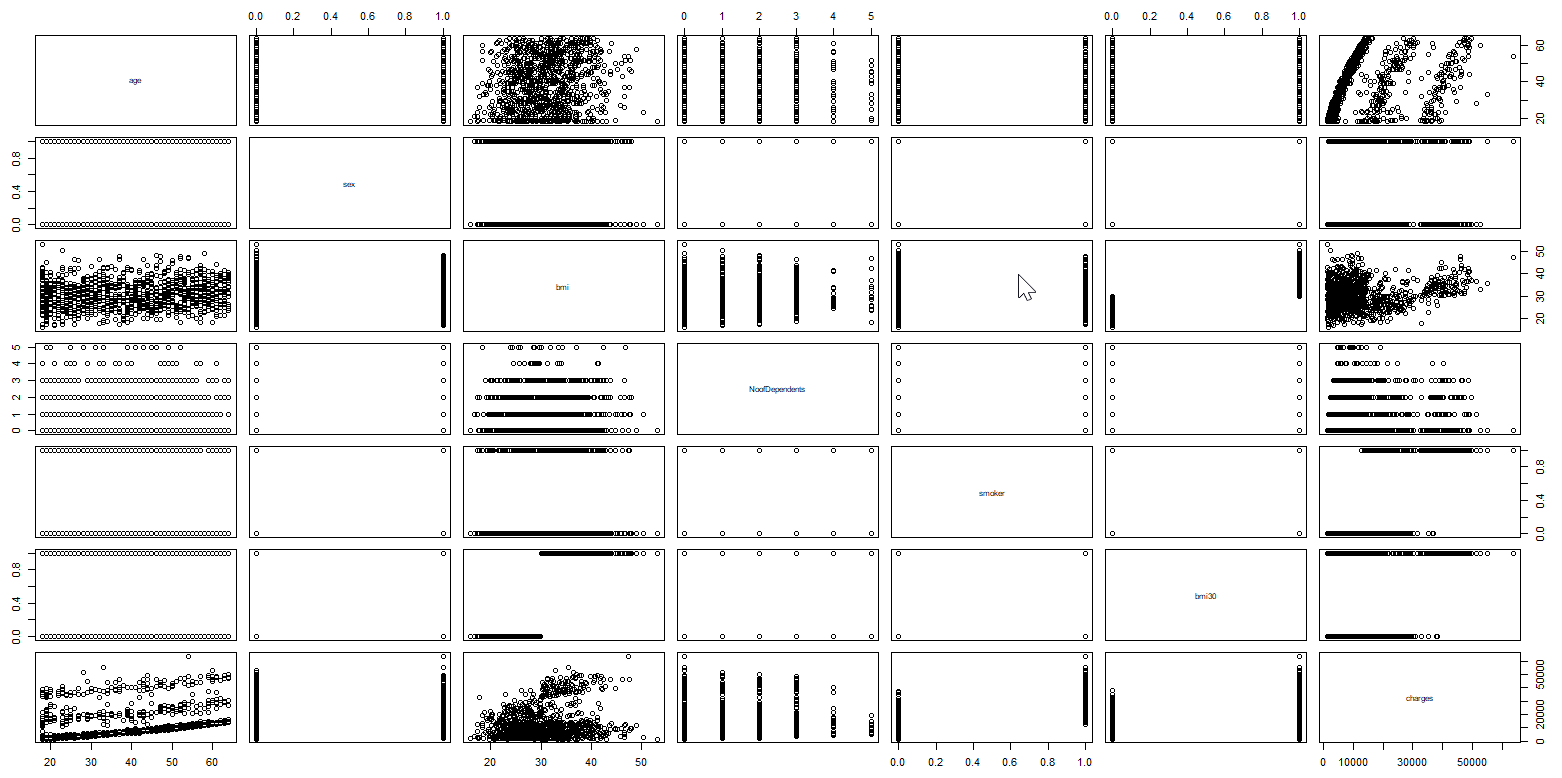


This proves that residuals do not follow normal distribution. So, we should go for transformation on y or both x and y to resolve this issue.

Let’s check the **equal variance** assumption by plotting Residuals vs Fitted value. Variance seems to be mostly equal.



