Fitting the linear regression model for medical expenses and checking the model adequacy for the fitted model

INDUMATHI S

26/03/2021

# Abstract:

Now days, medical expenses is leading the main role in our life. We are spending more money on medical. We know that, life style and physical factors of the person impact the health condition. So, the medical expenses will depend on the some factors like age, bmi, smoking habit. In this analysis, we aims to find the linear relation between the medical expenses and others independent factors by using insurance dataset from kaggle which have & variables with 1338 observations and fit the model for them. The regression model can predict with 84% accuracy the charges. Also we aims to check the model adequacy to know whether the model is trust worthy or not. In analysis, we found that the normality assumption is violating which is an indication of an inadequate model. The violation of normality tells the non-linearity or outliers of the data. Therefore, the non-linear model may be suitable for this data. But this linear model is not adequate.

# Introduction:

The expenditure of medical uses of the person is depended by many other factors of the person which include smoking, age and etc. To predicting the medical expenses based on other factors, we need to fit the linear model. Once we fit the model we have to check the goodness of fit and model adequacy. For checking the model adequacy, we should verify the existence of multicollinearity. If we model the multiple linear regression with more independent variables, then the independent variables might be dependent to each other i.e) one independent variable is explained by other independent variables. This is called ‘multicollinearity’. If there exist the multicollinearity then we have to remove some highly dependent regressors. Variance influence factor(vif) is the measure of the amount of multicollinearity in a set of multiple regression variables. So using this vif, the highly dependent regressors can be removed. Then we should do the residual analysis. Residual analysis is verifying all the assumptions of residuals. The assumptions of residuals is the residuals are uncorrelated with zero mean and constant variance with 1 and they are normally distributed. There is a test called studentized Breusch-Pagan test (bptest) for checking the homoscedasticity(Constant variance). We have the test called Shapiro-Wilk normality test for testing normality. There is a test called dwtest which is used to test the autocorrelation. Using these test and some plots, we can verify the assumptions of residuals. If the residuals are satisfied the assumptions of the residuals then the model will be trust worthy. So, we can predict the dependent variable by using the model which satisties the all assumptions of residuals.

# About Dataset:

The dataset we use here is insurance dataset from kaggle which have 7 variables with 1338 observations. The dataset includes information about the insurance policy holder, their dependents, and their medical expenses throughout a year.

Age: Age of primary policyholder.

Sex: Sex of the policy policyholder.

BMI: Body Mass Index of policyholder, defined as the body mass divided by the square of the body height (kg/m2).

Smoker status: Whether the policyholder is a smoker or a non-smoker.

Children: Number of children/dependents covered in the policy.

Region of residence: Residential areas of the policy holder (in the US) - North East, South East, South West, North West.

Charges: Yearly medical expenses billed by the medical insurance provider ($).

Here, charges(Y) is depended on other variables.

# Importing the dataset:

library(readxl)  
data <- read\_excel("D:/Second Sem/Regression Analysis/Dataset/insurance.xlsx")

attach(data)  
head(data)

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

# Dummy coding for categorical variables:

Dummy coding is probably the most commonly used coding scheme. It basically consists of creating dichotomous variables(nominal variable with 2 levels) where each level of the categorical variable is contrasted to a specified reference level.

# creating the factor variable sex  
data$sex = factor(data$sex, labels=c("female","male"))

sex variable has 2 levels which are female and male. Automatically R itself choose first level as reference level. Here, female is the first level which is fixed as reference level and compare the mean of dependent variable(charges) for other level of sex male to the reference level of female as follows:

sexmale = 0 if sex indicates female sexmale = 1 if sex indicates male (sexmale)

# creating the factor variable children  
data$children = factor(data$children, labels=c(0,1,2,3,4,5))

# The contrast matrix for categorical variable with six levels  
contrasts(data$children)

## 1 2 3 4 5  
## 0 0 0 0 0 0  
## 1 1 0 0 0 0  
## 2 0 1 0 0 0  
## 3 0 0 1 0 0  
## 4 0 0 0 1 0  
## 5 0 0 0 0 1

children variable has 6 levels which are number of children in the family 0, 1, 2, 3, 4, 5. Here, 0 is the first level which is fixed as reference level and compare the mean of dependent variable(charges) for each level of children 1, 2, 3, 4 and 5 to the reference level of 0 as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Level of children | children.1 (0vs.1) | children.2 (0vs.2) | children.3 (1vs3) | children4 (0vs4) | children5 (0 vs 5) |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 | 0 |
| 5 | 0 | 0 | 0 | 0 | 1 |

# Creating the factor variable smoker  
data$smoker = factor(data$smoker, labels=c("no", "yes"))

# The contrast matrix for categorical variable with two levels  
contrasts(data$smoker)

## yes  
## no 0  
## yes 1

smoker variable has 2 levels which are no and yes. Here, no is the first level which is fixed as reference level and compare the mean of dependent variable(charges) for other level of yes to the reference level of no as follows:

smokeryes = 0 if sex indicates no smokeryes = 1 if sex indicates yes

# creating the factor variable region  
data$region = factor(data$region, labels=c("northeast","northwest","southeast","southwest"))

region variable (region of insurance holder) has 4 levels which are northeast, northwest, southeast, southwest. Here, northeast is the first level which is fixed as reference level and compare the mean of dependent variable(charges) for each level of region southeast, northwest, southwest to the reference level of northeast as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Level of region | region.northwest (NE vs. NW) | region.southeast (NE vs. SE) | region.southwest (NE vs. SW) |
| Northeast (NE) | 0 | 0 | 0 |
| Northwest (NW) | 1 | 0 | 0 |
| Southeast (SE) | 0 | 1 | 0 |
| Southwest (SW) | 0 | 0 | 1 |

