

15<sup>th</sup> May, 2022

# Predictive Analysis of Insurance Premium Data using Multiple Linear Regression Model

BY Saanvi Jain

BS21DMU005



Statistical Data Analysis  
Dr. Suchismita Das

# Contents

---

## 1. Introduction

## 2. Descriptive statistic

2.1. Data description

2.2 Descriptive Statistics for Dependent and Independent variables

2.3 Histogram

2.4 Box Plots

2.5 Graphs and fitted lines

2.6 Correlation

## 3. Multiple Linear Regression Prediction Model

3.1 Model

3.2 Regression statistics table

## 4. Model Validation

4.1 Variance Inflation Factor (VIF):

4.2 Residuals and QQ plot

## 5. Hypothesis Testing

## 6. Prediction Analysis

## 7. Summary and Conclusion

## 8. References

# 1. Introduction

---

## About the data set –

Health insurance is a way to pay for a person's or individuals' medical bills.

In the United States, the majority of individuals have private health insurance, which is generally obtained via their current work, while the minority is covered by government-sponsored schemes.

The insurer determines rates for their insurance policies based on two key factors: the cost the insurer expects to pay under their policies and the cost of operating certain policies or plans. Medical expenditures are determined in a variety of methods, including the policyholder's health status, region of residence, job status, and salary.

Regression analysis is a statistical tool that is frequently used to predict healthcare expenditures and compute insurance premiums. The medical expenditures of a policyholder can be impacted by a variety of factors, including their habits, chronic diseases, age, economic circumstances, occupational hazards, area of residence, and so on. Regression analysis may be used to find factors that have a substantial impact on medical expenditures. Regression analysis may also be used to forecast the real cost of an insurance policy, allowing insurers to establish competitive rates.

Setting the same price for all policyholders is not a competitive approach, since those with low expenditures would overpay and maybe abandon the service, while those with high expenses would continue to use the service and cause the insurance business to lose money. Regression models are tools that may be used to create effective categorization systems that provide clients a fair pricing while increasing the company's profitability.

## 2. Descriptive Statistics

---

### 2.1. Data description

Training Data consists of 200 rows and 6 columns (attributes). The 6 attributes:

- **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/m<sup>2</sup>).
- **Smoker status:** Whether the policyholder is a smoker or a non-smoker.
- **Children:** Number of children/dependents covered in the policy.
- **Charges:** Yearly medical expenses billed by the medical insurance provider (\$).

**Dependent variable (Y):** Charges

**Independent variable (Xi):** Age, Sex, BMI, Smoker status, Children

First 6 rows of the dataset:

```
> head(training_data)
```

	age	sex	bmi	children	smoker	charges
1	19	0	27.900	0	1	16884.924
2	18	1	33.770	1	0	1725.552
3	28	1	33.000	3	0	4449.462
4	33	1	22.705	0	0	21984.471
5	32	1	28.880	0	0	3866.855
6	31	0	25.740	0	0	3756.622

Table 2.1 First 6 rows of the dataset.

## 2.2 Descriptive Statistics for Dependent and Independent variables

Descriptive statistics are used to characterize or summarize the properties of a sample or data collection, such as the mean, standard deviation, or frequency of a variable. Inferential statistics can assist us comprehend the collective qualities of a data sample's constituents.

```
> print(data_descriptive)
vars  n    mean    sd  median  trimmed   mad   min   max   range  skew  kurtosis
age    1 200   38.13  14.49   37.00   37.66  19.27  18.00  64.00   46.00  0.21   -1.27
sex    2 200    0.48   0.50    0.00    0.48   0.00   0.00   1.00    1.00  0.08   -2.00
bmi    3 200   30.63   5.66   30.21   30.68   5.79  15.96  49.06   33.10  0.02   -0.17
children 4 200    1.06   1.23    1.00    0.90   1.48   0.00   5.00    5.00  0.95    0.12
smoker 5 200    0.22   0.42    0.00    0.16   0.00   0.00   1.00    1.00  1.31   -0.29
charges 6 200 13098.09 12356.10 8527.56 10902.67 7499.60 1137.01 51194.56 50057.55 1.40    0.97
```

Table 2.1 Descriptive statistic

	age	sex	bmi	children	smoker	charges
count	200	200	200	200	200	200
mean	38.135	0.48	30.63418	1.065	0.225	13098.0894
std	14.48988	0.500854	5.655182	1.23222	0.41863	12356.0981
min	18	0	15.96	0	0	1137.011
25%	26	0	26.81375	0	0	4368.34386
50%	37	0	30.2075	1	0	8527.55873
75%	52	1	34.8	2	0	16933.963
max	64	1	49.06	5	1	51194.5591

Table 2.3 descriptive statistics

Table 2.1 and Table 2.2 depicts the descriptive stats like count (number of observations), mean, standard deviation, minimum value, maximum value, and the quartiles.