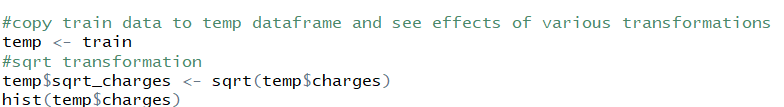
Let us check for the linearity by plotting each covariate with response variable.



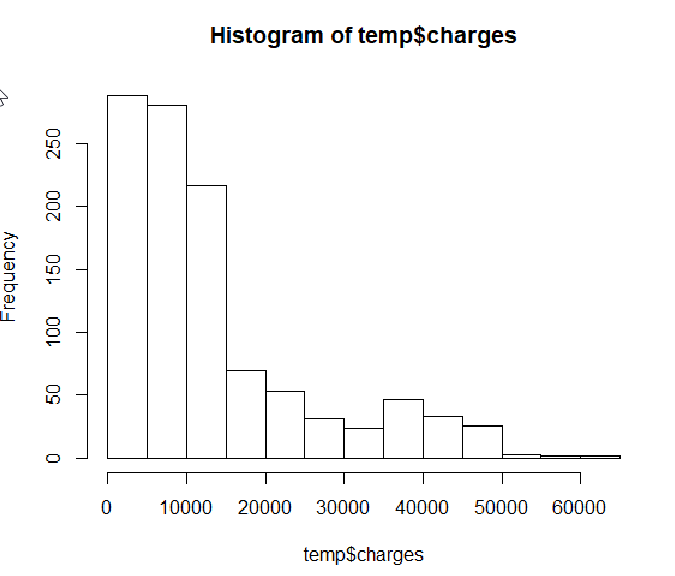
We can see slight non-linearity between “age” and “charges”. Rest of the covariates seems fine.

We do not need to check for independence of errors as our data is not time-series data.

In order to resolve non-normality and non-linearity issues, let’s try square-root, cube-root and log transformations on y.



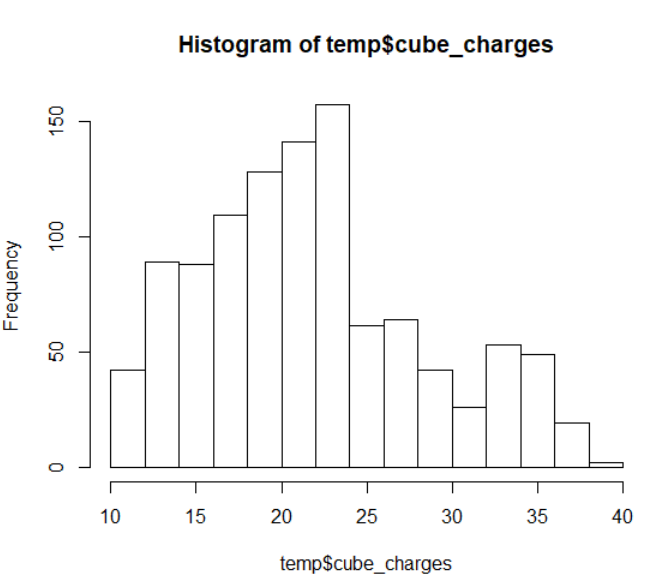
Square-root transformation on y reduced skewness slightly, but not much.



Cube-root transformation on y



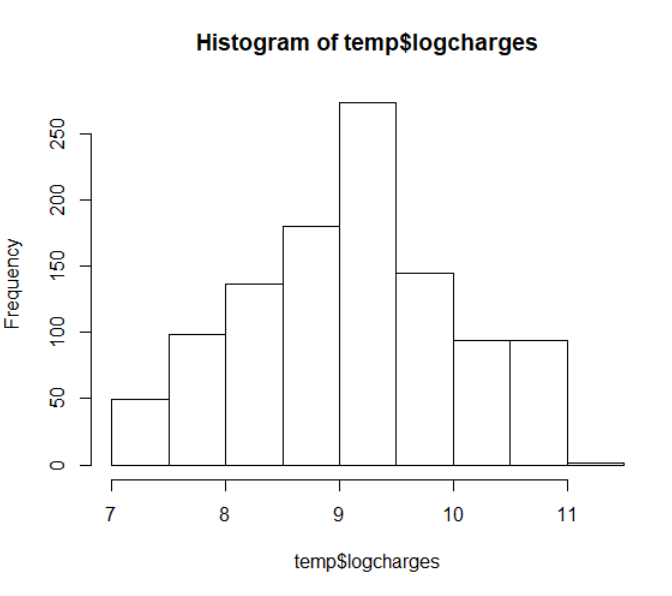
We can see that cube-root transformation has normalized response variable to a great extent



