

**Insurance Forecast by using Linear Regression**

**PH 700A: Data Analysis Using R**

**Project Report**

**San Diego State University**

**Dec 18, 2018**

**Team: Mean Squares**

Kanchan Pathak

Mayuri Kudale

Preethi Narayanan

Mayank Kapoor

**Background**

Insurances are used to cover the medical costs a person bear. Insurance is provided based on multiple variables associated to an individual. We picked our dataset from Kaggle:

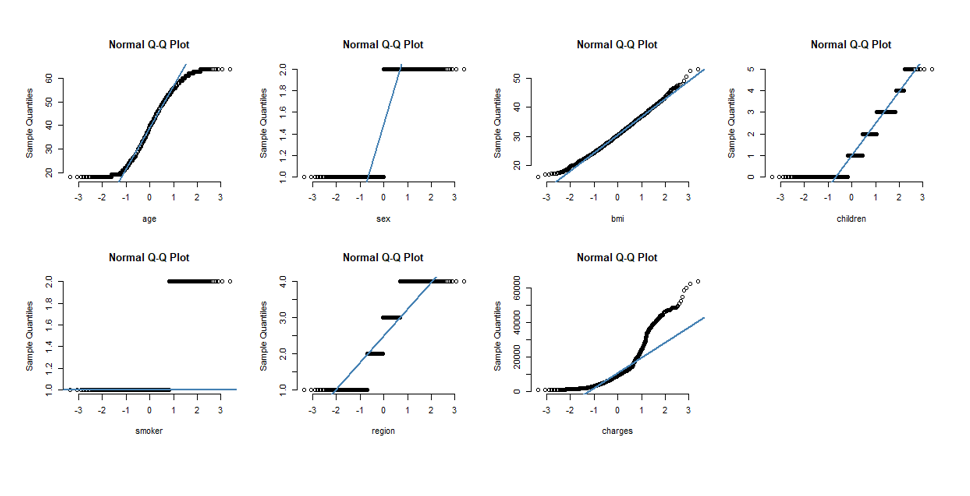
<https://www.kaggle.com/mirichoi0218/insurance/home>

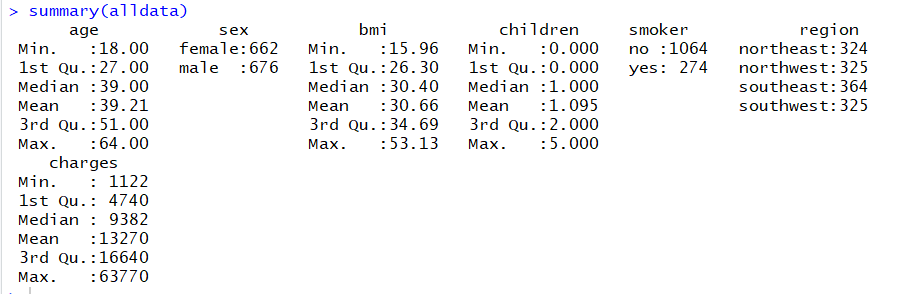
Dataset used as part of our project is an Insurance dataset. We can use this dataset for building a linear regression model for inference, determining if there is relationship between medical charges and predictors.

Our dataset consists of 1338 observations and 7 columns:

* age: age of primary beneficiary
* sex: insurance contractor gender, female, male
* bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
* children: Number of children covered by health insurance / Number of dependents
* smoker: If a person smokes or not
* region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
* charges: Individual medical costs billed by health insurance

Distribution of the data seen as below.





From summary of the data we can determine that

1. The age of participants varies from 18 to 64.
2. Around 49.48% of participants are female.
3. The bmi of participants ranges from 15.96 to 53.13.
4. Only 20.48% of the participants are smokers.

