HOW EXPENSIVE HEALTHCARE IS -

A Regression Case Study

DEPARTMENT OF MATHEMATICS AND STATISTICS

1. Rajdeep Saha (201380)
2. Sagnik Dey (201397)
3. Saumyadip Bhowmick (201408)
4. Shuvam Gupta (201421)
5. Soumik Karmakar (201428)

Course Project of **Regression Analysis (MTH416A)**

SUPERVISED BY DR. SHARMISHTHA MITRA

STUDENT'S DECLARATION

We, hereby declare that, the project work entitled "**How Expensive Healthcare is- A Regression Case Study**" is a record of an original work done by us under the guidance of **Dr. Sharmishtha Mitra** and this project work is submitted in the partial fulfillment of the requirements for the paper **MTH416A- Regression Analysis.**

We affirm that we have identified all our sources and that no part of our dissertation paper uses unacknowledged materials.

ACKNOWLEDGEMENT

We take this opportunity to express our sincere gratitude and deep regards to our project guide *Dr. Sharmishtha Mitra* for her exemplary guidance, monitoring and constant encouragement throughout the course of this project. Her blessings, help, time to time guidance will carry us a long way in the journey of life which we are about to embark.

We would also like to thank the institute authorities for giving us the chance to do the project and for providing the environment and necessary facilities required for the completion of our project. We would also like to thank PK Kelkar Library for having provided various reference books related to our project.

We would like to thank our department professors for teaching all the necessary topics with immense care which was needed to make the project fruitful.

Finally, we would like to extend our gratitude and thanks to our parents. Without their constant support and encouragement, it would not have been possible for us to proceed with our effort.

CONTENTS

➢ Introduction:

Healthcare is the maintenance or improvement of the health by the prevention, treatment, recovery of diseases, illness and other physical and mental impairments in people, making it an important aspect in day to day life. It is heartening to see how healthcare system has advanced in recent years across the world, but despite these positive developments, healthcare remains to be expensive to a lot of families or individuals, specially belonging to the lower middle-class category. As a result, a number of corporate firms are trying to observe the factors that can influence healthcare cost and if a factor influences the healthcare cost, whether its influence is significant or negligible in order to set up a model that can calculate estimated healthcare cost depending upon the factors that have impact on healthcare. The main objective behind this study is to check whether we can find any relationship between healthcare cost and the factors that are supposed to influence the healthcare cost and use this model to predict the future healthcare cost beforehand.

➢ Data Description:

We have obtained the dataset from https://www.kaggle.com/ .

Our data consists of 1338 observations with the information on the following variables, -

- **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:** Smoking
- **region:** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges:** Individual medical costs billed by health insurance.

Our data is of the following form, -

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622
46	female	33.44	1	no	southeast	8240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	northeast	6406.411
60	female	25.84	0	no	northwest	28923.14
25	male	26.22	0	no	northeast	2721.321
62	female	26.29	0	yes	southeast	27808.73
23	male	34.4	0	no	southwest	1826.843
56	female	39.82	0	no	southeast	11090.72
27	male	42.13	0	yes	southeast	39611.76
19	male	24.6	1	no	southwest	1837.237
52	female	30.78	1	no	northeast	10797.34
23	male	23.845	0	no	northeast	2395.172
56	male	40.3	0	no	southwest	10602.39

Here, **charges** is the **response variable.**

The regressors we have used are of two types, out of which, **age, bmi, children** are **numeric regressors** and **sex, smoker and region** are **categorical** regressors.

(a) Regressor "sex" has 2 categories, - male and female. we indicate "male" by 1 and "female" by 0.
(b) Regressor "smoker" has two categories, - yes and no. Here, the category "yes" is indicated by 1 and the category "no" is indicated by 0.
(c) Regressor "region" has four categories, - northeast, northwest, southeast, southwest. In "region", we have used pairwise indicators.

Here, the objective of our study is to determine how these regressors are able to explain **charges** by building up a **multiple linear regression model** of charges on the regressors.

➢ Checking for Missing Values:

For successful analysis of the data, firstly we have to check whether there are any missing observations in the dataset or not, as, improper handling of missing observation can lead to inaccurate inference about the data.

If there are missing values in the data set, we may leave the data or do data imputation to replace the them. Suppose the number of cases of missing values is extremely small; then, we may drop or omit those values from the analysis. In statistical language, if the number of the cases is less than 5% of the sample, then we can drop them.

To check for the presence of missing values, we have run the appropriate code in R and obtained the following output, -

```
```{r}
ins2=read.csv("insurance_project.csv",na.strings=c("n.a.","?","NA","n/a", "na", "--","<NA>"))
sum(is.na(ins2))
```

[1] 0
```

Since, there are no missing observations in the dataset, we can safely proceed into further analysis.

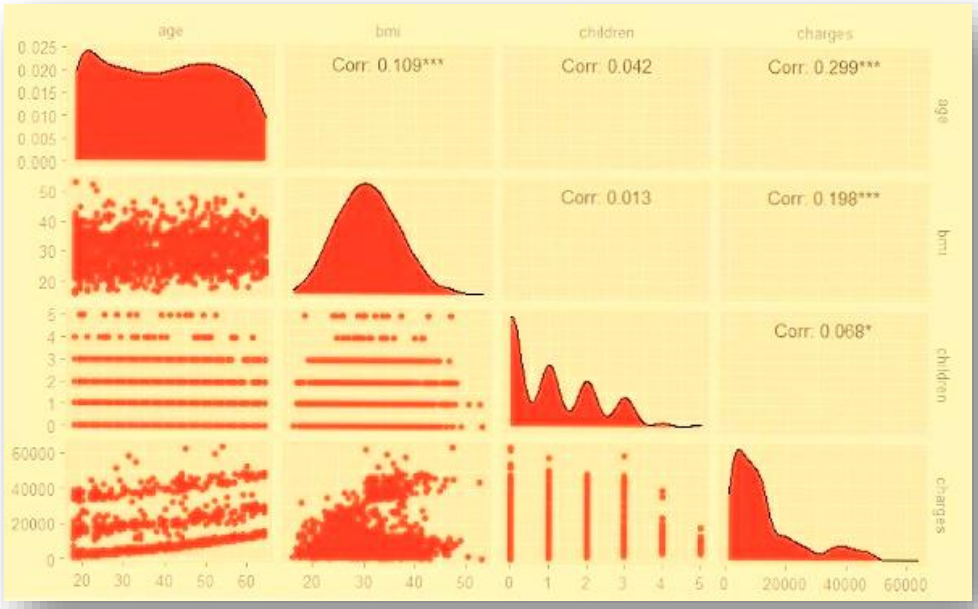
➢ Exploratory Data Analysis:

Before proceeding into theoretical analysis, we use a method to analyze the characteristics of the variables visually, named as Exploratory Data Analysis.

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

Exploring the regressors proposed to be used in the model and taking clue from the pictorial analysis, we can make a guess about the lineup of our analysis other than the formal modelling or hypothetical testing.

Since, in our dataset we have both numeric and categorical variables, first, we explore the numeric variables graphically.



We observe that,

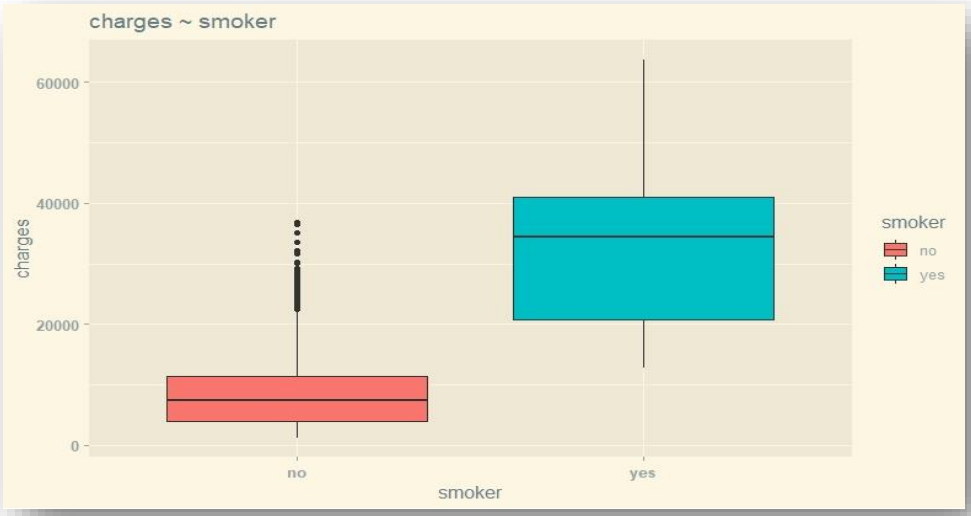
(i) Charges is positively skewed, which implies presence of outliers.

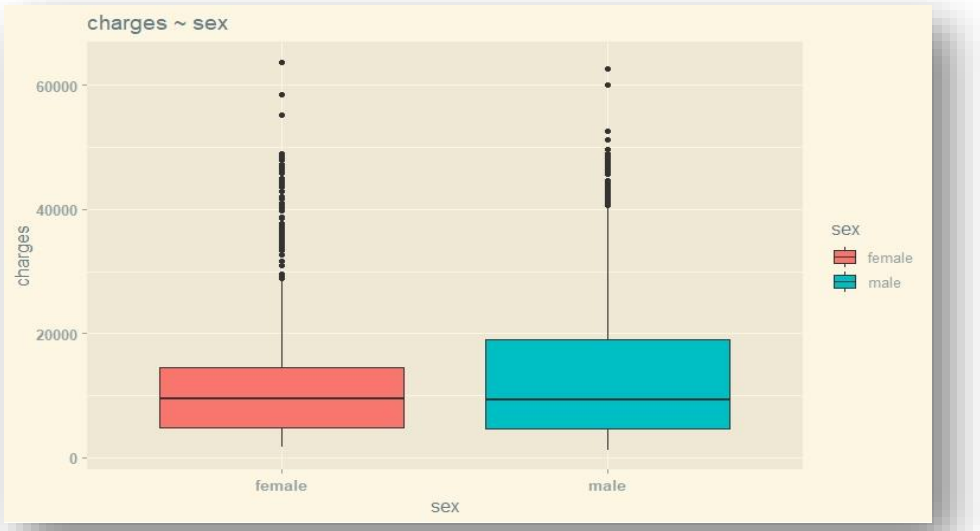
(ii) Among the regressors, only number of children seems to be positively skewed, indicating presence of high leverage points.

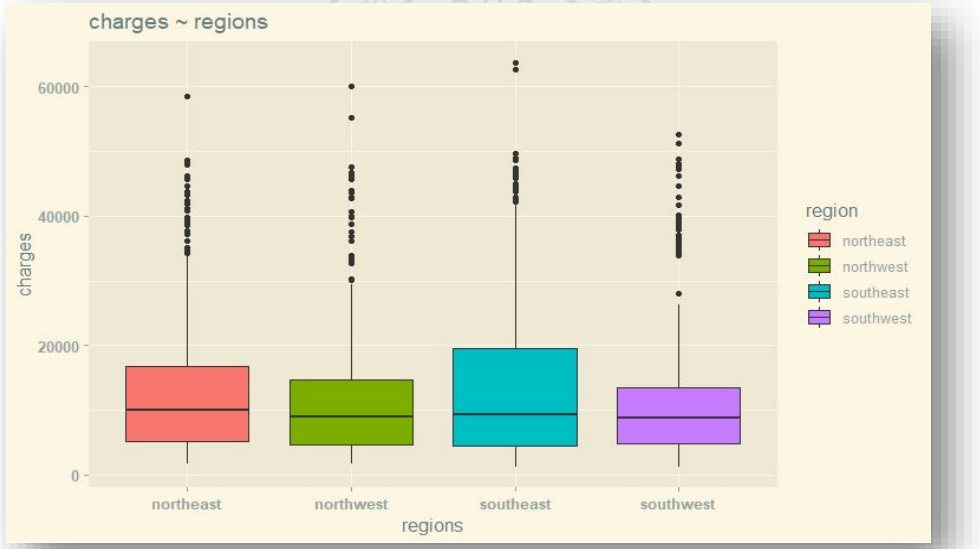
(iii) Correlation between the response variable and the regressors are not very high.

(iv) The scatterplot of charges with age and bmi does not show any specific pattern, but there is some clustering in both the scatterplots.

Now, we shall look at the **categorical variables.** We have used **Boxplot** for the visual analysis.







From the boxplots we have obtained, we observe that,

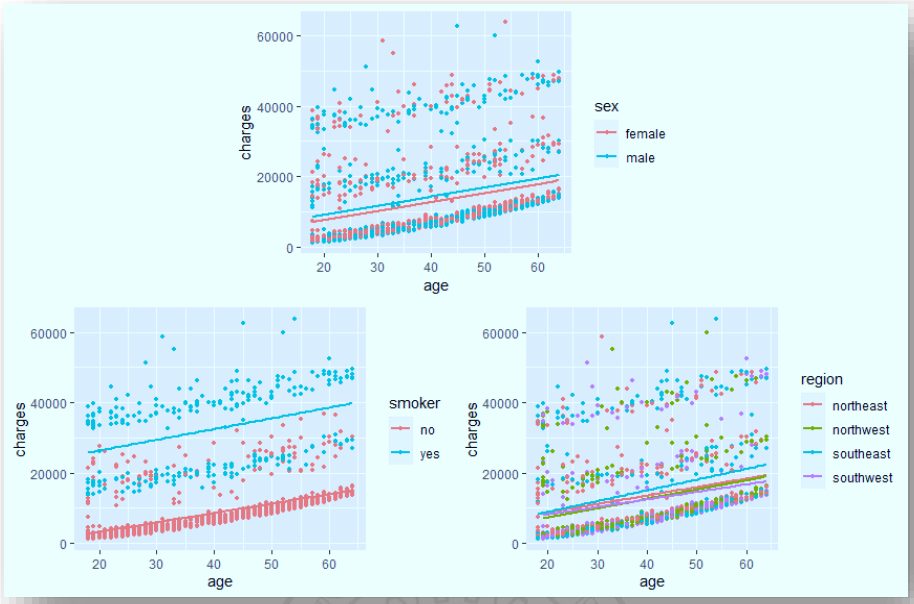
(i) Persons who smoke pay way more healthcare charges than the persons who don't smoke.