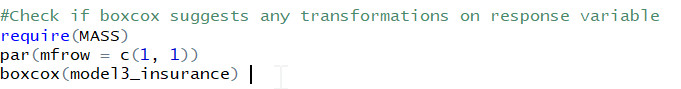
Let us see effects of log transformation on y



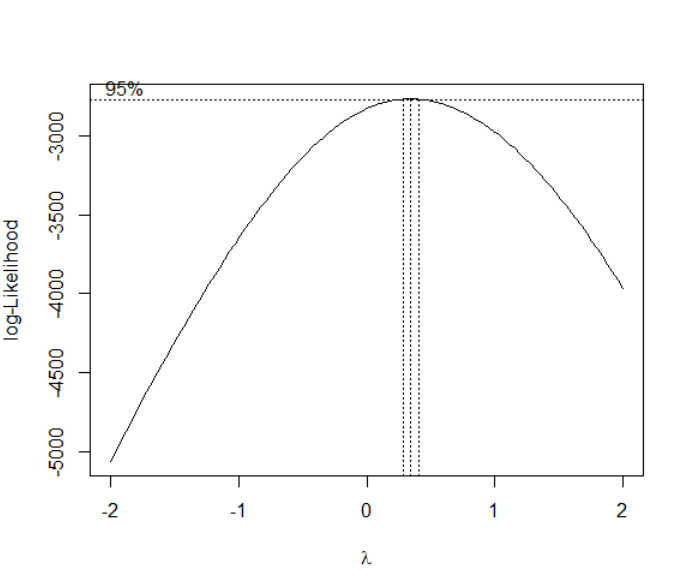
We can see that with log transformation, response variable is completely normalized.



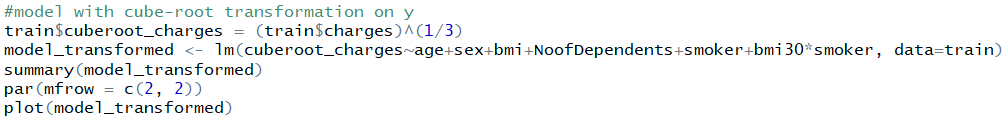
To find out the optimum transformation, let us use power transformation - boxcox method.

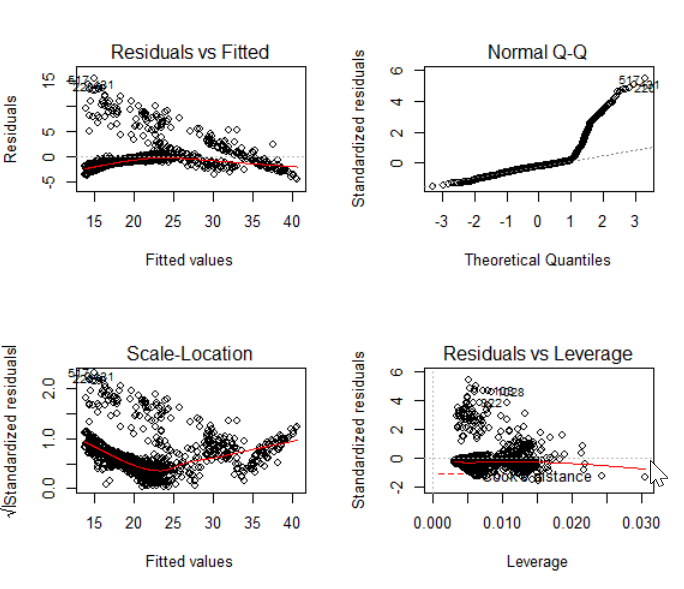


Boxcox suggests lambda = 0.35, which means cube-root transformation.



Let’s create a new model with cube-root transformation.



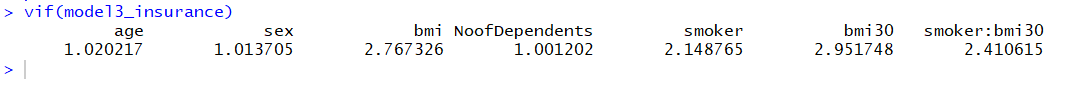


We can see that this transformation solves some normality issue but highly added unequal variance. These transformations on y is also decreasing the value of adjusted R-square. So, we will not go with these transformations.