5. Mix of all regions are included.

**Research Question**

***Null Hypothesis:***  There is no relation between any of the predictors and the response

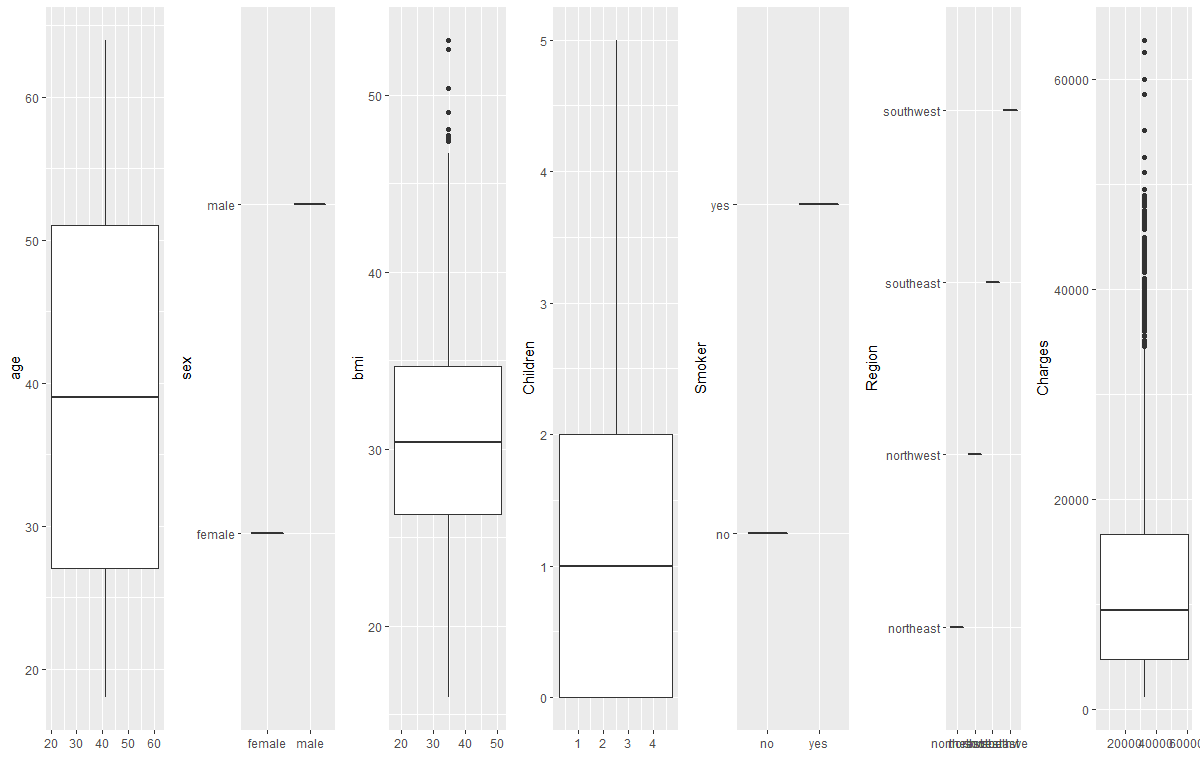
***Alternative Hypothesis:*** There is a relationship between any of the predictors and response

This question can be tested by computing tests like F-statistics or t-test. P-value of F statistics can be used to determine whether to accept or reject the null hypothesis.

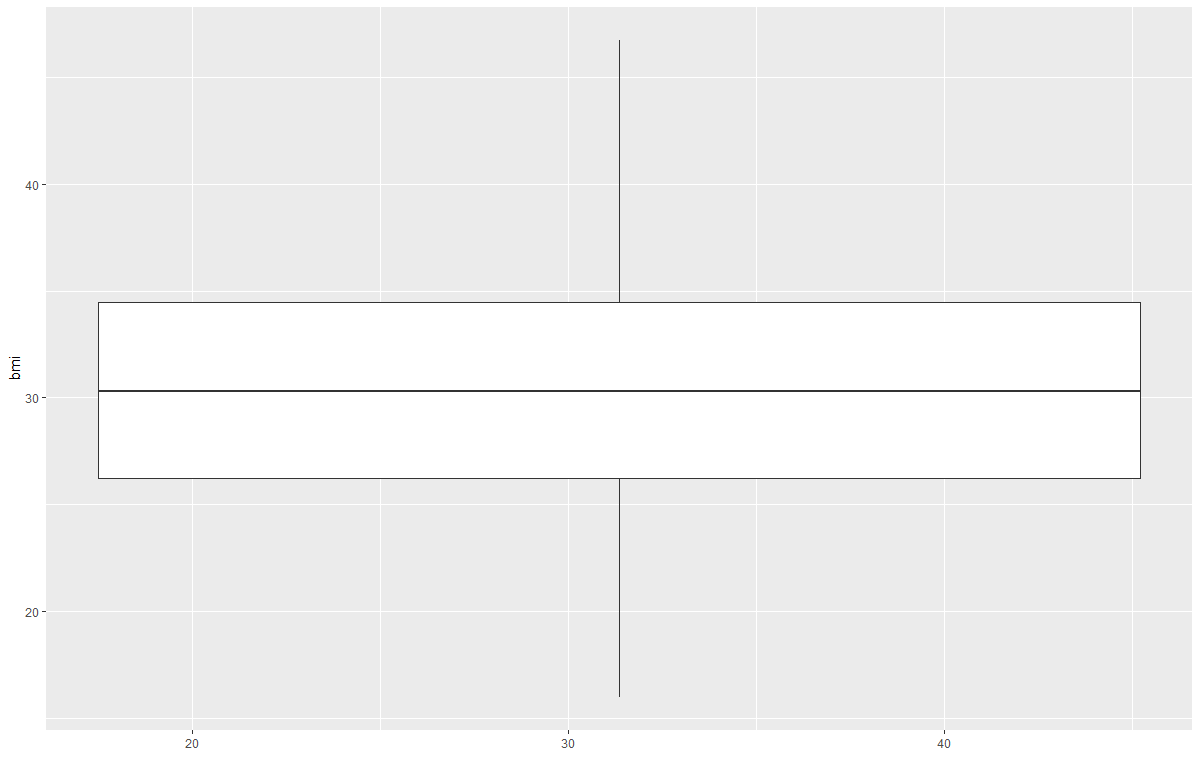
**Investing and Findings in Data:**

1. **Outliers**

In order to get a better understanding of the data, we analyze data and see if there exist any outliers in the dataset