(ii) Median value of healthcare charges paid by males and females are more or less same.

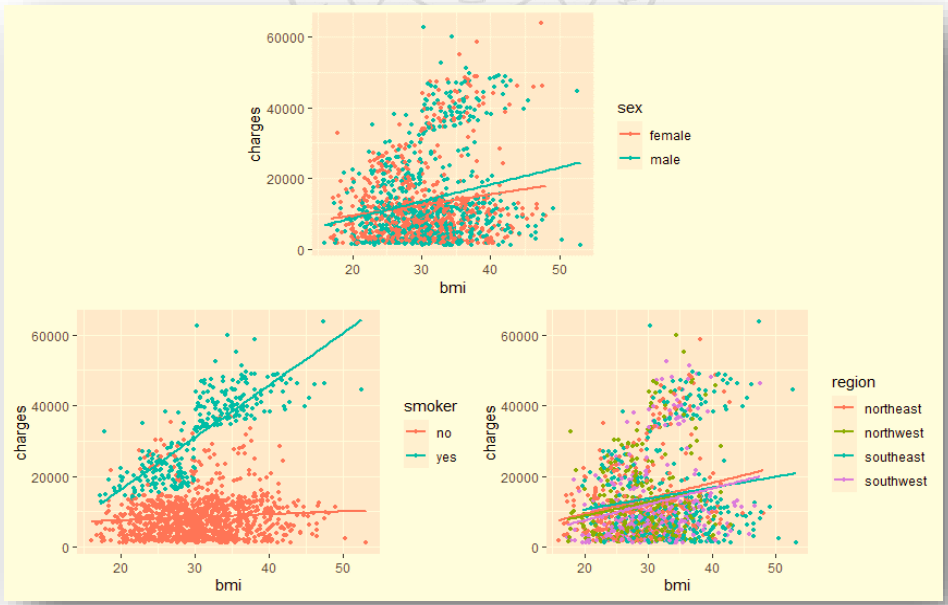
(iii) There are outliers in almost all the boxplots. So, before proceeding into further analysis, we have to take necessary action for the removal of them.

Next, we will observe the possible **interaction** between the variables.

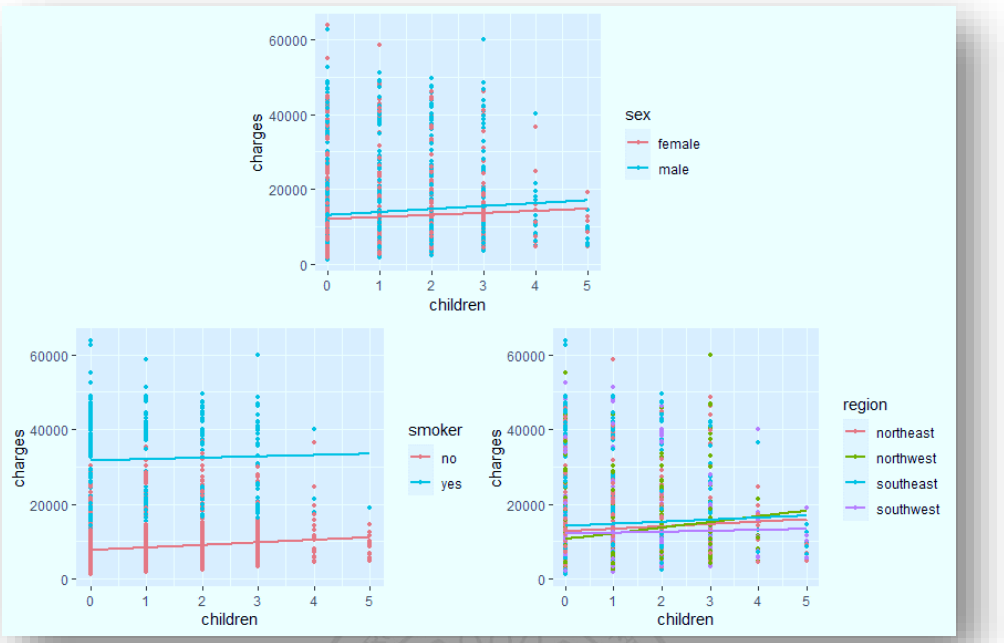
(i) Sex, Smoker and Region with age:



(ii) Sex, Smoker and Region with bmi:



(iii) Sex, Smoker and Region with Children:



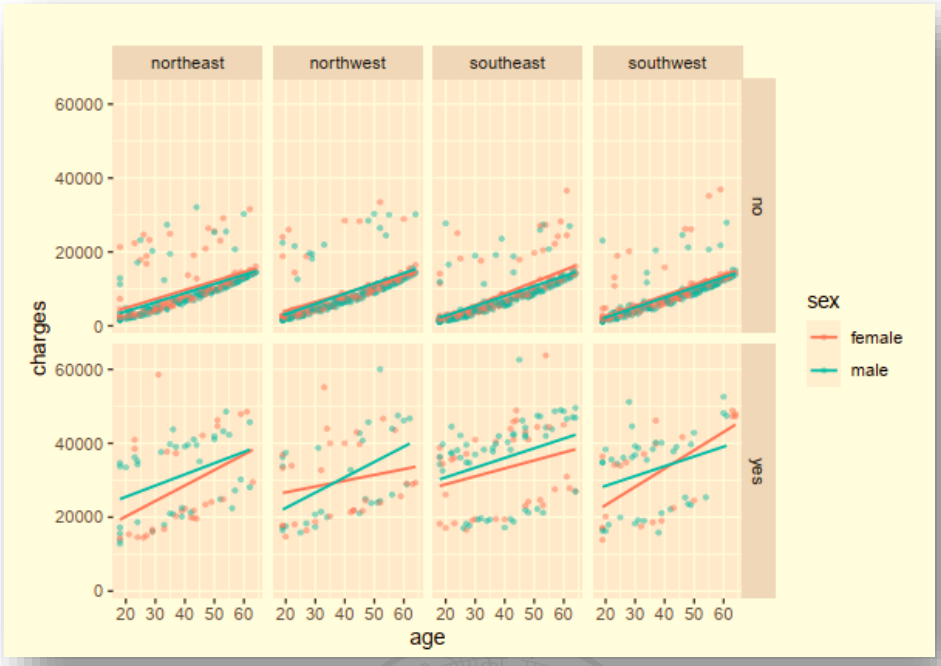
Observations:

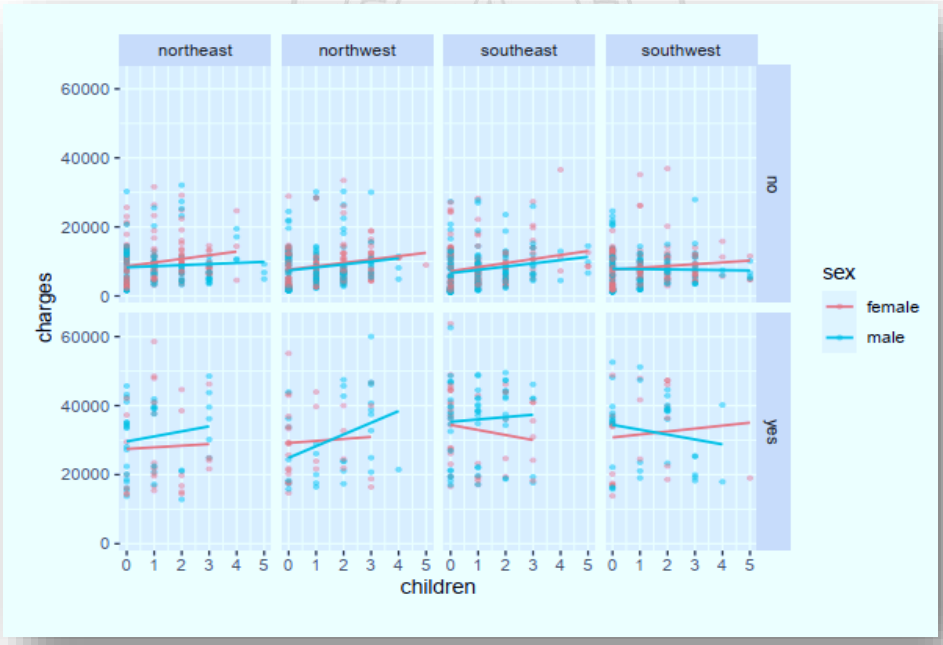
(a) Among the same age group, male persons have to pay more healthcare charges than females and as usual, smokers have to pay more healthcare charges than non smokers.
(b) Among the persons of same body mass index(bmi) males and smokers have to pay more healthcare charges. In case of children, only smoking habits cause significant rise in healthcare charges.

It is to be noted that region doesn't play any significant role in determining healthcare charges.

In general, looking at the three interaction plots, it is clear that, the difference in healthcare charges is large between smokers and non- smokers, i.e., the category smokers plays a much significant role in determining healthcare charges.

In order to observe these interactions properly, we made another set of plots, -


northeast
northwest
southeast
southwest
no
yes
charges
60000
40000
20000
0
20 30 40 50 60
age
sex
female
male


northeast
northwest
southeast
southwest
no
yes
charges
60000
40000
20000
0
0 1 2 3 4 5
children
sex
female
male

Observations:

(a) Among the smokers, males of all ages of northeast and southeast region pay more charges than the females. However, in northwest region, after a certain age, males pay more than females and in southwest region, after a certain age, females pay more than males.
(b) Among the non smokers, having same bmi, males and females pay almost same healthcare charges and we can also observe clustering of charges for both males and females. Among the smokers, having same bmi, we can also observe there is not so much difference in healthcare charges, paid by males and females, but, the amount paid rises too high for both males and females.
(c) For non smokers, number of children doesn't cause significant change in the healthcare charges across all the region.

Our aim is to fit an appropriate model that best describes the relationship between the response variable, and the other regressors. We will start by assuming a suitable linear model with some underlying assumptions. We will proceed by verifying whether the assumptions are reasonable or not. Thereafter we will estimate the parameters of the model and check for outliers. We aim to determine whether all of the regressors are significant or not. If evidence directs us we may drop some regressors from the model depending on their significance.

➢ Multiple Linear Regression Model:

We define a Multiple Linear Regression (henceforth, MLR) Model as, -

$$Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots . + \beta_p x_{ip} + \epsilon_i \text{ ,i=1,2,..,.n}$$

under the assumptions,

(a) $E(\epsilon_i) = 0 \; \forall i = 1,2,..,n$
(b) $V(\epsilon_i) = \sigma^2 \forall i = 1,2,..,n$
(c) $cov(\epsilon_i, \epsilon_j) = 0 \; \forall i \neq j$
(d) $\epsilon_i \sim N(0, \sigma^2) \forall i = 1,2,\ldots,n$

For the ease in calculation, we represent the model in matrix notation, -

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\boldsymbol{Y} = (Y_1, Y_2, \ldots., Y_n)'$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots., \beta_p)'$

$\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots., \epsilon_p)'$

$$X=\begin{pmatrix}1 & x_{11} & \cdots & x_{1p}\\ 1 & x_{21} & \dots & x_{2p}\\ \vdots & \dots & \ddots & \vdots\\ 1 & x_{n1} & \cdots & x_{np}\end{pmatrix}$$

We assume, X is of full column rank, i.e., $\rho(X) = p + 1$

We get the estimate of $\boldsymbol{\beta}$ by Least Square Method and denote it by, $\widehat{\boldsymbol{\beta}}$, which is given by,

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'Y$$

We have 8 regressors and our response variable is charges. Hence, the regression equation is given by,

charges= β_o+$\beta_1$age+ β_2 male+ β_3 bmi+ β_4 children+ β_5 smokes+ β_6 northwest+ β_7 southeast+ β_8 southwest+ error term