# Check for multicollinearity

We will use the Variance Inflation Factor to diagnose the multicollinearity between covariates.





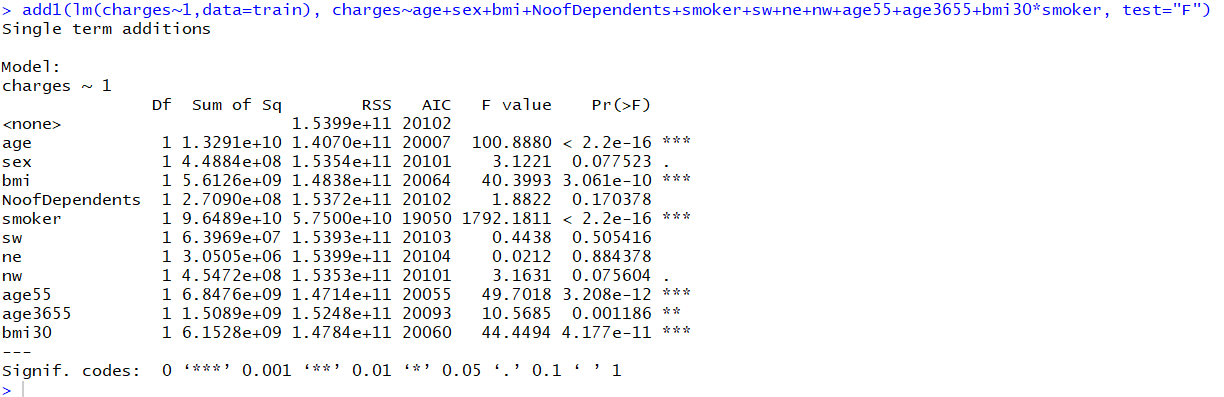
Comparing with the benchmark VIF of 10, the values here are small and so there is no multicollinearity in this model.

# Variable Selection in Model building

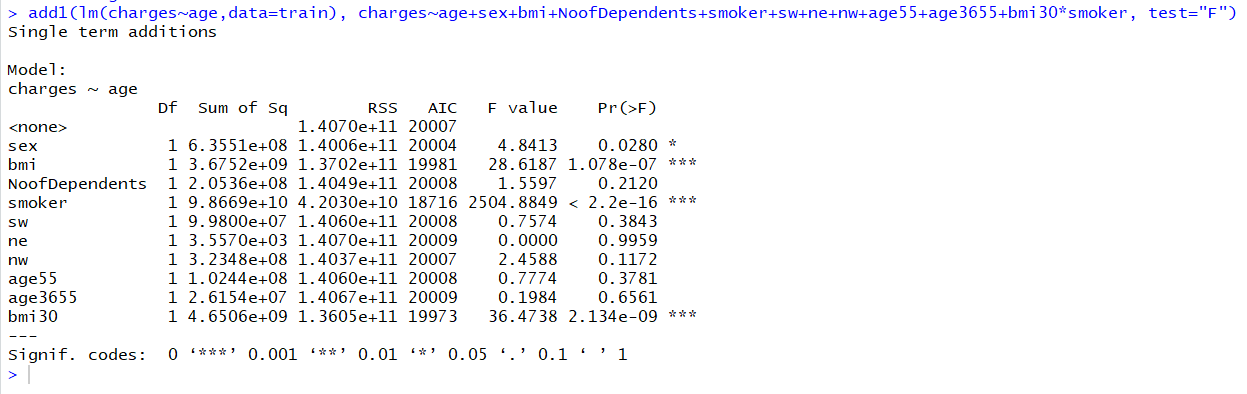
We will use **stepwise regression** method to find the optimum covariates that should be included in the model.

We will first add covariates one by one till the covariates are significant.