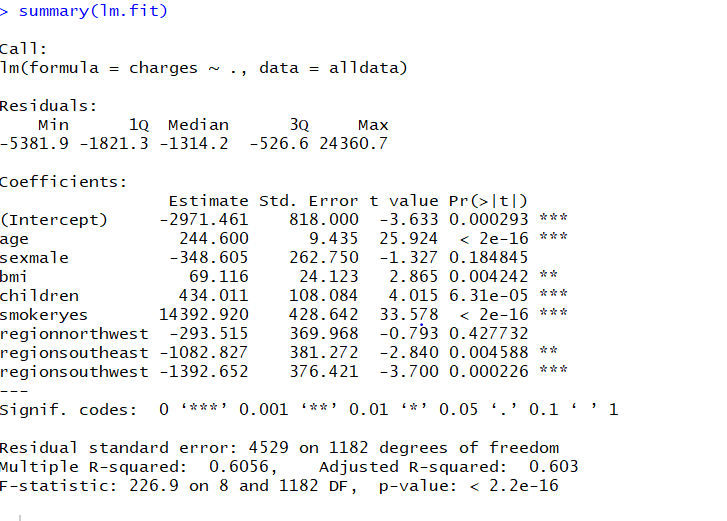


From the above plot we understand that BMI and Charges have some outliers in them and we need to preprocess this data in order to have a clean dataset before running our model on it.



*BMI data after removing Outliers*

We run Linear Regression on our dataset in order to determine if there is any relationship between predictors.



F-statistics has a very high value (226.9) and a very low p-value(<2.2e-16), this indicates that we can reject null hypothesis meaning there is a relationship between predictors.

We can further study from the summary that the RSE has a large value (4529). RSE (Residual Standard Error) is the estimate of the standard deviation of irreducible error i.e. it is the average deviation between actual response and true response. The large RSE value means that our model is deviated highly from true regression line.

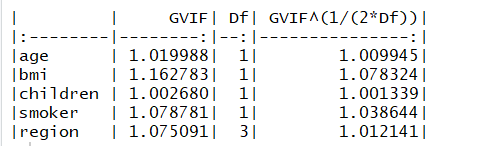
R2 (R-Square) measure the proportion of variability as explained in the model and the value of R2 always ranges between 0 and 1. The closer to 1 the better. Our model has a R2 value of 0.603 meaning 60% of variance in the data is being explained by the model.

The t-value of the predictor here tells us that how many standard deviations it is away from the mean.

As we have now determined that there is a relationship between the predictors our next step would be to determine the variable which has a strong relation to Charges.

1. **Multicollinearity:**

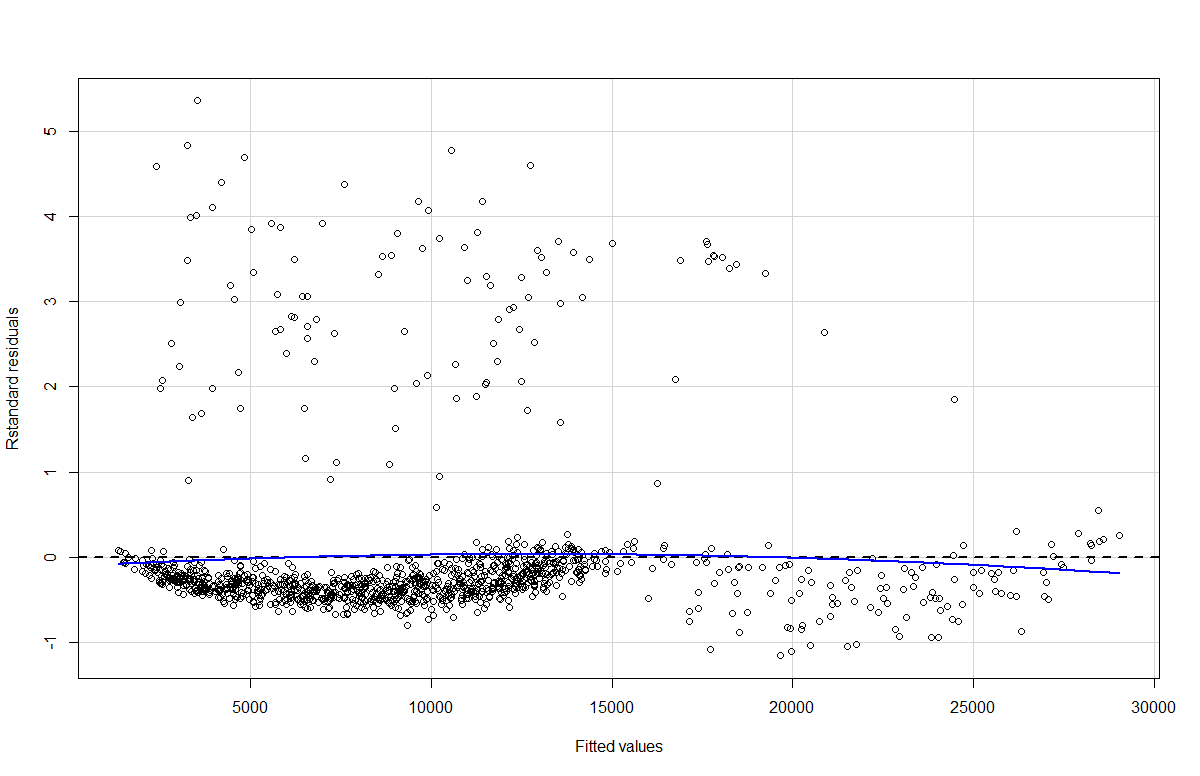
With multiple predictors comes the problem of multicollinearity. We can determine if our dataset has any multicollinearity.