By fitting the regression line with the data in our hand, we have obtained the following result, -

```
Call:
lm(formula = charges ~ ., data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-11304.9  -2848.1   -982.1   1393.9  29992.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11938.5      987.8 -12.086  < 2e-16 ***
age            256.9       11.9  21.587  < 2e-16 ***
bmi            339.2       28.6  11.860  < 2e-16 ***
children       475.5      137.8   3.451 0.000577 ***
male          -131.3      332.9  -0.394 0.693348
smokes       23848.5      413.1  57.723  < 2e-16 ***
southwest     -960.0      477.9  -2.009 0.044765 *
southeast    -1035.0      478.7  -2.162 0.030782 *
northwest     -353.0      476.3  -0.741 0.458769
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7509,    Adjusted R-squared:  0.7494
F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

By looking at the model summary, the R-squared value of 0.75 is moderately good for a dataset with 1338 observations. However, coming to the individual regression coefficients, it is seen that sex(male) and region(northwest) are not statistically significant.

➢ Detection of Unusual Observations:

There are 3 types of unusual observations in the dataset, -

- Outlier: A data point whose response y does not follow the general trend of the rest of the data, is called outlier.
- Leverage Point: A Data point is said to be high leverage point if it has "extreme" or "unusual" regressor values for one or more regressors.
- Influential Point: A data point is said to be influential point if it unduly influences any part of a regression analysis, i.e. affecting the regression coefficients.

All these three kind of unusual observations are problematic since they will affect the fitted response. Out of them, we will go on detection of **influential points**.

❖ Detection of Influential Point- Cook's distance:

Cook's distance or Cook's D is a commonly used estimate regarding influential points. Cook's distance measures the change in the fitted regression coefficient if an observation is deleted from the regression equation. **It therefore combines the outliers and the leverage point diagnostics of a measure.** Points with a large Cook's distances are considered to further examination in the analysis.

Cook's Distance D_i of the observation i (i=1, 2, …, n) is defined as the sum of all the changes in the regression model when observation I is removed from it.

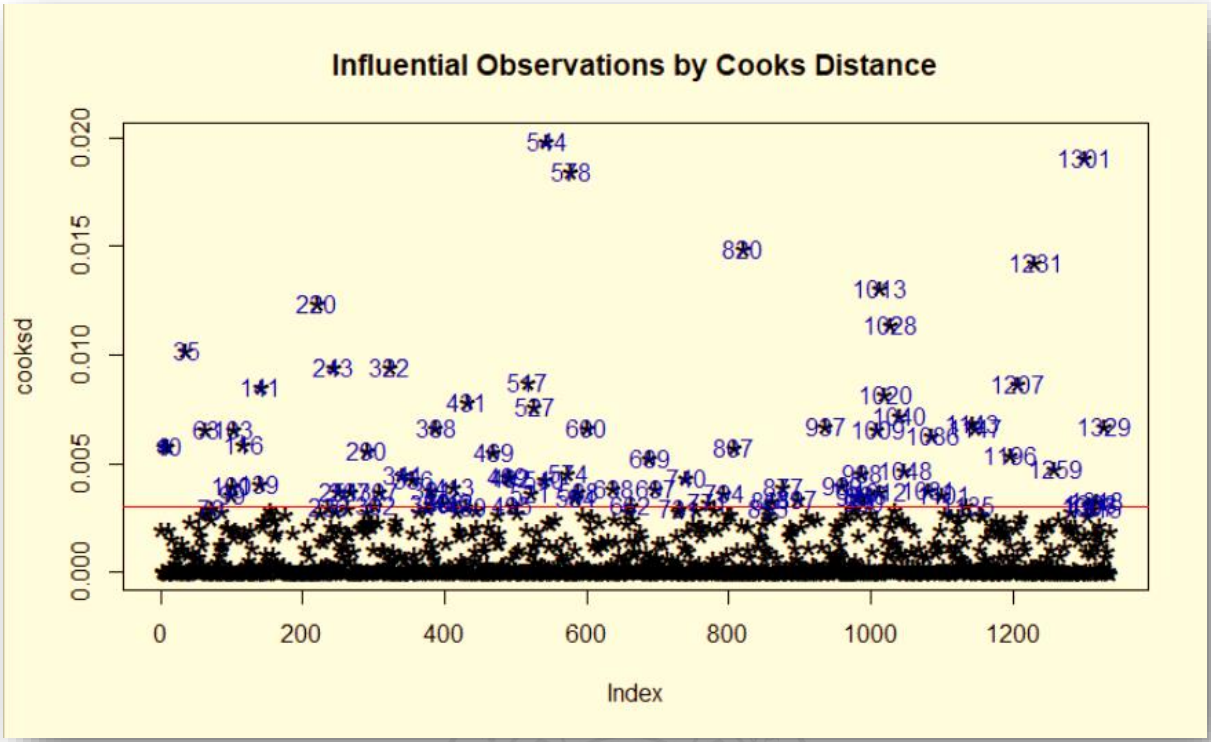
$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \widehat{y_{j(i)}}\right)^2}{ps^2}$$

where, $\widehat{y_{j(i)}}$ is the fitted response value obtained excluding observation i and s^2 is the Mean Square Error (MSE).

If the number of observations is 'n', then a point is said to be an influential point if,

$$D_i > \frac{4}{n}$$

Considering our dataset, we have obtained the following plot, -



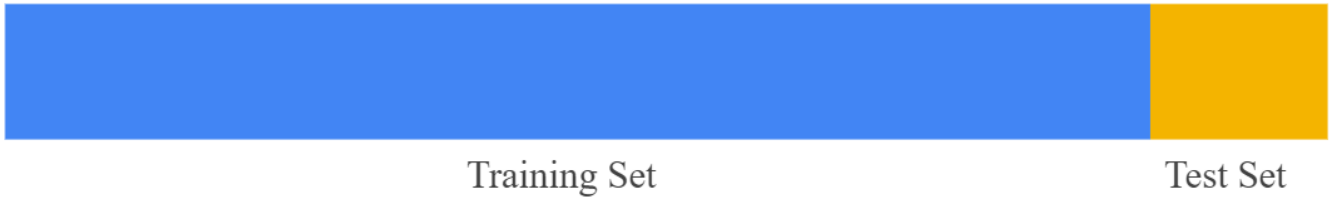
The points, having Cook's Distance more than $\frac{4}{n}$ is considered to be as Influential points. The blue coloured points in the dataset are influential points. We have 1338 observations, so, we consider removing influential points.

➢ Train and Test Data:

To find the efficacy of the model more briefly, we divide the whole dataset into two parts, -

- training set—a subset to train a model.
- test set—a subset to test the trained model.

We could imagine slicing the whole dataset as follows, -



We have started with 1338 observations and detected 87 outliers. We consider removing the outliers. After removing the outliers, we have (1338-87) = 1251 data points. Out of those, we have sample 80% observations for training set, i.e., we have a total of 1001 observations in training set.

First, we fit a multiple linear regression model with all our regressors using the train dataset. We have obtained the following result.

```
Call:
lm(formula = formula_0, data = Data_train)

Residuals:
     Min       1Q   Median       3Q      Max
-11284.6  -1870.6   -317.4   1540.3  15027.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11422.68     864.16 -13.218  < 2e-16 ***
age            260.32      10.30  25.268  < 2e-16 ***
male           156.72     289.43   0.541 0.588301
bmi            290.27      25.42  11.419  < 2e-16 ***
children       610.03     121.78   5.009 6.47e-07 ***
smokes       24494.10     368.14  66.534  < 2e-16 ***
southwest     -962.39     414.65  -2.321 0.020489 *
southeast    -1469.10     421.16  -3.488 0.000508 ***
northwest     -999.85     415.73  -2.405 0.016352 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 4542 on 992 degrees of freedom
Multiple R-squared:  0.8451,    Adjusted R-squared:  0.8439
F-statistic: 676.6 on 8 and 992 DF,  p-value: < 2.2e-16
```

From the regression equation, we are getting the coefficient of determination(r^2) 85%, i.e., 85% of the total variation can be explained by fitted regression equation. However, coming into individual regression coefficients, we are seeing that, not all the regressors are wothwhile statistically significant.

➢ **Autocorrelation:**

The term autocorrelation may be defined as correlation between the members of a series of observations ordered in time or space. The Classical Linear Regression Model (CLRM) assumes that, the disturbance term related to any observation is not influenced by the disturbance term related to any other observation. However, if there is such a dependence, autocorrelation is said to present in the model and we have, $cov(\epsilon_i, \epsilon_j \neq 0)$, for some $i \neq j$.

- <u>Durbin- Watson Test for detecting Autocorrelation:</u>

The Durbin- Watson '**d**' statistic is given by,

$$d = \frac{\sum_{i=2}^{n}(\epsilon_i - \epsilon_{i-1})^2}{\sum_{i=1}^{n}\epsilon_i^2}$$

where, the residuals ϵ_i are obtained through Ordinary Least Square (OLS) method. The underlying assumptions for using '**d**' statistic are listed below, -

- The regression model includes intercept term.
- The explanatory variables are non stochastic.
- ϵ_i's are generated by AR(1) scheme, i.e., $\epsilon_i = \epsilon_{i-1} + u_i$. Hence, d-statistic can't be used for detecting higher order AR schemes.
- ϵ_i's are assumed to be normally distributed.
- The regression model doesn't include the lagged value(s) of the dependent variable as one of the explanatory variables.
- There are no missing observation.

An estimate of sample correlation coefficient is given by, -

$$\hat{\boldsymbol{\rho}} = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i\,\widehat{\epsilon_{i-1}}}{\{\sum_{i=2}^{n}.\widehat{\epsilon_{i-1}^2}\sum_{i=1}^{n}.\widehat{\epsilon_i^2}\}^{\frac{1}{2}}}$$

Assuming, $\sum_{i=1}^{n}\widehat{\epsilon_i^2} \approx \sum_{i=2}^{n}\widehat{\epsilon_{i-1}^2}$, we have,

$$d=2(1-\hat{\rho})$$

We want to test, H_0: $\hat{\boldsymbol{\rho}} = \mathbf{0}$ against H_1: $\hat{\boldsymbol{\rho}} \neq 0$

We calculate the value of d. If it comes out to be 2, we conclude that, **there exists no serial autocorrelation in the error terms.**

Using R, we got the following result,-

```
        Durbin-Watson test

data:  mq
DW = 2.0079, p-value = 0.9001
alternative hypothesis: true autocorrelation is not 0
```

We got the value of Durbin Watson test statistic **'d'** as 2.0079 and p-value 0.9001(>0.05). Thus, we accept the null hypothesis at 5% level and conclude that, there is **no serial auto correlations in error terms** as expected, because it is not spatial or time series data and there is no strict rule of selecting any data as in the i^{th} data.

➢ Multicollinearity:

- What is Multicolliearity?

The word "**Multicollinearity**" means the existence of exact or perfect relationship among some or all explanatory variables in a regression model.

In a model with 'p' explanatory variables $x_1, x_2, \ldots, x_p$, an exact relationship is said to exist if the following condition is satisfied, -

$$\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots .. + \beta_p x_{ip} = 0$$

where not all the coefficients are simultaneously 0.

We can explore an issue of multicollinearity while obtaining estimates of multiple linear regression equation. If there exists exact linear relationship among the regressors, then atleast one column of X (as defined earlier) will be linear combination of the others and $\rho(X)$ will be deficit of full column rank and as a result $X'X$ will not be invertible.

Again, if we work out with a dataset with a large number of regressors, then it may highly be possible that, $\sum_{j=1}^{p} \beta_j x_{ij} \approx 0$. In this case, we say the regressors are nearly perfect multicollinear. In this case, $det(X'X) \approx 0$. As a result, $X'X$ will be invertible, but will have large condition number (condition number of a matrix= $\frac{maximum\ eigen\ value}{minimum\ eigen\ value}$). In such case, we may be able to compute $(X'X)^{-1}$ but the result will be too much sensitive to small variation in dataset.

- Why may we have multicollinearity?

Some major causes of multicollinearity are,

(i) Sampling over a limited range of values of the explanatory variables.
(ii) Wrong specification of model.
(iii) Presence of more explanatory variables in the model than needed.

- How to detect multicollinearity?

A standard measure for the detection of multicollinearity is, **Variance Inflation Factor** (henceforth, **VIF**). In the model,

$$Y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots . + \beta_p x_{ip} + \epsilon_i \text{ ,i=1,2,..,.n}$$

the VIF of the j^{th} regressor is defined as, -

$$\boldsymbol{VIF_j = \frac{1}{1 - R^2_{(j)}}}$$

where, $\boldsymbol{R^2_{(j)}}$ is the coefficient of determination from the regression equation of X_j on (X_1, X_2, ….., X_{j-1}, X_{j+1}, …., X_p)

$\boldsymbol{VIF_j}$ measures the combined effect of the dependence of X_j on all other (p-1) regressors.

A large value of VIF indicates the presence of multicollinearity in the model.

As a thumb rule, if, $\boldsymbol{R^2_{(j)} > 0.8}$, i.e., VIF>5 then we can expect the presence of multicollinearity in the model.

In case, of near exact linear relationship among the regressors, $\boldsymbol{R^2_{(j)} \approx 1}$ and as a result, we get, $\boldsymbol{VIF_j \to \infty}$, which indicates high multicollinearity between X_j and other regressors present in the model.

In our model, we have obtained the following value of VIF's

```
 vif(mq)
     age      male       bmi  children    smokes southwest southeast northwest
1.022107  1.016224  1.128644  1.007575  1.011467  1.567594  1.703435  1.546131
```

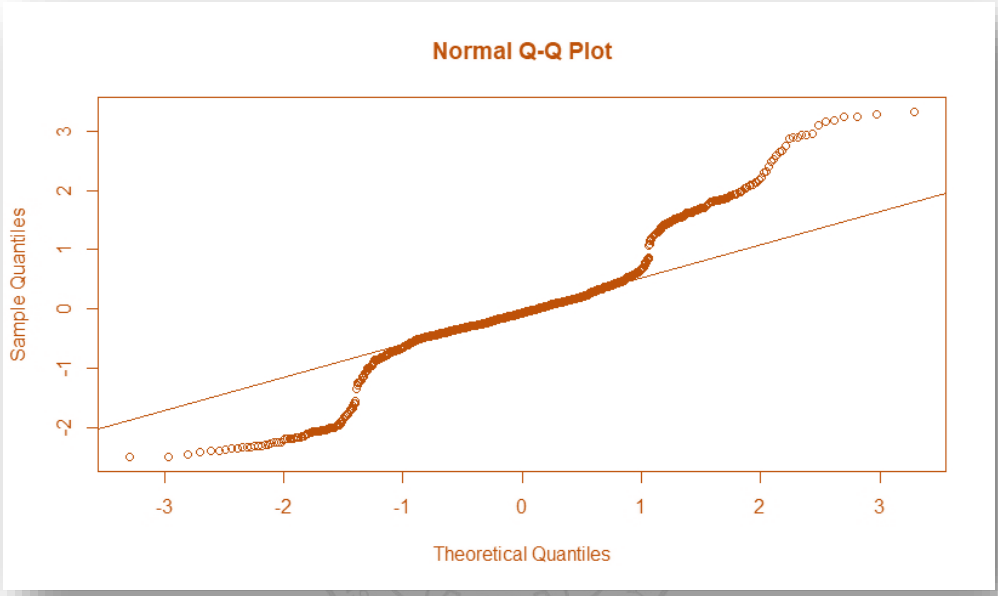
We can see that, none of the VIF's are greater then 5. Hence, we can conclude, there is **no multicollinearity** present in our model.

➢ Checking of Model Assumptions:

We have obtained the MLR equation taking all the assumptions true. However, while working with real life data, it may often happen, the data in hand doesn't meet the assumptions required for an appropriate MLR model. Hence, we first have to check the validity of assumptions in our case. If any of the assumptions gets violated, we have to take necessary remedial measures and have to rebuild the model in such a way that in the reformed model, all the model assumptions remain valid.

- **Q-Q Plot to check Normality of Errors:**

Assumption of Normality of errors is one of the vital assumptions in an MLR model. To check the validity of such an assumption, we use Q-Q plot. The Q-Q plot, or quantile-quantile plot, is a graphical tool to help us assess if errors plausibly came from Normal Distribution. A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that is roughly straight.



From the Q-Q plot, we can fairly observe that the line is not straight line and conclude that, assumption of normality of errors is violated. Still we do a formal test to assure the violation of normality.

- **Shapiro- Wilks Test:**

The Shapiro-Wilks test is a useful test of normality. In regression analysis, this test is often used to detect whether the errors came from a normal distribution.

Here, the null hypothesis of interest is,

H_0: The errors are generated from a normal distribution.

The appropriate test statistic for testing H_0 is given by,

$$W = \frac{\left(\sum_{i=1}^{n} a_i \epsilon_{(i)}\right)^2}{\sum_{i=1}^{n} (\epsilon_i - \bar{\epsilon})^2}$$

Here,

$\epsilon_{(i)}$ is the ith ordered error term in the model.

a_i's are calculated using the means, variances and covariances of the $\epsilon_i's$.

W is compared against the tabulated values of this statistic's distribution. Too small value of W will lead to the rejection of null hypothesis.

If we are using p-value, then the p-value smaller than the desired level will lead to the rejection of null hypothesis.

In our model, we have obtained the following result, -