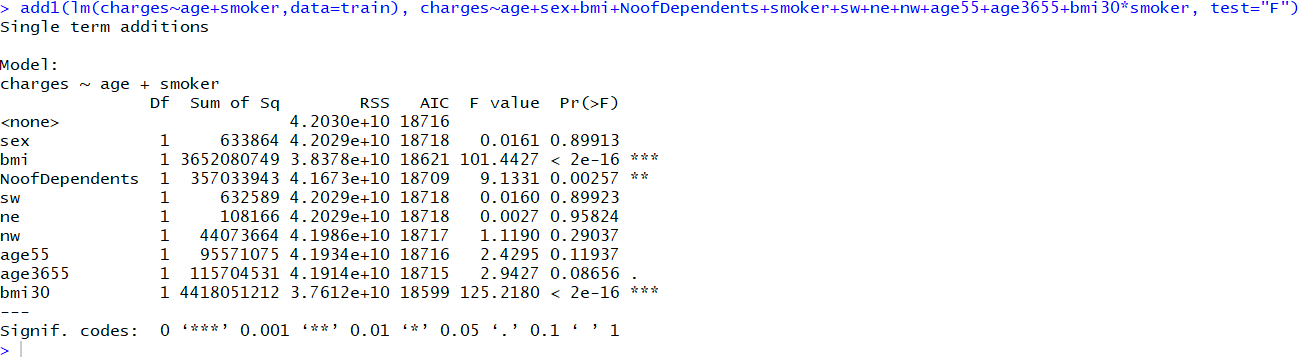
#1



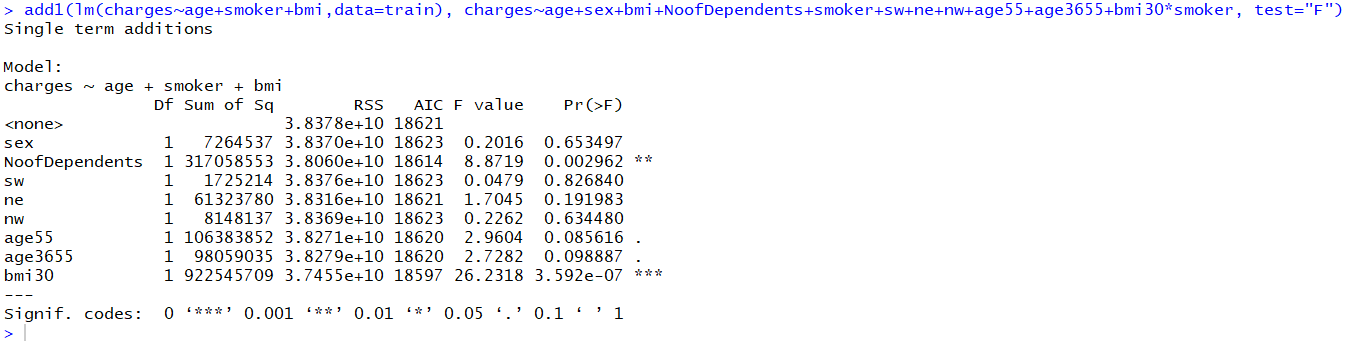
#2 Adding “age” since it is one of the most significant covariate



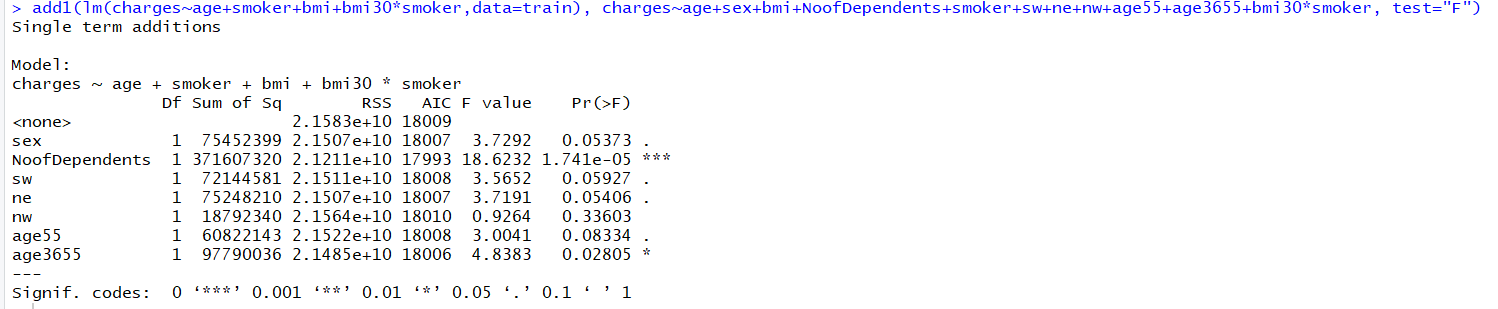
#3 Adding “smoker”