# Backward elimination procedure for selecting the best model:

fitall= lm(charges~., data=data) # regression model with all variables

reg=step(fitall,direction="backward")

## Start: AIC=23318.92  
## charges ~ age + sex + bmi + children + smoker + region  
##   
## Df Sum of Sq RSS AIC  
## - sex 1 5.4429e+06 4.8645e+10 23317  
## <none> 4.8639e+10 23319  
## - region 3 2.2641e+08 4.8865e+10 23319  
## - children 5 6.3800e+08 4.9277e+10 23326  
## - bmi 1 5.0897e+09 5.3729e+10 23450  
## - age 1 1.7106e+10 6.5745e+10 23720  
## - smoker 1 1.2161e+11 1.7025e+11 24993  
##   
## Step: AIC=23317.07  
## charges ~ age + bmi + children + smoker + region  
##   
## Df Sum of Sq RSS AIC  
## <none> 4.8645e+10 23317  
## - region 3 2.2613e+08 4.8871e+10 23317  
## - children 5 6.3668e+08 4.9281e+10 23325  
## - bmi 1 5.0853e+09 5.3730e+10 23448  
## - age 1 1.7131e+10 6.5775e+10 23719  
## - smoker 1 1.2220e+11 1.7084e+11 24996

reg

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + smoker + region,   
## data = data)  
##   
## Coefficients:  
## (Intercept) age bmi children1   
## -11977.3 257.3 336.4 388.7   
## children2 children3 children4 children5   
## 1635.2 963.0 2938.6 1106.4   
## smokeryes regionnorthwest regionsoutheast regionsouthwest   
## 23824.2 -379.4 -1032.4 -952.2

The less significantly contributing independent variable(sex) has been removed by backward elimination procedure.

**The final fitted model is *charges = -11977.3 + 257.3age + 336.4bmi + 388.7children1 + 1635.2children2 + 963.0children3 + 2938.6children4 + 1106.4children5 + 23824.2smokeryes + (-379.4)regionnorthwest + (-1032.4)regionsoutheast + (-952.2)\*regionsouthwest.***

The sign of regression coefficient tells about the linear relationship. Here, the coefficient of all variables except region are positive which tells that there exist the positive linear relationship between these independent variables & dependent variable(charges). the coefficient of region is negative which tells that there exists the negative linear relationship.

The regression coefficient of age tells that the one unit of age increases and other independent variables is fixed then the average of charges will increase by 257.3 dollars.

The regression coefficient of bmi tells that the one unit of bmi increases and other independent variables is fixed then the average of charges will increase by 336.4 dollars.

The regression coefficient of children1 compares the mean of the charge for level of children 0 and 1. The expected difference in charges variable between level of children 0 and 1 is 388.7 dollars in other words, the difference between the mean of the charges for level 0 and the mean of the charges for level 1. A person with one children will increase the charges by 388.7 dollars when other variables are fixed. Like this, the estimated regression coefficient of children 2 is 1635.2 dollars which gives the expected difference in charges variable between level of children 0 and 2. A person with two children will increase the charges by 1635.2 dollars when other variables are fixed, The estimated regression coefficient of children 3 is 963.0 dollars which gives the expected difference in charges variable between level of children 0 and 3. A person with three children will increase the charges by 963.0 dollars when other variables are fixed, The estimated regression coefficient of children 4 is 2938.6 dollars which gives the expected difference in charges variable between level of children 0 and 4. A person with four children will increase the charges by 2938.6 dollars when other variables are fixed and The estimated regression coefficient of children 5 is 1106.4 dollars which gives the expected difference in charges variable between level of children 0 and 5. A person with five children will increase the charges by 1106.4 dollars when other variables are fixed

The regression coefficient of smoker yes compares the mean of the charges for level of no and yes. The expected difference in charges variable between level of yes and no is 23824.2 dollars in other words, the difference between the mean of the charges for level no the mean of the charges for level yes. A person be a smoker will increase the charges by 23847.5 dollars.

The regression coefficient of region northwest compares the mean of the charge for level of region northeast and northwest. The expected difference in charges variable between level of region northeast and northwest is -379.4. A person living in the northwest might decrease the charges by 379.4 dollars. Like this, the estimated regression coefficient of region southeast is -1032.4 which gives the expected difference in charges variable between level of region northeast and southeast. A person living in the southeast might decrease the charges by 1032.4 dollars., The estimated regression coefficient of region southwest is -952.2 which gives the expected difference in charges variable between level of region northeast and southwest. A person living in the southwest might decrease the charges by 952.2 dollars.