The following major observations may be drawn from the above table output:

- The Charges (Y variable) has a mean of 13098.089, which is higher than its median of 8527.56, showing a positive skew of 1.40.
- The column Children and charges have a high kurtosis (4th derivative of the moment generating function) 0.12 of and 0.97, respectively.

## 2.3 Histogram

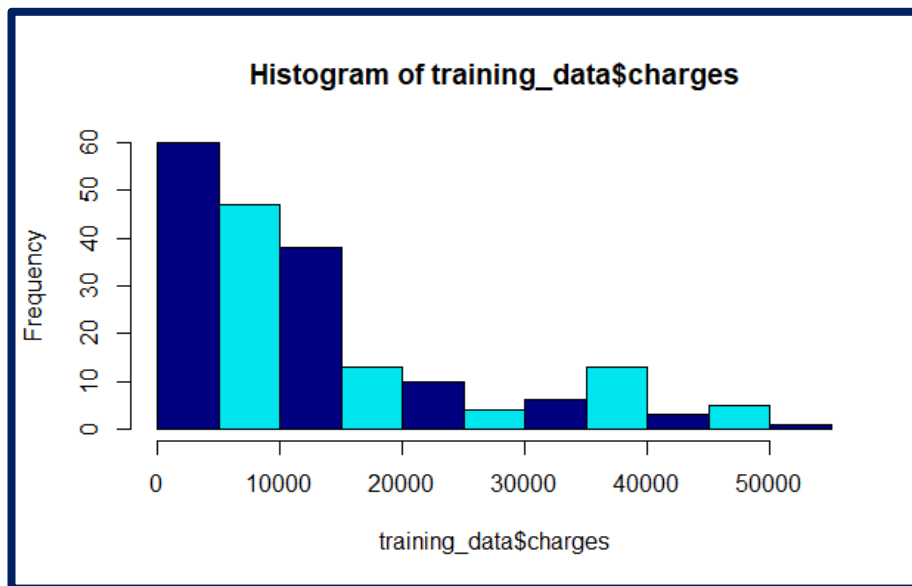


Figure 2.1 Histogram against charges

As the above figure is a normal approximation but positively skewed.

