Medical Insurance Analysis

Stat 3340

Dalhousie University

Group 16

Meng Lyu B00722473

Dongxu Lyu B0079817

Chuhan Wang B00774998

Abstract

Medical expenses are hard to estimate since there will always be random conditions. But some situations are more prevalent among the population covered by the people. For example, from everyday observation, smokers are more likely to get lung cancers than non-smokers. Our goal is to estimate average medical expenses base on multiple linear regression using the medical insurance dataset.

Introduction:

This project predicts Insurance costs according to different factors to help companies to make business decisions and help to predict cost. Moreover, various visualization methods to help companies to explore the Medical Insurance dataset and obtain accurate results.

Objective:

The objective is to predict the medical insurance costs accurately. Ideally, we want to find various relationships between medical expenses and other variables from the medical insurance dataset. We are curious about the following questions, how would key variables affect medical charges and predict medical insurance costs for a different population.

Data Description:

We have the Medical insurance Dataset from LIONBRIDGE.AI. This dataset consists of 1338 records with seven variables: population' age, sex, BMI, children, smoker, region and medical charges. In this dataset, Charges are the dependent variable or response variable, and the rest of the variables are predictor variables.

Data Source:

From project dataset page: LIONBRIDGE.AI

There are 7 variables in this dataset:

Age, Se, BMI, Children, Smoker, Region and Charges for medical insurance

We can see the first few records from the dataset below.

	age	sex	bmi	children	smoker	region	charges
1	19	female	27.900	0	yes	southwest	16884.924
2	18	male	33.770	1	no	southeast	1725.552
3	28	male	33.000	3	no	southeast	4449.462
4	33	male	22.705	0	no	northwest	21984.471
5	32	male	28.880	0	no	northwest	3866.855

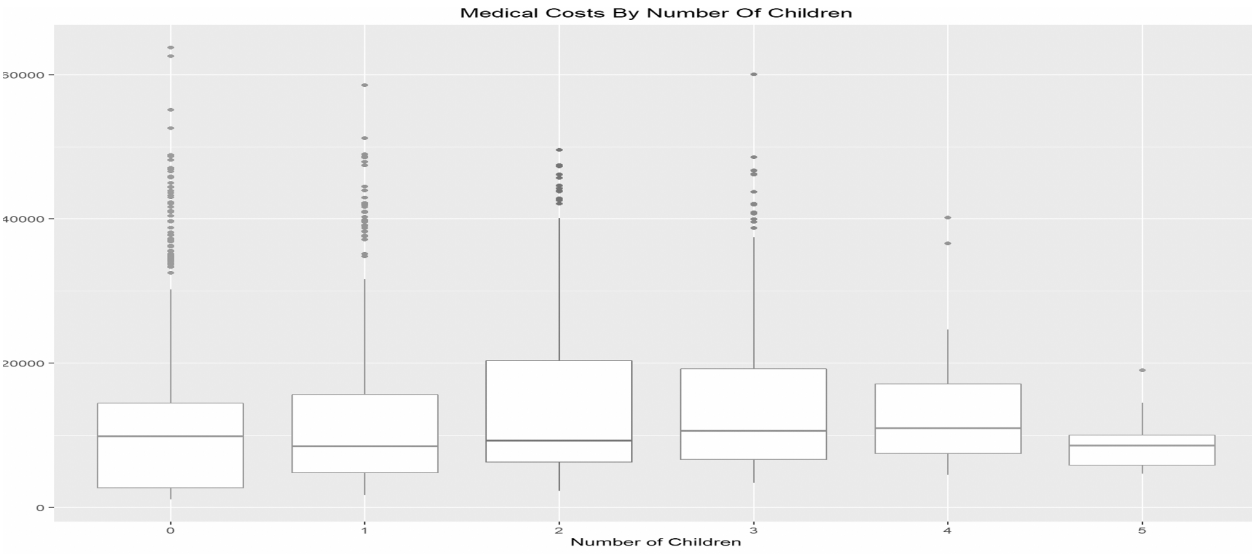
In this Medical Insurance dataset, the insurance costs are the charges variable in dollars. Age, BMI and children are numerical variables; sex, smoker, and region are categorical variables.