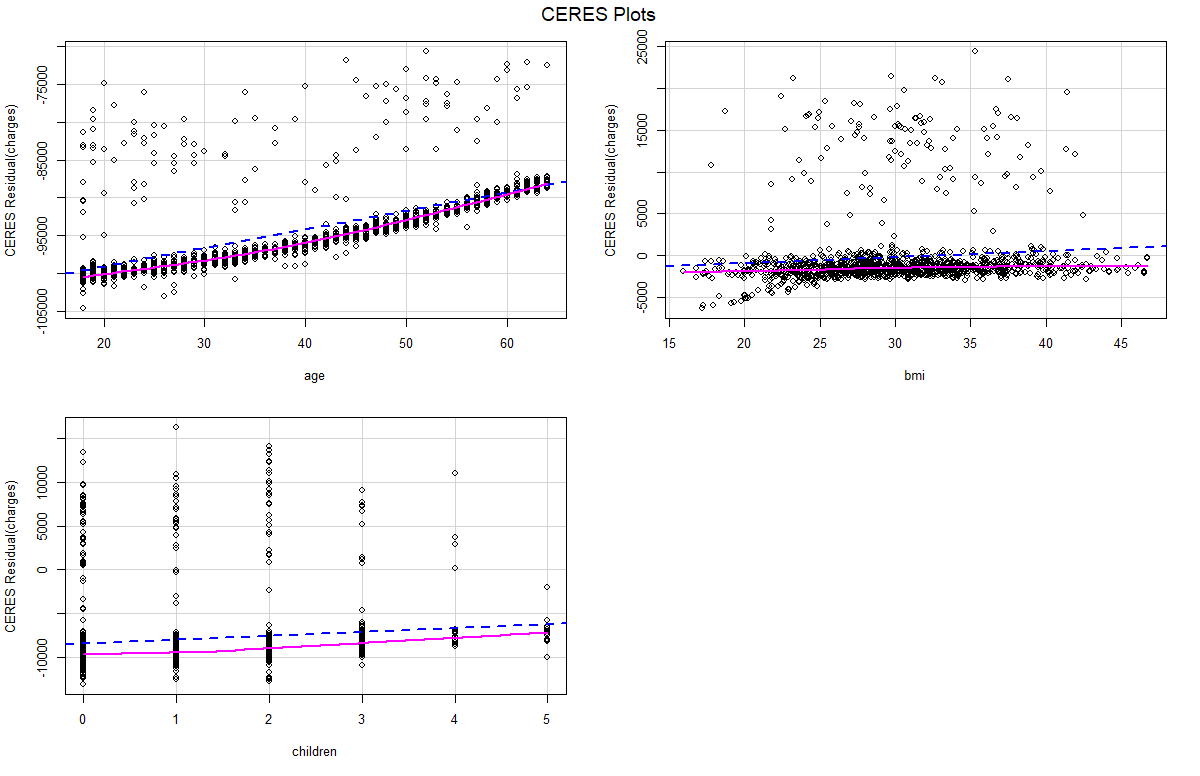
In case of our dataset we don’t have any variable with a higher VIF value, so we are ensured that multicollinearity doesn’t exists in our data.

1. **Linear vs Non-Linear Data:**

Initially to determine if our hypothesis is right or not, we used Linear Regression. However, we assumed that our data has a linear relationship. Let’s review if our assumption was right or wrong.