## 2.4 Box Plots

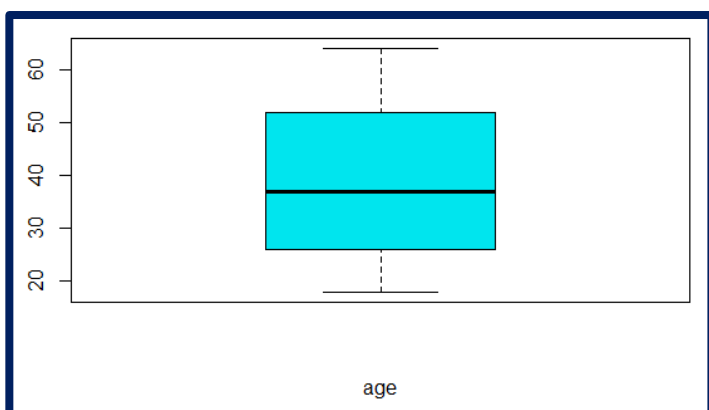


Figure 2.2 Boxplot of charges against age

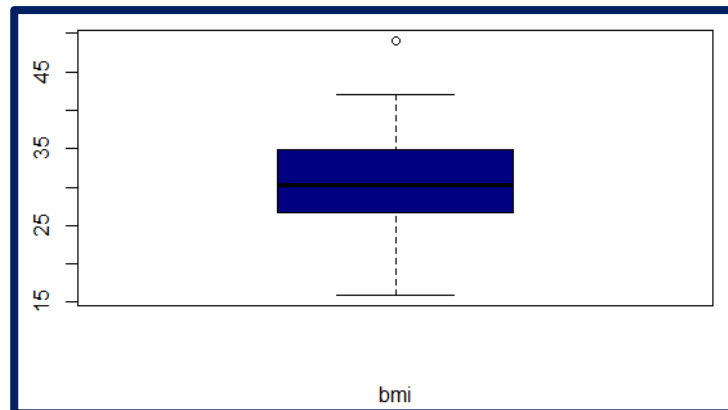


Figure 2.3 Boxplot of charges against BMI

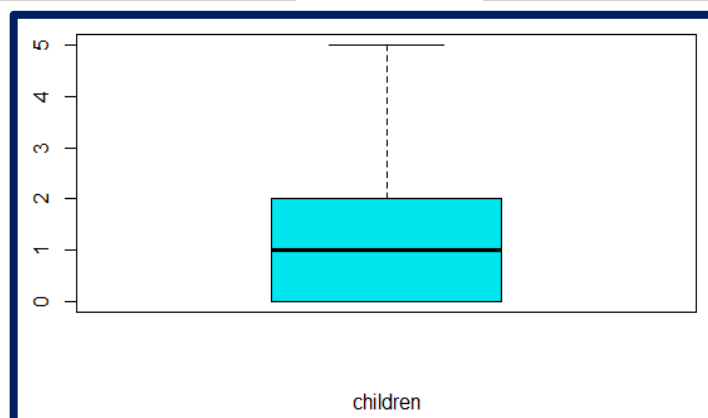


Figure 2.4 Boxplot of charges against BMI

## 2.5 Graphs and fitted lines

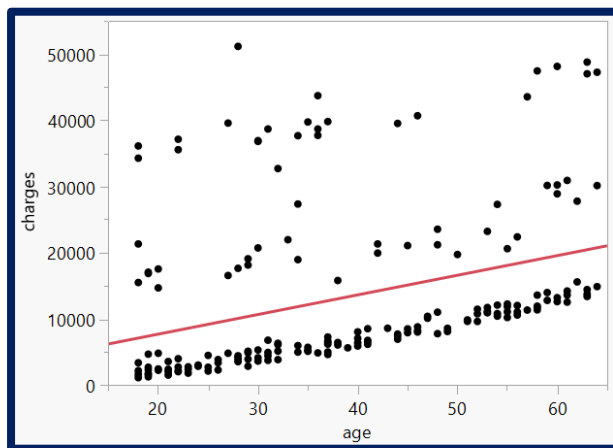


Figure 2.5 Scatter plot of Charges against Age

$$\text{Charges} = 1787.0574 + 296.60501 \cdot \text{Age}$$

### Charges v/s Age

As the points appear to fall around the line, it looks to be significantly positively connected. There is a chance of a linear relationship.

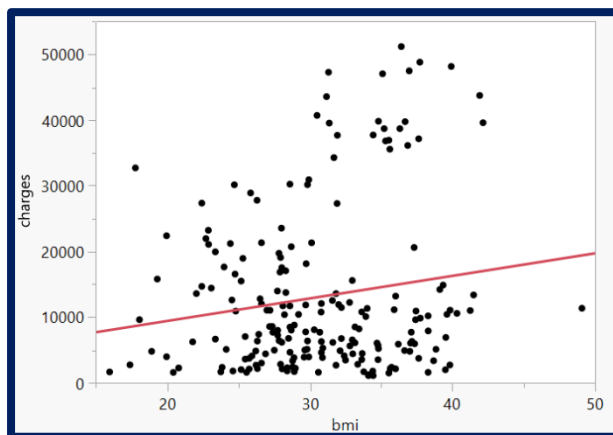


Figure 2.6 Scatter plot of Charges against BMI

$$\text{Charges} = 2579.7429 + 343.35334 \cdot \text{BMI}$$

### Charges v/s BMI

Even though the relationship between Charges and BMI is positive, we can see and conclude that the gap between certain points and the line is too high.