Exploratory Data Analysis: Now we will make plots, diagrams on Data to get some basic information

Descriptive statistics:

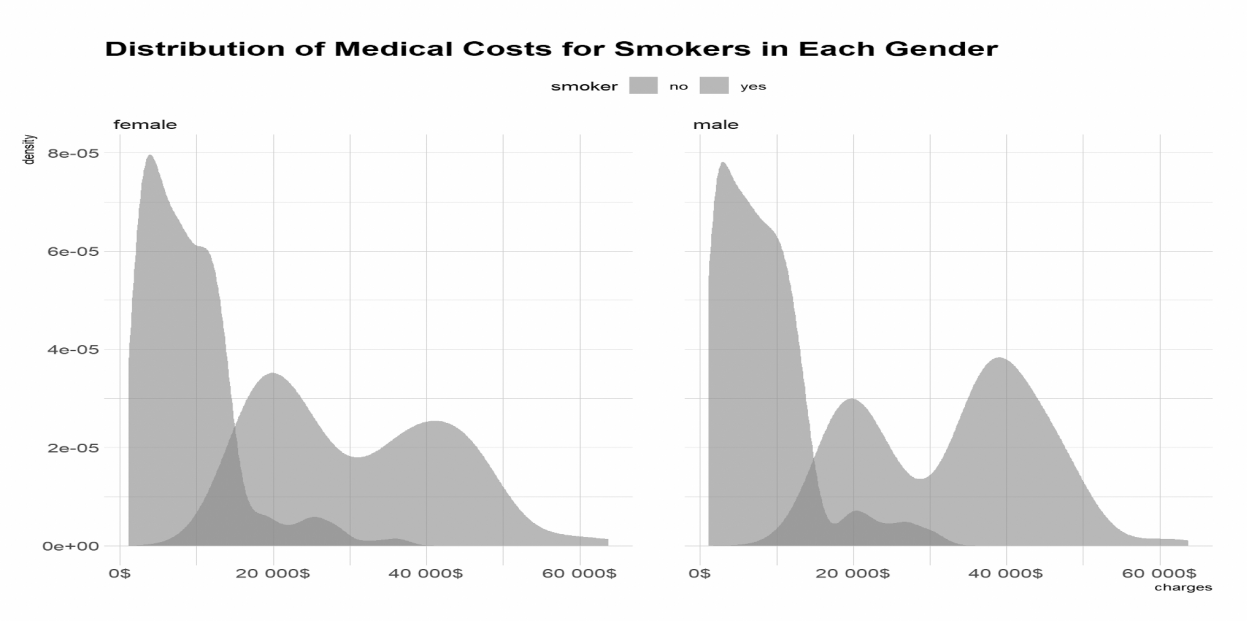
Below is a summary of the basic statistics for our variables. Looking at the response variable, the minimum value is 1122, while the maximum value is 63770. Most points cluster between 4740 and 16640. This variance of the response variable indicates various extremely, which tell us that there might be some potential outliers.

Medical Costs and children covered by health insurance



From our day to day experience, we shall see an increase in the medical cost for large families, But in reality, it is not the case. A severe disease for one person would cost more than a healthy family combined.

Distribution of Medical Costs for Smokers in Each Gender:



But this is not necessarily the case. A single beneficiary with chronic disease would have higher medical expenses than that of a healthy family combined. Nevertheless, the sample sizes for 4 and 5 children groups are very small, so the plot might not reflect the true nature of their distribution.

There is no apparent relationship between gender and medical insurance cost. We are interested in children as well. We are curious about whether having children or not would affect family members' smoking habit as well. There are a lot of interesting questions we can consider.

Methods:

We use linear regression by equation below:

$$charges=\beta0+\beta1*age+\beta2*bmi+\beta3*sex+\beta4*children+\beta5*region+\beta6*smoker+e$$

We would predict the medical costs according to the rest of the variables in our dataset. There are various approaches to choose from; we use AIC and R-squared, and RMSE. Based on these two values, our model performance is if we get greater R-squared and smaller RMSE values. We have created Multiple Linear regression and use the Stepwise model to predict medical insurance costs.