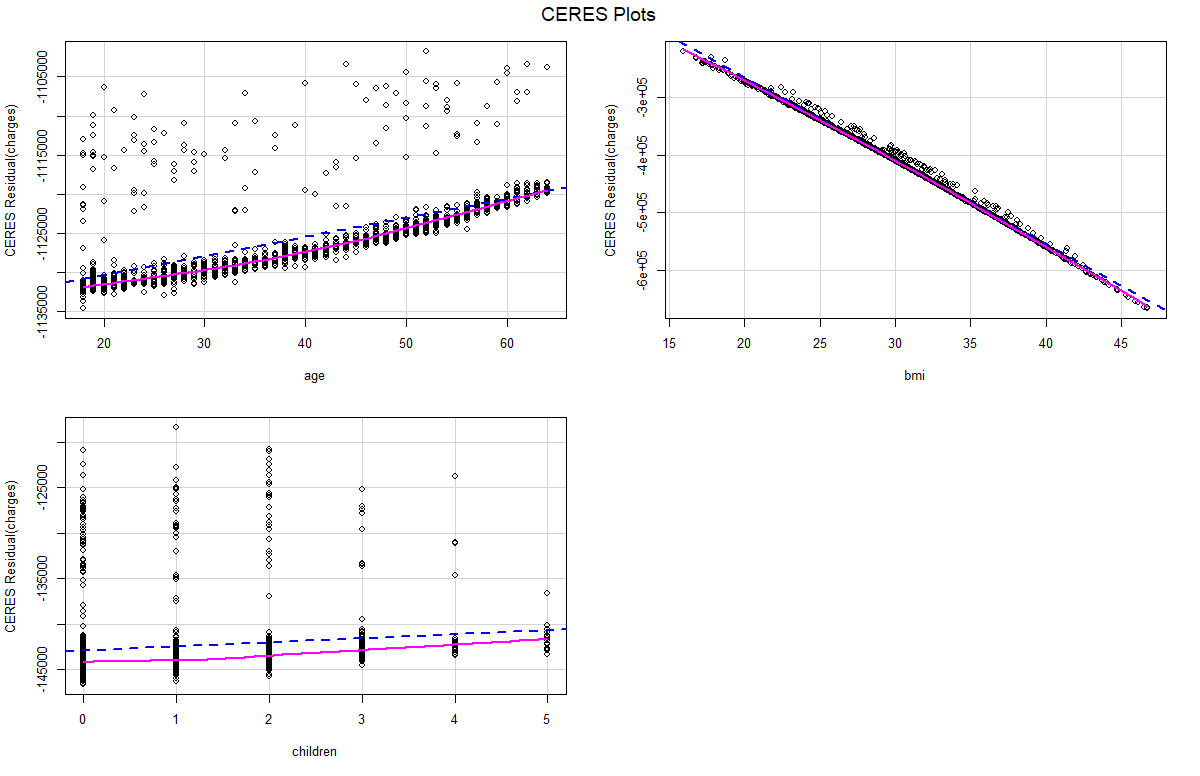


The blue line here indicates the smooth pattern of the fitted data. In our case the line is slightly non-linear meaning that there is some non-linearity existent in our data. We can further verify this non-linearity.



The pink line in the above graph indicates a residual line which determines the relation between predictor and residuals. Blue dashed line indicates component line which determines the line of best fit. And if the difference between the two lines is significant then it indicates that the relationship between the predictors and the outcome is not linear.

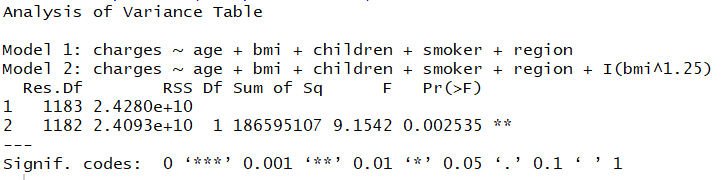
By studying the graphs, we find that bmi shows some inconsistency. This can be fixed by performing non-linear transformations of the predictor.



**Analyzing with Statistical Method:**

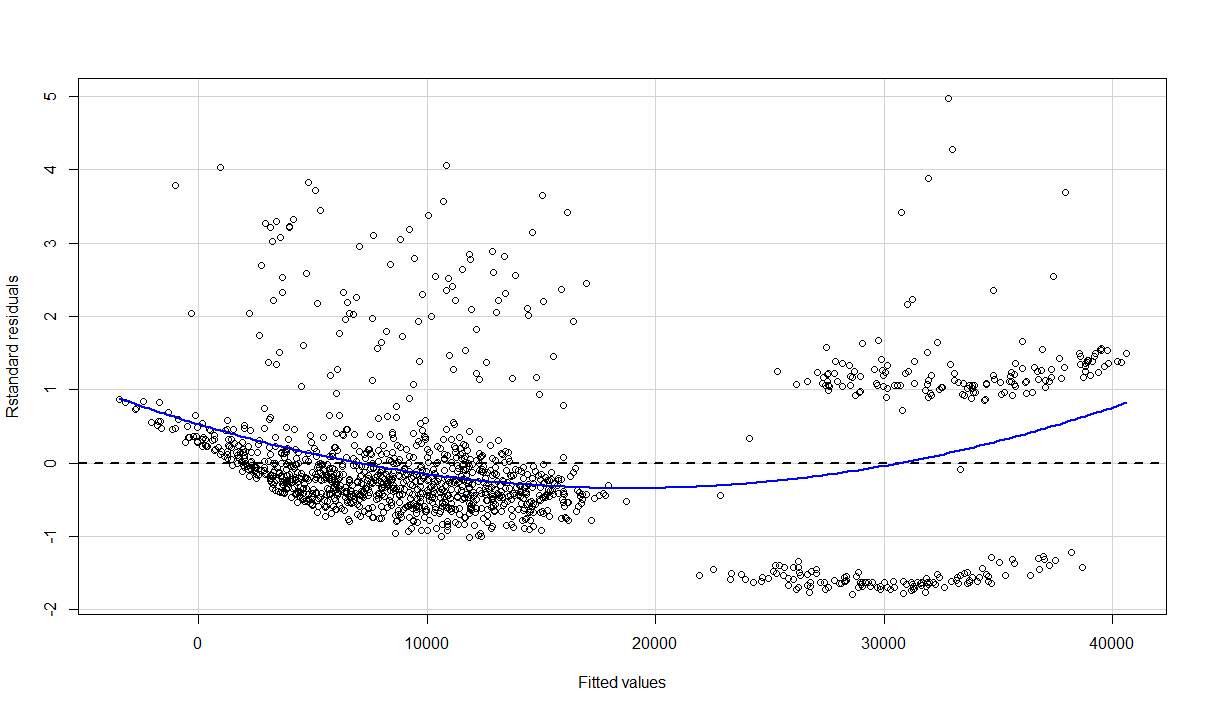
**ANOVA**

We can use ANOVA to study if our new model is better than our previous model, this test will help us determine their significance.