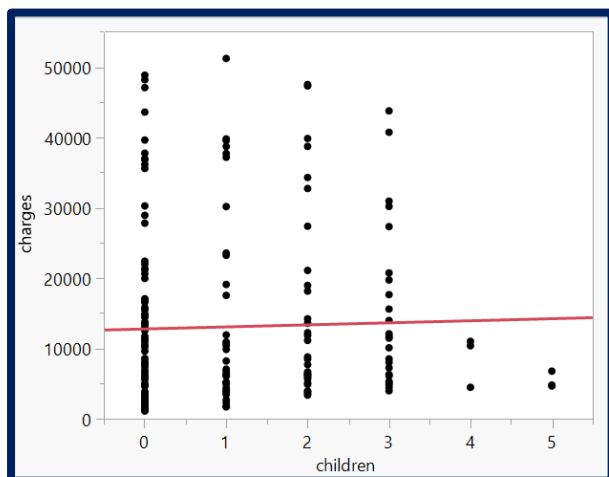


Figure 2.7 Scatter plot of Charges against Children

$$\text{Charges} = 12787.227 + 291.88912 \cdot \text{Children}$$

### Charges v/s Children

Charges and Children appear to be less related since the points on both sides are separated from the line. There is no relationship.

## 2.6 Correlation chart

Correlation is a statistical term that describes how closely two variables are connected linearly (meaning they change together at a constant rate). It's a typical method for explaining simple interactions without stating a cause-and-effect link.

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2} \sqrt{\sum(Y - \bar{Y})^2}}$$

Where,  $\bar{X}$  = mean of X variable  
 $\bar{Y}$  = mean of Y variable

Matrix of correlation coefficients (r) for six numerical variables:

The stronger the link in Figure 2.8, the darker the orange.

	age	sex	bmi	children	smoker	charges
age	1.000000	-0.084448	0.100379	0.035531	0.007394	0.347826
sex	-0.084448	1.000000	-0.057830	-0.042666	0.153386	0.138598
bmi	0.100379	-0.057830	1.000000	0.042107	-0.018176	0.157147
children	0.035531	-0.042666	0.042107	1.000000	-0.038236	0.029109
smoker	0.007394	0.153386	-0.018176	-0.038236	1.000000	0.799030
charges	0.347826	0.138598	0.157147	0.029109	0.799030	1.000000

FIGURE 2.8 – Correlation matrix

Correlation coefficient r provides a measure of linear relationship between X and Y.

Strongly positively linearly correlated variables - Dark orange

Strongly negatively linearly correlated variables – Light orange (white)



### 3. Multiple Linear Regression Prediction Model

---

Multiple linear regression (MLR), often known as multiple regression, is a statistical approach that predicts the result of a response variable using numerous explanatory factors.

#### 3.1 Model

Multiple linear regression prediction model for Points(Y) and  $X_i$  (Age, Sex, BMI, Children, Smoker):

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

Where,

$\hat{Y}$  is the predicted value of dependent variable Y.

$X_i$  is the actual value of independent/explanatory variable:

$X_1$ : Age

$X_2$ : Sex

$X_3$ : BMI

$X_4$ : Children

$X_5$ : Smoker

$\beta_i$  is the regression coefficient of respective  $X_i$ .