# Multicollinearity:

# Variance influence factor(vif):  
  
library(car)

## Loading required package: carData

vif(reg)

## GVIF Df GVIF^(1/(2\*Df))  
## age 1.020006 1 1.009954  
## bmi 1.106363 1 1.051838  
## children 1.024128 5 1.002387  
## smoker 1.012092 1 1.006028  
## region 1.109896 3 1.017530

variance influence factor (vif) is used to measure the amount of multicollinearity in the set of independent variables. The vif > 5 indicates that the associated independent variable is highly correlated with other independent variables in the model. If vif lies between 1 to 5 indicates that there is moderate correlation between the corresponding variable and other independent variables. The vif is equal to 1 indicates that there is no correlation between the corresponding variable and other independent variables.

Here, **vif of all the independent variables are less than 5. So, the final fitted model has the negligible multicollinearity. Therefore, the assumption of multicollinearity is satisfied**.

The final fitted model is Sale price of the charges ~ age + bmi + children + smoker + region.

# Goodness of fit:   
  
summary(reg)

##   
## Call:  
## lm(formula = charges ~ age + bmi + children + smoker + region,   
## data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11620.3 -2883.5 -945.6 1513.0 29986.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11977.26 984.79 -12.162 < 2e-16 \*\*\*  
## age 257.30 11.91 21.609 < 2e-16 \*\*\*  
## bmi 336.39 28.57 11.774 < 2e-16 \*\*\*  
## children1 388.71 421.17 0.923 0.356211   
## children2 1635.23 466.52 3.505 0.000471 \*\*\*  
## children3 962.98 547.91 1.758 0.079055 .   
## children4 2938.65 1238.56 2.373 0.017804 \*   
## children5 1106.45 1455.33 0.760 0.447227   
## smokeryes 23824.24 412.80 57.714 < 2e-16 \*\*\*  
## regionnorthwest -379.44 476.40 -0.796 0.425908   
## regionsoutheast -1032.43 478.98 -2.155 0.031304 \*   
## regionsouthwest -952.16 478.00 -1.992 0.046577 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6057 on 1326 degrees of freedom  
## Multiple R-squared: 0.7519, Adjusted R-squared: 0.7498   
## F-statistic: 365.3 on 11 and 1326 DF, p-value: < 2.2e-16

we have obtained the adjusted R-square is 0.7498 which tells that **74% of charges is predicted by the independent variables(age, bmi, children, smoker, region). Therefore the model is good fit.**

Now, we have to verify the model assumptions of the final fitted model.

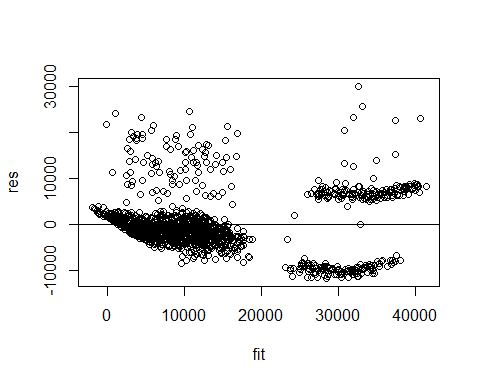
# Residual analysis:

In residual analysis, we have to verity all the properties or assumptions of residuals. The properties are: i) The expected value of residual should be zero. ii) The variance of residual should be 1 and constant. iii) The residuals are normally distributed and uncorrelated.

fit<-fitted.values(reg) # Fitted values  
res<-residuals(reg) # Residuals of model with interaction term

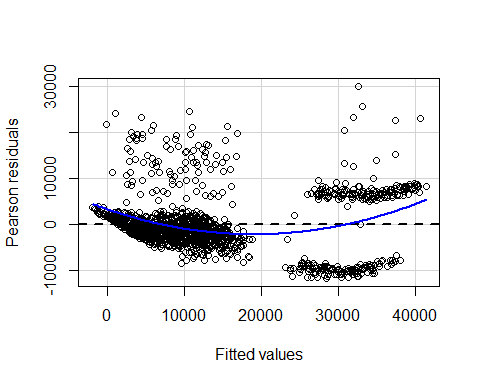
The plot of fitted value against residuals is used to find the nature of the variance. Here, we are going to find the nature of variance by using normal residual, standardized residual and studentised residual.

#(i) Scatter plot of fitted values against the residuals  
plot(fit,res)  
abline(0,0)



We know that If the plot of fitted vs residual is scatter then the variance is constant and linear or if the plot has funnel shape then the variance is non constant but linear. If the plot has oval shape (curve) then the variance is non constant and also non-linear. Here, We can clearly see that, **the points are following curve pattern. Therefore the variance of the residuals is not constant and the curve plot indicates the non-linearity**.

# To see the curve pattern  
residualPlot(reg)



The presence of a curve in the residual plot imply a problem with the linear assumption of the model. Since the residual plot of fitted values versus the residuals looks like curve, the model can be the non-linear. There are a various of reasons why a model can have this problem. The possibilities include a missing of:

1. Independent variable.
2. Polynomial term to model a curve.
3. Interaction term - An interaction effect exists when the effect of an independent variable on a dependent variable changes, depending on the value(s) of one or more other independent variables.