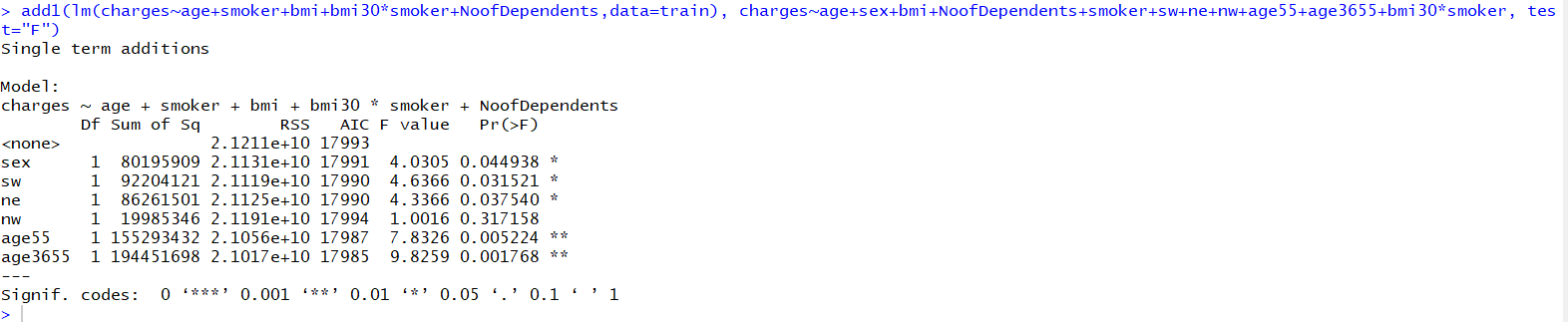
#4 Adding “bmi”