## 3.2 Regression statistics table

### Model – 1

Regression Output Coefficients and p-value:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -13397.49    2549.69  -5.255 3.89e-07 ***
training_data$age      282.28     29.03   9.724 < 2e-16 ***
training_data$sex     1356.13    846.98   1.601  0.111
training_data$children 442.31    339.25   1.304  0.194
training_data$bmi      305.11     74.25   4.109 5.85e-05 ***
training_data$smoker   23387.38   1008.67  23.186 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5881 on 194 degrees of freedom
Multiple R-squared:  0.7791,    Adjusted R-squared:  0.7734
F-statistic: 136.9 on 5 and 194 DF,  p-value: < 2.2e-16
```

Figure 3.1 Regression output of model 1

The above result show that for attributes **sex** and **children** has not significant p-value. This gives clear indication that these parameters have less or no impact on the output.

Output of the model:

Statistic	Value
Residual standard error	5881
Multiple R-squared	0.7791
Adjusted R-squared	0.7734

Table 3.1 Model 1 output

Model	Df	F	p value
Regression	5	136.9	2.2e-16
Residual	194		
Total	199		

Table 3.2 Model 1 output

P value < 0.05 of the above F test indicates that the Model 1 holds good for predicting the output. Moreover, value of R-square is also 0.77 which indicates the percentage of the variance in the dependent variable that the independent variables explain collectively and measures the strength of the relationship between our model and the dependent variable.

Let us proceed further and improve the model in the following steps.

### Model-2 : (After removing the insignificant attributes)

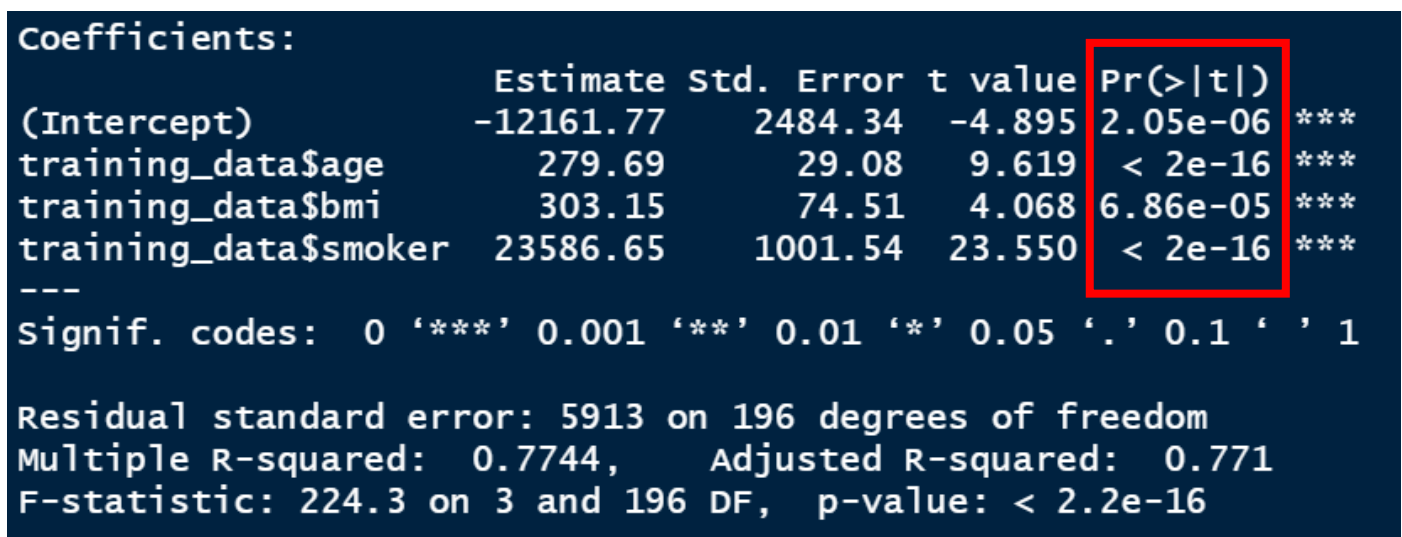
X variables  $\beta_1$ ,  $\beta_2$  etc. ... in the test are as follows:

1. Age
- ~~2. Sex~~
- ~~3. Children~~
4. bmi
5. smoker

Y variable for the model is:

1. Charges

Regression Output Coefficients and p-value After removing the insignificant attributes:



Coefficients:	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-12161.77	2484.34	-4.895	2.05e-06	***
training_data\$age	279.69	29.08	9.619	< 2e-16	***
training_data\$bmi	303.15	74.51	4.068	6.86e-05	***
training_data\$smoker	23586.65	1001.54	23.550	< 2e-16	***

---  
signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5913 on 196 degrees of freedom  
Multiple R-squared: 0.7744, Adjusted R-squared: 0.771  
F-statistic: 224.3 on 3 and 196 DF, p-value: < 2.2e-16

Figure 3.2 Model 2 output

The above result show that for attributes have significant p-value. This gives clear indication that the Model holds good for predicting the output. A p value of 2.2e-16 would suggest a significant result, implying that the actual p value is significantly lower (a typical threshold is 0.05, anything smaller counts as statistically significant).

## 4. Model Validation

---

We'll use the VIF test (Variance Inflation Factor) and step AIC to evaluate if the model is optimal

### Variance Inflation Factor (VIF):

In least squares regression models, variance inflation factors (VIFs) evaluate the connection between independent variables. Multicollinearity is the statistical term for this sort of association. Excessive multicollinearity can cause regression models to fail.

It measures the correlation (linear relationship) between each x variable and other x variables.

$$VIF = \frac{1}{1 - R_i^2}$$

Where  $R_i$  is the coefficient for regressing  $x_i$  on another  $x$ 's

**Criteria:  $VIF > 5$  can be an indication of multi collinearity.**

