Factor Befitting for Insurance Charges

Shivi Nigam, Mihir Gandhi, Mahita Koduru, Ruchi Parab

Northeastern University

Week 6 Final Project

02/17/2019

ALY6015- Intermediate Analytics [CRN: 21011]

Instructor: Prof. Valeriy Shevchenko

Abstract

The report talks about the analysis which was performed on a dataset which contained the insurance charges for 1070 individuals. Parameters like age, sex, region, smoker etc. were some of the variables considered for the determining the cost. The first section gives a general introduction about how the project was envisioned and what are the datasets and methods used for the analysis. The second section gives a detailed explanation of the analysis which was carried out in R and the predictive model which was created for determining the cost of a random individual. The third section gives a brief conclusion of the project while answering some of the research questions mentioned before and the future scope of this analysis.

Keywords:

U.S.: United States of America

RMSE: Root Mean Squared Error

BMI: Body Mass Index

DW: Durbin-Watson test

age\_bmi: square root of the ratio of age and BMI

age.bmi: square root of the product of age and BMI

Contents

[Abstract 2](#_Toc1318786)

[Introduction 5](#_Toc1318787)

[Background and History 5](#_Toc1318788)

[Dataset 6](#_Toc1318789)

[Methods used 7](#_Toc1318790)

[Research Questions 8](#_Toc1318791)

[Analysis 9](#_Toc1318792)

[Output and Results 9](#_Toc1318793)

[Checking normality of the plot 9](#_Toc1318794)

[Checking the linearity and multicollinearity 10](#_Toc1318795)

[Creating the linear model and RMSE value 12](#_Toc1318796)

[Residual testing 14](#_Toc1318797)

[Durbin-Watson test 15](#_Toc1318798)

[Determining the multicollinearity between variables and outliers 16](#_Toc1318799)

[Fitting the train model 18](#_Toc1318800)

[Fitting the testing model 21](#_Toc1318801)

[Building a predictive model 24](#_Toc1318802)

[Conclusion and Future Scope 26](#_Toc1318803)

[Based on research questions 26](#_Toc1318804)

[Applications of Linear Regression 27](#_Toc1318805)

[Future Scope 27](#_Toc1318806)

[Reference 29](#_Toc1318807)

[Appendix 31](#_Toc1318808)

[Source Code 31](#_Toc1318809)

# Introduction

## Background and History

The insurance industry is going through a rapid phase of transformation and embracing changes is the only way to move forward in this competitive market. Insurance industry at-a-glance. U.S. insurance industry net premiums written totalled $1.2 trillion in 2017, with premiums recorded by life/health (L/H) insurers accounting for 52 percent. The implementation of new healthcare reforms and cost structures is making it difficult for health insurance companies to meet the rising demand for insurance coverage while keeping their costs low. Consequently, many health insurance providers have started leveraging strategic tools such as insurance analytics to reduce costs and streamline processes. When we look at the statistics of the health insurance in USA, in 2016, the per­centage of people without health insurance coverage for the entire calendar year was 8.8 percent, or 28.1 million. The major barrier to coverage for the uninsured is the high cost. We have focused on point of high insurance cost barrier for the uninsured in this project and worked on building an accurate cost predictor model by conducting analysis on health insurance data set.

Medical expenses are tricky to estimate because the costliest conditions are rare and seemingly random. Still, some conditions are more prevalent for certain segments of the population that helped us in the analysis. For instance, lung cancer is more likely among smokers than non-smokers, and heart disease may be more likely among the obese. [BusinessWire, 2018]

We have done analysis on the healthcare dataset to identify the influence of each factor on cost of the insurance and build a model to estimate accurate healthcare costs using prior demographics and diagnoses. The goal of this analysis is to build an accurate health insurance rate-setting model. This model can help reduce the count of the uninsured people due to high insurance cost and can be used to create actuarial tables that set the price of yearly premiums higher or lower, depending on the expected treatment costs.

## Dataset

The dataset we have used in this project was retrieved from Kaggle which contained 1070 cases of the patients and the insurance charges for each one of them. Few basic factors like age, sex, region etc. are also mentioned. Below mentioned are all the variables of the dataset and a brief description about them.

Our dataset has 7 variables, which consists of 6 predictor variables and 1 target variable. The 6 Predictor variables are:

1. Age: The age of the primary beneficiary. It is a continuous variable

2. Sex: The insurance contractor’s gender. It is a categorical variable with two categories which are male and female.

3. BMI: The Body mass index, provides an understanding of the body, weights that are relatively high or low relative to height. It is a continuous variable.

4. Children: The number of children covered by health insurance or we can also consider it as the number of dependents. It is a continuous variable.

5. Smoker: This variable depends upon whether the person is a smoker or not. It is a categorical variable which has two categories, namely, Yes which means that the person smokes and No which means that the person does not smoke.

6. Region: The beneficiary’s residential area in the US. It is a categorical variable with four categories, which are, northeast, southeast, southwest, northwest.

The dependent or the target variable is:

* Charges: The Individual medical costs billed by health insurance

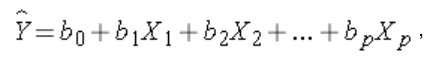
## Methods used

We have worked our way through “Factor Befitting for Insurance charges” by implementing one of the most recurrently used regression analysis technique named Linear Regression Analysis. Linear Regression is a technique of modeling a target value based on the independent predictors. This method is mostly used for forecasting and discover the out cause and effect relationship between variables. Regression techniques mostly differ based on the number of independent or predictor variables and the type of relationship between the independent or predictor and dependent or target variables. Since we have more than one predictor variable we will use the Multiple Linear Regression Technique. Multiple Linear Regression the most common form of linear regression analysis.

As a predictive analysis, the multiple linear regression is used to describe the relationship between one continuous dependent variable and two or more independent variables.

The independent variables can be continuous or categorical.

The mathematical representation for multiple linear regression is –



Where,

Y is the target variable.

X1 to XP are the predictor variables.

B1 to Bp are the estimated regression coefficients.

The Assumptions which must be considered while creating the linear regression model are:

1. There should be a linear and additive relationship between the dependent (response) variable and the independent (predictor) variable(s). Once the model has been created we will plot a graph between “charges” and the suitable predictor variables to justify the linearity.
2. There should be no correlation between the residual (error) terms. The absence of this phenomenon is known as Autocorrelation.
3. The independent variables should not be correlated. The absence of this phenomenon is known as multicollinearity.
4. Heteroskedasticity, thepresence of outliers or extreme leverage values should be avoided while modeling a linear regression.

Thus, considering these assumptions we carried out the analysis of the data set which is explained in the next section.

## Research Questions

With the help of this analysis, we were able to answer some research questions like:

1. Which factors contribute to the reduction of insurance costs?
2. Does splitting the dataset into two parts: training model and testing model affect the efficiency of the model?
3. How will the model benefit insurance companies and consumers?
4. What if the input values in the predictive model are out of range?

These questions are answered by the end of the report in the conclusion of the report and supporting evidence is also provided for the explanation of these answers.

For further reference, the R file, the dataset and the source code text file are attached below:



# Analysis

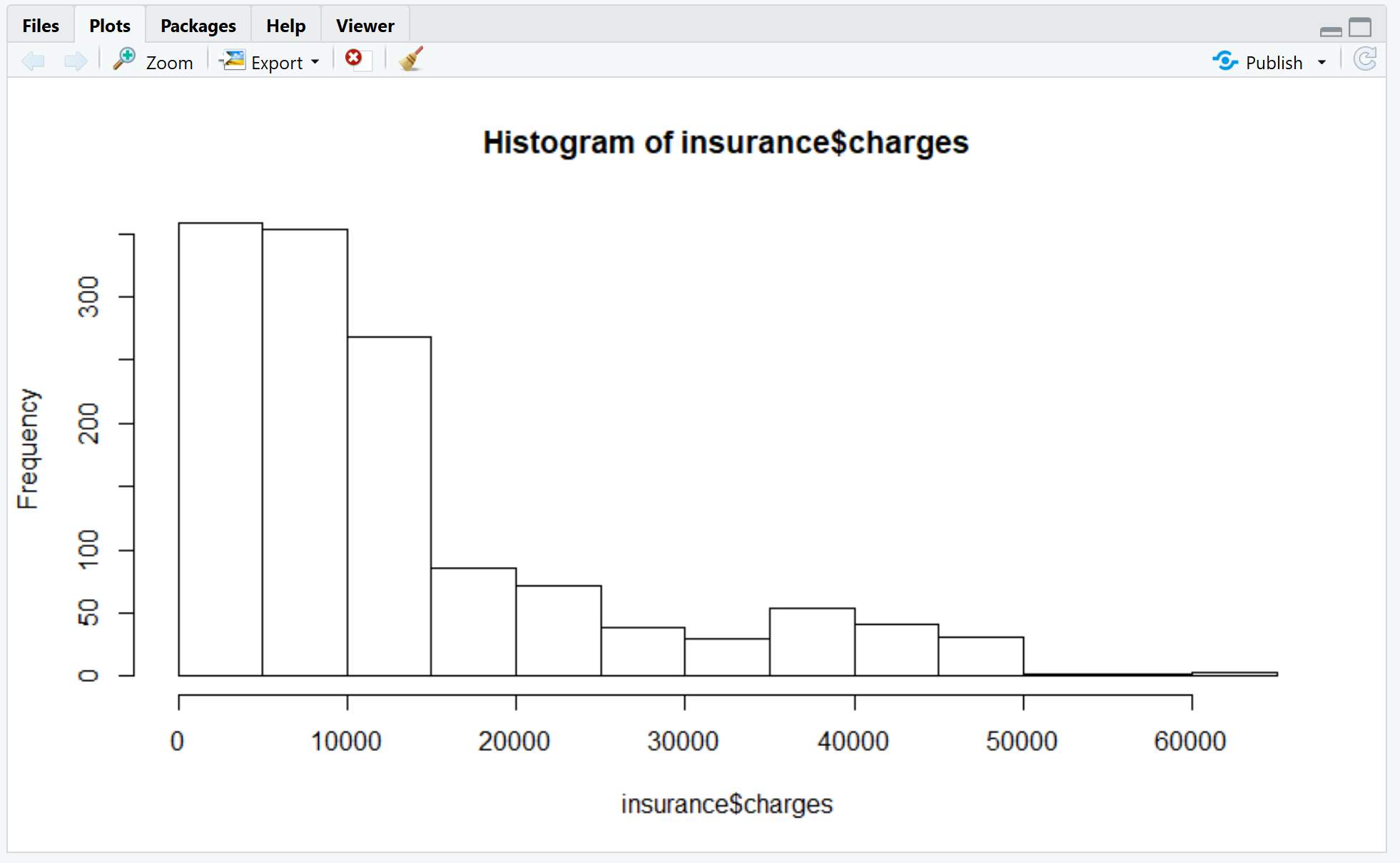
The libraries used in our analysis included:

1. “ggplot2” – used for graphical and data visualization techniques.
2. “MASS” – used to calculate the residuals of the given data.
3. “lmtest” – used for testing the linear regression model.
4. “car” – used for applied regression techniques.

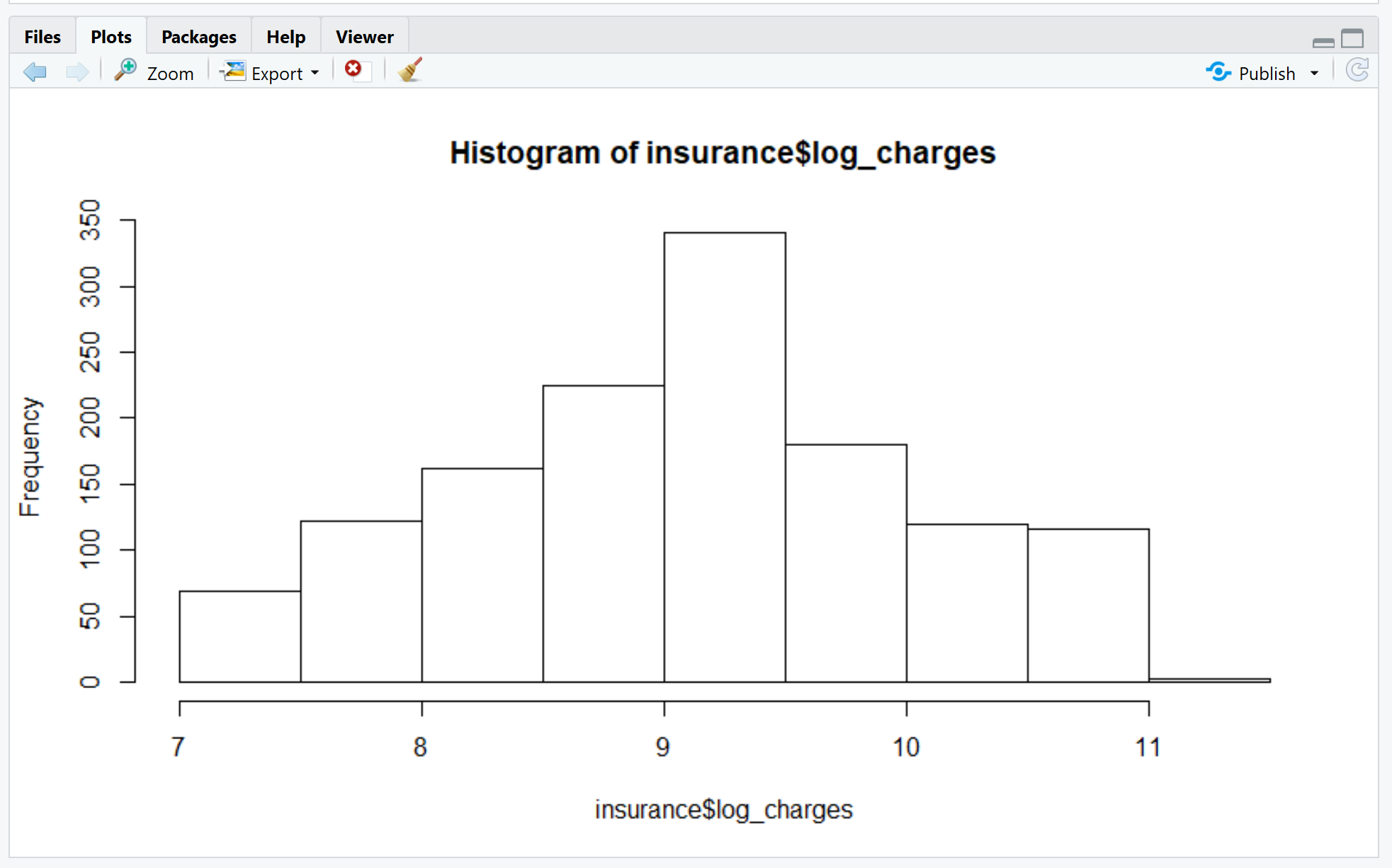
## Output and Results

### Checking normality of the plot

The data when initially plotted with the help of a histogram showed that the data was right skewed (i.e. most of the values were concentrated towards 0). The histogram obtained was:



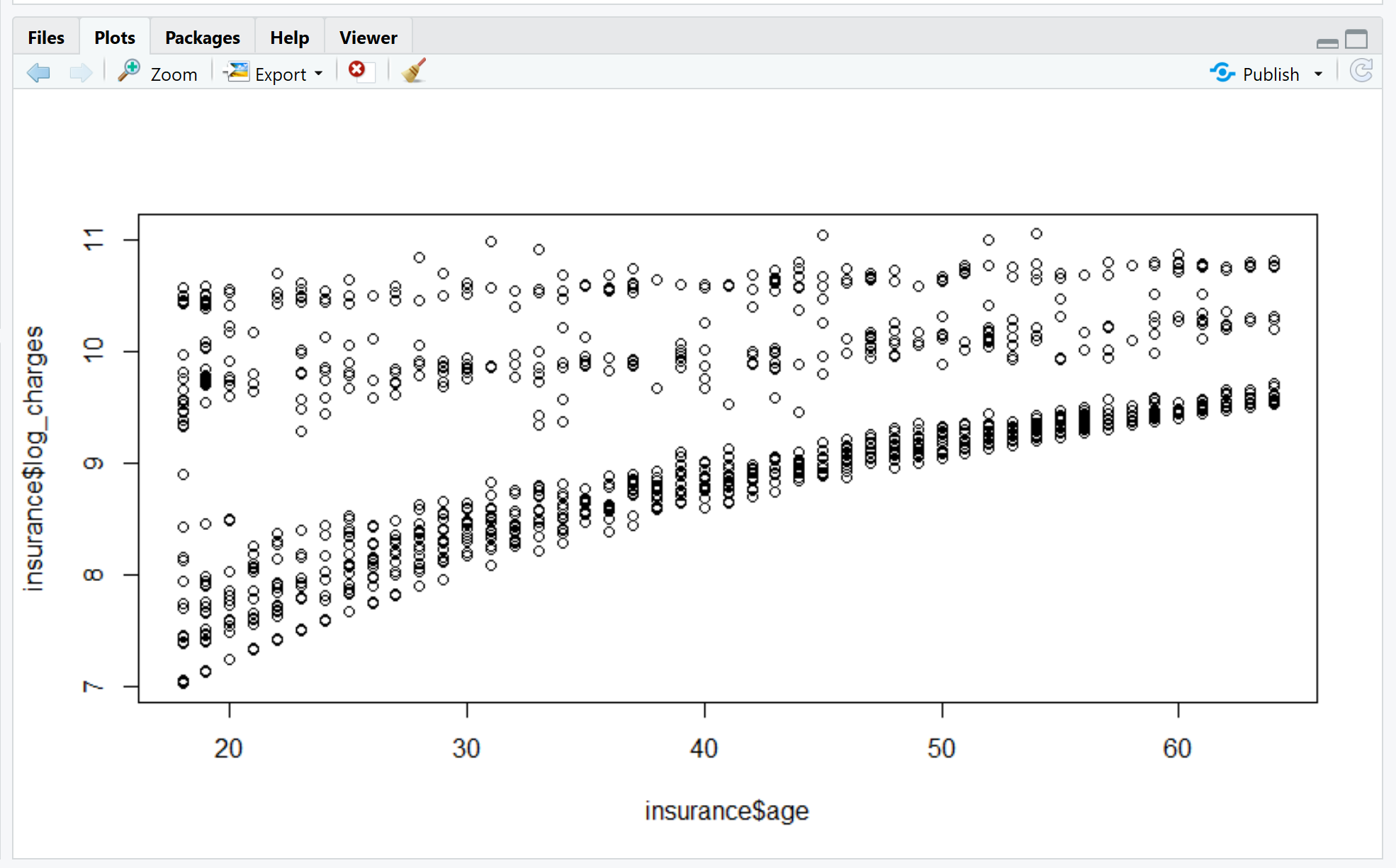
Thus, to normalize the data, log transformation was applied to each observation and the results obtained are shown in the plot below:



Thus, with the help of log transformation, we were able to normalize the data around the center which would be essential for us in the further analysis.

### Checking the linearity and multicollinearity

We checked the linearity of the two parameters age (independent variable) and the log of the charges (dependent variable). The output obtained was as follows:

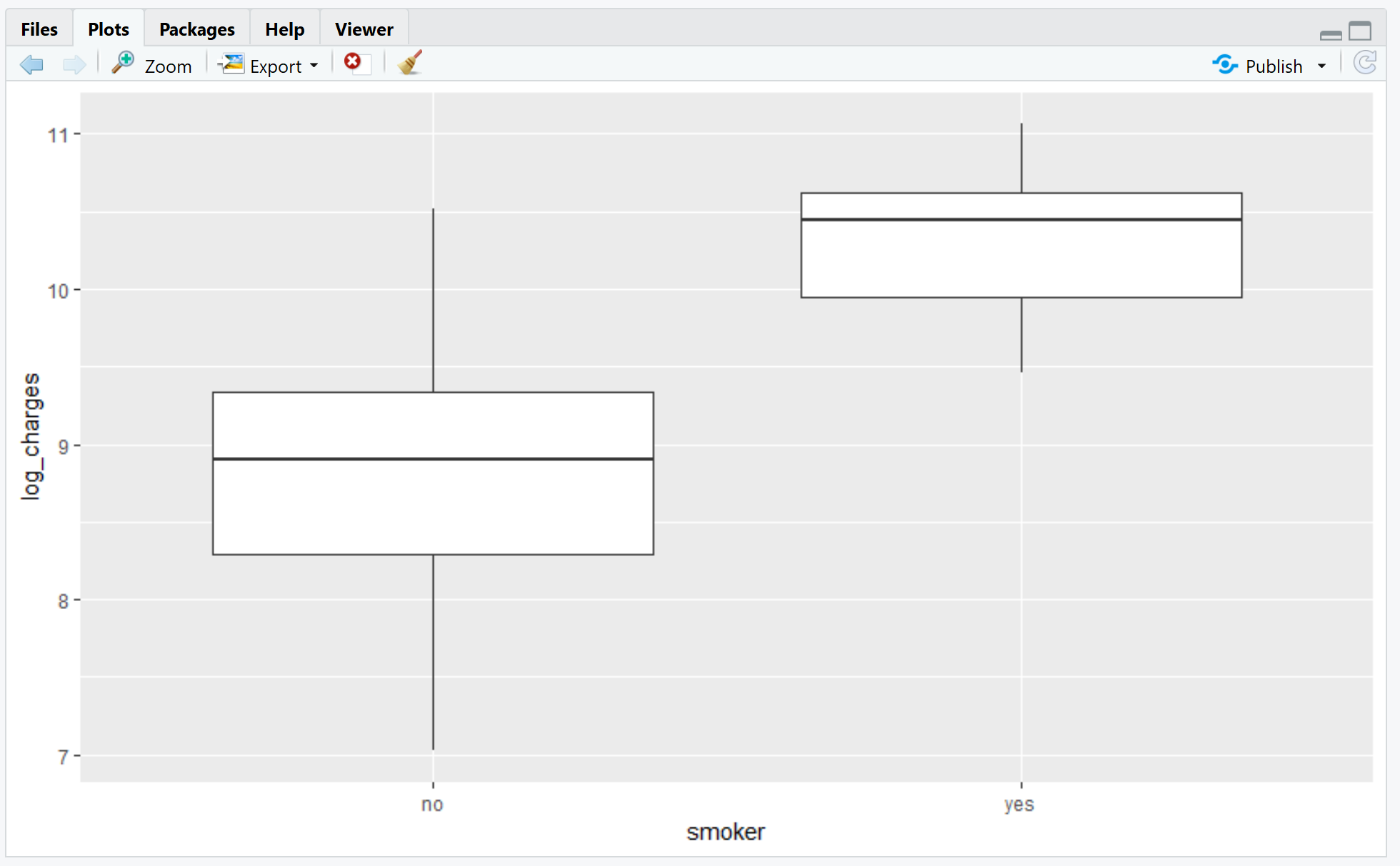


Thus, we inferred that there was some positive collinearity between the age and the charges. We observed that with an increase in age, the charge of insurance also increased. The correlation coefficient was found out to be 0.30 which suggests that the parameters are moderately correlated.

Now, to check for the multicollinearity, we considered two variables age and BMI. Age was considered as the input variable. The value obtained by computing the correlation was 0.1092.

This showed that the two values on their own had close to no correlation as the value was tend towards 0.

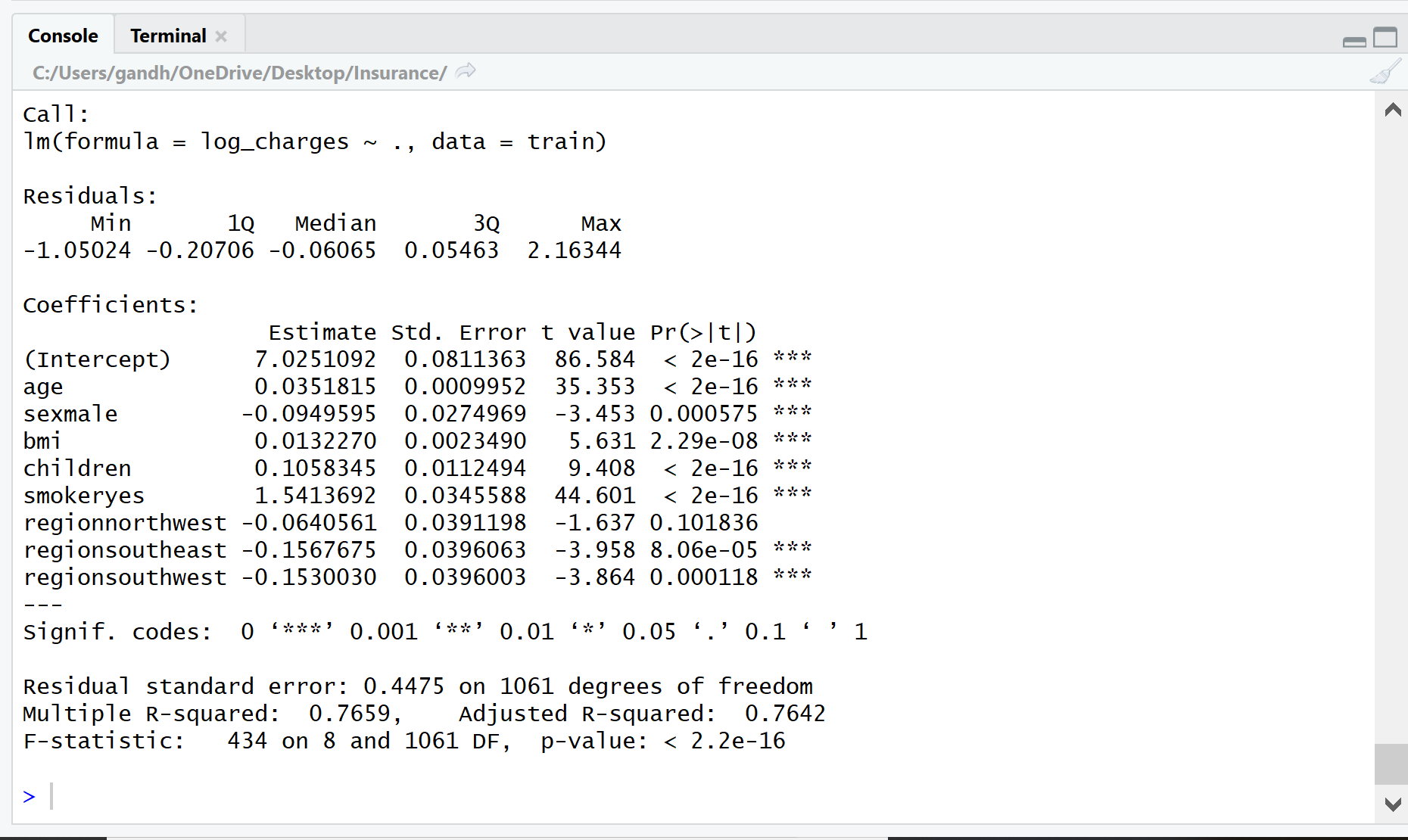
Furthermore, we decided to check the effect of smoking on the insurance charges. The analysis gave rise to a boxplot and the plot is shown below:



Thus, from the graph we can conclude that the median of charges for non-smokers is comparatively lower than the median of charges for the smokers. This means that, a person who has a habit of smoking will generally end up paying more than the person who is not.

### Creating the linear model and RMSE value

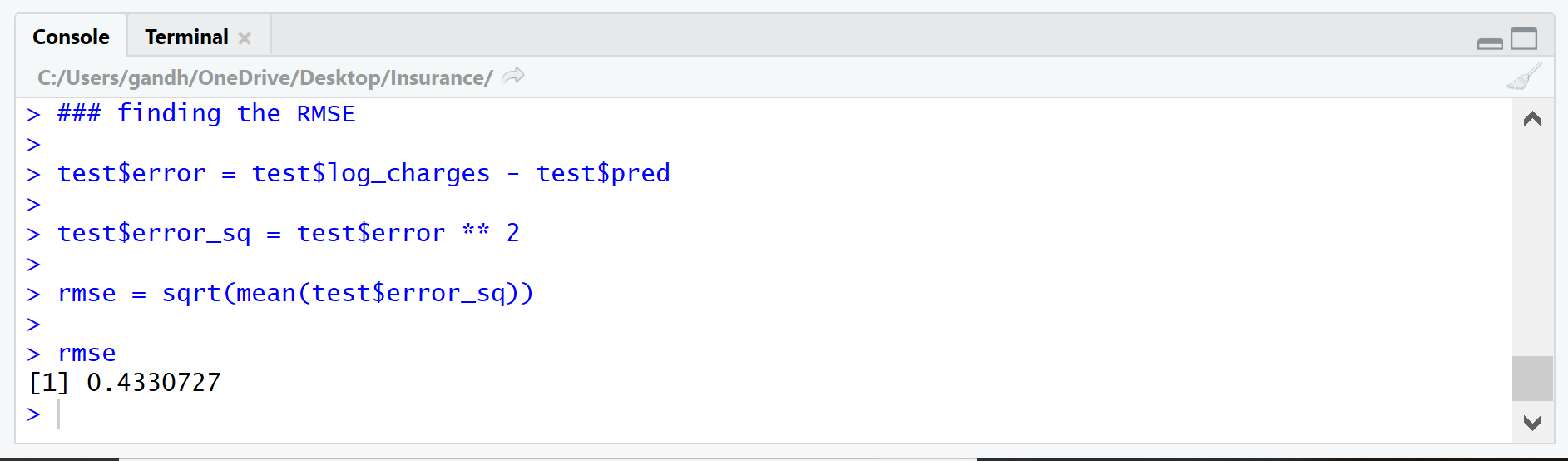
The linear model was created by dividing the datasets into two parts: Train model and the test model. Train model consisted of 1070 observations where the test model consisted of 268 observations. The summary of the linear model was created with respect to the train model and the outputs are shown below:



On observing the values of residual standard error, F-statistic and the p-value, we interpreted that the variables have a statistical significance with each other as the residual standard error and the p-value are tending towards 0 and the F-statistic is positive and greater than 0. The R-squared value was 0.7659 which meant that the accuracy was 76.59%.