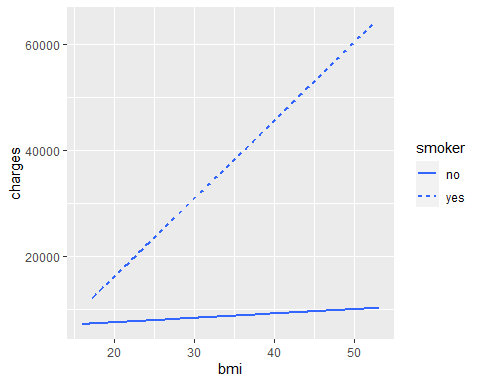
To fix the problem, we need to identify the missing information, variable, or higher-order term and include it in the model. After we correct the problem and refit the model, the residuals should look random.

In our dataset, **we can see the interaction that a person who smokes and has a high bmi may have higher charges as compared to a person who has lower bmi and is a non-smoker that is, the relationship between charges and bmi depends on the smoking habit**. To see this interaction visually:

# Interaction Plot  
  
library(ggplot2)

ggplot(data,aes(bmi,charges,linetype=smoker))+  
 geom\_smooth(method="lm",se=FALSE)

## `geom\_smooth()` using formula 'y ~ x'



To understand interaction effects, we have to compare the lines from the interaction plot. If the lines are parallel, there is no interaction. If the lines are not parallel, there is an interaction. Here, **we can clearly see that the lines are not parallel that is, there is interaction between bmi and smoker.**

**Since there is interaction between bmi and smoker, we have to refit the model with interaction term.**

# Refitting the linear model with interaction term  
  
reg\_int = lm(charges ~ age + children + bmi + smoker + region + bmi\*smoker, data=data) # bmi\*smoker is the interaction term

# Confidence intervel for regression coefficient:  
  
confint(reg\_int,level = 0.95)

## 2.5 % 97.5 %  
## (Intercept) -4106.50933 -724.87812  
## age 245.98211 283.36660  
## children1 -345.53945 976.32504  
## children2 795.33720 2259.55070  
## children3 144.29993 1863.88365  
## children4 1453.38610 5341.06513  
## children5 -479.02168 4089.53659  
## bmi -29.43991 71.01874  
## smokeryes -23523.39493 -17060.76938  
## regionnorthwest -1343.25421 152.22286  
## regionsoutheast -1954.73811 -451.29158  
## regionsouthwest -1975.46632 -474.80008  
## bmi:smokeryes 1334.46171 1540.78499

This confidence intervel tells that 95% sure about the true value of regression coefficient of indepandent variables lie in the respective intervals.

# Predicting the dependent variable(Charges)  
  
pred\_var = predict(reg\_int, data)  
head(pred\_var)

## 1 2 3 4 5 6   
## 21785.620 2162.881 5482.316 6195.068 6058.768 5121.316

head(data)

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

This values give the predicted values of response values of medical charges with respect to independent variables of first six observations.

pi<-predict(reg\_int,data,interval="confidence",level=0.95)  
head(pi)

## fit lwr upr  
## 1 21785.620 20868.193 22703.047  
## 2 2162.881 1360.790 2964.973  
## 3 5482.316 4540.502 6424.130  
## 4 6195.068 5484.835 6905.301  
## 5 6058.768 5423.084 6694.452  
## 6 5121.316 4415.012 5827.620

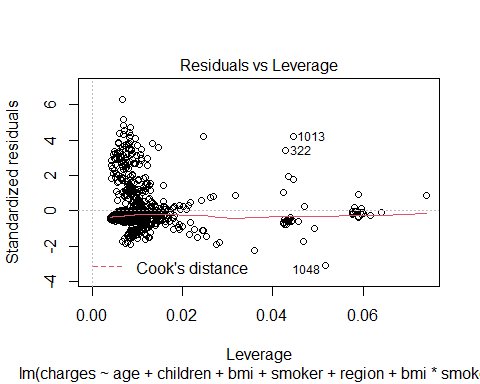
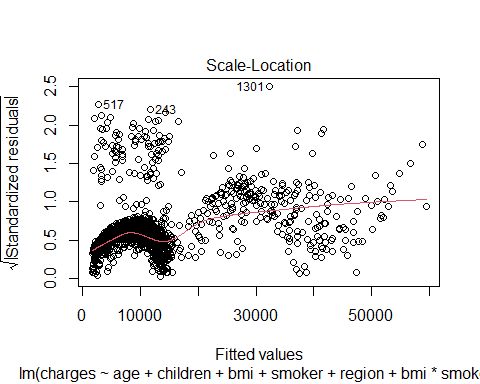
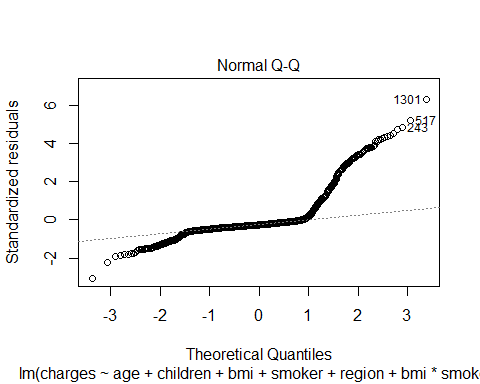
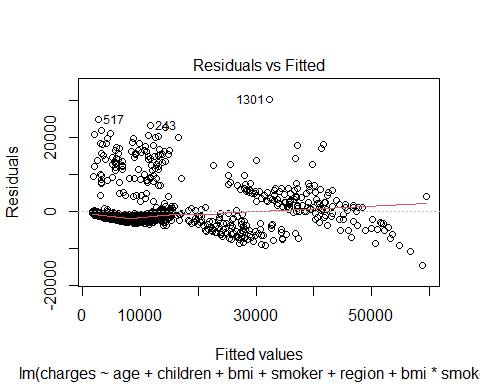
We know that, our prediction is based on fitted regression model. lower bound and upper bound is about the response variable. The fitted value is the response variable (charges) with independent variables for first six observations. This interval tells that 95% confidence that the true value of the fitted value lie between this intervals.