Results

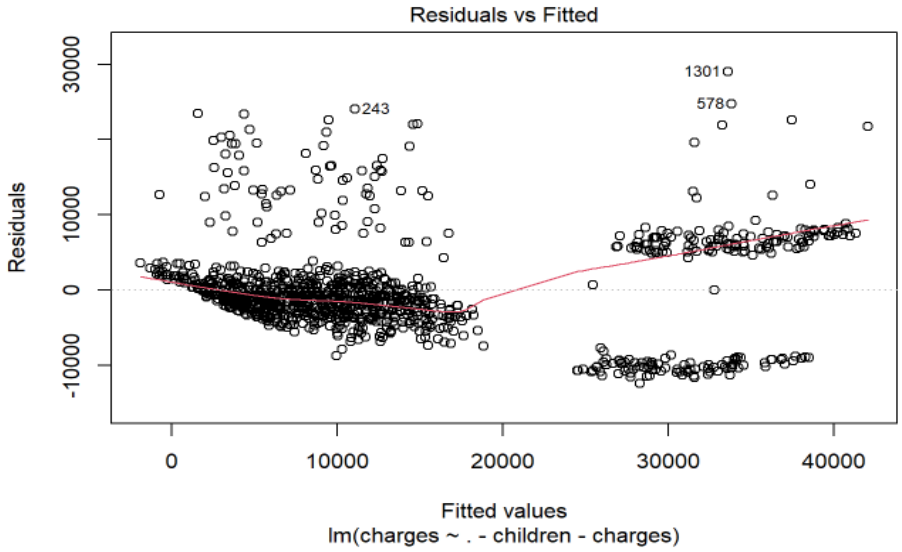
Multiple Linear Regression:

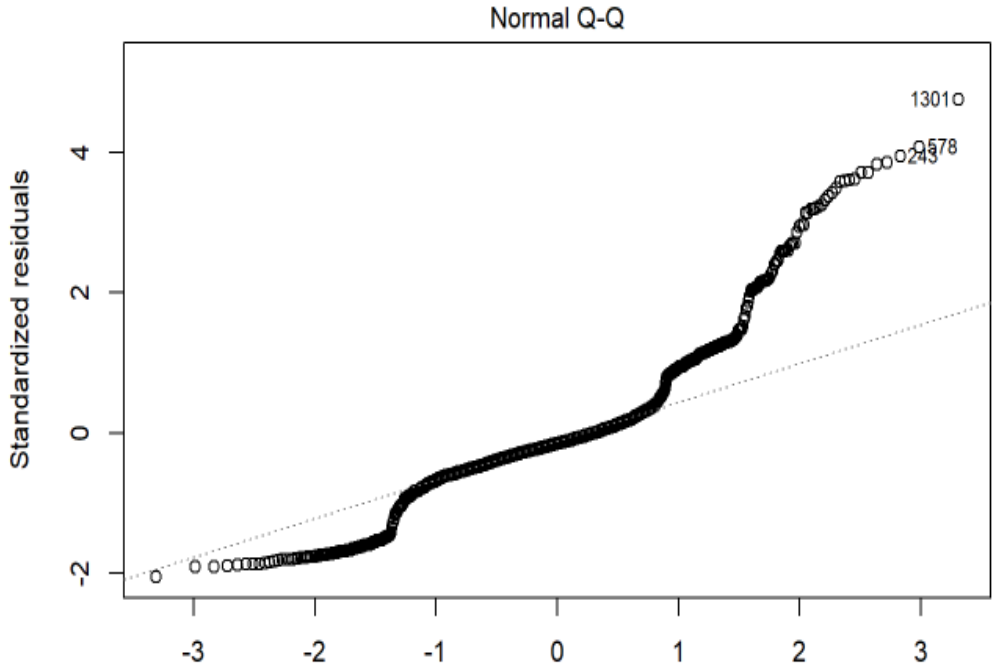
```
model<-lm(formula = charges ~ . - children - charges, data = insurance)
summary(model)
```