```
> vif(model)
      training_data$age      training_data$sex training_data$children
           1.017892           1.035258           1.005274
      training_data$bmi      training_data$smoker
           1.014172           1.025742
```

*Figure 4.1 Model 2 VIF Test output*

Since our VIF values are in the criteria we will not conduct Step AIC method.

## Residuals and QQ plot

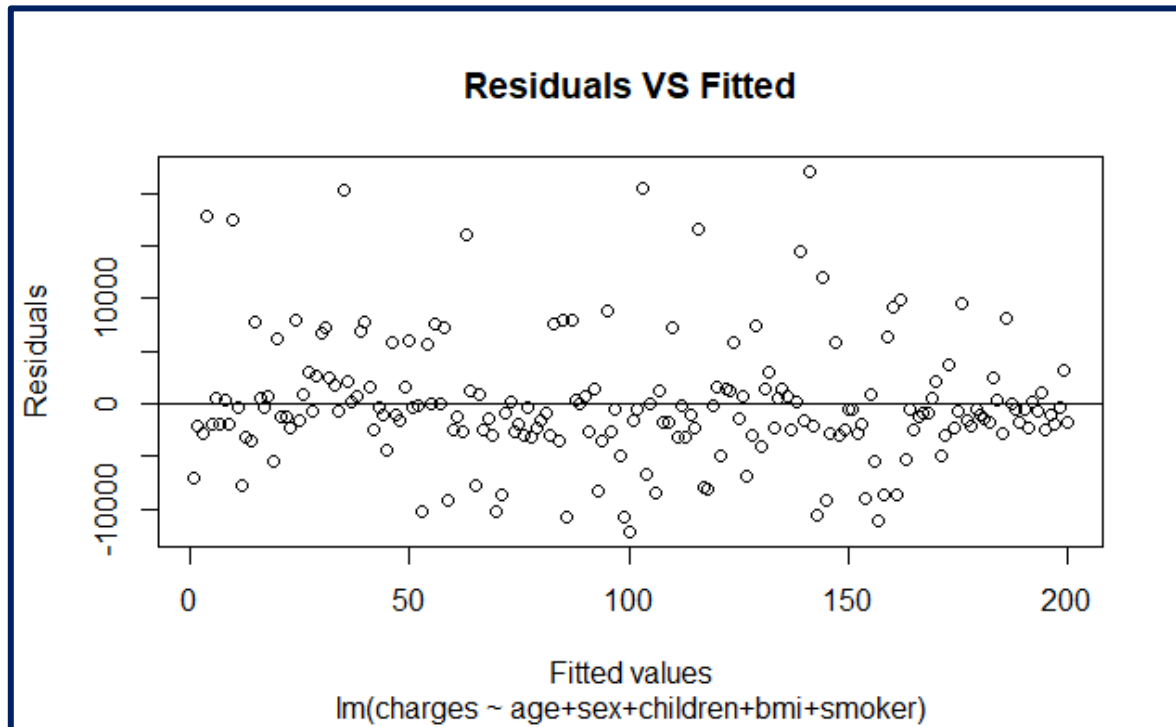


Figure 4.2 Residual plot

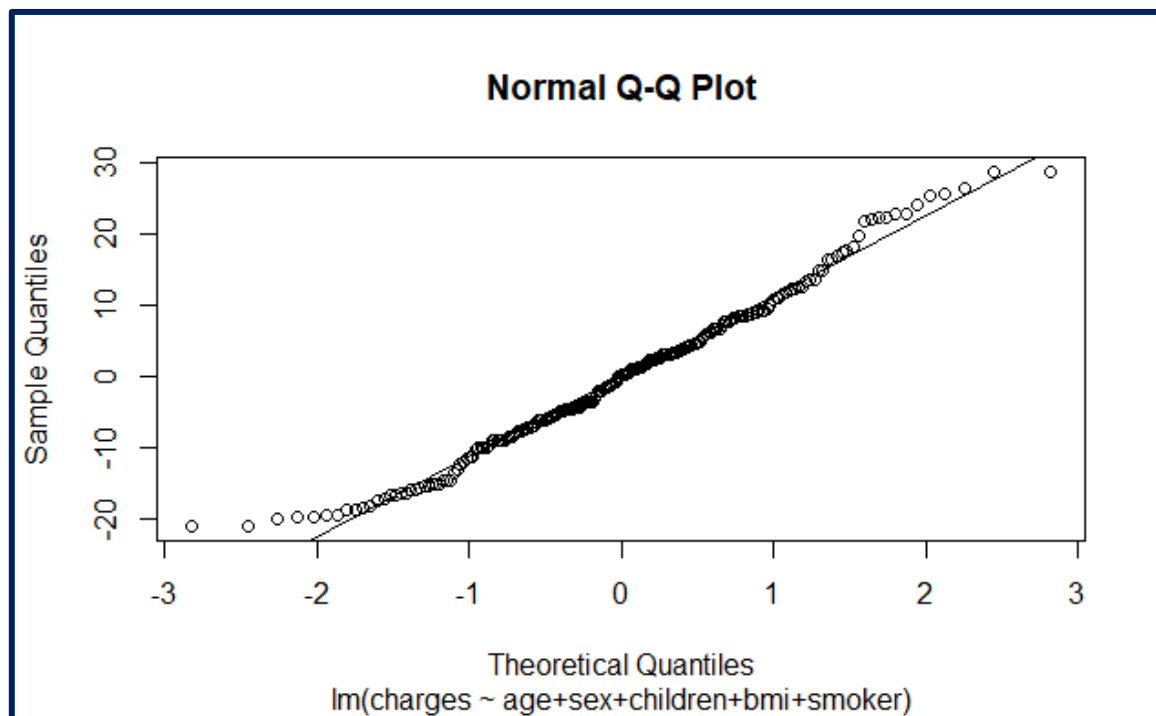


Figure 4.3 QQ plot

We can see that data is linearly distributed and is accurate from both the residual vs fitted plot and the Normal QQ plot.

## 5. Hypothesis Testing

To check model utility.

### 5.1 Hypothesis -

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

There is no linear relationship between the dependent and independent variables.

$$H_A: \beta_j \neq 0 \text{ where } j = 1, 2, 3, 4, 5$$

There is at least one independent variable which has a linear relationship with dependent variable.

### 5.2 ANOVA Output –

The ANOVA table separates the components of variation in the data into treatment variation and error or residual variation. ANOVA tables are produced as part of the standard output for ANOVA by statistical computing software.