# Goodness of fit with interaction term  
summary(reg\_int)

##   
## Call:  
## lm(formula = charges ~ age + children + bmi + smoker + region +   
## bmi \* smoker, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14409.4 -1935.8 -1252.4 -370.2 30315.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2415.694 861.889 -2.803 0.005140 \*\*   
## age 264.674 9.528 27.778 < 2e-16 \*\*\*  
## children1 315.393 336.908 0.936 0.349373   
## children2 1527.444 373.190 4.093 4.52e-05 \*\*\*  
## children3 1004.092 438.277 2.291 0.022120 \*   
## children4 3397.226 990.867 3.429 0.000625 \*\*\*  
## children5 1805.257 1164.405 1.550 0.121292   
## bmi 20.789 25.604 0.812 0.416966   
## smokeryes -20292.082 1647.153 -12.319 < 2e-16 \*\*\*  
## regionnorthwest -595.516 381.158 -1.562 0.118436   
## regionsoutheast -1203.015 383.189 -3.139 0.001730 \*\*   
## regionsouthwest -1225.133 382.480 -3.203 0.001392 \*\*   
## bmi:smokeryes 1437.623 52.586 27.338 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4845 on 1325 degrees of freedom  
## Multiple R-squared: 0.8414, Adjusted R-squared: 0.8399   
## F-statistic: 585.7 on 12 and 1325 DF, p-value: < 2.2e-16

After fitting the model with interaction term, we have obtained the multiple R-square is 0.8397 which is higher than the R square of model without interaction term and it tells that 84% of medical charges of the patient is predicted by the independent variables. Therefore, the model is good fit.

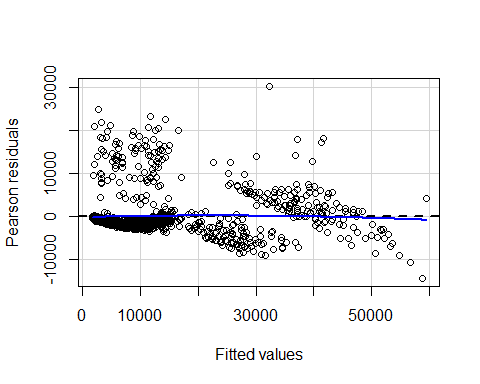
# Plots for the regression model with interaction term  
plot(reg\_int)



1.Residuals vs fitted -> used to check the linear relationship assumptions. Red line is horizontal in the graph i.e) there are no distinct patterns. Therefore, this is an indication for linear relationship

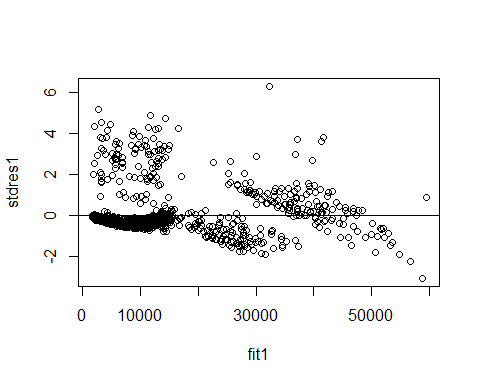
1. Normal Q-Q -> used to determine whether the residuals are normally distributed. We can see that the residuals points don’t follow straight line. Therefore, the residuals are not normally distributed.
2. Scale-location(or spread-location) is used to to check whether the residuals are spread along with the range of predicted value that is called homoscedasticity. If the plot has the horizontal line which indicates that the points are distributed randomly above and below the line then we can say that there is homoscedasticity. But in our plot, there is horizontal line so that means the points are equally spared. So, there is homoscedasticity. Therefore, the assumption on constant variance of residuals is satisfied.
3. Residuals vs Leverage -> is used for identifying influential points, there are some points with number in this plot. that points might impact the regression results when included or excluded from the analysis.

residualPlot(reg\_int)

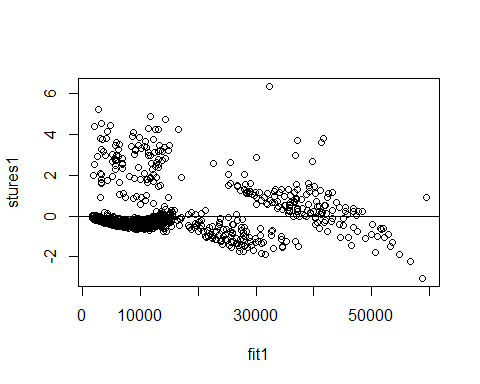


fit1<-fitted.values(reg\_int) # Fitted values of the model with interaction term  
res1<-residuals(reg\_int) # Residuals of the model with interaction term

#(ii)standardized residuals  
stdres1<-rstandard(reg\_int)  
  
#scatter plot of fitted values against standardized residual   
plot(fit1,stdres1)  
abline(0,0)

 Here, we didn’t see any pattern followed by the residuals. Therefore, there is linear relationship. Also, we know that if any standardized residuals is greater than absolute value of 3, then it might indicate an outlier. That particular value is outlier. From this graph, we can see that there are many points are greater than absolute value of 3. So they may be the outliers.