```
> library(stats)
> shapiro.test(rstudent(mq))

        Shapiro-Wilk normality test

data:  rstudent(mq)
W = 0.93578, p-value < 2.2e-16
```

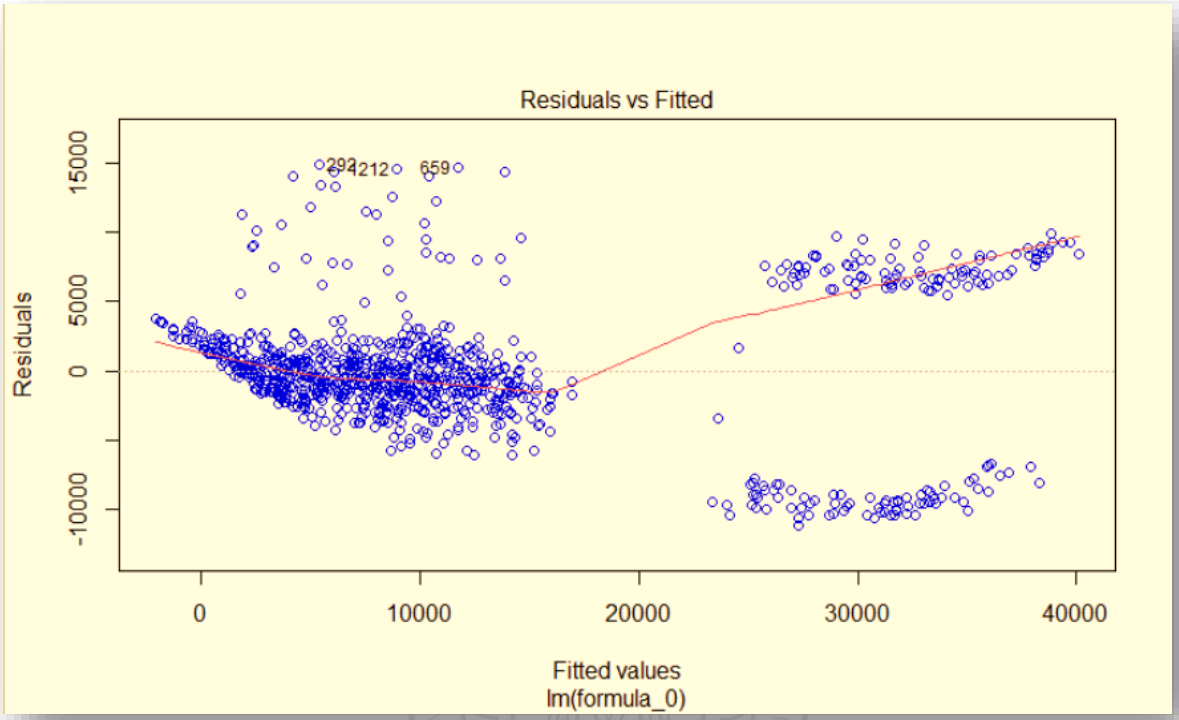
Since, the p-value is coming out to be very small(<2.2e-16), we reject the null hypothesis and conclude that, **the assumption of normality of errors has been violated in the model.**

- **Residual Plot:**

The residuals are defined as the difference between the observed value and the predicted value. If, $\hat{Y}_i$ denotes the predicted value, then the residual is defined by,

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i \; \forall i = 1,2, \dots, n$$

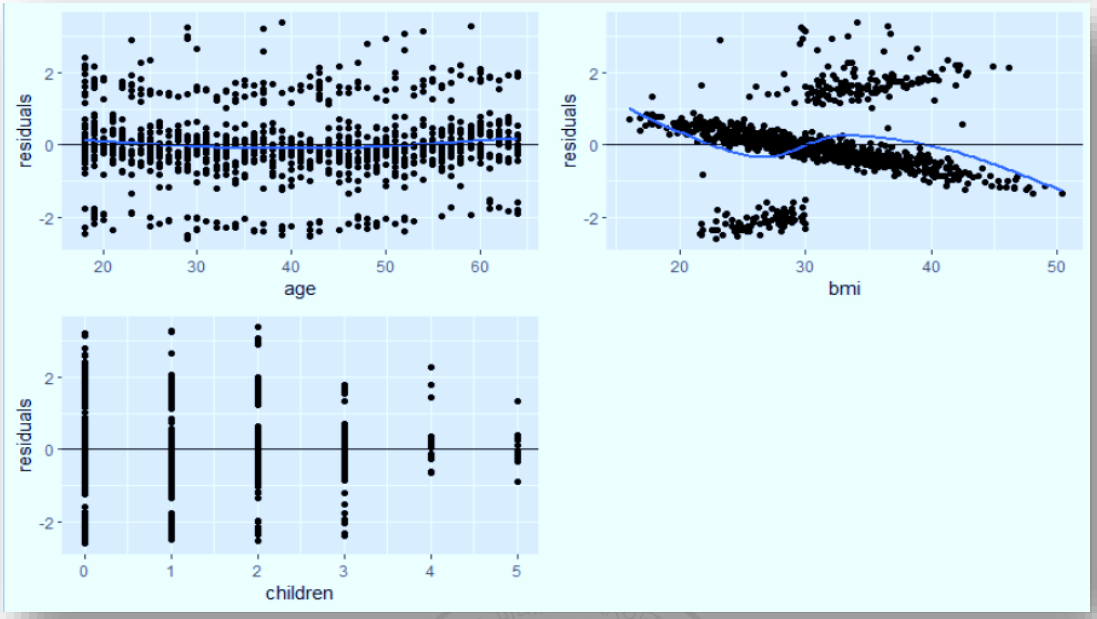
First, we have obtained the residuals vs fitted value plot.



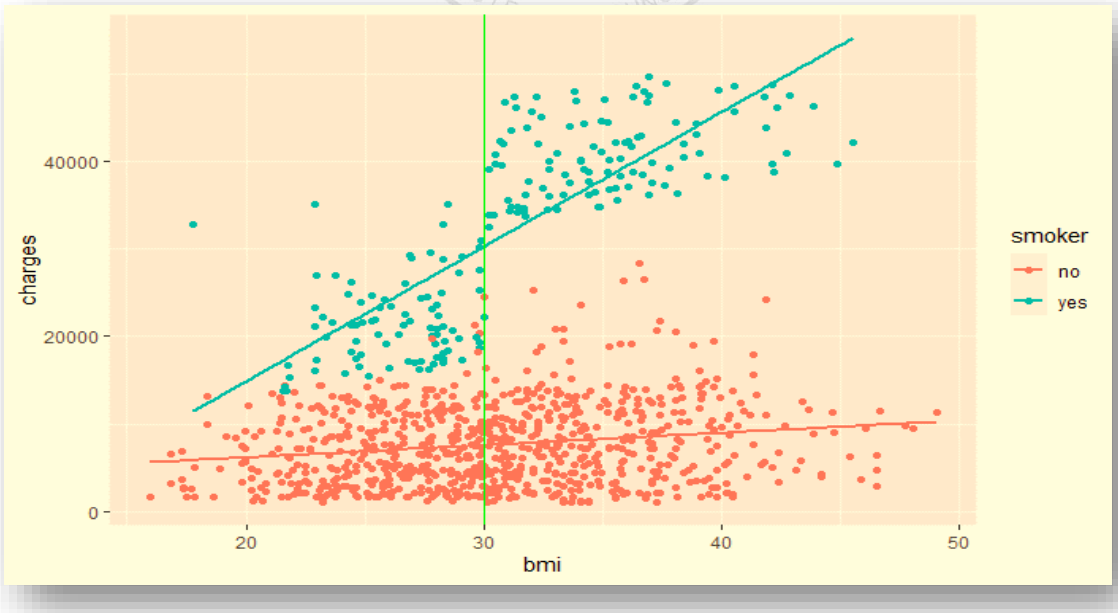
From the residual plot, we can see that the residuals are not spread randomly, rather, we can see a clear pattern. More specifically, there are three clustering in the plot, making the plot looking like an outward opening funnel. So, the residual plot indicates, the variances are not non constant, **i.e., fairly observing the residual plot, we can conclude that, assumption of homoscedasticity is violated.**

Now, we are interested to check, which regressor(s) is(are) responsible for such a pattern in the residual plot. We have 6 regressors in our model and out of them 3 are of numeric type (age, bmi, number of children). So, we obtained separate residual plot, where in each case, we have obtained respective regressor versus reidual plot.

From the regressor versus residual plot, we will observe, whether the residual plot of a particular regressor matches with the actual fitted value versus residual plot. On the basis of that observation, we shall carry out our further analysis.



We can see that, bmi vs residual plot is giving similar pattern that we have obtained in the residual plot. Again, from the interactions among the regressors we have seen that, the category "person who are smoker" results in a high healthcare charges. Also, we have observed the following plot.



From the above plot, we can see that, the points that include the category "charges of smokers in all bmi level" have two clear clusters. More specifically, there is one cluster, that is below 30 bmi and one cluster that is above 30 bmi.

➢ Splitting and Analysis of Train Dataset:

Combining all our observations, we divide the dataset into 3 subsets, -

(a) Smokers with high bmi (bmi>30)
(b) Smokers with low bmi (bmi<30)
(c) Non Smokers

Smokers with high bmi:

Here we proceed with the observartions indicating persons who smoke and having high bmi(>30).

First, we fit initial MLR model taking all assumptions true. We have obtained following result, -

```
> m2=lm(charges~age + male + bmi + children  +southwest+southeast+northwest,data=Data2_smokes_highbmi)
>
> summary(m2)

Call:
lm(formula = charges ~ age + male + bmi + children + southwest +
    southeast + northwest, data = Data2_smokes_highbmi)

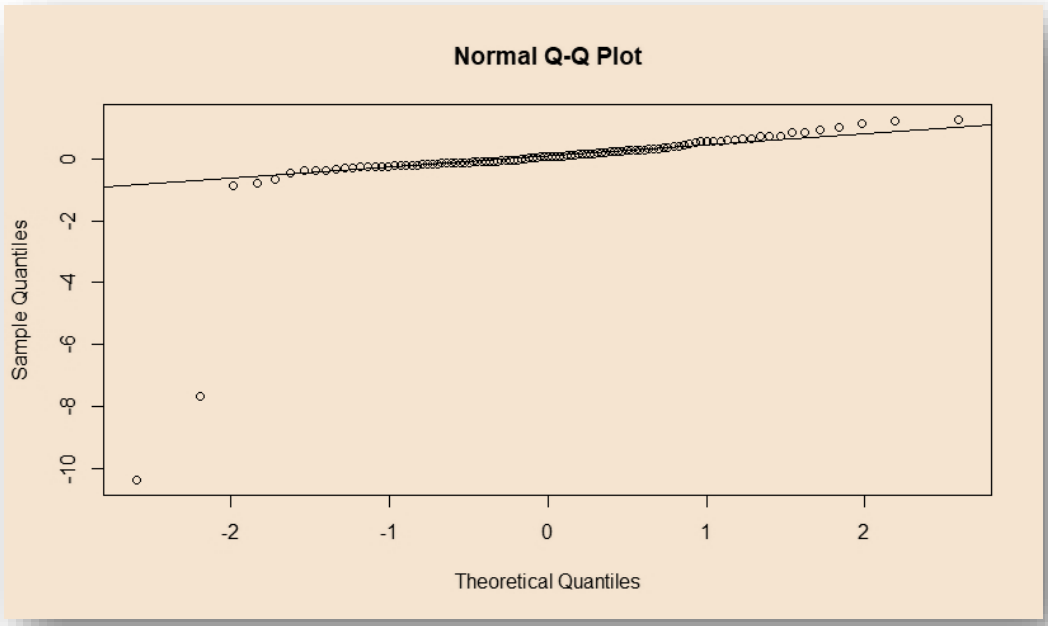
Residuals:
     Min       1Q   Median       3Q      Max
-15067.1   -312.9    166.0    695.5   2632.9