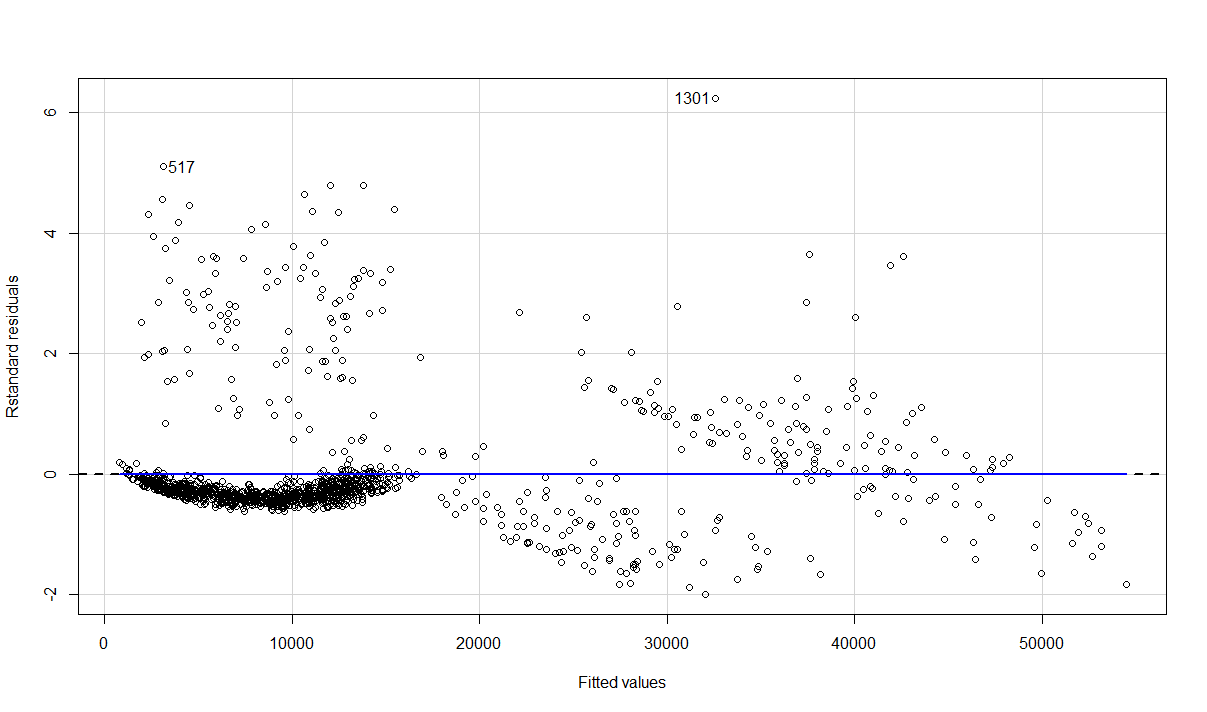


As the model with non-linear transformation of bmi has a significantly low bmi , we can conclude that this model is better than the previous model.

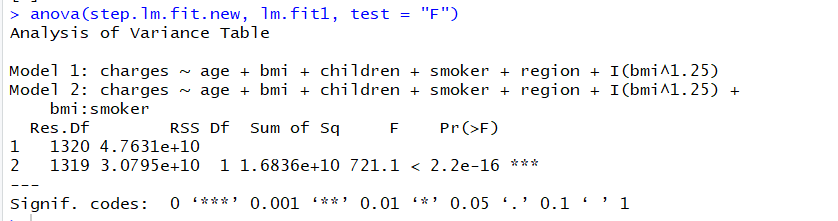
Looking at the residual plot of the new model, we can figure out that there is no difference between the between the overall pattern of the standard residuals.



We can further modify the code in order to fit the data well, by fixing the problem of non-linearity by introducing an interaction term. We introduced the term smoker and visualize the results.



We can see that there is significant difference in the data now and data is now in line with the regression line.

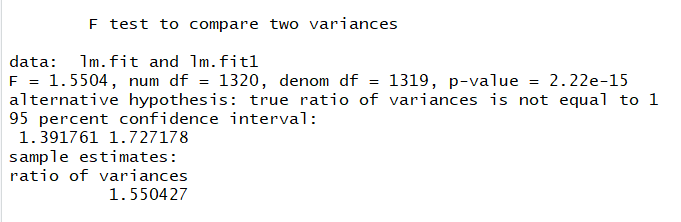


R2 value for this model improves and increases to 0.8377



**F-Test**

F-tests are used for comparisons of variances between two cases. In our case we created a initial linear regression model and later performed feature selection on it and again ran a linear regression on this dataset. In order to verify if the newly created linear regression model is better than the previous one, we ran F-tests to ensure the significance of these models.