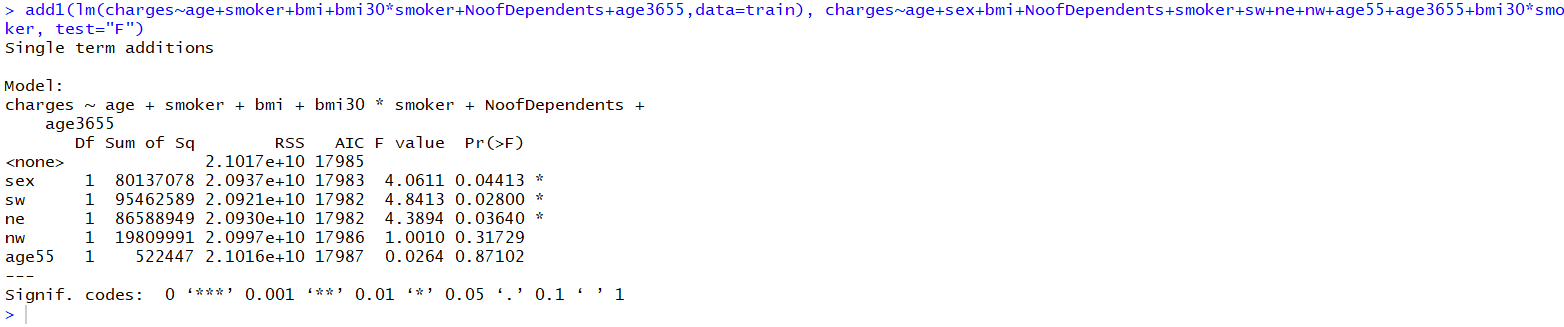


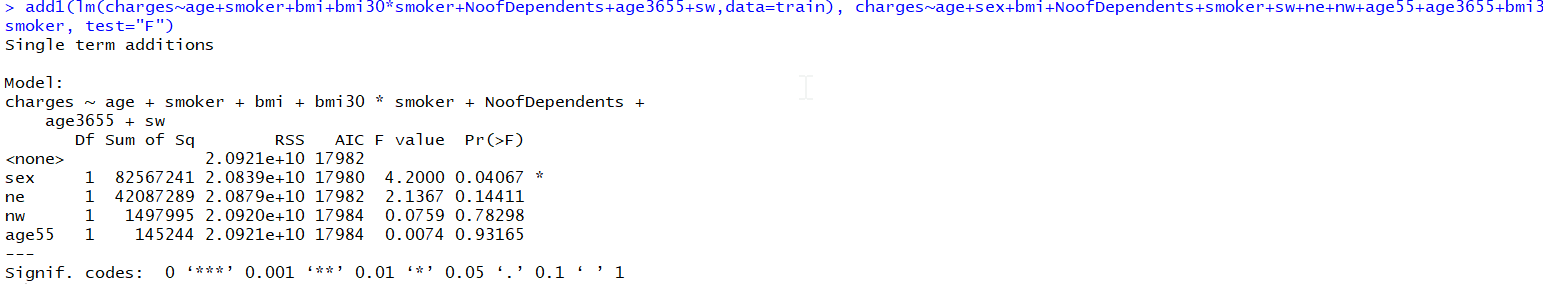
#5 Adding interaction between bmi30 and smoker



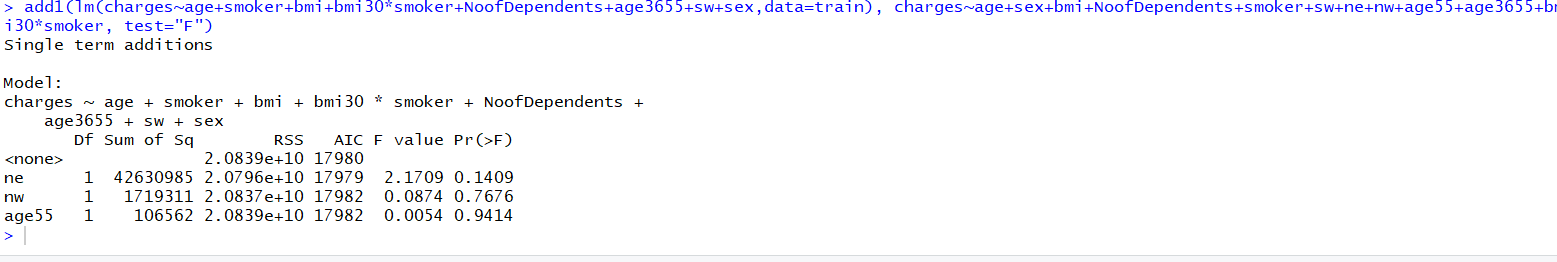
#6 Adding Noofdependents



#7 Added age3655 

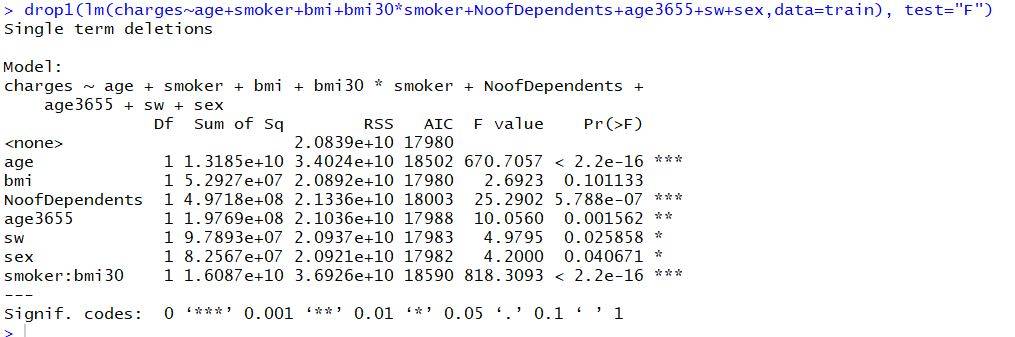
# Added sw

#Added “sex”

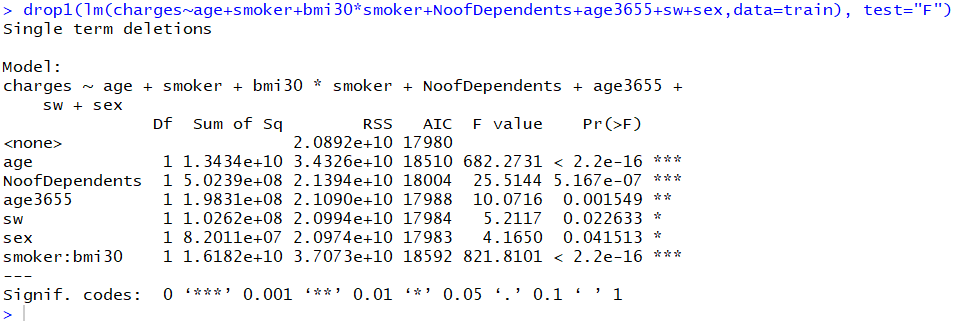


Here, we will stop adding covariates as none of them are significant.