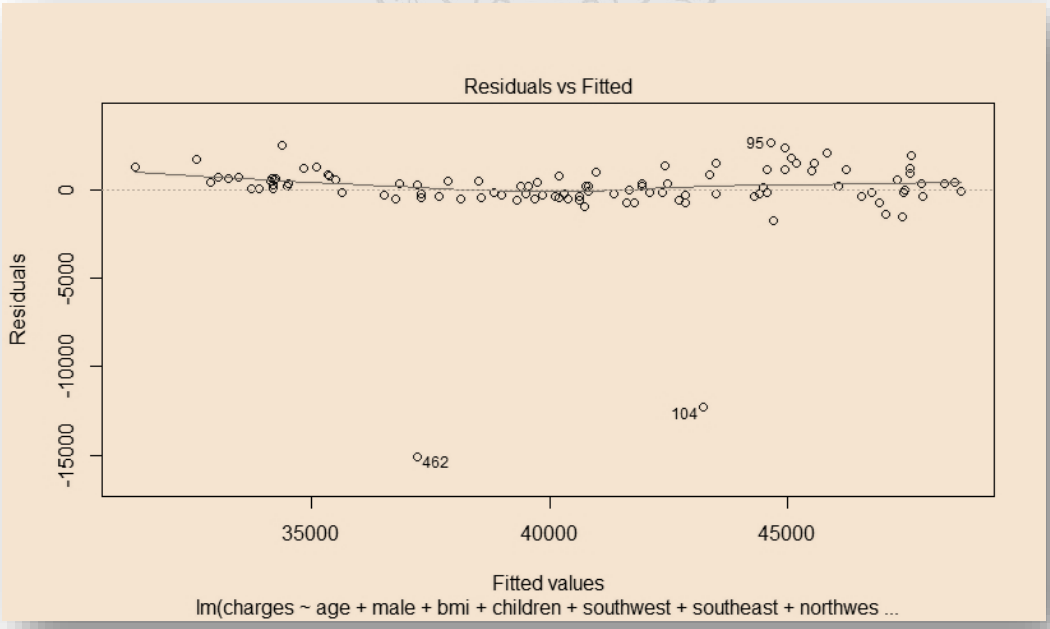
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9550.71    2155.14   4.432 2.46e-05 ***
age           258.17      14.59  17.698  < 2e-16 ***
male         -311.67     464.62  -0.671   0.5039
bmi           603.56      59.00  10.230  < 2e-16 ***
children      335.46     191.37   1.753   0.0828 .
southwest    -977.97     621.16  -1.574   0.1186
southeast   -1115.77     603.92  -1.848   0.0677 .
northwest     304.54     711.51   0.428   0.6696
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2171 on 97 degrees of freedom
Multiple R-squared:  0.8343,    Adjusted R-squared:  0.8223
F-statistic: 69.77 on 7 and 97 DF,  p-value: < 2.2e-16
```

From the summary table, we observe that, not all the regressors are statistically significant to explain charges. Hence, we further proceed for variable selection.

Next, we shall fit Q-Q plot and Residual plot. We fit Q-Q plot to check whether we can apply AIC for variable selection or not.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
0
-2
-4
-6
-8
-10
-2
-1
0
1
2


Residuals vs Fitted
Residuals
Fitted values
lm(charges ~ age + male + bmi + children + southwest + southeast + northwes ...
0
-5000
-10000
-15000
35000
40000
45000
95
104
462

❖ Variable Selection Using AIC Method:

The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. In statistics, AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from:

- the number of independent variables used to build the model.
- the maximum likelihood estimates of the model (how well the model reproduces the data).

The best-fit model according to AIC is the one that explains the greatest amount of variation using the fewest possible independent variables. AIC is most often used for model selection.

By calculating and comparing the AIC scores of several possible models, we can choose the one that is the best fit for the data. A good way to find out is to create a set of models, each containing a different combination of the independent variables we have measured.

Once we have created several possible models, we can use AIC to compare them. Lower AIC scores are better, and AIC penalizes models that use more parameters. So, if two models explain the same amount of variation, the one with fewer parameters will have a lower AIC score and will be the better-fit model.

The AIC is calculated by using the following formula,

$$\boldsymbol{AIC = n\ ln\ SSRes(p) + 2p}$$

where,

n= total number of observation

SSRes=Sum of Squares due to Residuals

p= number of regresors

We choose that model, for which AIC is minimum.

For finite sample size, we may also use corrected AIC, which is given by,

$$\boldsymbol{AIC(Corrected) = AIC + \frac{2p(p+1)}{n-p-1}}$$

We have obtained the following result, -

```
> stepAIC(m2, direction = "both")
Start:  AIC=1621.1
charges ~ age + male + bmi + children + southwest + southeast +
    northwest

             Df  Sum of Sq        RSS    AIC
- northwest   1     863482  458063663 1619.3
- male        1    2121003  459321185 1619.6
<none>                     457200182 1621.1
- southwest   1   11683698  468883880 1621.8
- children    1   14482983  471683164 1622.4
- southeast   1   16089033  473289215 1622.7
- bmi         1  493271253  950471435 1695.9
- age         1 1476297185 1933497366 1770.5

Step:  AIC=1619.3
charges ~ age + male + bmi + children + southwest + southeast

             Df  Sum of Sq        RSS    AIC
- male        1    2351132  460414795 1617.8
<none>                     458063663 1619.3
- children    1   15836473  473900136 1620.9
+ northwest   1     863482  457200182 1621.1
- southwest   1   19858615  477922278 1621.8
- southeast   1   26508168  484571831 1623.2
- bmi         1  492408156  950471820 1693.9
- age         1 1480683047 1938746710 1768.8

Step:  AIC=1617.84
charges ~ age + bmi + children + southwest + southeast

             Df  Sum of Sq        RSS    AIC
<none>                     460414795 1617.8
- children    1   15043289  475458085 1619.2
+ male        1    2351132  458063663 1619.3
+ northwest   1    1093611  459321185 1619.6
- southwest   1   19880361  480295156 1620.3
- southeast   1   28601945  489016740 1622.2
- bmi         1  491999981  952414776 1692.2
- age         1 1507422803 1967837598 1768.4

Call:
lm(formula = charges ~ age + bmi + children + southwest + southeast,
    data = Data2_smokes_highbmi)

Coefficients:
(Intercept)          age          bmi     children    southwest    southeast
     9505.0        258.7        602.3        337.5      -1109.7      -1284.3
```

In the third model, we are getting the value of AIC minimum. Hence, we further proceed with that model and we take age, bmi, children, southwest and southeast as our regressors.

❖ Model Fitting after Variable Selection:

We have selected age, bmi, children, southwest and southeast as our variables and based on that we obtained our new model, and obtained the following result, -

```
> m5=lm(charges~age + bmi+children  +southwest+southeast,data=Data2_smokes_highbmi)
> summary(m5)

Call:
lm(formula = charges ~ age + bmi + children + southwest + southeast,
    data = Data2_smokes_highbmi)

Residuals:
     Min       1Q   Median       3Q      Max
-15183.5   -380.4    175.3    688.6   2911.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9505.01    2072.08   4.587 1.32e-05 ***
age           258.67      14.37  18.004  < 2e-16 ***
bmi           602.27      58.55  10.286  < 2e-16 ***
children      337.48     187.64   1.799   0.0751 .
southwest   -1109.71     536.73  -2.068   0.0413 *
southeast   -1284.35     517.90  -2.480   0.0148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2157 on 99 degrees of freedom
Multiple R-squared:  0.8331,    Adjusted R-squared:  0.8247
F-statistic: 98.86 on 5 and 99 DF,  p-value: < 2.2e-16
```

From the summary table, we can see that, all the variables are statistically significant in explaining charges.

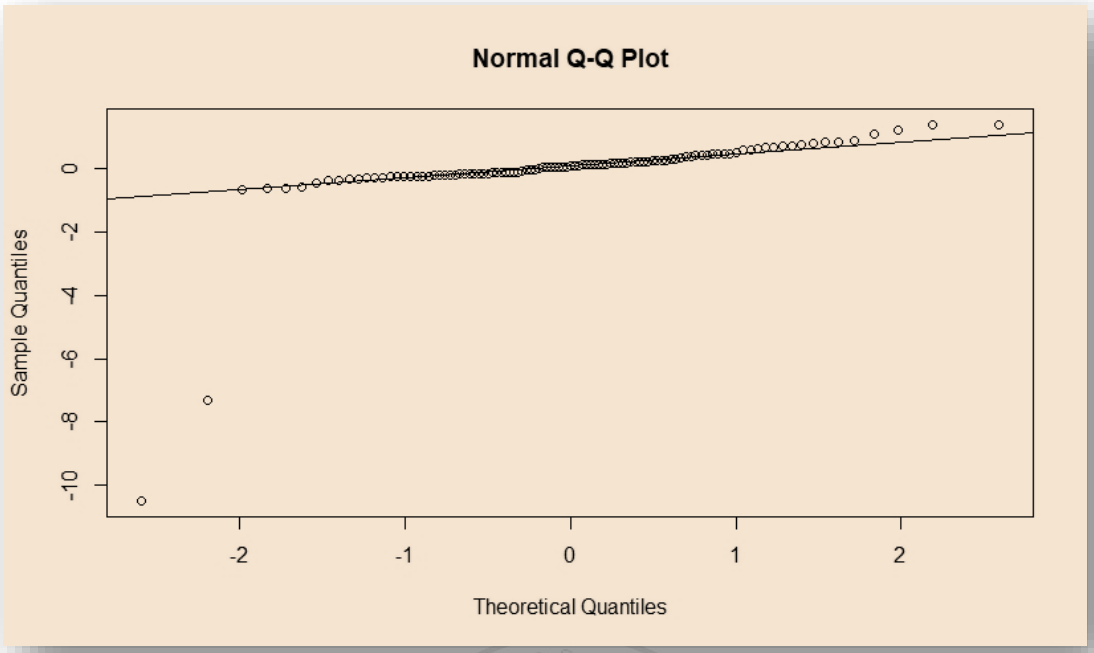
Next, we proceed to check the validity of error assumption, i.e., we will check whether the assumption of normality and the assumption of homoscedaticity remains valid in our selected model.

Before going into checking error assumptions, we first check whether there is autocorrelation between the errors. We have followed Durbin- Watson Test to check this. We have obtained the following result, -

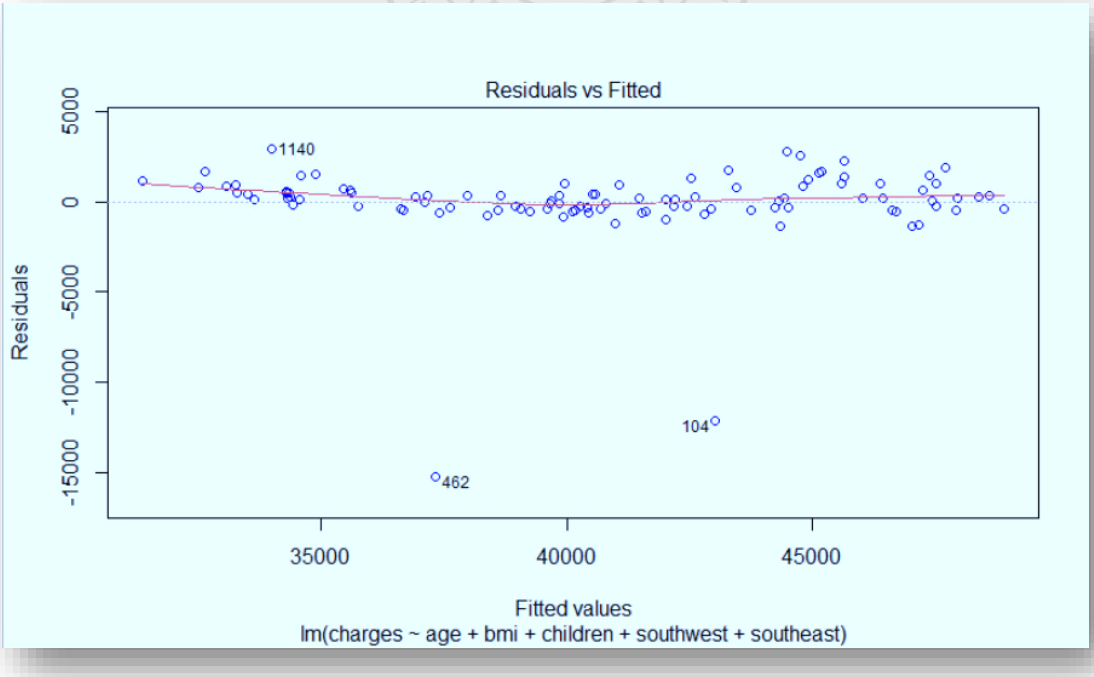
```
> durbinWatsonTest(m5)
 lag Autocorrelation D-W Statistic p-value
   1      0.01371381      1.971509   0.728
 Alternative hypothesis: rho != 0
```

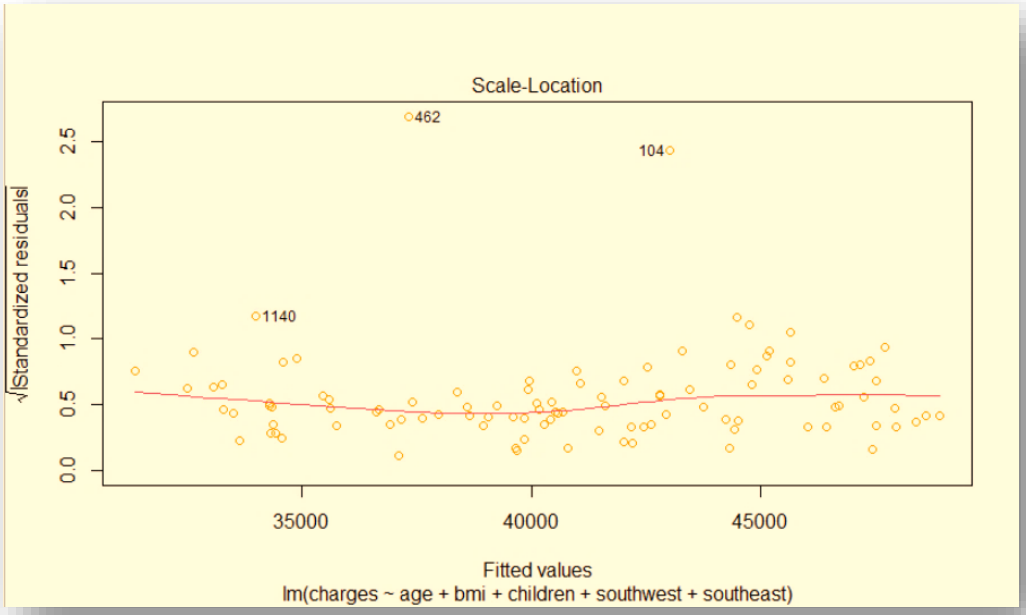
We have obtained p-value 0.728(>0.05). Hence, we accept the null hypothesis and conclude that, there is no autocorrelation between the errors.

Next, we check for normality of errors. We used Q-Q plot to check this.



From the Q-Q plot, we can conclude that the error terms have been generated from normal distribution. Next, we proceed for checking **homoscedasticity assumption**. We have used residual plot to observe this.





From the residual plot, we can see the errors are randomly spread asround 0 and the standardized residual plot also shows randomness of error. Hence, we can conclude that, **the errors are homoscedastic.**

Smokers with low bmi:

Here we proceed with the observartions indicating persons who smoke and having low bmi(<30).

First, we fit initial MLR model taking all assumptions true. We have obtained following result, -