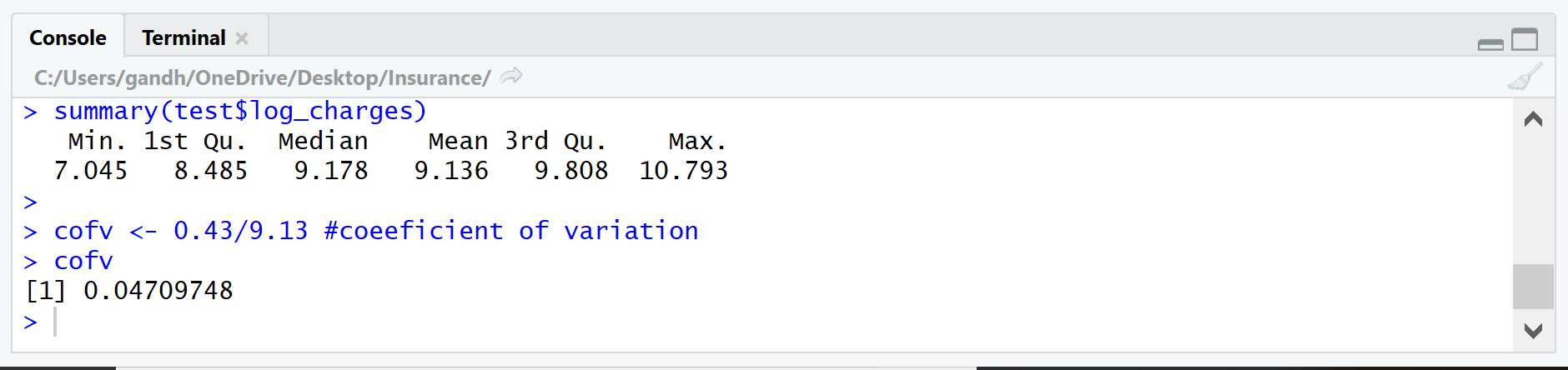
RMSE value (Root mean squared error):

RMSE value is used to determine how far away the residuals are from the regression line. It also states how concentrated is the data around the line of best fit of our model. RMSE value was calculated by taking the square root of the mean of the error value and the results obtained are shown below.



Thus, as we obtained a value of 0.433, we inferred that, the data was quite concentrated around the line of best fit as the value of RMSE is quite small and tending towards 0.

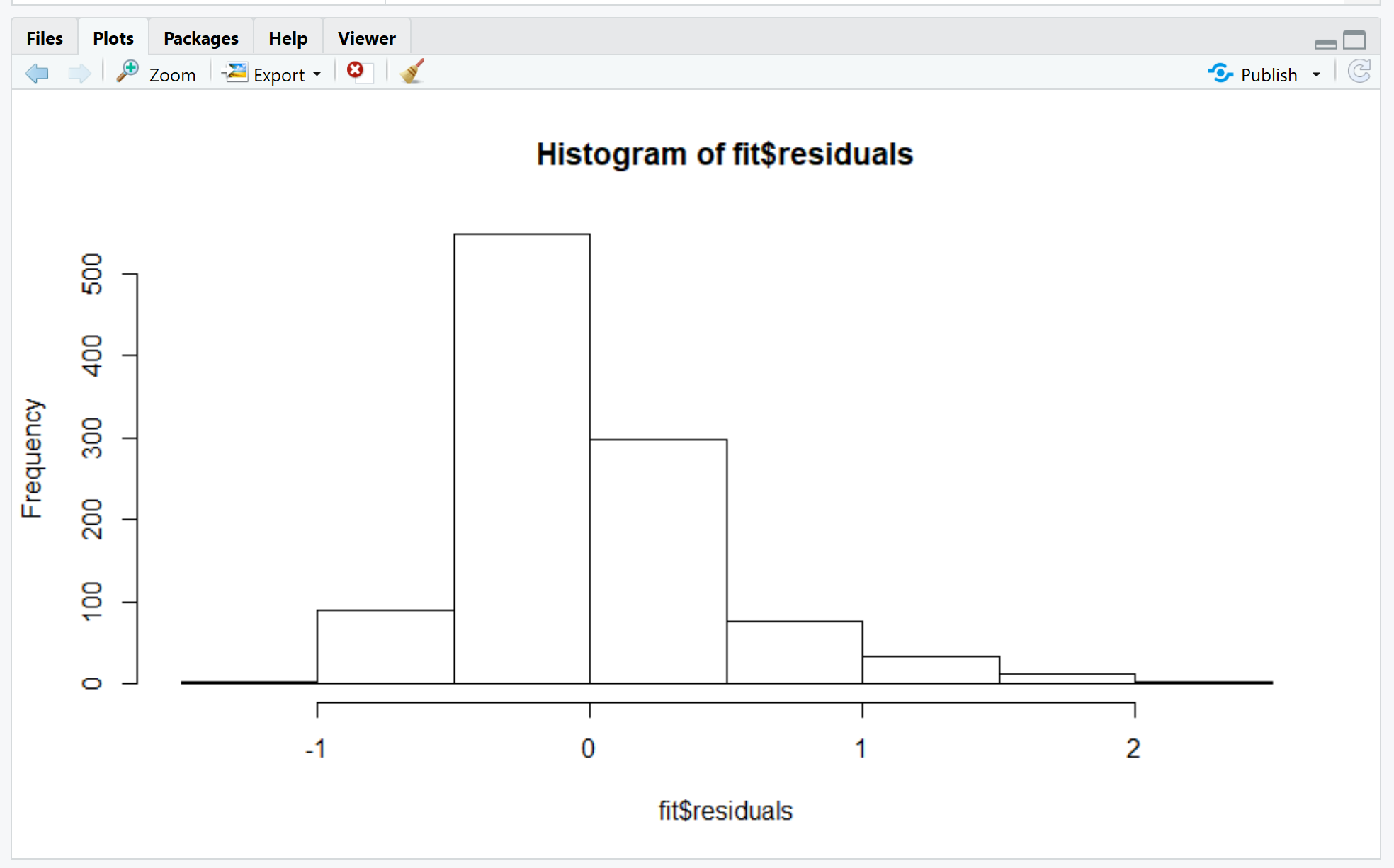
Moreover, we calculated the coefficient of variation by dividing the RMSE value to the mean of the data. The results obtained were:



Thus, we concluded that, there is quite low dispersion of the variables in the dataset as the coefficient of variation was 4.7%.

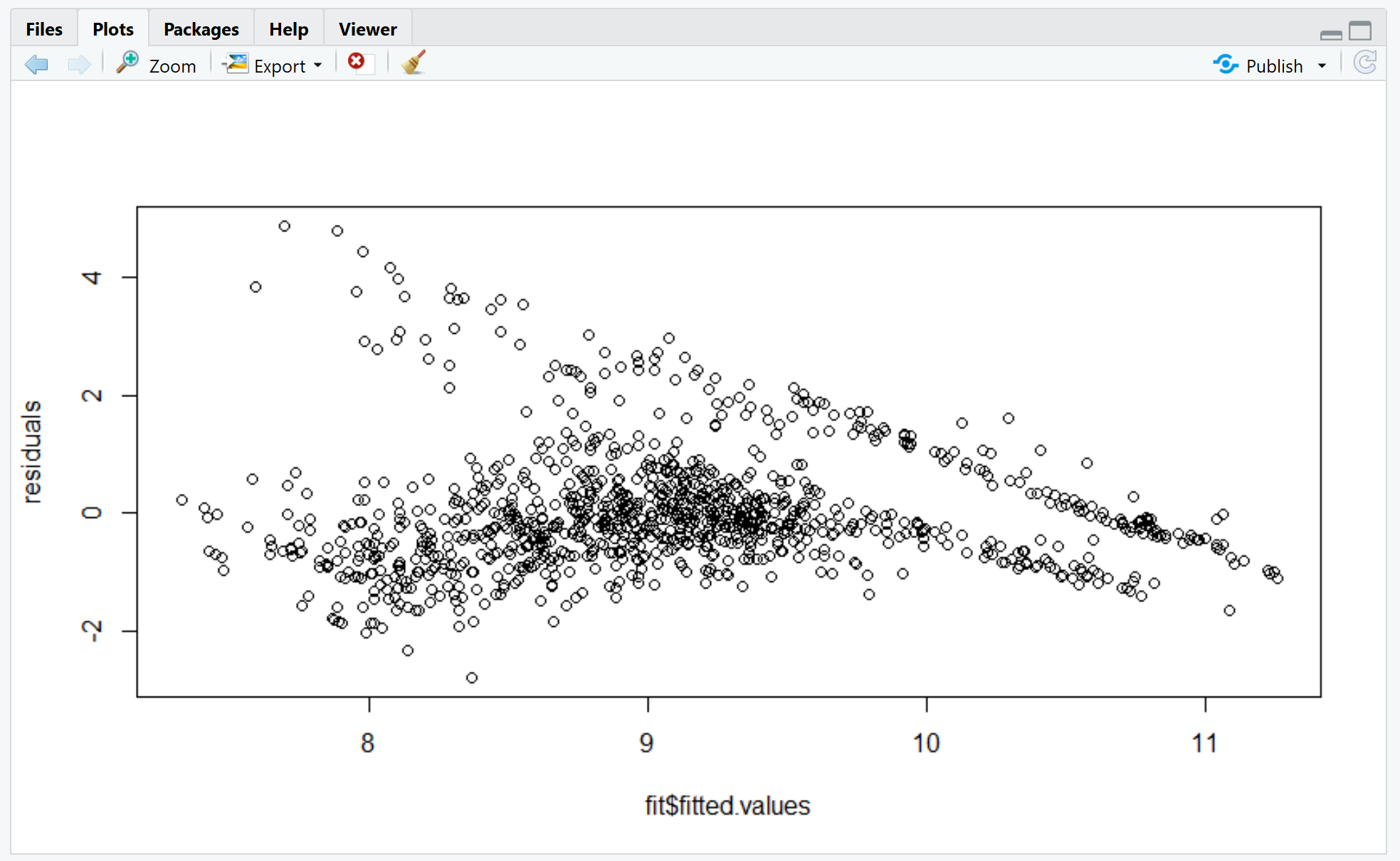
### Residual testing

From the fitted model of the data, the residuals were calculated and the graph was plotted as shown below:



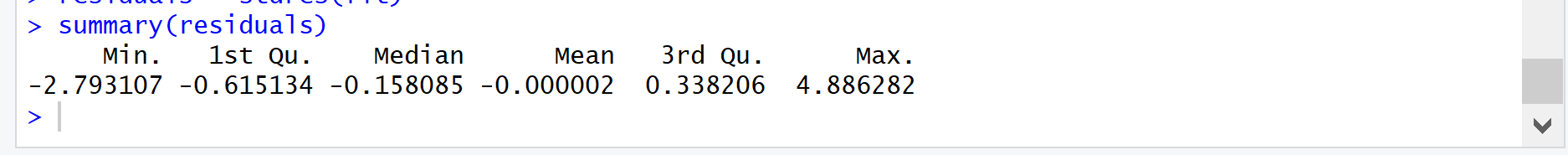
We observed from the plot that, most of the data is concentrated near the value 0 which showed that, there is not much variability in the given data. We also gained evidence on the low variability of the data by calculating the coefficient of variance in Point 3 of the report.

A graph was created which shows a relation between the fitted values and the residuals of the observations. The plot is shown below:



As we can see, most of the fitted values lie close to the 0 value of the residuals which suggests that the data has low variability and all the predicted values tend to lie close to the best fit line of the model.

The summary of the residuals was calculated as shown below:

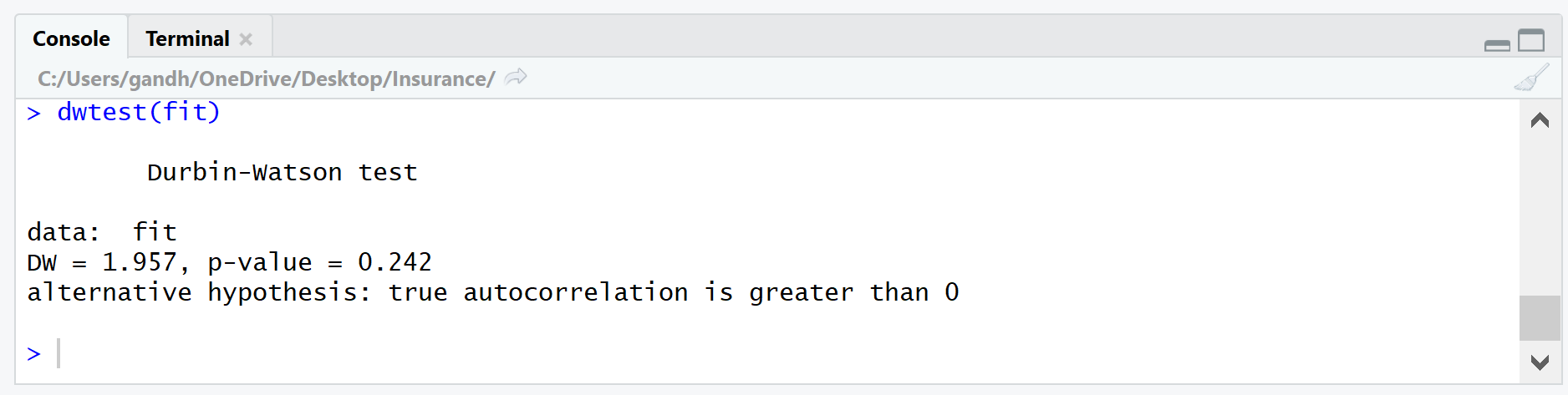


The values of mean and median go on to show the same inference we got from the histogram and the scatterplot of the residuals.

### Durbin-Watson test

The Durbin-Watson test was carried out to test the autocorrelation of the residuals of the given observations.