#(iii)studendized residuals  
stures1<-rstudent(reg\_int)  
  
#scatter plot of fitted values against the studendized residual   
plot(fit1,stures1)  
abline(0,0)

 Here, we didn’t see any pattern followed by the residuals. Therefore, there is linear relationship. Also, we know that if any studendized residuals is greater than absolute value of 3, then it might indicate an outlier. That particular value is outlier. From this graph, we can see that there are many points are greater than absolute value of 3. So they may be the outliers.

From scale location plot, we can find that variance is constant. But for more clarity, we can use bptest for homoscedasticity.

# Test for homoscedasticity (constant variance):

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

bptest(reg\_int)

##   
## studentized Breusch-Pagan test  
##   
## data: reg\_int  
## BP = 20.127, df = 12, p-value = 0.06472

Null hypothesis H0 : There is an homoscedasticity (constant variance) Alternative hypothesis H1 : There is no homoscedasticity

Since p value is greater than the significant level 0.05, we can accept the null hypothesis that is there is an homoscedasticity. Therefore, residuals has the constant variance.

# variance of standardized residuals  
var(stdres1)

## [1] 1.001331

# variance of studendized residuals  
var(stures1)

## [1] 1.008393

Here, the variance of the standardized and studendized residuals are approximately 1. Therefore, the assumption of variance that is the variance of residual should be 1 have been verified.

# mean of standardized residuals  
mean(stdres1)

## [1] -9.217809e-05

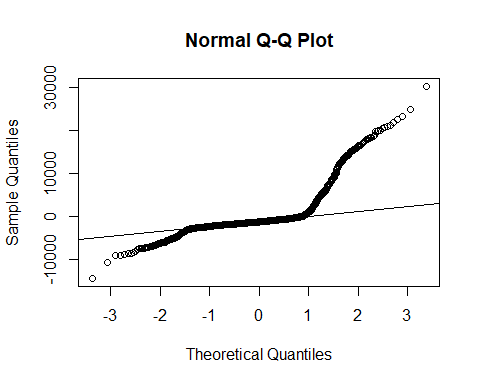
#mean of studendized residuals  
mean(stures1)

## [1] 0.0008559649

Here the expected values of standardized and studendized residuals are approximately 0 which satisfies the mean assumption of residuals.

# To check the normality assumption:

# Normality Plot  
qqnorm(res1)  
qqline(res1)

 The normal QQ plot is used to check the normality of residuals. The most of the points are not lying on the straight line which indicates the residuals are not normally distributed.

# normality test  
shapiro.test(res1)

##   
## Shapiro-Wilk normality test  
##   
## data: res1  
## W = 0.67255, p-value < 2.2e-16

Null hypothesis H0: The residuals are normally distributed

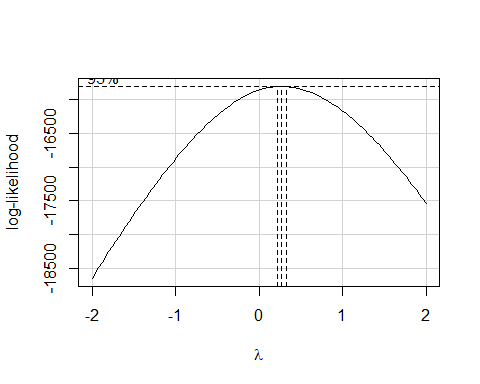
Alternative hypothesis H1: The residuals are not normally distributed

Since p value is less than the significant level 0.05, there is no evidence to accept the null hypothesis. Therefore by the alternative hypothesis, the residuals are normally distributed.

# Box cox transformation  
  
library(MASS)

## Warning: package 'MASS' was built under R version 4.0.4

b = boxCox(reg\_int)

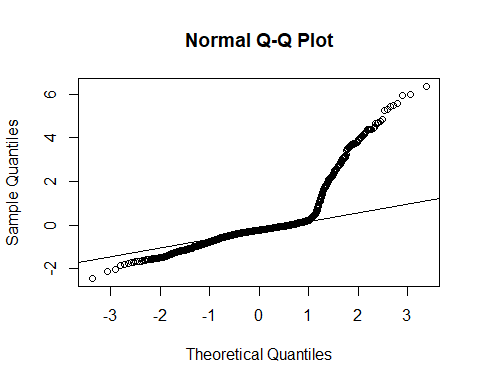


lambda <- b$x # lambda values  
  
lik <- b$y # log likelihood values for SSE  
  
bc <- cbind(lambda, lik) # combine lambda and likelihood values  
  
sorted\_bc <- bc[order(-lik),] # values are sorted to identify the lambda value for the maximum log likelihood for obtaining minimum SSE  
  
head(sorted\_bc, n = 10)