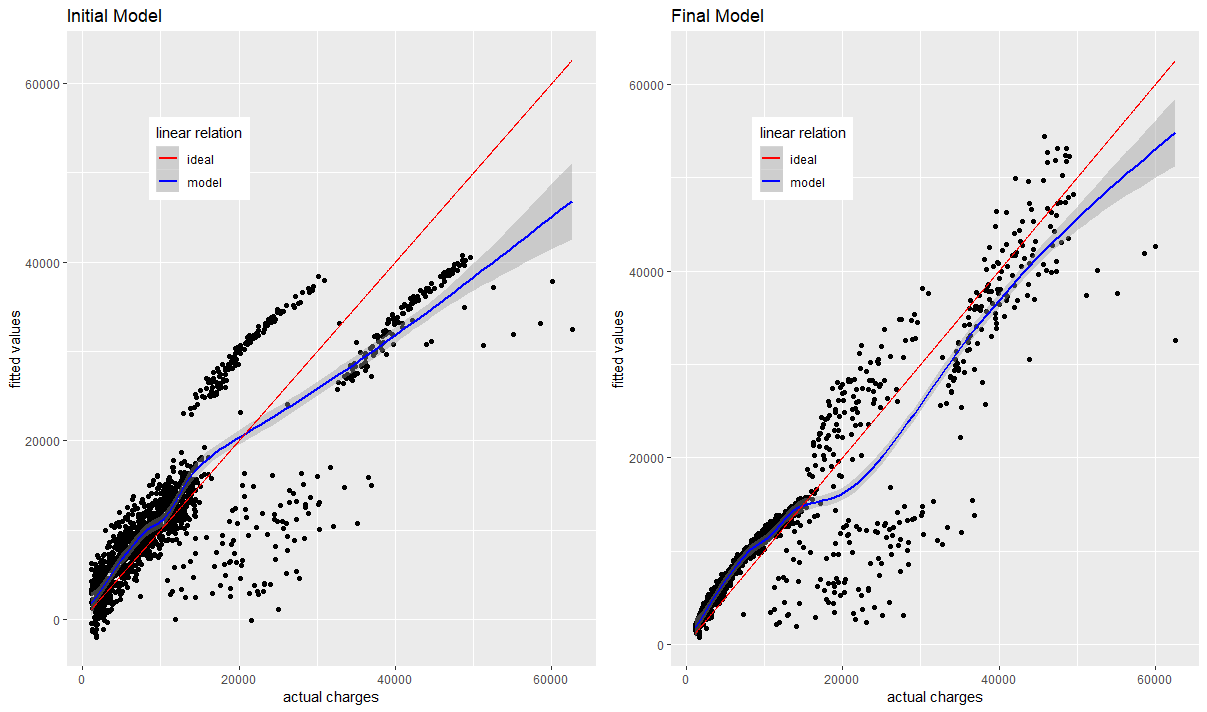


As we can see in the results of F-tests the p-value is 2.22e-15 which is less than 0.05 meaning that we can reject the Null hypothesis which states that both the initial and final model have no difference.

The p-value and the estimate indicate that our both models have significant difference between them and we can state that the feature selection helped us better our initial model.

**Interpreting Fitted Values vs Actual Values:**

Let us look at the fitted vs Actual values and see how our model persons.



As per the plots the actual value is close to the predicted value in our updated model as compared to our initial model.

**Conclusion**

The model we have built helps in interpreting predictors influence of predictors on the outcome. We determined the significance of the data that we have using tests like ANOVA and t and F tests and they helped us understand the importance of each predictors. These estimations can be used to summarize data in a useful way.

**Code**

######## INSURANCE PREDICTION DATA MODELS ##############################

######################### LOAD LIBRARIES ##########################

library(magrittr)

library(car)

library(broom)

library(ggplot2)

library(gridExtra)

library(knitr)

library(dplyr)

getwd()

alldata <- read.csv(file = "insurance.csv")

summary(alldata)

str(alldata)

####################### Data Distribution #######################

#Plotting raw data to see the distribution of the data

attach(alldata)

par(mfrow=c(3,4))

qq\_plot1 <- qqnorm(alldata$age,pch = 1, frame = FALSE,xlab = "age")

qq\_plot1 <- qqline(alldata$age, col = "steelblue", lwd = 2)

qq\_plot2 <- qqnorm(as.numeric(alldata$sex), pch = 1, frame = FALSE,xlab = "sex")

qq\_plot2 <- qqline(as.numeric(alldata$sex), col = "steelblue", lwd = 2)

qq\_plot3 <- qqnorm(alldata$bmi, pch = 1, frame = FALSE,xlab = "bmi")

qq\_plot3 <- qqline(alldata$bmi, col = "steelblue", lwd = 2)

qq\_plot4 <- qqnorm(alldata$children, pch = 1, frame = FALSE,xlab = "children")

qq\_plot4 <- qqline(alldata$children, col = "steelblue", lwd = 2)

qq\_plot5 <- qqnorm(as.numeric(alldata$smoker), pch = 1, frame = FALSE,xlab = "smoker")

qq\_plot5 <- qqline(as.numeric(alldata$smoker), col = "steelblue", lwd = 2)

qq\_plot6 <- qqnorm(as.numeric(alldata$region), pch = 1, frame = FALSE,xlab = "region")

qq\_plot6 <- qqline(as.numeric(alldata$region), col = "steelblue", lwd = 2)

qq\_plot7 <- qqnorm(alldata$charges, pch = 1, frame = FALSE,xlab = "charges")

qq\_plot7 <- qqline(alldata$charges, col = "steelblue", lwd = 2)