The results from the test are shown below:

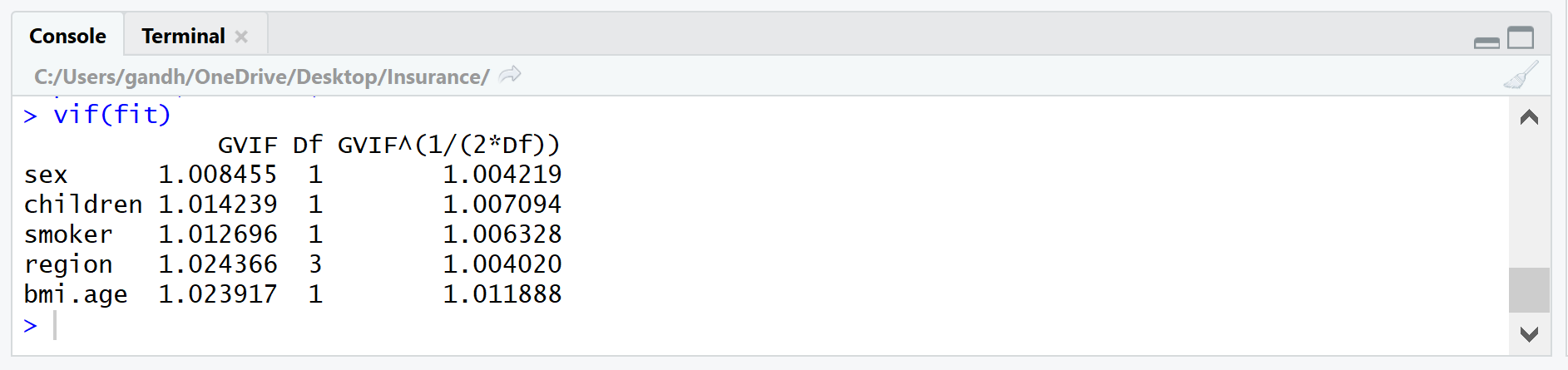


We had performed the test on the linear model named “fit” which we had created. Our DW value which was 1.95 (very close to 2), which was the level of acceptance to justify the absence of auto-correlation in data. The default null hypothesis for the DW test was that there was no auto-correlation between the independent variables. Since, our P-value which was 0.242, was greater than 0.05(default level of significance) we failed to reject the null hypothesis and thus we can state that our independent variables were not correlated. [Samuels, 2019]

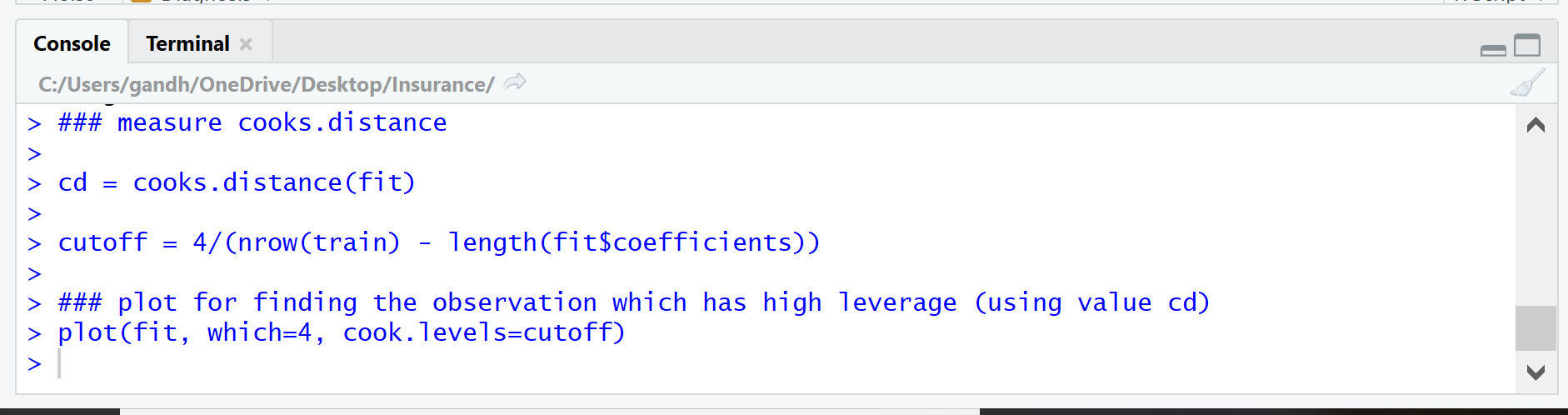
### Determining the multicollinearity between variables and outliers

The independent variables should not be correlated. The absence of this phenomenon is known as multicollinearity. To prove the multicollinearity between the 6 independent variables, we used the vif () function in R. This function determines how much the variance of the estimated regression gets effected due to the multicollinearity amongst the predictor variables. If the VIF factor had a value greater than 10, meant that there were high chances of multicollinearity and hence we cannot proceed with the desired regression model. [Heckman, 2019]

The results obtained were:

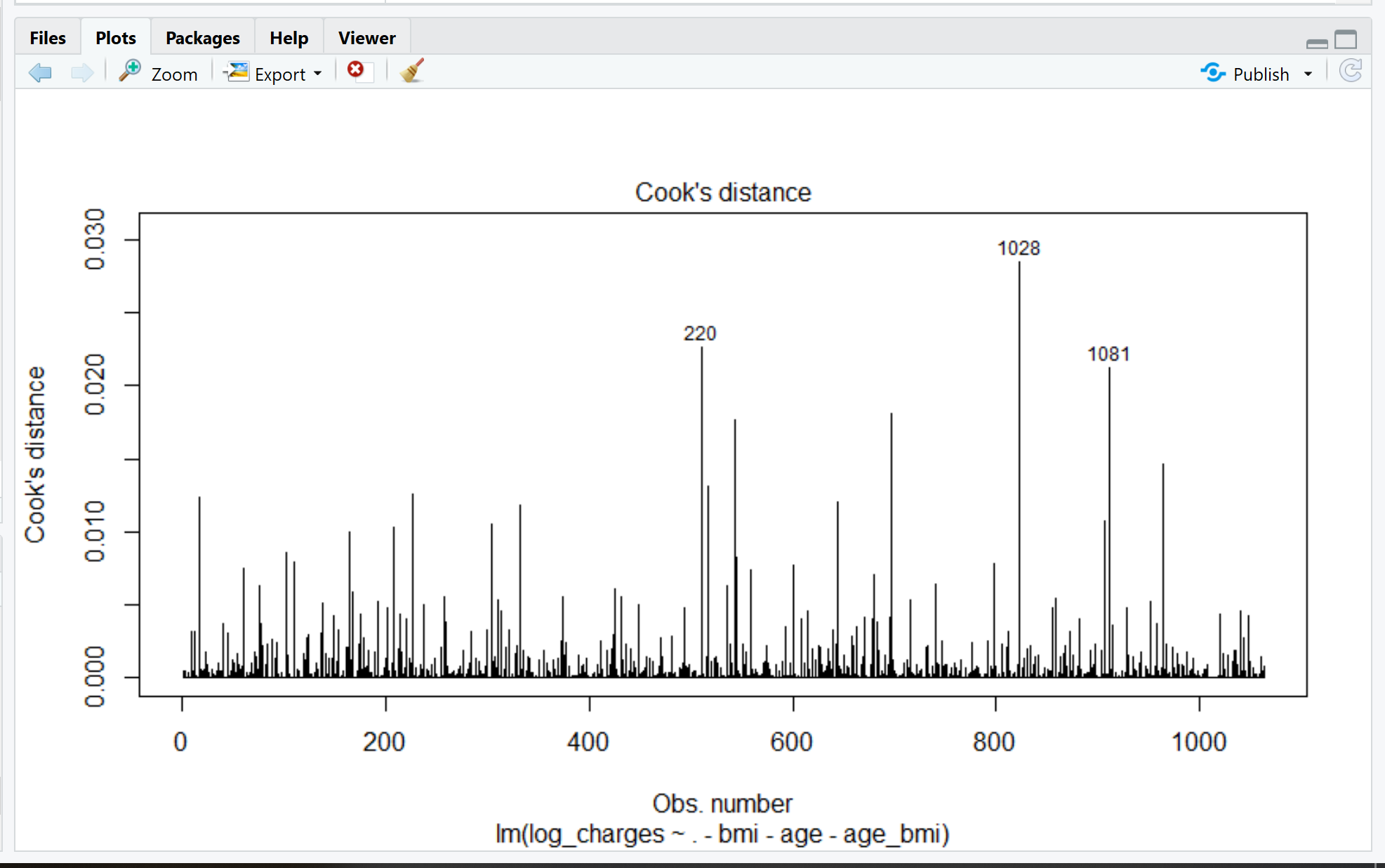


From the output, we can infer that all the predictor variables have their VIF factors ranging from 1.008 to 1.02, which denotes that our variables do not depict the presence of multicollinearity and hence we can proceed with the linear regression model.

Heteroskedasticity, the presence of outliers or extreme leverage values should be avoided while modeling a linear regression. We can test the phenomenon of heteroskedasticity by performing the Breusch-Pagan / Cook – Weisberg test or White general test. For our dataset, we had performed the Cook – Weisberg test. The outputs from test is shown below: 

We had calculated the cook's distance for the model which we had created named “fit”. The cut-off for the outliers was calculated by using the formula shown in the R-snippet. If the values go above the cutoff they were considered as outliers.

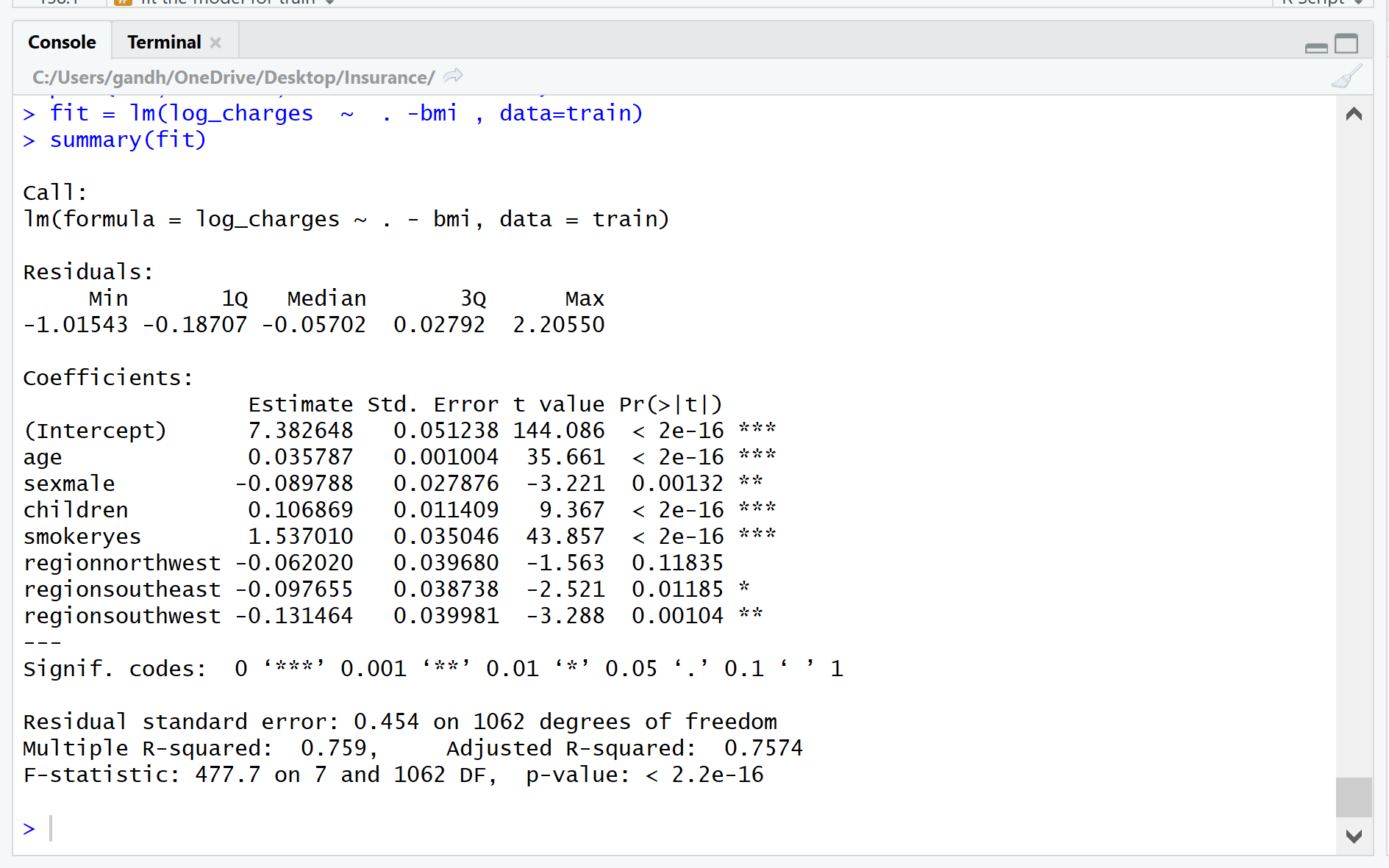
Cook’s plot:



The value of the cutoff was 0.003, thus the values which have the cook’s distance greater than 0.003 were considered as outliers. From the graph, it can be inferred that there were a few observations which were considered as the outliers. Thus, the observations were removed from the training dataset to get the accurate results.