## lambda lik  
## [1,] 0.2626263 -15798.30  
## [2,] 0.3030303 -15798.87  
## [3,] 0.2222222 -15800.16  
## [4,] 0.3434343 -15801.86  
## [5,] 0.1818182 -15804.45  
## [6,] 0.3838384 -15807.25  
## [7,] 0.1414141 -15811.18  
## [8,] 0.4242424 -15815.01  
## [9,] 0.1010101 -15820.34  
## [10,] 0.4646465 -15825.13

# Model fitting (After transformation)  
  
lambda = 0.2626263  
reg\_box = lm(charges^(lambda) ~ age + children + bmi + smoker +region + bmi\*smoker)

res\_b =resid(reg\_box) # residuals of the model after transformation  
qqnorm(res\_b)  
qqline(res\_b)

 After transformation also, there is no changes in the QQ plot that is the most of the points are not lying on the straight line which indicates the residuals are not normally distributed.

# normality test after box cox transformation  
  
shapiro.test(res1)

##   
## Shapiro-Wilk normality test  
##   
## data: res1  
## W = 0.67255, p-value < 2.2e-16

Null hypothesis H0: The residuals are normally distributed

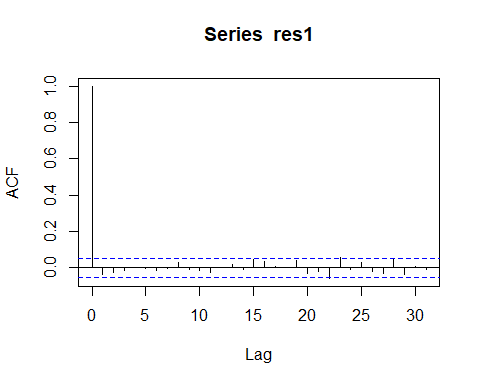
Alternative hypothesis H1: The residuals are not normally distributed

Since p value is less than the significant level 0.05, we reject the null hypothesis. Therefore by the alternative hypothesis, the residuals are not normally distributed.

We can see that after transformation also we are getting same p value. So the residuals are not normally distributed which is violating the normality assumption on residuals.

# To check whether the residuals are uncorrelated:

acf(res1)



In this graph, X axis is lags, Y axis is correlation and blue line is threshold line. The correlation of residual in different lags are plotted here. If the correlation of any lag is less than the threshold line then that correlation is negligible i.e) there is no correlation in that lag. But the correlation in lag 0 is always one because that indicates correlation of any lag itself. In our graph, There are no correlation lies outside the threshold line that is, correlation of all lags are negligible. Therefore the residuals are uncorrelated which satisfies the assumption on uncorrelated residuals.

# Test for autocorrelation:  
dwtest(reg\_int)

##   
## Durbin-Watson test  
##   
## data: reg\_int  
## DW = 2.0685, p-value = 0.8954  
## alternative hypothesis: true autocorrelation is greater than 0

Null hypothesis H0 : true autocorrelation is equal to 0 Alternative hypothesis H1 : true autocorrelation is greater than 0

Since the p value is greater than the significant level 0.05, we have to accept the null hypothesis. So by null hypothesis, we can say that the residuals are uncorrelated that is assumption on uncorrelated residuals is satisfied.

# Interpretation:

Since medical expenses depended on some factors of the person, we had analyzed the medical expenses based on their age, region, bmi, smoking habit and the number of children they have using multiple linear regression. In our dataset, we had four categorical variables. Before fitting the model, we have converted the those variables to factors, because R may not understand which are the categorical variables. After converting the categorical variables to factors, we have selected the significant independent variables for fitting the best model using backward elimination. Only variable **sex has been removed from model**. The fitted model is ***charges = -11977.3 + 257.3age + 336.4bmi + 388.7children1 + 1635.2children2 + 963.0children3 + 2938.6children4 + 1106.4children5 + 23824.2smokeryes + (-379.4)regionnorthwest + (-1032.4)regionsoutheast + (-952.2)regionsouthwest*.** Here, the coefficient of all variables except region are positive which tells that there exist the positive linear relationship between these independent variables & dependent variable(charges). the coefficient of region is negative which tells that there exists the negative linear relationship. For estimated model coefficients, it tells how much the amount of medical cost varies with one coefficient while other variables are held constant.

1. A unit increase in age will increase the charges by 257.3 dollars when other variables are fixed.
2. A unit increase in bmi will increase the charges by 336.4 dollars when other variables are fixed.
3. A person make one children will increase the charges by 388.7 dollars when other variables are fixed. A person make two children will increase the charges by 1635.2 dollars when other variables are fixed. A person make three children will increase the charges by 963.0 dollars when other variables are fixed. A person make four children will increase the charges by 2938.6 dollars when other variables are fixed. A person make five children will increase the charges by 1106.4 dollars when other variables are fixed.
4. A person be a smoker will increase the charges by 23824.2 dollars when other variables are fixed.
5. A person living in the southwest might decrease the charges by 952.2 dollars when other variables are fixed. A person living in the northwest might decrease the charges by 379.4 dollars when other variables are fixed. A person living in the southeast might decrease the charges by 1032.4 dollars when other variables are fixed.