Let’s start removing variables in descending order of significance.

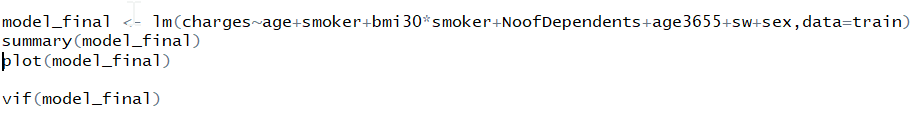


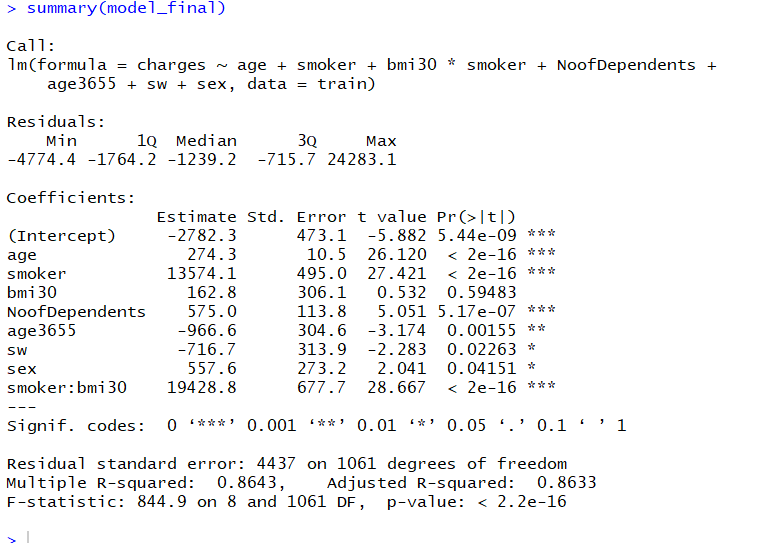
Removed “bmi” as it is significant no more



Rest all the covariates are significant and so we will not remove any more covariates.

So, let’s create the final model based on the above selection of covariates

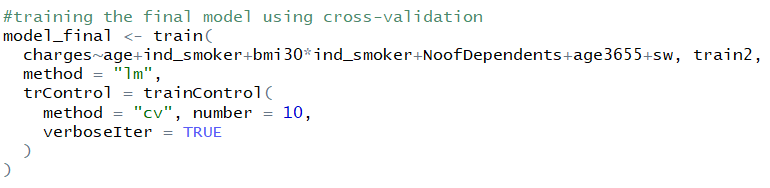


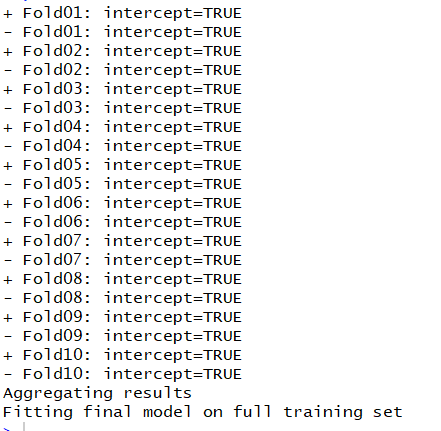


In this final model after variable selection, our adjusted R-square increased slightly from 0.8619 to 0.8633.

# Training the model using cross-validation

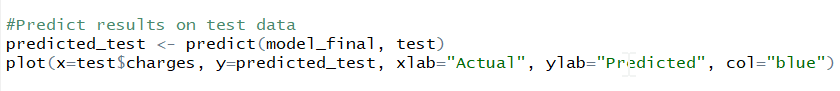
We will use cross-validation with 10 folds to train our final model on the training data.





# Predicting “charges” on test data

After all the efforts, we used the final model we created to test the “charges” from the test data we had kept aside



Let us plot Actual values vs Predicted values