```
> m3=lm(charges~age + male + bmi + children  +southwest+southeast+northwest,data=Data2_smokes_lowbmi)
> summary(m3)

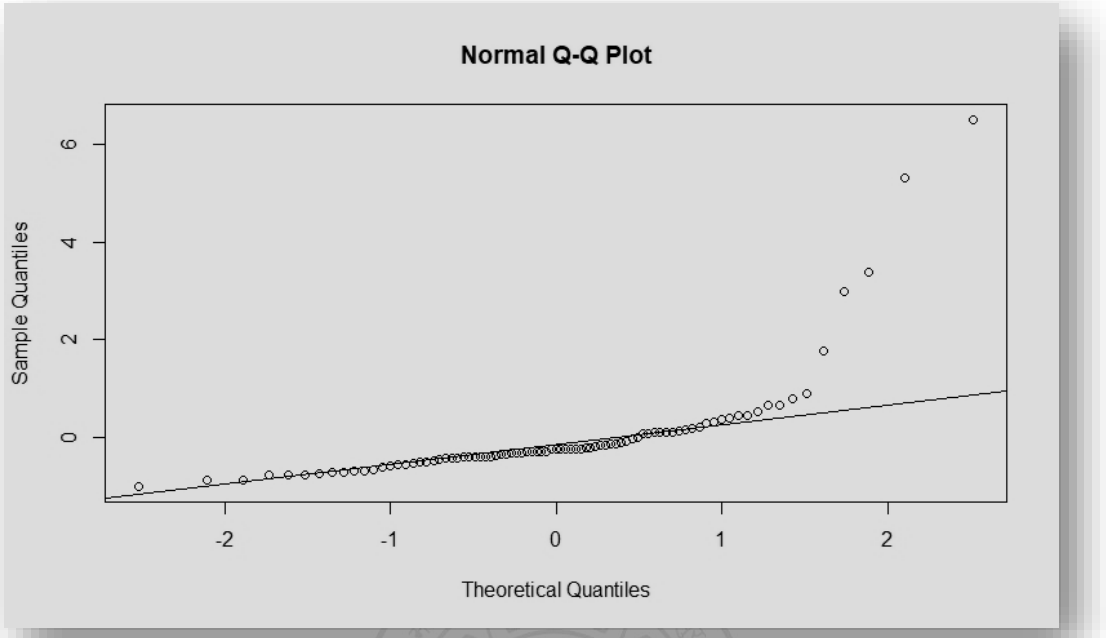
Call:
lm(formula = charges ~ age + male + bmi + children + southwest +
    southeast + northwest, data = Data2_smokes_lowbmi)

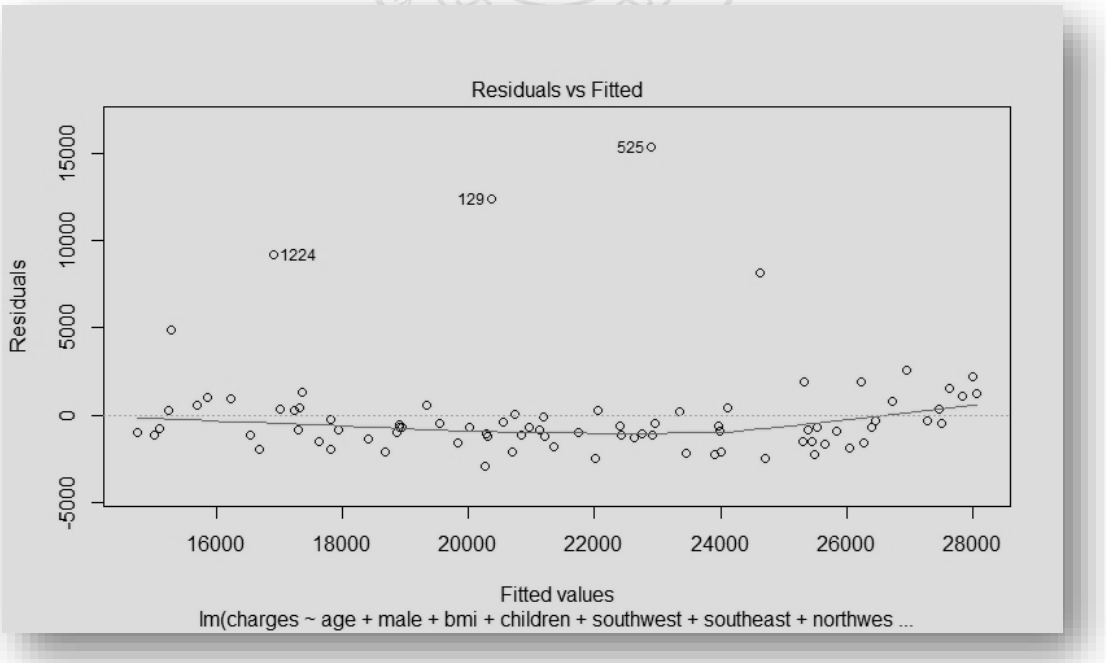
Residuals:
    Min      1Q  Median      3Q     Max
-2901.8 -1239.8  -704.3   337.4 15348.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7712.4     3627.4   2.126   0.0367 *
age            245.0       23.5  10.427 2.23e-16 ***
male          -327.4      695.1  -0.471   0.6389
bmi            137.0      139.5   0.983   0.3289
children       693.3      324.9   2.134   0.0361 *
southwest     -324.7     1068.5  -0.304   0.7620
southeast      956.6      932.7   1.026   0.3083
northwest      992.0      895.3   1.108   0.2713
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3020 on 77 degrees of freedom
Multiple R-squared:  0.6342,    Adjusted R-squared:  0.6009
F-statistic: 19.07 on 7 and 77 DF,  p-value: 1.58e-14
```

Next, we shall fit Q-Q plot and Residual plot. We fit Q-Q plot to check whether we can apply AIC for variable selection or not.





Here, we have also used AIC method for model selection.

We have obtained the following result, -

```
Step:  AIC=1366.12
charges ~ age + bmi + children + southeast + northwest

            Df  Sum of Sq        RSS    AIC
- bmi        1    7185165  712037144 1365.0
<none>                    704851980 1366.1
- southeast  1   17254264  722106244 1366.2
- northwest  1   20552193  725404173 1366.6
+ male       1    1832938  703019041 1367.9
+ southwest  1     651709  704200271 1368.0
- children   1   39060127  743912106 1368.7
- age        1 1030944920 1735796900 1440.7

Step:  AIC=1364.98
charges ~ age + children + southeast + northwest

            Df  Sum of Sq        RSS    AIC
<none>                    712037144 1365.0
- southeast  1   19024163  731061307 1365.2
- northwest  1   21973026  734010170 1365.6
+ bmi        1    7185165  704851980 1366.1
+ male       1     534245  711502899 1366.9
+ southwest  1     452926  711584218 1366.9
- children   1   40175328  752212472 1367.7
- age        1 1040655525 1752692669 1439.5

Call:
lm(formula = charges ~ age + children + southeast + northwest,
    data = Data2_smokes_lowbmi)

Coefficients:
(Intercept)          age     children    southeast    northwest
    10856.8        247.8        662.2       1193.3       1208.4
```

In the second model, we are getting the value of AIC minimum. Hence, we further proceed with that model and we take age, children, southwest and southeast as our regressors.

Next, we shall fit Q-Q plot and Residual plot. We fit Q-Q plot to check whether we can apply variable selection or not.

❖ Model Fitting after Variable Selection:

We have selected age, bmi, children, southwest and southeast as our variables and based on that we obtained our new model, and obtained the following result, -

```
> m6=lm(charges~age + children  +southeast+northwest,data=Data2_smokes_lowbmi)
> summary(m6)

Call:
lm(formula = charges ~ age + children + southeast + northwest,
    data = Data2_smokes_lowbmi)

Residuals:
    Min      1Q  Median      3Q     Max
-3460.4 -1448.3  -653.9   580.2 15127.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10856.82     993.81  10.924   <2e-16 ***
age           247.75      22.91  10.813   <2e-16 ***
children      662.18     311.68   2.125   0.0367 *
southeast    1193.30     816.21   1.462   0.1477
northwest    1208.43     769.10   1.571   0.1201
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2983 on 80 degrees of freedom
Multiple R-squared:  0.629,     Adjusted R-squared:  0.6105
F-statistic: 33.91 on 4 and 80 DF,  p-value: < 2.2e-16

>
```

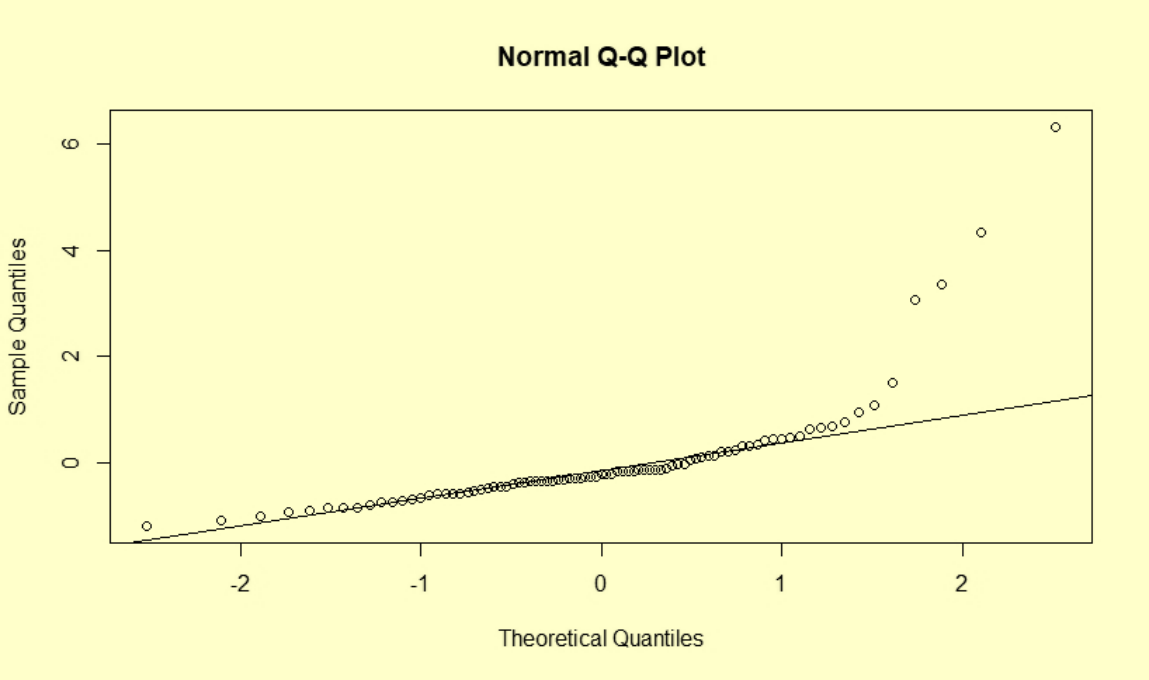
Next, we proceed to check the validity of error assumption, i.e., we will check whether the assumption of normality and the assumption of homoscedaticity remains valid in our selected model.

Before going into checking error assumptions, we first check whether there is autocorrelation between the errors. We have followed Durbin- Watson Test to check this. We have obtained the following result, -