```
> anova(model)
Analysis of Variance Table

Response: training_data$charges
              Df      Sum Sq   Mean Sq F value    Pr(>F)
training_data$age      1 3.6757e+09 3.6757e+09 105.115 < 2.2e-16
training_data$bmi      1 4.5855e+08 4.5855e+08  13.113 0.0003733
training_data$smoker    1 1.9394e+10 1.9394e+10 554.615 < 2.2e-16
Residuals             196 6.8538e+09 3.4968e+07
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5.1 ANOVA Table

As the above results show all the p values are significant. We can reject the NULL hypothesis. Model can be used for prediction.

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	2.353E+10	7842726431	224.2812702	4.16539E-63
Residual	196	6.854E+09	34968263		
Total	199	3.038E+10			

Table 5.1 ANOVA TABLE

**F – critical (df1 = 3, df2 = 196,  $\alpha$  = 0.05) = 2.65; F – statistic = 224.28**

Since **F – statistic > F – critical**

**We reject the null hypothesis.** Hence, there is at least one independent variable which has a linear relationship with the dependent variable.

### 5.3 Regression coefficient table and final model

A simple summary of the above output is that the fitted line is (By substituting coefficients in equation)

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
<b>Intercept</b>	-12161.76523	2484.336792	-4.895377016	2.05E-06	-17061.22823	-7262.302237	-17061.22823	-7262.302237
<b>age</b>	279.6901738	29.07792814	9.618641757	3.48E-18	222.344393	337.0359545	222.344393	337.0359545
<b>bmi</b>	303.1540527	74.51461223	4.068383953	6.86E-05	156.2007143	450.1073911	156.2007143	450.1073911
<b>smoker</b>	23586.64686	1001.54465	23.55026995	4.68E-59	21611.45937	25561.83435	21611.45937	25561.83435

*Table 5.2 Regression Coefficient table*

Analysis of the coefficient table:

- $\beta_1, \beta_2, \beta_3$  are positive.
- A negative intercept in a regression model indicates that the model is overestimating the y values on average, necessitating a negative adjustment in the projected values.

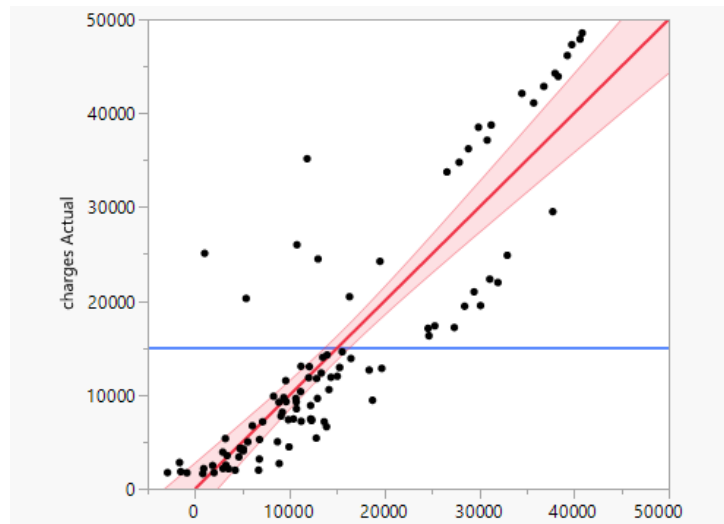
A simple summary of the above output is that the fitted line is:

$$\hat{Y} = -12161.76 + 279.69X_1 + 303.15X_2 + 23586.65X_3$$

## 6. Prediction Analysis

---

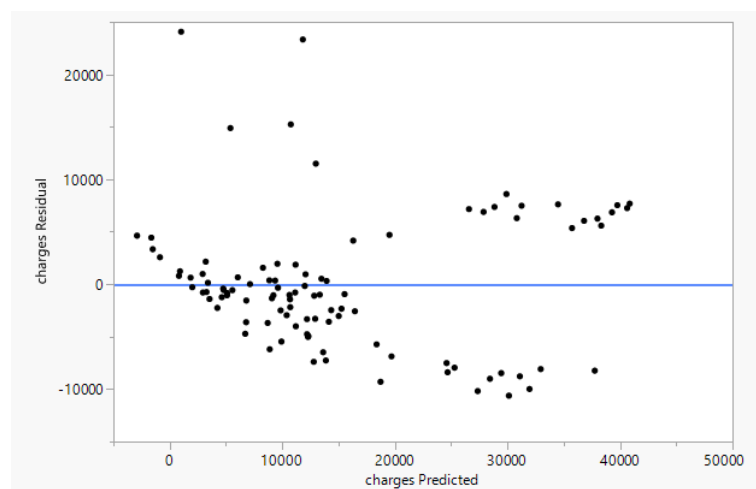
### Actual by Predicted plot :



*Figure 6.1 - Actual by Predicted plot*

Our model fits well because the points are near to the fitted line and the confidence bands are narrow. Points farthest from the mean on the left or right of the plot have the most leverage and can effectively pull the fitted line toward the point. Possible outliers are points that are vertically away from the line. Both types of points might have a negative impact on the fit.

### Residual by charges predicted plot :



*Figure 6.2 - Residual by charges predicted plot*



## 7. Summary and Conclusion

---

Smoking having the strongest effect on medical expenses is quite expected. It's natural for medical costs to rise as the number of dependents grows.

However, having three dependents covered by insurance appears to be less expensive than having two, while having five dependents increases prices less than having four. This could be explained by the fact that each group has a different number of observations. There are 574 observations in the no dependents group, but just 18 in the five dependents group.

---

## 8. References:

- <https://www.kaggle.com/datasets/mirichoi0218/insurance>