### Fitting the train model

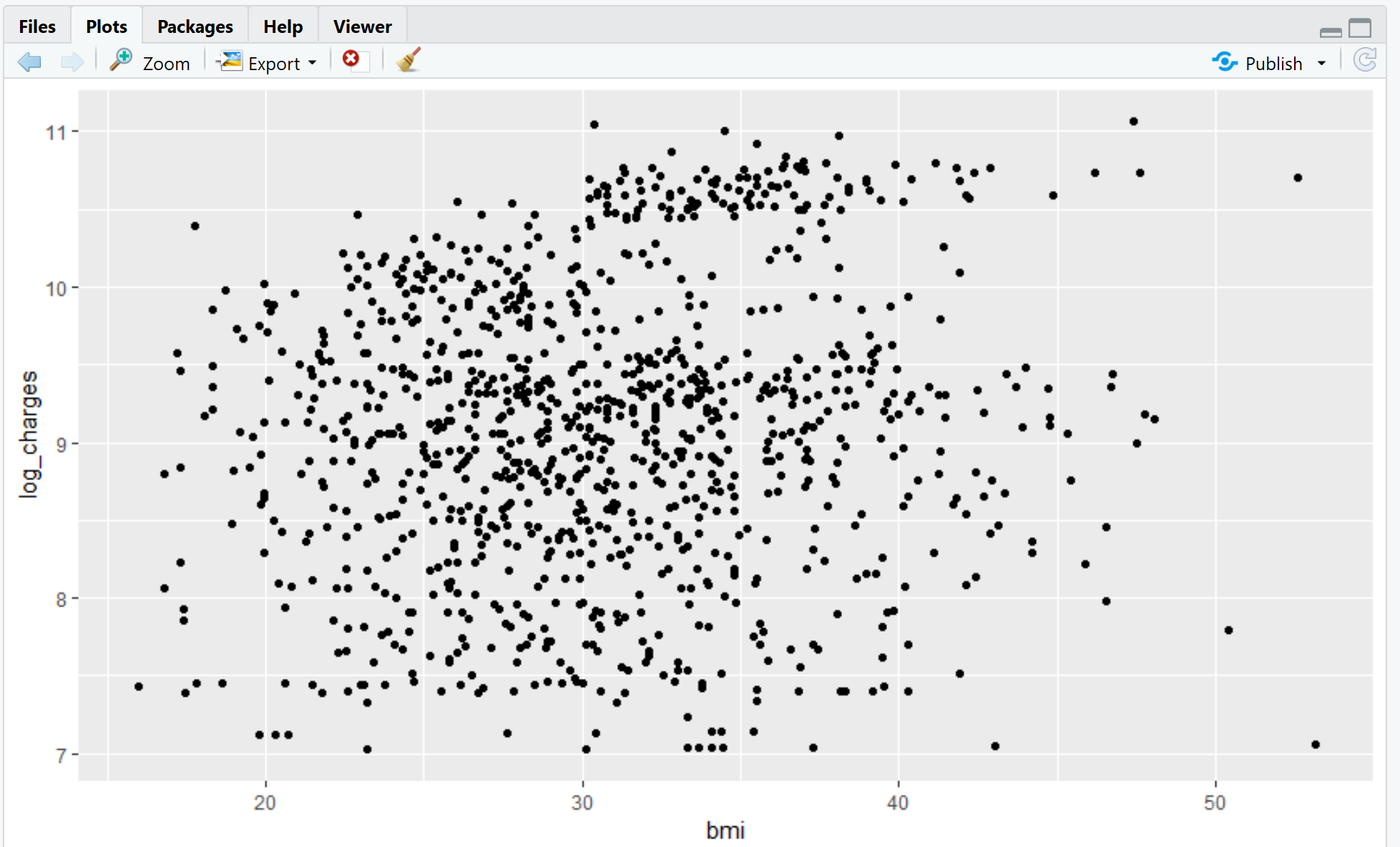
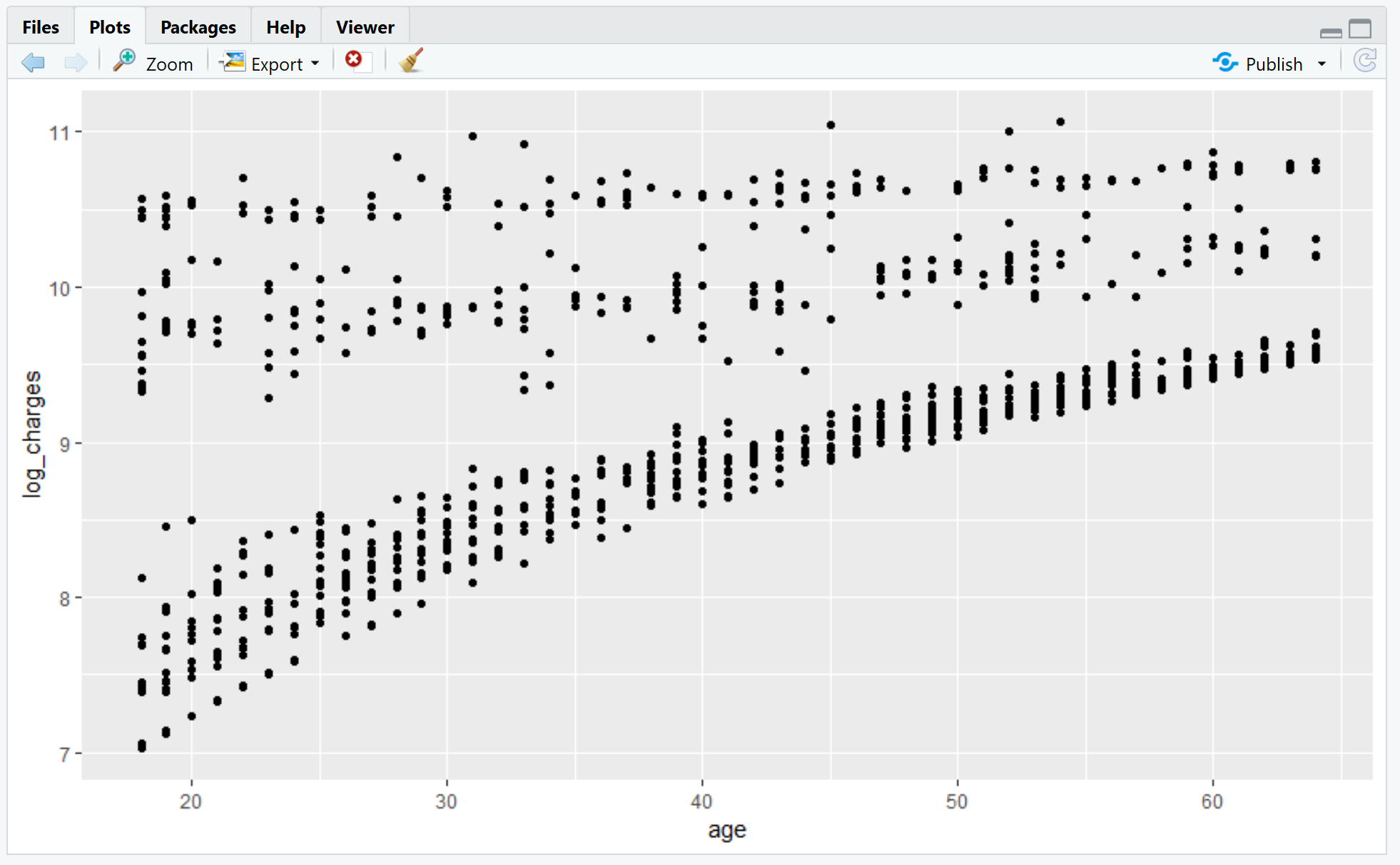
The train model was created with 1070 observations and the outliers were removed from the dataset. The summary of the dataset is given below:



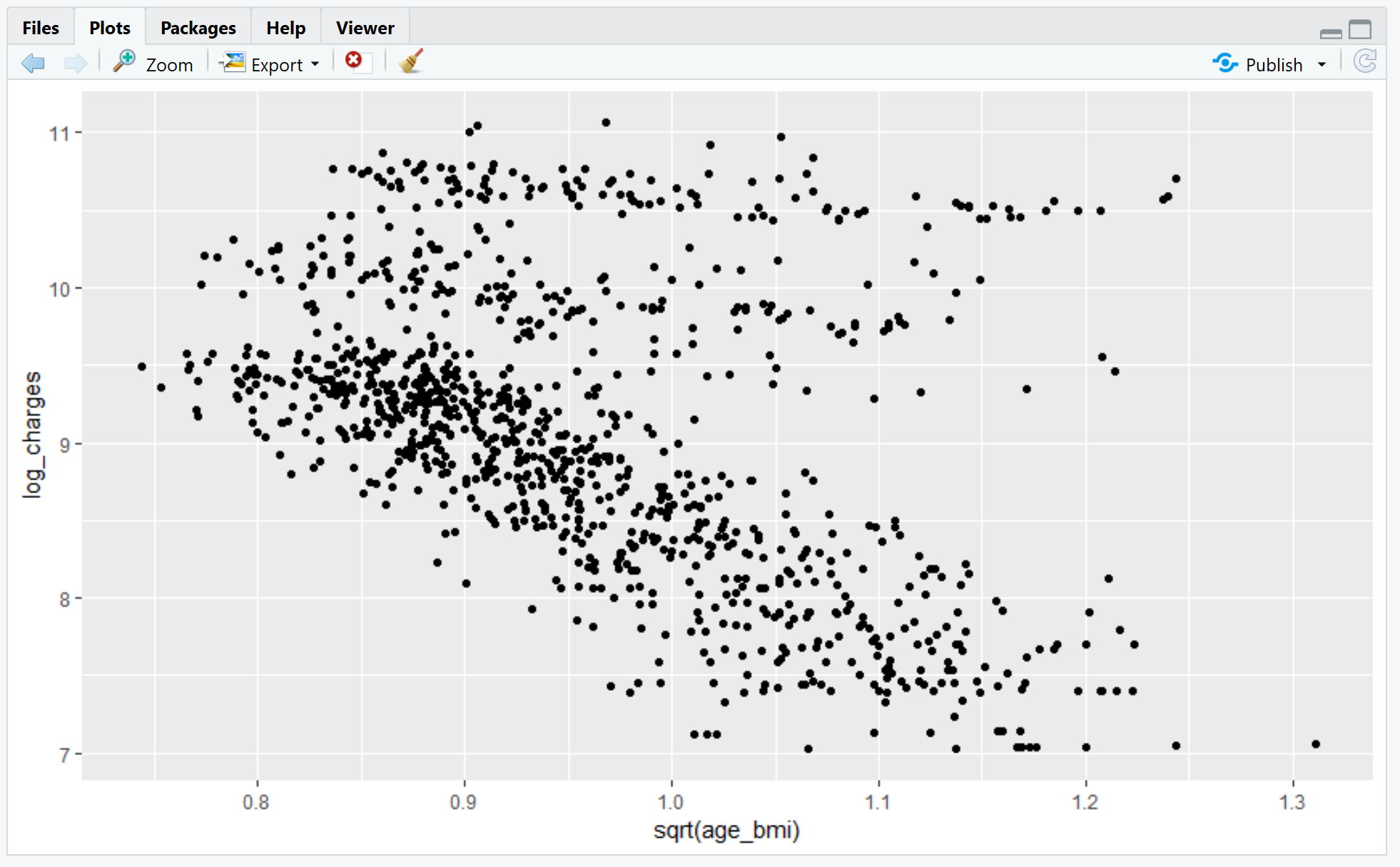
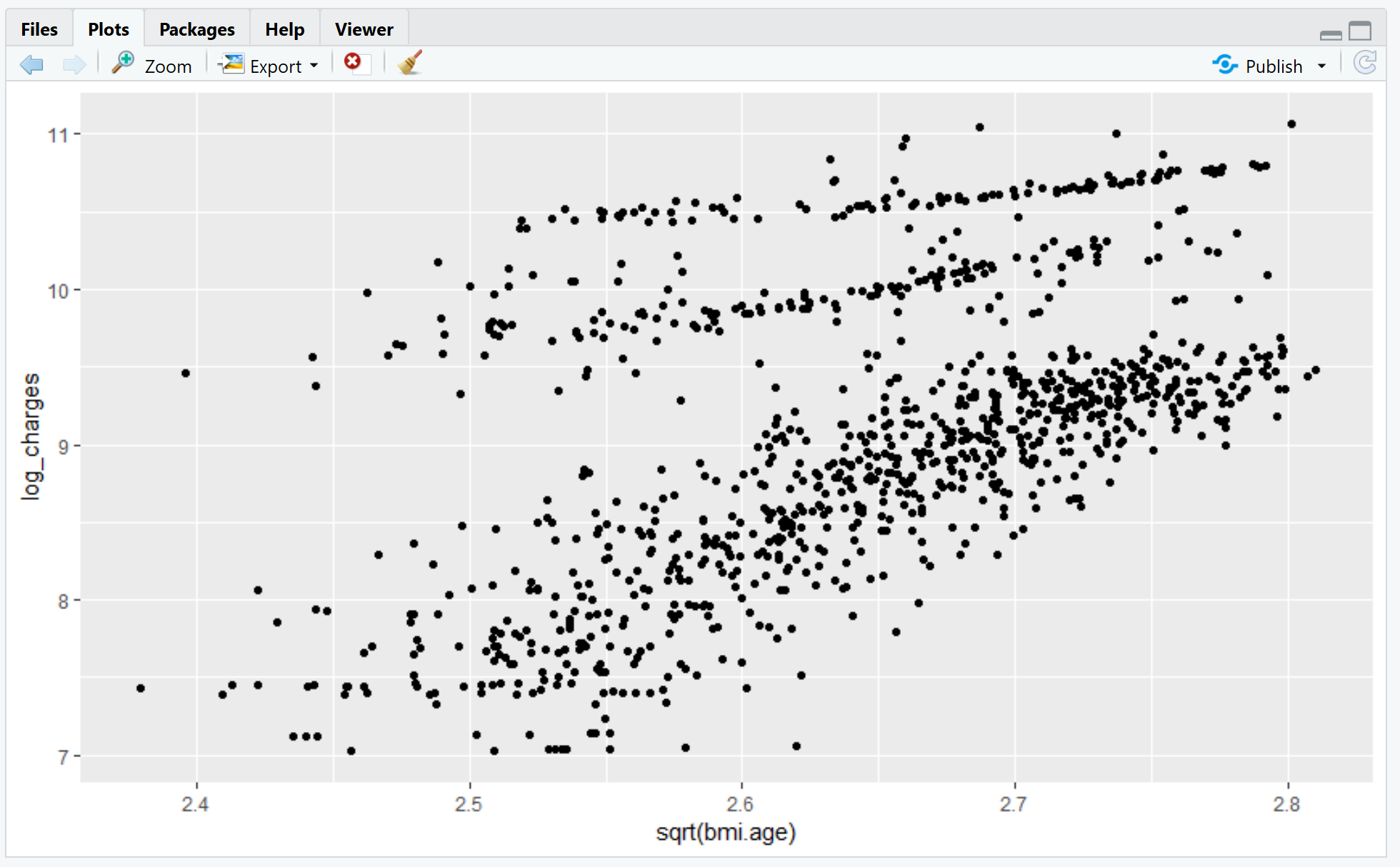
Thus, we observed that, the parameters had significant relationship between them given the low residual standard error and p-value. Also, since the F-statistic obtained was greater than 0, we had enough evidence to conclude that, there parameters were significant. Moreover, R-squared value was 0.759 which means that, the accuracy of the model was 75.9%.

Now, for adjusting the collinearity for age and BMI, we knew that, there was no correlation between BMI and charges and there was a moderate correlation between age and charge. This is proved by the graph below:

For BMI v/s log of insurance charge and age v/s log of insurance charge

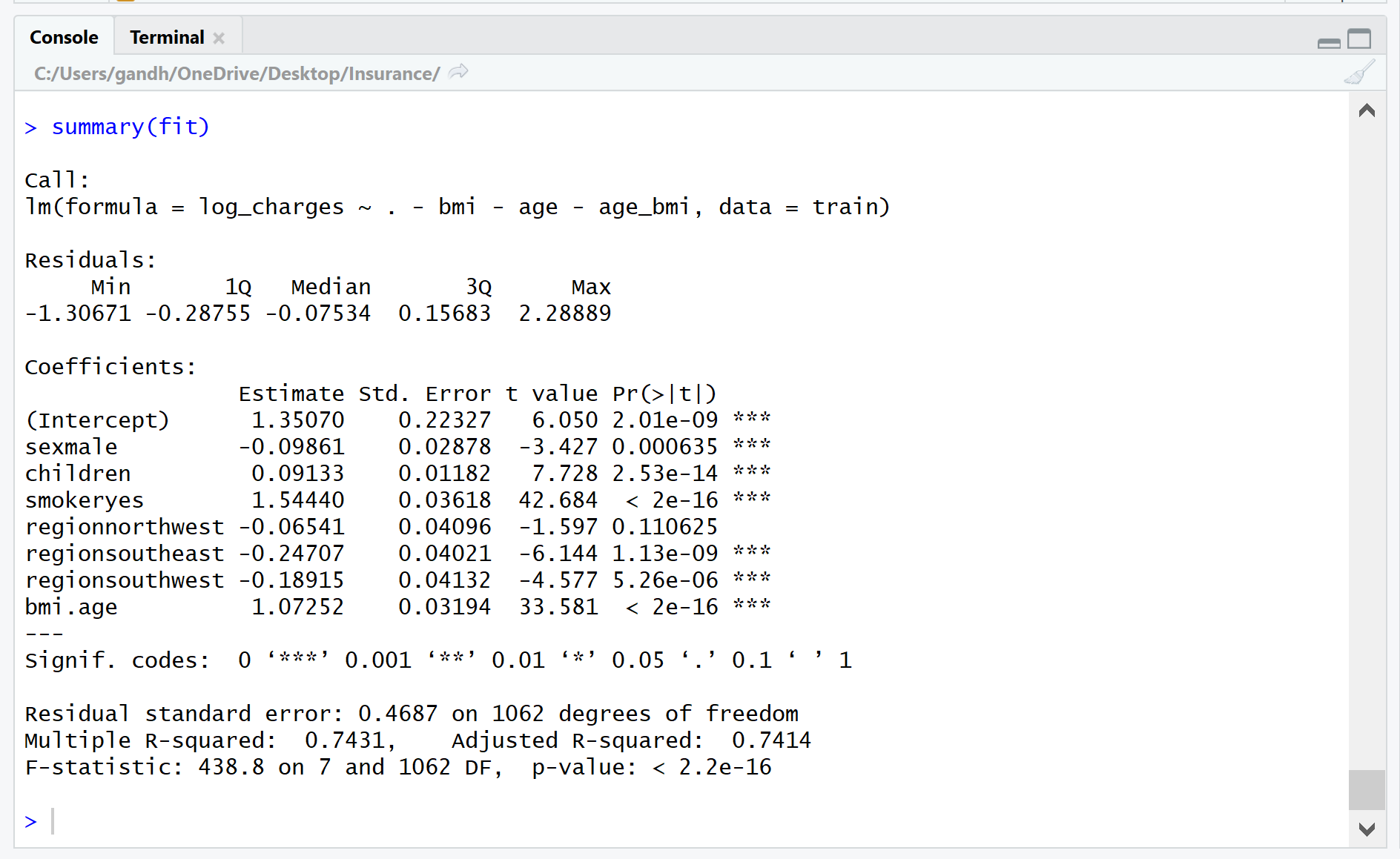
 

Because of this phenomenon, the two new variables “age\_bmi” and “age.bmi” were introduced. The “age\_bmi” was the square root of the ratio of age and BMI and “age.bmi” was the square root of the product of age and BMI. The graphical representation is as shown below:

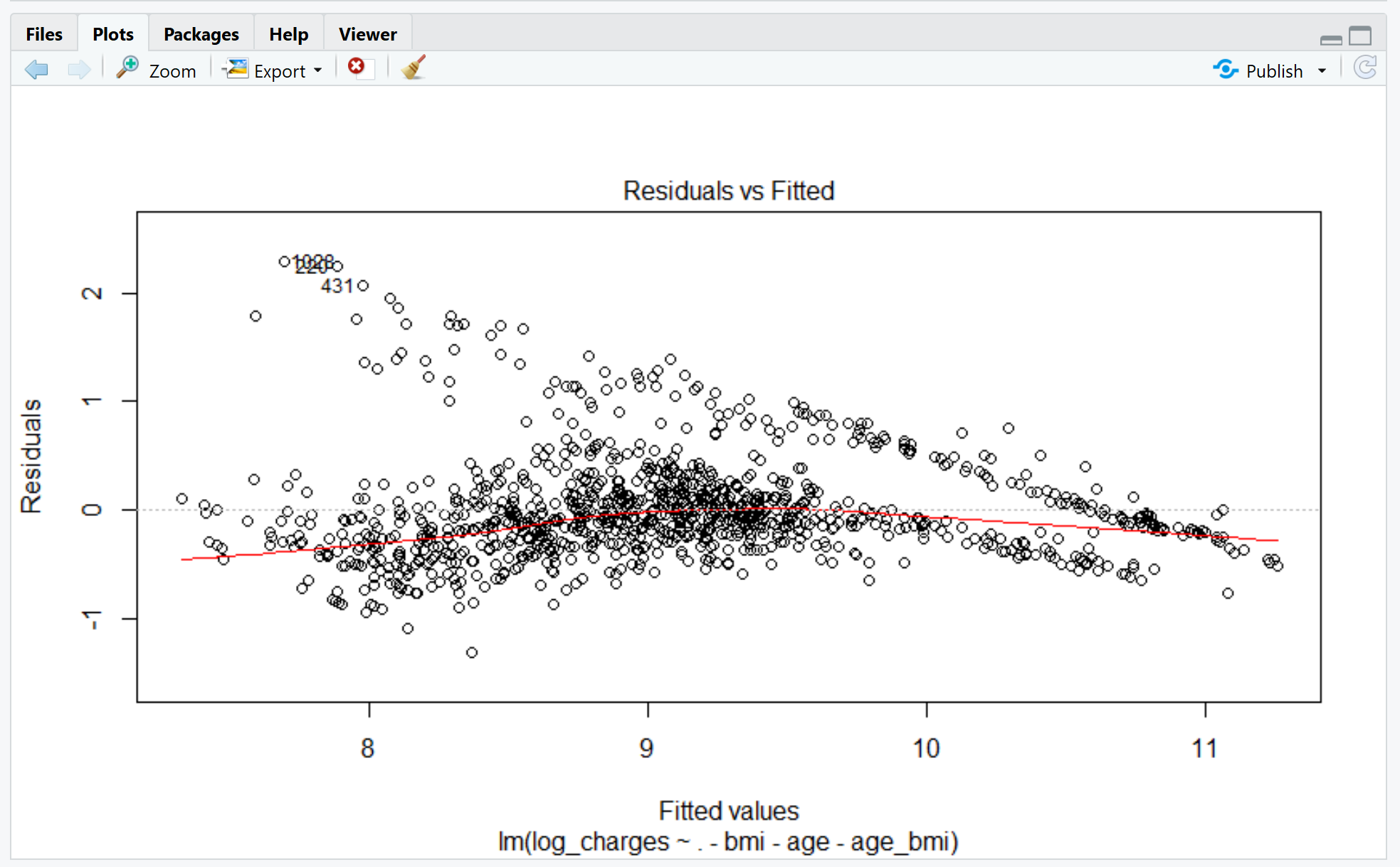
 

Both variables were introduced in the training model and fitted into the linear model with the other variables. The summary of the fitted linear model for the training data set is as shown below:

Output:



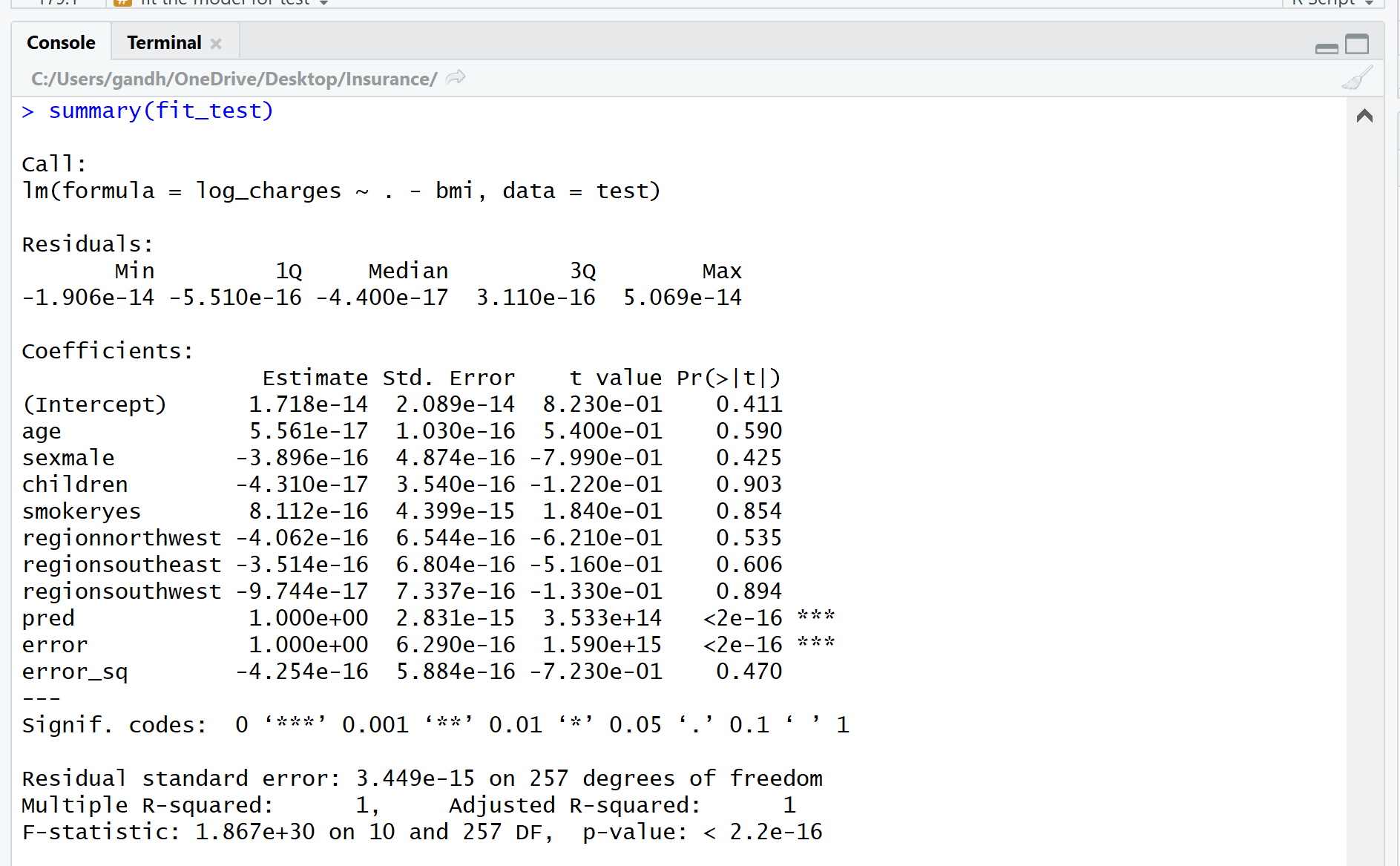
The graphical representation of the fitted values as compared to the residual values was obtained as:



Thus, we can see that, most of the values lie close to the line of best fit. There are a few outliers which can be observed in the plot. On observation, we can conclude that, the model is quite accurate, and the further analysis can be computed on the testing data set.

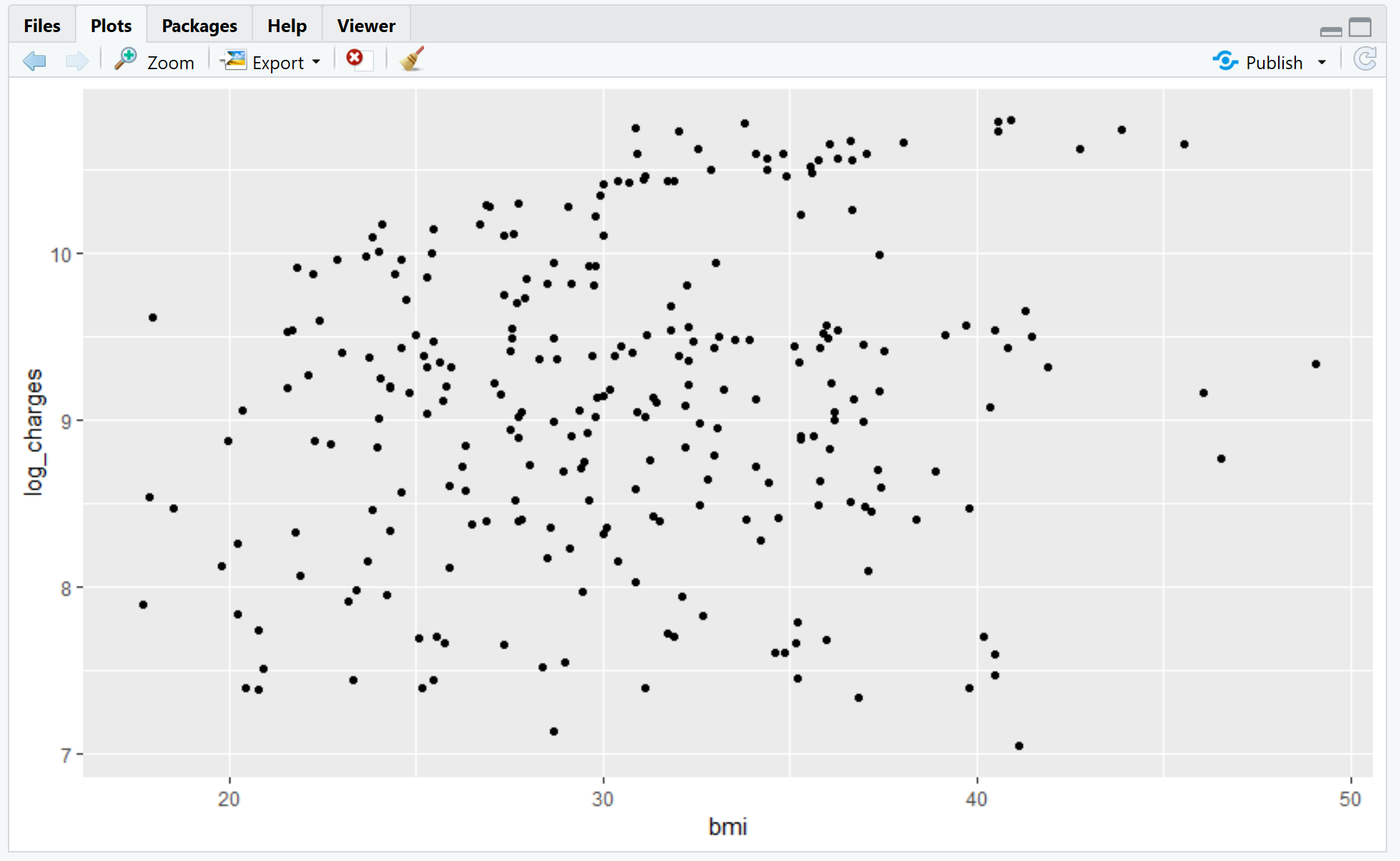
### Fitting the testing model

The testing model was created with 268 observations. The summary of the data set is given below:

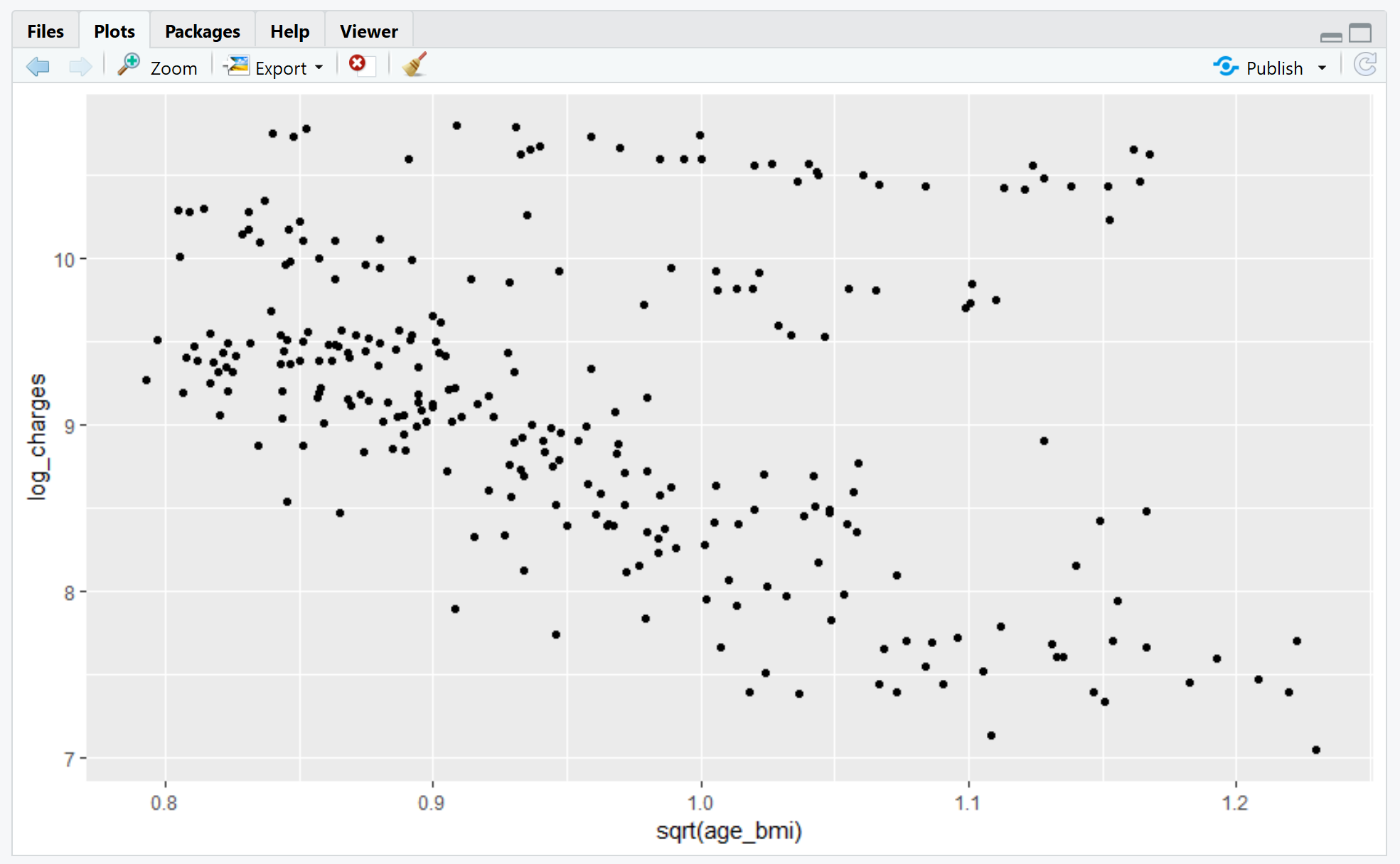
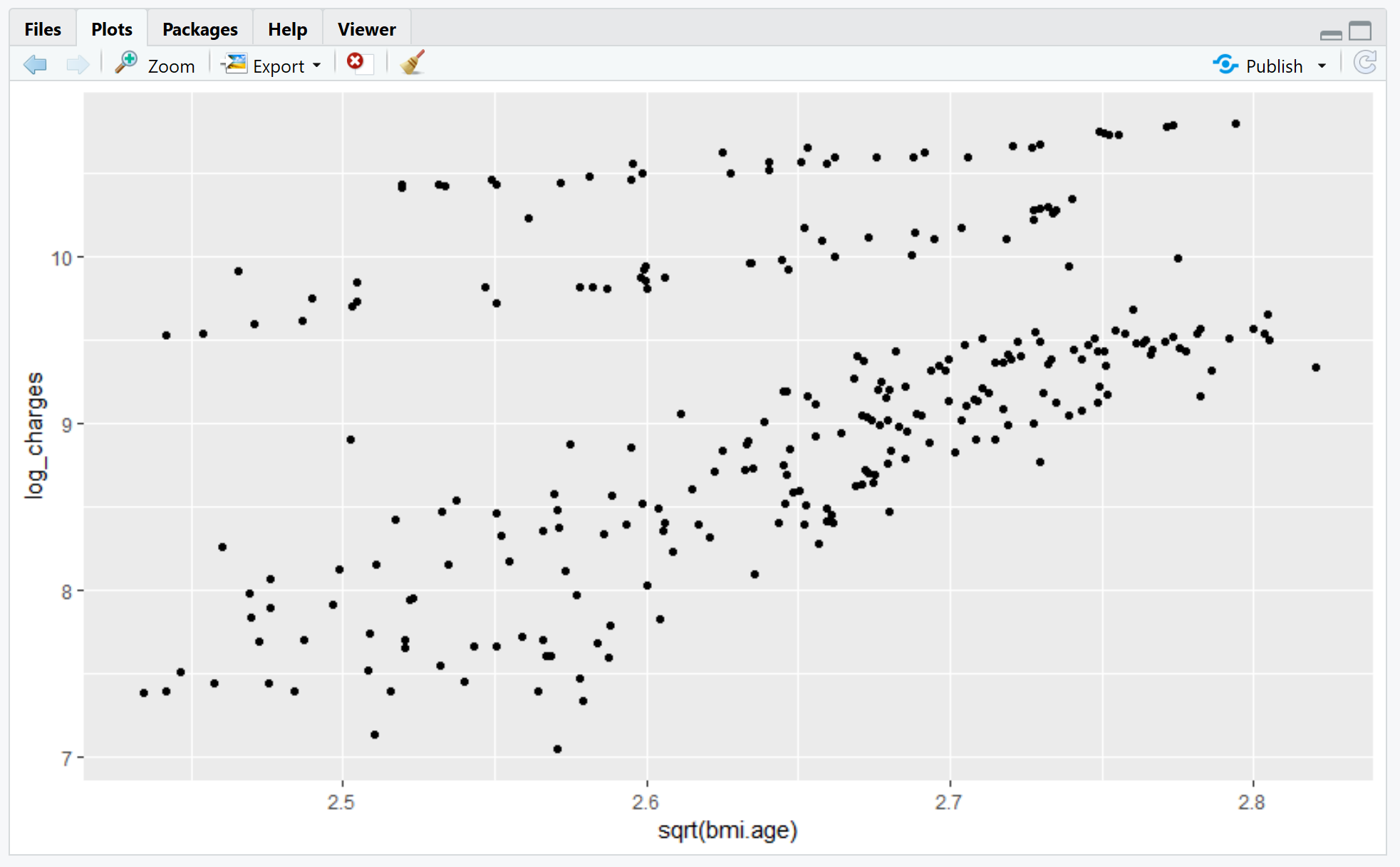


Similar parameters were assumed for age and BMI to improve the efficiency of the model it had been done for the training dataset. The obtained plots are shown below:

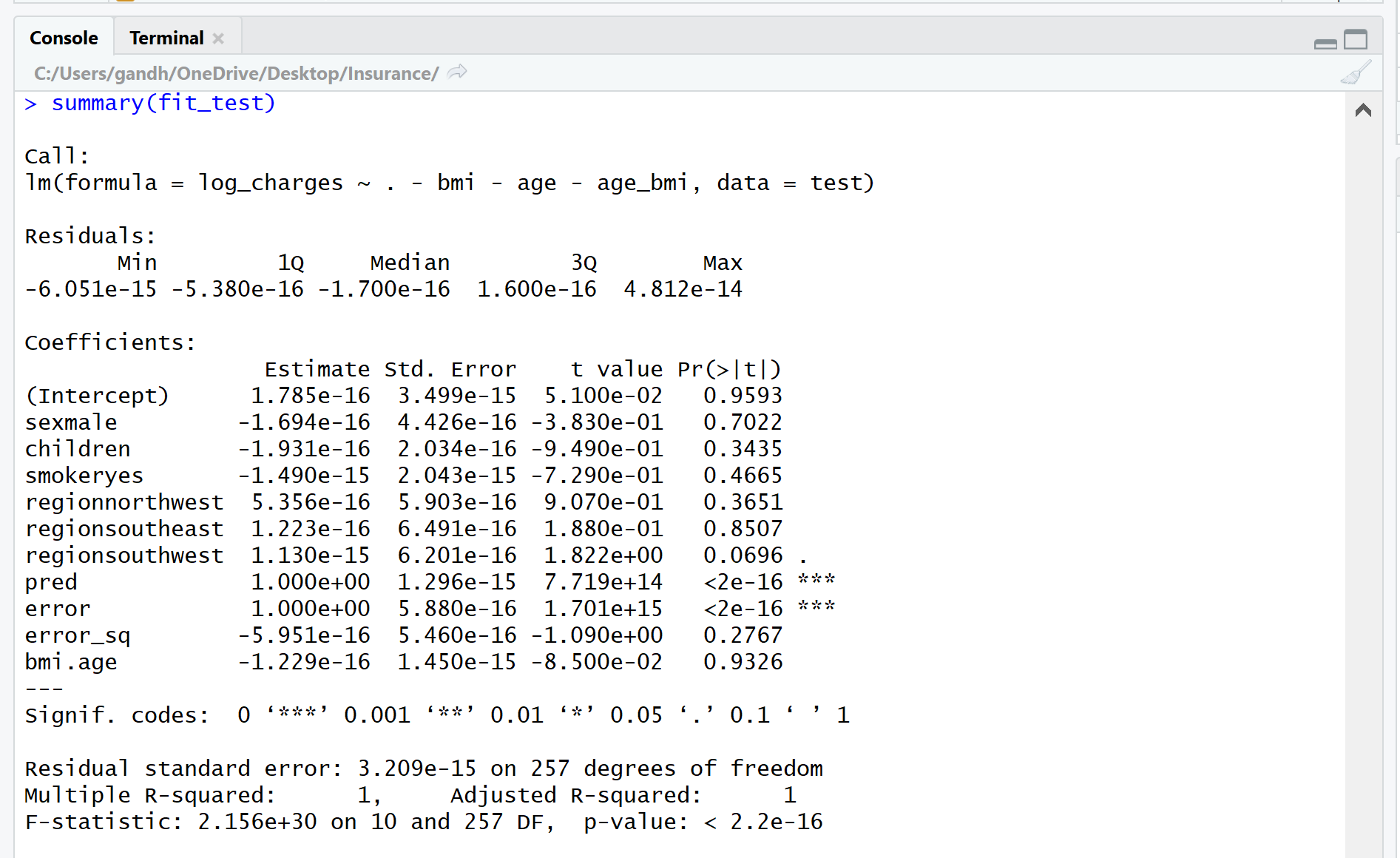
For BMI v/s log of insurance charges and age v/s log of insurance charges:

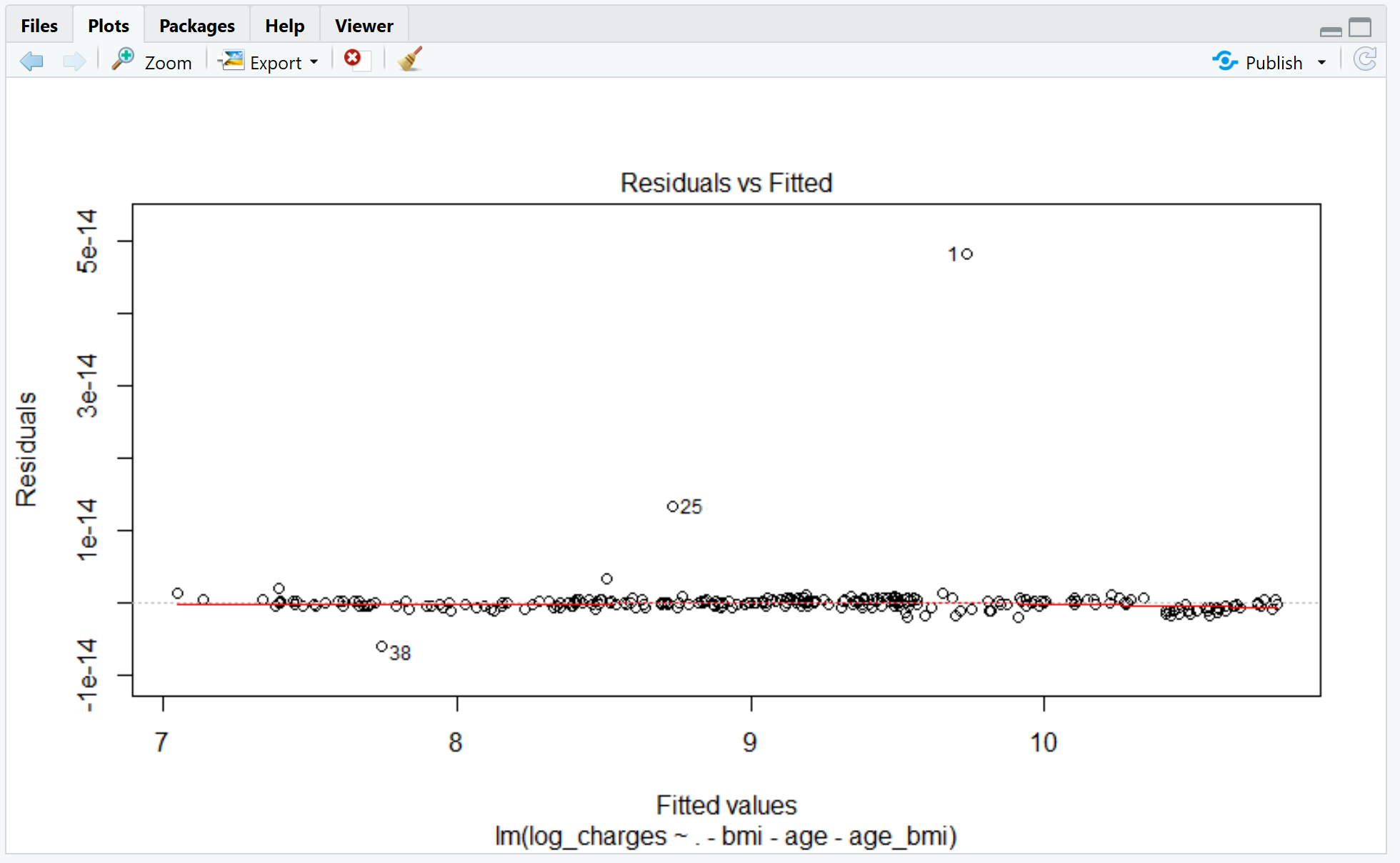
Computing the “age\_bmi” and “age.bmi” parameters, we obtain the following plots:

We created a new linear model by fitting these values into the original parameters of the test model. The summary of the new fitted model is shown below:



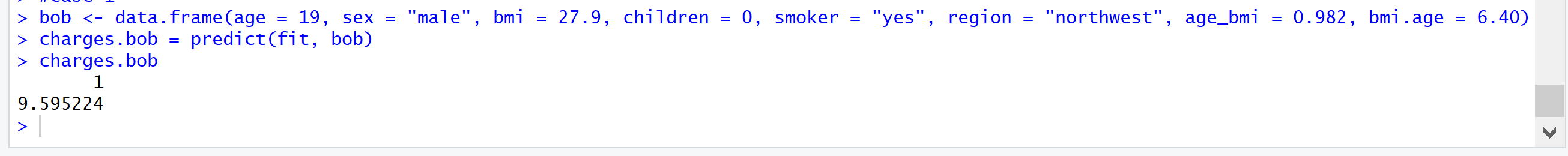
From the model, we obtain a significantly high value of F-statistic of 2.156e+30 which says that the model is good fit. Moreover, we obtain a significantly low value of residual standard error and p-value which further suggests that the parameters are highly significant, and the model is highly accurate. Also, the value of R-squared is 1 which suggested that the model was a perfect fit and the accuracy is 100%. We obtained graph of the fitted value v/s the residuals which gives us an estimate of how well fit the model is. The plot is shown below:



From the graph, we can infer that, almost all the values lie on the line of best fit. The line of best fit can be observed at the 0 value of the plot. A handful of outliers were seen scattered across the plot which do not vastly affect the accuracy of the model. Thus, we can conclude that, the model created by us was a perfect fit and can be implemented on a variety of relevant data. [Devasthali, 2018]

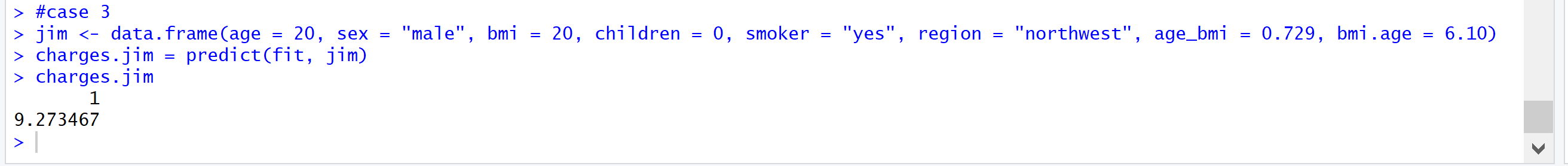
### Building a predictive model

As we observed that, the model created by us was a perfect fit, we decided to implement our linear model in a predictive model which will predicted the insurance charge of a customer by factoring in a few basic variables which we had included in the linear model. The result obtained was obtained as a log value of the original charge. To retrieve the original value, antilog of the value was taken by keeping the base as 2.718 i.e. the value of “e”. We considered a few cases by changing different variables and observed changes with respect to the insurance charge. Some of the cases are shown below:



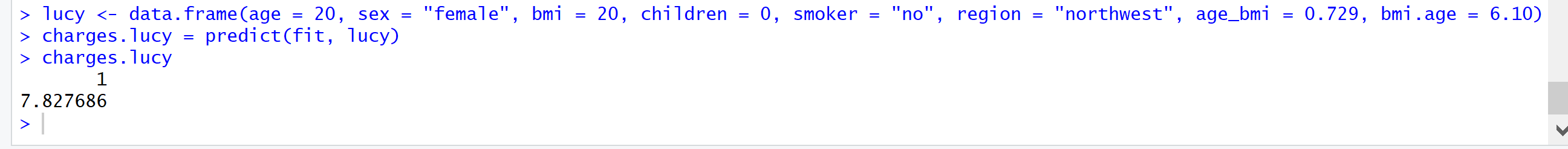
We created a data frame for an individual named Bob and input in the desired parameters which were present in the fitted model of our testing data set. We then used the predict function to predict the insurance charge for him. The value obtained was the log of the original value. Thus, to retrieve the original value, we took the antilog of the value with the base 2.718. Thus, the value obtained by us was 14,676.53 dollars.

Now, we considered the case of an individual with the same parameters except BMI. Thus, subsequently the age\_bmi and age.bmi would change. The results obtained were:



Now, for Jim the log value of the insurance charge was obtained as 9.273 even though all the parameters were same as Bob’s except age, BMI and their related parameters. The antilog value of 9.273 came out to be 10604.55 dollars. With a reduction of BMI and its relationship with the age factor, Jim ended up getting an estimate of close to 4000 dollars less for the insurance charge.

Taking a third case where the age and BMI and the factors around them were kept constant whereas a few parameters were tweaked, and the results obtained were:



The log value obtained was 7.83 which was significantly lower than Jim’s given the same age and BMI. The only change was she was a non-smoker. Taking the antilog of the obtained value, we got the insurance charge for Lucy as 2512.88 dollars. Thus, we concluded that smoking was also one of the major factors which contributed to the reduction in insurance cost.

# Conclusion and Future Scope

## Based on research questions

1. The reduction in insurance cost would be majorly possible by maintaining an optimal BMI given the age of an individual. BMI alone wouldn’t contribute to a change in the cost but when combined with age makes an impact in determining the insurance charge of an individual. Also, maintaining a non-smoking lifestyle will contribute majorly in reducing the insurance costs as from our project there is a clear and wide difference between the insurance costs of smokers and non-smokers. Thus, maintaining a healthy lifestyle is the key to lower insurance costs for any individual living in any given area.
2. The predictive model is quite accurate in determining the insurance cost of a given individual as it takes the values from the fitted testing model created by us and predicts the values based on that. The predicted value would be the most accurate when the values which are input into the predictive model lie inside or close to the values in the fitted model. When the values input lies far outside the range of the fitted linear model, the accuracy of the predicted value would be affected. This can be resolved by widening the range of the dataset by adding more cases to it. Bigger the data set higher would be the accuracy of the predicted value.
3. It helps in increasing the cost efficiency as the insurance company and the customers get a transparent picture of the insurance charges. The aim of the model was also to aid insurance companies in reducing the risk of insurance fraud and to be better equipped to fight fraud.
4. Increase in age will lead to an increase in cost of insurance premium as with age, health complications increase. So, to conclude we can say that, maintaining a healthy lifestyle is the best way to keep insurance costs as low as possible.
5. Splitting the data into a training and testing model proved quite useful in the project as by training the data with 80% of the observations, the model became quite accurate and once the testing model was fitted, the accuracy of the model peaked to 100%.
6. Another benefit of the model is that it is time saving and easily operational, i.e. it reduces the manual labor of assessing each factor for computing the insurance charges. For example: If a customer goes to buy any insurance, he or she can get an estimate of the insurance premium on the spot as the insurance companies can input the customer’s data in the model to get an optimal premium for the customer.

## Applications of Linear Regression

Linear Regression is a very powerful statistical technique that has myriad applications. It can be used to understand business, generate insights on consumer behavior and factors that influence profitability. It can also be used in business to evaluate trends and make estimates or forecasts. Moreover, it can also be used to analyze the marketing effectiveness along with the pricing and promotions on sales of a product. A major benefit of linear regression is it can also assess risk in financial services or insurance domain. [Shitut, 2018]

## Future Scope

1. Predictive analytics plays a critical role in appraising and controlling the risk in ratings, claims, pricing, underwriting and marketing in the insurance industry. It assists insurance companies in understanding and modelling the future behavior of their customers thereby permitting them to create financial product models that are pertinent and engaging to them.  It can also map trends enabling the organization to be ahead in the market when predicting risks.
2. According to the “Coalition Against Insurance Fraud”, US lost at least $80bn annually in insurance fraud. Therefore, insurance companies made it a high priority to identify fraud when processing claims. They need the ability to create models that are increasingly intricate that can access federated and governed data sources and easily be deployed through an organization from one analytical hub.
3. Assessing the role played by data in business performance and fraud prevention, insurance companies need to have powerful analytical platforms consisting of easy to use visualization tools and multiple devices allowing real time reporting to enable the experts to derive crucial insights and take data informed business decisions.

# Reference

Samuels, J. S. (2019). Durbin Watson test statistic. Retrieved from <https://stats.stackexchange.com/questions/109234/durbin-watson-test-statistic/109239>

Heckman, E. (2019). What in the World Is a VIF? Retrieved from <http://blog.minitab.com/blog/starting-out-with-statistical-software/what-in-the-world-is-a-vif>

BusinessWire. (2018). Analyzing the Impact of Health Insurance Analytics: Quantzig’s Take on How the Insurance Industry is Evolving in 2018. Retrieved from: <https://www.businesswire.com/news/home/20181023005754/en/Analyzing-Impact-Health-Insurance-Analytics-Quantzig>

Berchick, E., Hood, E., Barnett, J. (2018, September 12). Health Insurance Coverage in the United States: 2017. Retrieved from: <https://www.census.gov/library/publications/2018/demo/p60-264.html>

Analytics Vidhya Content Team, & Analytics Vidhya Content. (2017, January 31). Going Deeper into Regression Analysis with Assumptions, Plots & Solutions. Retrieved February 14, 2019, from [https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions](https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/)

Bobs guide. (2018, July 24). Predictive analytics is key to insurance sector competitiveness. Retrieved from: <https://www.bobsguide.com/guide/news/2018/Jul/24/predictive-analytics-is-key-to-insurance-sector-competitiveness/>

Shitut, N. (2018, November 18). Popular Applications of Linear Regression For Businesses. Retrieved from: <https://analyticstraining.com/popular-applications-of-linear-regression-for-businesses/>

Rego, F. (2015, October 23). QUICK GUIDE: INTERPRETING SIMPLE LINEAR MODEL OUTPUT IN R. Retrieved from: <https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R>

Loginov, M., Marlow, E., Potruch, V. (2018). PREDICTIVE MODELING IN HEALTHCARE COSTS USING REGRESSION TECHNIQUES. Retrieved from: <https://umanitoba.ca/faculties/management/academic_depts_centres/centres_institutes/warren/media/P2B_3_Michael_Loginov_Predictive_Modeling_Regression(3).pdf>

Devasthali, R. (2018, April 8). Coefficient of Determination ( R-squared) Explained. Retrieved from: <https://towardsdatascience.com/coefficient-of-determination-r-squared-explained-db32700d924e>

Li, E. (2017, May 6). Linear Regression Analysis of Insurance Data. Retrieved from: <http://rstudio-pubs-static.s3.amazonaws.com/274113_c958eac1bbe846969eb560e5ab3694b8.html>

# Appendix

## Source Code

## initializing and loading the dataset and looking at the sample dataset ##

insurance = read.csv("insurance.csv")

head(insurance)

##loading the libraries

library(ggplot2)

library(MASS)

library(lmtest)

library(car)

### pre checking the values as target variable should follow normal distribution ###

hist(insurance$charges)

### target variable is right skewed

## apply log transformation to make it a normal distribution

insurance$log\_charges = log(insurance$charges)

hist(insurance$log\_charges)

### finding the linear relationship b/w input and target for both the original and log values

plot(insurance$age, insurance$log\_charges)

plot(insurance$age , insurance$charges)

### multicollinearity (input variables are correlated with each other)

cor(insurance$age, insurance$bmi) ### cor with input variable (not desirable as value tends to 0)

#checking the effect of smokers on the log\_charge values

ggplot(insurance, aes( smoker, log\_charges)) + geom\_boxplot()

### Model bulding for linear regression analysis ###

insurance$charges = NULL

## splitting the test and train sets of data

set.seed(675)

test\_train = sample(nrow(insurance), nrow(insurance)\*0.8)

train = insurance[test\_train,]

test = insurance[-test\_train,]

## model

lin\_model = lm( log\_charges ~ . , data=train )

summary(lin\_model)

## Test the model

test$pred = predict(lin\_model, newdata=test)

### finding the RMSE

test$error = test$log\_charges - test$pred

test$error\_sq = test$error \*\* 2

rmse = sqrt(mean(test$error\_sq))

rmse

summary(test$log\_charges)

cofv <- 0.43/9.13 #coefficient of variation

cofv

################ Diagnosis #####################

### select only a few variable

fit = lm(log\_charges ~ ., data=train)

#### correlation check or Multicollinearity

summary(fit)

names(fit)

fit$coefficients

head(fit$residuals)

### check for normality of errors by plotting the residuals plot

hist(fit$residuals)

### check for auto correlation and heteroscedasticity

## Standardized residuals

residuals = stdres(fit)

summary(residuals)

### predicted values vs. fitted values

plot(fit$fitted.values, residuals)

### statistical test for autocorelation

## to check auto correlation

dwtest(fit)

### Outliers test

outlierTest(fit)

### measure cooks.distance

cd = cooks.distance(fit)

cutoff = 4/(nrow(train) - length(fit$coefficients))

### plot for finding the observation which has high leverage (using value cd)

plot(fit, which=4, cook.levels=cutoff)

#Outliers and leverage observations

### Variance inflation factor to check multicollinearity

vif(fit)

## rebuild the model by removing outlier observations

train = train[ -c(431, 220, 1028, 1040, 103,527, 1020), ]

####### fit the model for train ######

fit = lm(log\_charges ~ . -bmi , data=train)

summary(fit)

## check if Autcorrelation or Heteroscedasticty has improved

plot(fit, which=3)

##explore the relationship between target and input

ggplot(train, aes( bmi, log\_charges)) + geom\_point()

ggplot(train, aes( age, log\_charges)) + geom\_point()

###combine bmi and age by dividing them and taking the square root into a new variable

train$age\_bmi = sqrt(train$bmi/train$age)

ggplot(train, aes( sqrt(age\_bmi), log\_charges)) + geom\_point()

#### final model with ratio of age and bmi

fit = lm( log\_charges ~ . -bmi -age, data=train)

summary(fit)

### combine age and bmi by multiplying them and taking them as a new variable

train$bmi.age = log( train$bmi\*train$age)

ggplot( train, aes( sqrt(bmi.age), log\_charges)) + geom\_point()

#adding the new value to the fit model

fit = lm( log\_charges ~ . -bmi -age -age\_bmi, data=train)

fit\_pred = lm( log\_charges ~ . -bmi -age -age\_bmi, data=train)

summary(fit)

plot(fit, which = 1)

########### fit the model for test ##########

fit\_test = lm(log\_charges ~ . -bmi , data=test)

summary(fit\_test)

## check if Autocorrelation or Heteroscedasticty has improved

plot(fit\_test, which=3)

##explore the relationship b/w target and input

ggplot( test, aes( bmi, log\_charges)) + geom\_point()

ggplot( test, aes( age, log\_charges)) + geom\_point()

### combine bmi and age into a new variable (similar to train model)

test$age\_bmi = sqrt(test$bmi/test$age)

ggplot( test, aes( sqrt(age\_bmi), log\_charges)) + geom\_point()

#### final model with ratio of age and bmi

fit\_test = lm( log\_charges ~ . -bmi -age, data= test)

summary(fit\_test)

test$bmi.age = log( test$bmi\*test$age)

ggplot( test, aes( sqrt(bmi.age), log\_charges)) + geom\_point()

#fitting the model with the new added variables

fit\_test = lm( log\_charges ~ . -bmi -age -age\_bmi, data=test)

summary(fit\_test)

plot(fit\_test, which = 1)

##############predictive model test###################

#case 1

bob <- data.frame(age = 19, sex = "male", bmi = 27.9, children = 0, smoker = "yes", region = "northwest", age\_bmi = 0.982, bmi.age = 6.40)

charges.bob = predict(fit, bob)

charges.bob

#case 2

tracy <- data.frame(age = 30, sex = "female", bmi = 27.9, children = 2, smoker = "yes", region = "northwest", age\_bmi = 0.884, bmi.age = 5.32)

charges.tracy = predict(fit, tracy)

charges.tracy

#case 3

jim <- data.frame(age = 20, sex = "male", bmi = 20, children = 0, smoker = "yes", region = "northwest", age\_bmi = 0.729, bmi.age = 6.10)

charges.jim = predict(fit, jim)

charges.jim

#case 4

lucy <- data.frame(age = 20, sex = "female", bmi = 20, children = 0, smoker = "no", region = "northeast", age\_bmi = 0.729, bmi.age = 6.10)

charges.lucy = predict(fit, lucy)

charges.lucy