While checking the multicollinearity, vif of all the independent variables are less than 5. So, the fitted model has the negligible multicollinearity. Therefore, the assumption on non-multicollinearity of residuals is satisfied.

The final fitted model is the ***charges ~ age + bmi + children + smoker + region*. For this model, we have obtained the adjusted R-square is 0.7498 which tells that 74% of medical charges of the person is predicted by the independent variables(age, bmi, children, smoker, region). Therefore the model is good fit.**

we also checked the model adequacy. For checking the model adequacy, we did the residual analysis.

From the residual vs fitted plot, we have seen that, the points were following curve pattern(oval). Therefore the variance of the residuals is not constant and the curve plot also indicates the non-linearity. To fix the non-linearlity problem, we have needed to identify the missing information, variable, or higher-order term and include it in the model. In our dataset, we could see *the interaction that is a person who smokes and has a high bmi may have higher charges as compared to a person who has lower bmi and is a non-smoker that is, the relationship between charges and bmi depends on the smoking habit*. Since **there is interaction** **between bmi and smoker**, we have refitted the model with interaction term (bmi*smoker).* ***The refitted model is******charges ~ age + children + bmi + smoker + region + bmi \* smoker***. **After fitting the model with interaction term**, we have obtained the multiple R-square is 0.8397 which is higher than the R square of model without interaction term and it tells that **84% of medical charges of the patient is predicted by the independent variables. Therefore, the model is good fit.**

**After refitted the model with interaction term**, For checking the model adequacy, we did the residual analysis for checking the model adequacy.

1. **Non-linearity** -> From **residuals vs fitted plot red line is horizontal in the graph** i.e) there are no distinct patterns. Therefore, this is an indication for linear relationship. And from standardized residuals vs fitted values plot, we didn’t see any pattern followed by the residuals. Therefore, there is linear relationship. Also, **we could see that there are many points are greater than absolute value of 3. So that points might be the outliers.**
2. **Assumption on zero mean and unity variance** -> The expected values of standardized and studendized residuals are approximately 0 which ***satisfied the mean assumption of residuals***. The variance of the standardized and studendized residuals are approximately 1. Therefore, the ***assumption of variance that is the variance of residual should be 1 have been verified***.
3. **Assumption on constant variance** -> From scale-location(or spread-location), here is horizontal line so that means the points are equally spared. So, there is homoscedasticity. And also from bp test, since p value is greater than the significant level 0.05, we can accept the null hypothesis that is there is an homoscedasticity. Therefore, the ***assumption on constant variance of residuals is satisfied***.
4. **Assumption on normality** -> From the normal QQ plot, The most of ***the points didn’t follow the straight line which indicates the residuals are not normally distributed.*** And also from sharpio wilk normality test, since p value is less than the significant level 0.05, there is no evidence to accept the null hypothesis. So by the alternative hypothesis, the residuals are normally distributed. Therefore, the ***assumption on*** ***normality of residuals was violated***.
5. **Box cox transformation** -> For fixing the normality problem, we have done the box cox transformation and refitted the model. After box cox transformation also, there is no changes in the QQ plot that is the most of the points didn’t follow the straight line which indicates the residuals are not normally distributed. We also could see that we were getting the same p value. So the residuals are not normally distributed which is violating the normality assumption of residuals.
6. **Assumption on uncorrelated residuals** -> From autocorrelation plot, there is ***no correlation lies outside the threshold line*** that is, correlation of all lags are negligible. From dwtest, Since the p value is greater than the significant level 0.05, we have to accept the null hypothesis. So by null hypothesis, we can say that the residuals are uncorrelated. Therefore the residuals are uncorrelated which satisfies the ***assumption on uncorrelated residuals.***

# Conclusion:

From the analysis, we found that the regression model can predict the medical charges with 84% accuracy that is the goodness of fit is good but the normality assumption of the residuals of the model is violating which is an indication of an inadequate model that is the prediction of the model is not trust worthy. The violation of normality tells the non-linearity or outliers of the data or both. So, Further we can try the non-linear model for the getting better model for this medical expenses.

**References:**

[1*] R Library contrast coding systems for categorical variables.* Institute for Digital Research & Education for statistical consulting.

[https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/#dummy](https://stats.idre.ucla.edu/r/library/r-library-contrast-coding-systems-for-categorical-variables/%23dummy)

[2] *Check Your Residual Plots to Ensure Trustworthy Regression Results!* . Statistics by Jim

<https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/>

[3] *Understanding Interaction Effects in Statistics***.** Statistics by Jim

<https://statisticsbyjim.com/regression/interaction-effects/>

[4] *Interaction Effects in Regression*. Stat Trek

<https://stattrek.com/multiple-regression/interaction.aspx>

[5] Interaction plot with continuous and categorical variable. Stack overflow

<https://stackoverflow.com/questions/10457756/r-interaction-plot-with-a-continuous-and-a-categorical-variable-for-a-glmm-lme>