```
> durbinWatsonTest(m6)
 lag Autocorrelation D-W Statistic p-value
   1     -0.08309106      1.980843   0.856
 Alternative hypothesis: rho != 0
>
```

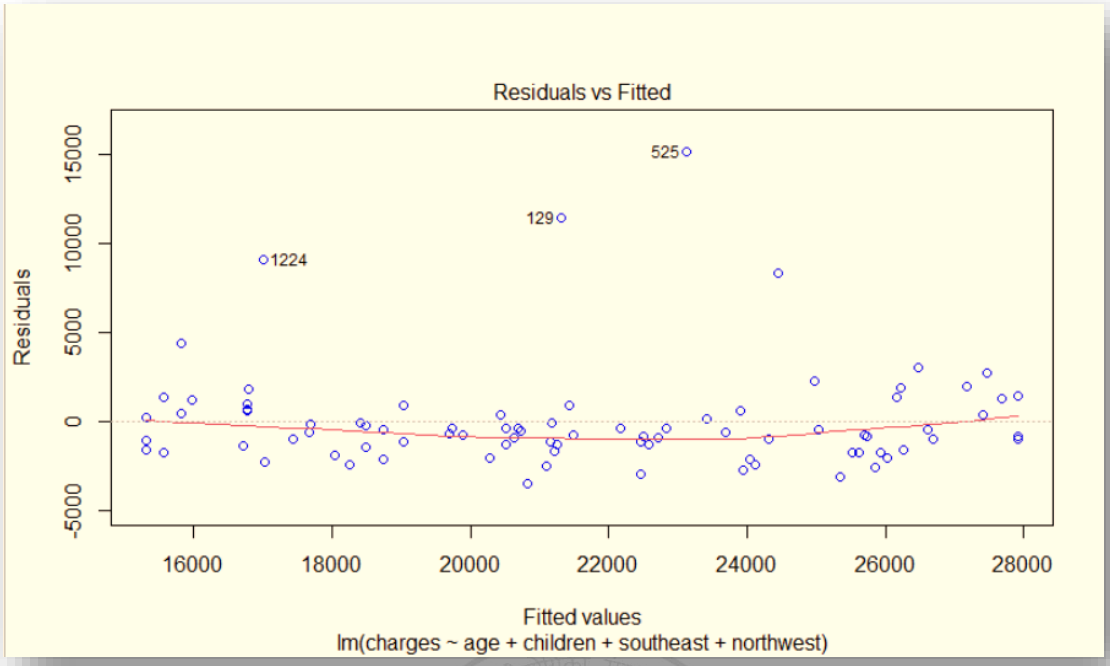
We have obtained p-value 0.856(>0.05). Hence, we accept the null hypothesis and conclude that, there is no autocorrelation between the errors.

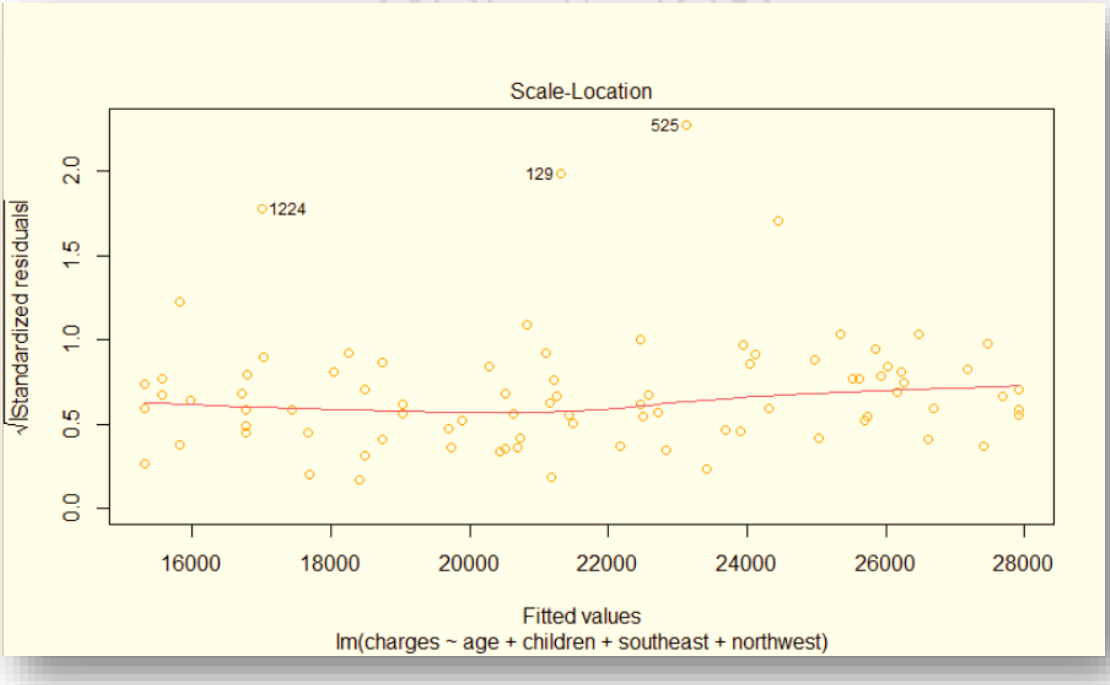
Next, we check for **normality of errors.** We used Q-Q plot to check this.



From the Q-Q plot, we can conclude that the error terms have been generated from normal distribution.

Next, we proceed for checking **homoscedasticity assumption.** We have used residual plot to observe this.


Residuals vs Fitted
Residuals
15000
10000
5000
0
-5000
525
129
1224
16000
18000
20000
22000
24000
26000
28000
Fitted values
lm(charges ~ age + children + southeast + northwest)


Scale-Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
525
129
1224
16000
18000
20000
22000
24000
26000
28000
Fitted values
lm(charges ~ age + children + southeast + northwest)

From the residual plot, we can see the errors are randomly spread asround 0 and the standardized residual plot also shows randomness of error. Hence, we can conclude that, **the errors are homoscedastic.**

Non Smokers:

Here we proceed with the observartions indicating persons who doesn't smoke.

First, we fit initial MLR model taking all assumptions true. We have obtained following result, -

```
> m4=lm(charges~age + male +bmi + children  +southwest+southeast+northwest,data=Data2_non_smokes)
> summary(m4)

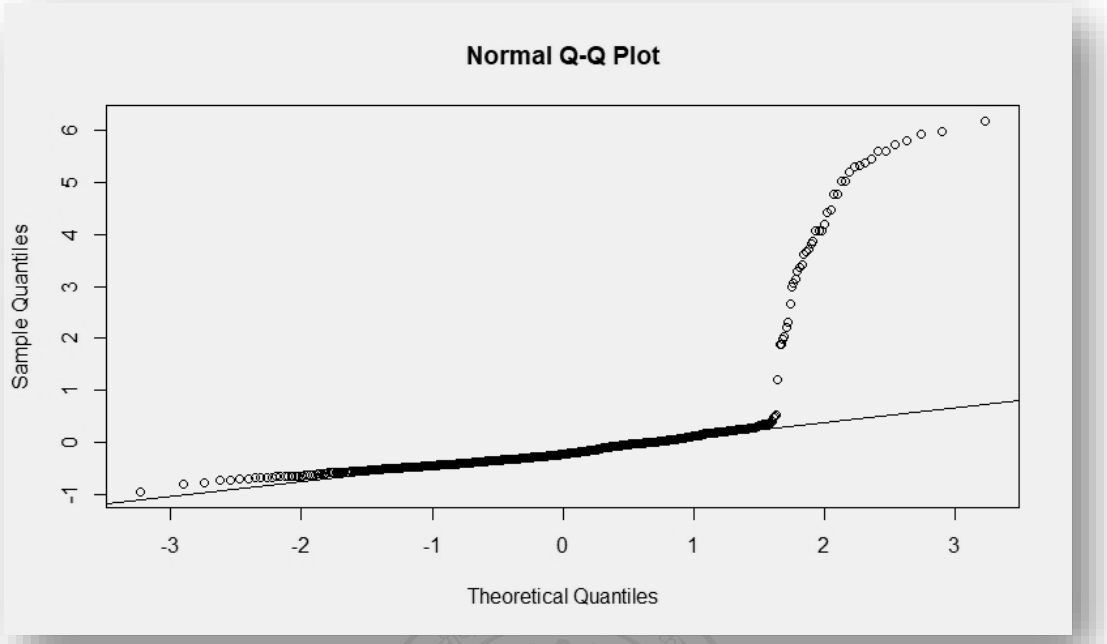
Call:
lm(formula = charges ~ age + male + bmi + children + southwest +
    southeast + northwest, data = Data2_non_smokes)

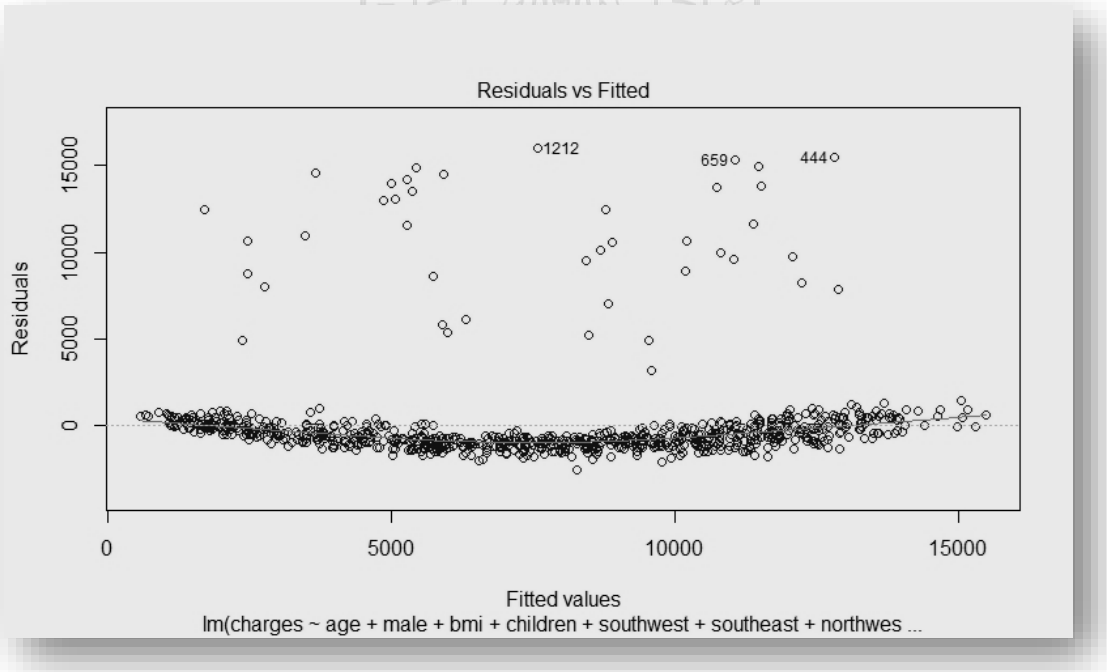
Residuals:
    Min      1Q  Median      3Q     Max
-2533.7 -1023.0  -605.9     0.4 15984.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3878.731    553.788  -7.004 5.26e-12 ***
age           262.184      6.755  38.814  < 2e-16 ***
male         -319.764    186.498  -1.715  0.08681 .
bmi            53.086     16.276   3.262  0.00115 **
children      455.528     77.771   5.857 6.86e-09 ***
southwest   -1163.433    266.624  -4.364 1.45e-05 ***
southeast   -1169.130    276.975  -4.221 2.71e-05 ***
northwest    -809.490    268.237  -3.018  0.00263 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2650 on 803 degrees of freedom
Multiple R-squared:  0.6748,    Adjusted R-squared:  0.672
F-statistic: 238.1 on 7 and 803 DF,  p-value: < 2.2e-16
```

Now, we fit Q-Q plot and Residual Plot.





Since, Normality of Errors is not achieved, we cannot apply AIC method here. Instead, we fit all possible models and compare R^2 values, since R^2 doesn't have any distributional assumption. Among them, we choose that model in which we get the highest R^2 value.

```
> ols_step_best_subset(m4)
                    Best Subsets Regression
--------------------------------------------------------------------
Model Index    Predictors
--------------------------------------------------------------------
     1         age
     2         age children
     3         age children southwest
     4         age bmi children southwest
     5         age bmi children southwest southeast
     6         age bmi children southwest southeast northwest
     7         age male bmi children southwest southeast northwest
--------------------------------------------------------------------

                                                                Subsets Regression Summary

-----------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------
                       Adj.        Pred

Model    R-Square    R-Square    R-Square     C(p)         AIC          SBIC          SBC             MSEP
           FPE           HSP         APC
-----------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------
  1        0.6477      0.6472      0.6461    63.0730    15149.5866    12847.7875    15163.6814    6124685790.
3525    7570640.9102    9346.5556    0.3541
  2        0.6616      0.6607      0.6595    30.7496    15118.9454    12817.2477    15137.7385    5890363874.
5409    7289943.3249    9000.0944    0.3409
  3        0.6638      0.6626      0.6611    27.2234    15115.5650    12813.8604    15139.0563    5858674525.
7980    7259620.8075    8962.7677    0.3395
  4        0.6659      0.6642      0.6625    24.1542    15112.5980    12810.9076    15140.7876    5830136219.
7697    7233111.3107    8930.1749    0.3383
  5        0.6699      0.6678      0.6656    16.2510    15104.8058    12803.2248    15137.6937    5767329163.
3567    7163947.8050    8844.9455    0.3351
  6        0.6736      0.6712      0.6685     8.9397    15097.4898    12796.0593    15135.0759    5708558892.
0545    7099613.9673    8765.7028    0.3320
  7        0.6748      0.6720      0.6689     8.0000    15096.5262    12795.1671    15138.8106    5694828324.
8175    7091184.9127    8755.5089    0.3316
-----------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

We can see the model that takes into account all the regressors, has the highest R^2 value.Hence, we choose that model and we don't have to exclude any regressors in this case.

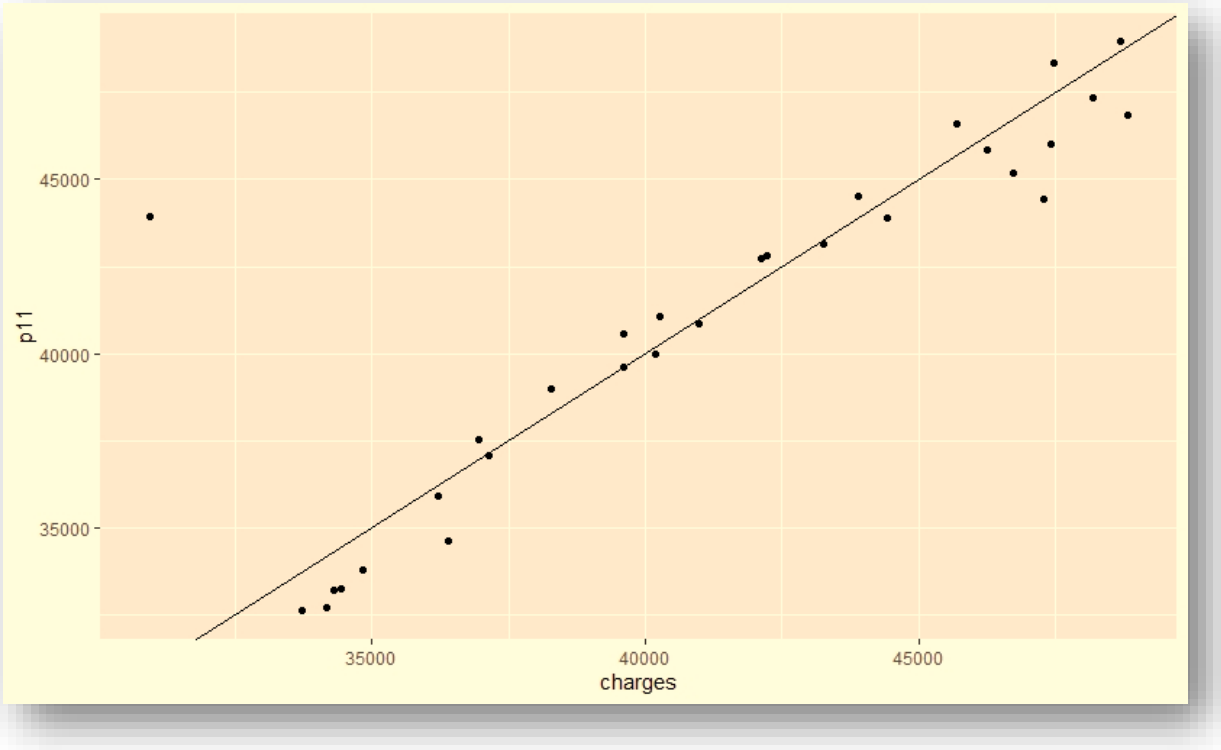
- **Model Fitting after Variable Selection:**

Since, we don't have to exclude any observation, model will be same after variable selection procedure. As a result, model summarty and observations will not change.