####################### Data Cleaning ##########################

#find columns which have NA and need to be imputed

x <- unlist(lapply(alldata, function(x)

any(is.na(x))))

x

# Plotting Boxplots to View Outliers in Important Variables

boxplot\_column1 <-

qplot(

x = alldata$age,

y = alldata[, 1],

geom = "boxplot" ,

xlab = "",

ylab = "age"

)

boxplot\_column2 <-

qplot(

x = alldata$sex,

y = alldata[, 2],

geom = "boxplot" ,

xlab = "",

ylab = "sex"

)

boxplot\_column3 <-

qplot(

x = alldata$bmi,

y = alldata[, 3],

geom = "boxplot" ,

xlab = "",

ylab = "bmi"

)

boxplot\_column4 <-

qplot(

x = alldata$children,

y = alldata[, 4],

geom = "boxplot" ,

xlab = "",

ylab = "Children"

)

boxplot\_column5 <-

qplot(

x = alldata$smoker,

y = alldata[, 5],

geom = "boxplot" ,

xlab = "",

ylab = "Smoker"

)

boxplot\_column6 <-

qplot(

x = alldata$region,

y = alldata[, 6],

geom = "boxplot" ,

xlab = "",

ylab = "Region"

)

grid.arrange(

boxplot\_column1,

boxplot\_column2,

boxplot\_column3,

boxplot\_column4,

boxplot\_column5,

boxplot\_column6,

ncol = 6

)

#Removing Ouliers determined using the Boxplot#

outliers <- boxplot(alldata$bmi, plot=FALSE)$out

print(outliers)

alldata<-alldata[-which(alldata$bmi %in% outliers),]

qplot(

x = alldata$bmi,

y = alldata[, 3],

geom = "boxplot" ,

xlab = "",

ylab = "bmi"

)

#Running initial Linear Regression model #

lm.fit <- lm(formula = charges~., data = alldata)

summary(lm.fit)

#Performing mixed selection for performing feature selection

step.lm.fit <- MASS::stepAIC(lm.fit, direction = "both",

trace = FALSE)

#Perform Multicollinearity check

vif(step.lm.fit) %>%

knitr::kable()

#Standardized residual plot to determine the improvement in the predicted model

residualPlot(step.lm.fit, type = "rstandard")

#Non-linearity check

ceresPlots(step.lm.fit)

#Transforming Non-linearity of BMI

step.lm.fit.new <- update(step.lm.fit, .~.+I(bmi^1.25))

ceresPlots(step.lm.fit.new)

####################### Significance Tests #######################

####### ANOVA #####

#ANOVA test to verify the statistcial importance of the variables used

anova(step.lm.fit, step.lm.fit.new, test = "F")

#Residual Plot verification

residualPlot(step.lm.fit.new, type = "rstandard")

#Modifying the step fit to fit the data better

lm.fit1 <- update(step.lm.fit.new, ~ .+bmi\*smoker)

residualPlot(lm.fit1, type = "rstandard", id=TRUE)

#ANOVA test

anova(step.lm.fit.new, lm.fit1, test = "F")

#checking adjusted R-square value

summary(lm.fit1)$adj.r.squared

####### F-Test #####

res.ftest <- var.test(lm.fit,lm.fit1, data = alldata)

res.ftest

####################fitted values of initial model##################################

#Creating function for applying ggplot on intial and predicted model

fitted\_vs\_actual <- function(predictions, title){

ggplot(predictions, aes(x=alldata$charges, y=fit))+

geom\_point()+

geom\_smooth(aes(color = 'model'))+

geom\_line(aes(x=seq(min(alldata$charges),max(alldata$charges), length.out = 1329),

y=seq(min(alldata$charges),max(alldata$charges), length.out = 1329),

color = 'ideal'))+

labs(x="actual charges", y="fitted values") +

scale\_color\_manual('linear relation', values = c('red', 'blue')) +

theme(legend.position = c(0.25, 0.8))+

ggtitle(title)

}

#Predicting the intitally fitted linear regressions model

fitted\_init <- predict(lm.fit, alldata,

interval = "confidence") %>%

tidy()

g1 <- fitted\_vs\_actual(fitted\_init, "Initial Model")

#Predicting the final fitted linear regressions model after feature selection

fitted\_final <- predict(lm.fit1, alldata,

interval = "confidence") %>%

tidy()

g2 <- fitted\_vs\_actual(fitted\_final, "Final Model")

#Plot initial and final predicted model

grid.arrange(g1,g2, ncol = 2)