➢ Actual vs Fitted Plot for Test Data:

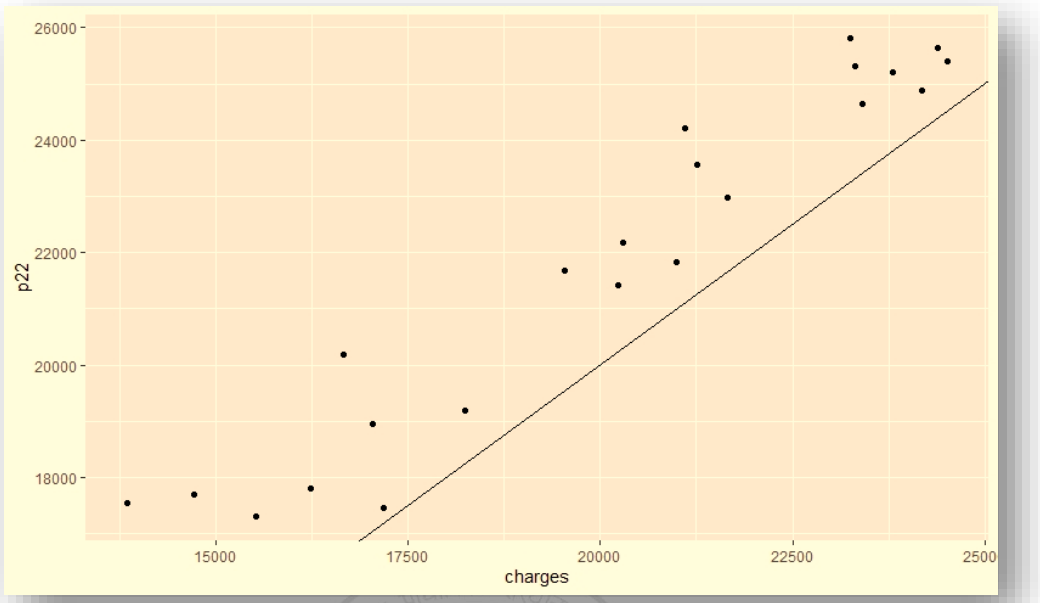
After successful selection of variables, and fitting model, based on the train data, we will test our model based on the test data. For this purpose, we obtain actual vs fitted plot for 3 defined subsets of test data.

- Smokers with High bmi:



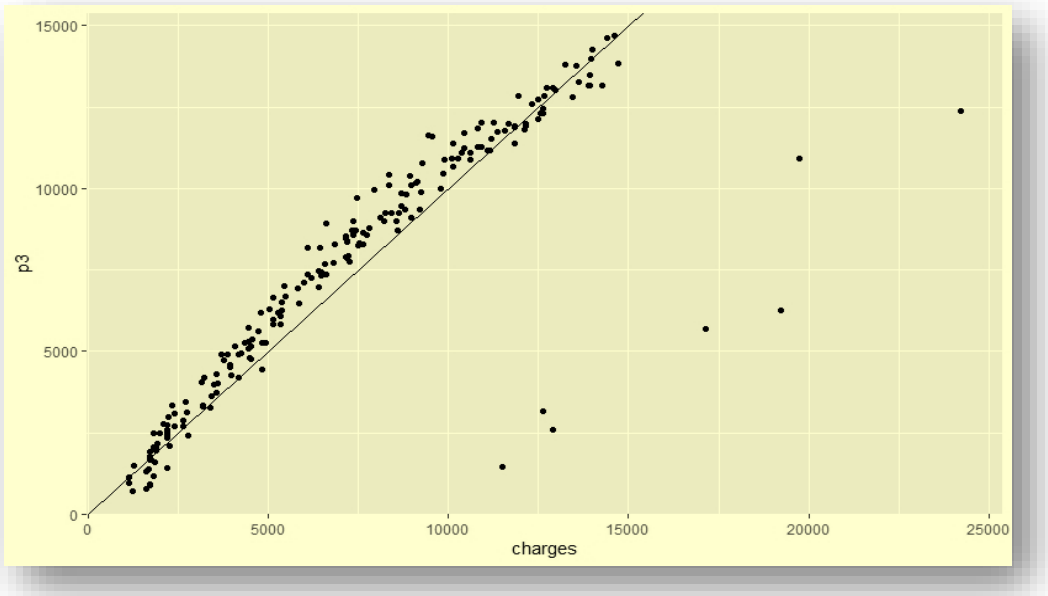
Here, "p11" denotes the predicted charges and "charges" denotes the actual charges for the category "smokers with high bmi"

- Smokers with Low bmi:



Here, "p22" denotes the predicted charges and "charges" denotes the actual charges for the category "smokers with low bmi"

- Non-Smokers:



Here, "p3" denotes the predicted charges and "charges" denotes the actual charges for the category "Non-Smokers"

➢ Conclusion:

After analyzing dataset and fitting 3 different models for smoker high bmi, smoker low bmi and nonsmoker we can conclude that,

- Smokers with High Bmi:

  Final Model:

  ```
  charges=9505.01+258.67*age+602.27*bmi+337.48*children-
  1109.71*southwest -1284.35*southeast
  ```

  For smokers with bmi > 30, most important attributes are age and bmi. As the model almost explains 82.47% of total variation, we can use this model for future prediction purpose.

- Smokers with Low Bmi:

  Final Model:

  ```
  charges=10856.82+247.75*age+662.18*children+1193.30*southeast+12
  08.43*northwest
  ```

  For smokers with bmi < 30, most important attribute is age. As the model explains 61.05% of total variation, this model is moderate to predict healthcare charges for smokers with low bmi population.

- Nonsmokers:

  Final Model:

  ```
  charges=-3878.731+262.184*age-319.764*male +53.086*bmi +
  455.528*children-1163.433*southwest -1169.130*southeast-
  809.490*northwest
  ```

  For nonsmokers, average healthcare charges is much less than that of smokers. Here, all the regressors are contributing significantly to explain healthcare charges for nonsmokers. As the model explains 67.2% of total variation, it is moderately good to predict healthcare charges for nonsmokers.

➢ Appendix:

R Code:

```
Data=read.csv("insurance_project.csv") head(Data)

male=vector() for (i in 1:nrow(Data)) { if (Data[,'sex'][i]=='male')
{ male[i]=1 } else { male[i]=0 } } smokes=vector() for (i in
1:nrow(Data)) { if (Data[,'smoker'][i]=='yes') { smokes[i]=1 } else {
smokes[i]=0 } } southwest=vector() for (i in 1:nrow(Data)) { if
(Data[,'region'][i]=='southwest') { southwest[i]=1 } else {
southwest[i]=0 } } southeast=vector() for (i in 1:nrow(Data)) { if
(Data[,'region'][i]=='southeast') { southeast[i]=1 } else {
southeast[i]=0 } } northwest=vector() for (i in 1:nrow(Data)) { if
(Data[,'region'][i]=='northwest') { northwest[i]=1 } else {
northwest[i]=0 } } Data$male = maleData$smokes=smokes Data$southwest =
southwestData$southeast=southeast Data$northwest=northwest head(Data)

#deleting columns no longer needed Data1=subset(Data,select=-
c(sex,smoker,region))

require(dplyr)

Data1 <- Data1 %>% relocate(charges, .after = northwest) head(Data1)
formula_0 <- as.formula("charges ~ age + male + bmi + children +
smokes + southwest+southeast+northwest") model_00 <- lm(formula_0,
data = Data1) cooksd <- cooks.distance(model_00) influential <-
as.numeric(names(cooksd)[(cooksd > (4/1338))]) Data3 <- Data1[-
influential, ]

sample_size <- 1338 plot(cooksd, pch="*", cex=2, main="Influential
Observations by Cooks Distance") # plot cook's distance abline(h =
4/sample_size, col="red") # add cutoff line
text(x=1:length(cooksd)+1,
y=cooksd,labels=ifelse(cooksd>4/sample_size,names(cooksd),""),
col="red")

n_train <- round(0.8* nrow(Data3)) train_indices <-
sample(1:nrow(Data3), n_train) Data2 <- Data3[train_indices, ]
Data_test <- Data3[-train_indices, ] Data_test Data_train=Data2

formula_0 <- as.formula("charges ~ age + male + bmi + children +
smokes + southwest+southeast+northwest") mq <- lm(formula_0, data =
Data_train) summary(mq) vif(mq) dwtest(mq,alternative = "two.sided")
library(stats) shapiro.test(rstudent(mq)) library(lmtest) bptest(mq)

p1=plot(mq, which=1, col=c("blue")) p2=plot(mq, which=2,
col=c("red")) ggarrange(p1,p2,nrow=2,ncol=2)
q1=ggplot(aes(x=Data_train$age,y=rstudent(mq)),data=mq)+geom_point()+
```

```
geom_smooth(se=FALSE)+geom_hline(yintercept =
0)+labs(x="age",y="residuals")

q2=ggplot(aes(x=Data_train$bmi,y=rstudent(mq)),data=mq)+geom_point()+
geom_smooth(se=FALSE)+geom_hline(yintercept =
0)+labs(x="bmi",y="residuals")

q3=ggplot(aes(x=Data_train$children,y=rstudent(mq)),data=mq)+geom_poi
nt()+geom_smooth(se=FALSE)+geom_hline(yintercept =
0)+labs(x="children",y="residuals") ggarrange(q1,q2,q3,nrow=2,ncol=2)
``` library(car) model_0 <- lm(formula_0, data = Data2) vif(model_0)


head(Data2)

Data2_smokes=Data2[Data2$smokes==1,] nrow(Data2_smokes)

Data2_non_smokes=Data2[Data2$smokes==0,] nrow(Data2_non_smokes)

Data2_smokes_highbmi=Data2_smokes[Data2_smokes$bmi>29.9,]
nrow(Data2_smokes_highbmi)

Data2_smokes_lowbmi=Data2_smokes[Data2_smokes$bmi<=29.9,]
nrow(Data2_smokes_lowbmi)


library(GGally)


library(ggthemes)


data(insurance)

ggpairs(Data2_smokes_highbmi[,c("age","bmi","children","charges")],
columns = 1:4, ggplot2::aes(colour='red')) + theme_solarized_2()

ggpairs(Data2_smokes_lowbmi[,c("age","bmi","children","charges")],
columns = 1:4, ggplot2::aes(colour='red')) + theme_solarized_2()

ggpairs(Data2_non_smokes[,c("age","bmi","children","charges")],
columns = 1:4, ggplot2::aes(colour='red')) + theme_solarized_2()

m2=lm(charges~age + male + bmi + children
+southwest+southeast+northwest,data=Data2_smokes_highbmi)

summary(m2) m3=lm(charges~age + male + bmi + children
+southwest+southeast+northwest,data=Data2_smokes_lowbmi) summary(m3)
m4=lm(charges~age + male +bmi + children
+southwest+southeast+northwest,data=Data2_non_smokes) summary(m4) #
summary(aov(m2)) # summary(aov(m3)) # summary(aov(m4)) stepAIC(m2,
```
```

```
direction = "both") stepAIC(m3, direction = "both") stepAIC(m4,
direction = "both")

m5=lm(charges~age + bmi+children
+southwest+southeast,data=Data2_smokes_highbmi) summary(m5)
m6=lm(charges~age + children
+southeast+northwest,data=Data2_smokes_lowbmi) summary(m6)
m7=lm(charges~age + male +bmi + children
+southwest+southeast+northwest,data=Data2_non_smokes) summary(m7) #
summary(m6) # summary(m7)

plot(m5, which=1, col=c("blue")) plot(m5, which=2, col=c("red"))
plot(m5, which=3, col=c("orange")) plot(m5, which=5, col=c("green"))

plot(m6, which=1, col=c("blue")) plot(m6, which=2, col=c("red"))
plot(m6, which=3, col=c("orange")) plot(m6, which=5, col=c("green"))

plot(m7, which=1, col=c("blue")) plot(m7, which=2, col=c("red"))
plot(m7, which=3, col=c("orange"))

durbinWatsonTest(m5) durbinWatsonTest(m6) durbinWatsonTest(m7)

library(lmtest) bptest(m5) bptest(m6) bptest(m7)

library(olsrr) ols_plot_resid_hist(m5) ols_plot_resid_hist(m6)
ols_plot_resid_hist(m7)

library(stats) shapiro.test(rstudent(m5)) shapiro.test(rstudent(m6))
shapiro.test(rstudent(m7))

Data_smoker_test=Data_test[Data_test$smokes==1,]
nrow(Data_smoker_test)

Data_nonsmoker_test=Data_test[Data_test$smokes==0,]
nrow(Data_nonsmoker_test)

Data_smoker_highbmi_test=Data_smoker_test[Data_smoker_test$bmi>29.9,]
nrow(Data_smoker_highbmi_test)

Data_smoker_lowbmi_test=Data_smoker_test[Data_smoker_test$bmi<=29.9,]
nrow(Data_smoker_lowbmi_test) #data5=Data_smoker_highbmi_test[,-
c(2,4)] p1=predict(m5,Data_smoker_highbmi_test)
ggplot(data=Data_smoker_highbmi_test,aes(x=charges,y=p1))+geom_point(
)+geom_abline(slope = 1, intercept = 0)

p2=predict(m6,Data_smoker_lowbmi_test)
ggplot(data=Data_smoker_lowbmi_test,aes(x=charges,y=p2))+geom_point()
+geom_abline(slope = 1, intercept = 0)

p3=predict(m7,Data_nonsmoker_test)
ggplot(data=Data_nonsmoker_test,aes(x=charges,y=p3))+geom_point()+geo
m_abline(slope = 1, intercept = 0)
```

➢ References:

(i) Introduction to linear regression analysis - Montgomery, D. C., Peck, E. A., and Vining, G. G., Wiley Inc. (Wiley Series in Probability and Statistics)

(ii) https://en.wikipedia.org/wiki/

(iii) https://stats.stackexchange.com/