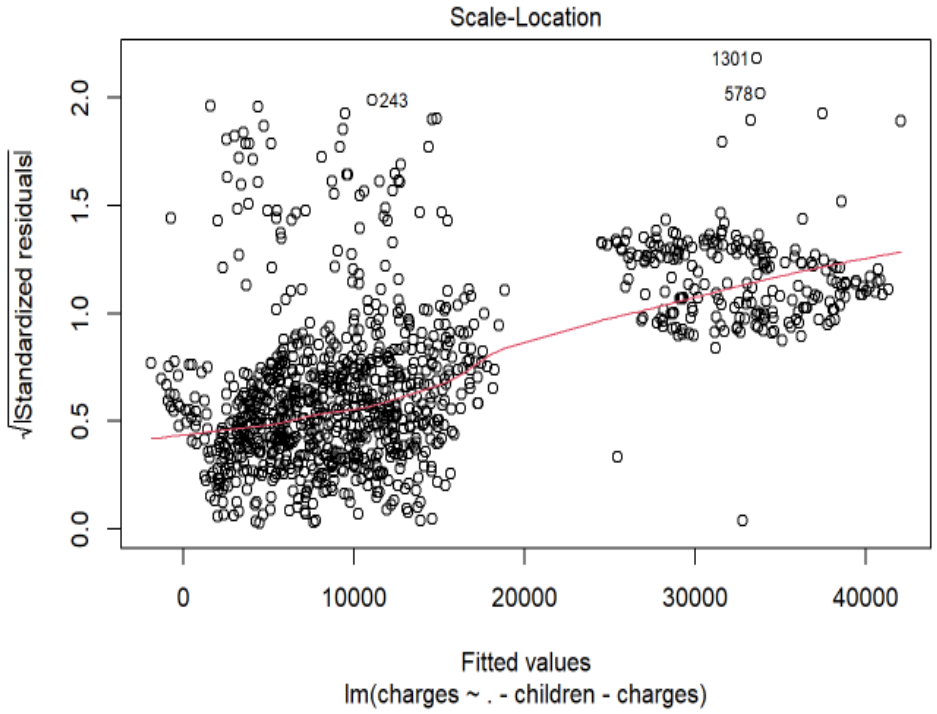
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12469.8  -2971.9   -903.8   1565.8  29007.8
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11668.93    1111.95 -10.494   <2e-16 ***
## age                257.25      13.35  19.269   <2e-16 ***
## sexmale           -320.86     375.85  -0.854    0.393
## bmi                341.64      32.10  10.642   <2e-16 ***
## smokeryes        24472.64     466.11  52.504   <2e-16 ***
## regionnorthwest   -182.21     536.10  -0.340    0.734
## regionsoutheast   -846.26     543.49  -1.557    0.120
## regionsouthwest   -561.61     545.92  -1.029    0.304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6125 on 1066 degrees of freedom
## Multiple R-squared:  0.7526, Adjusted R-squared:  0.751
## F-statistic: 463.3 on 7 and 1066 DF,  p-value: < 2.2e-16
```

Models show small p-values, R2 is 0.75. it indicates that there will be 75% variability. Another problem we need to consider is linearity between variables as well. We can confirm this by doing

residual analysis.







Use Stepwise approach to our model:

```
model_sp <- stats::step(lm(charges ~., data = insurance), direction = "backward", trace = 0)
summary(model_sp)
plot(model_sp)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12353.8  -2897.6   -940.8   1462.5  29044.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12191.62    1064.40  -11.45  < 2e-16 ***
## age            256.34      13.27   19.32  < 2e-16 ***
## bmi            323.67      30.53   10.60  < 2e-16 ***
## children       520.66     155.43    3.35 0.000837 ***
## smokeryes    24378.35     461.42   52.83  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6095 on 1069 degrees of freedom
## Multiple R-squared:  0.7543, Adjusted R-squared:  0.7534
## F-statistic: 820.6 on 4 and 1069 DF,  p-value: < 2.2e-16
```

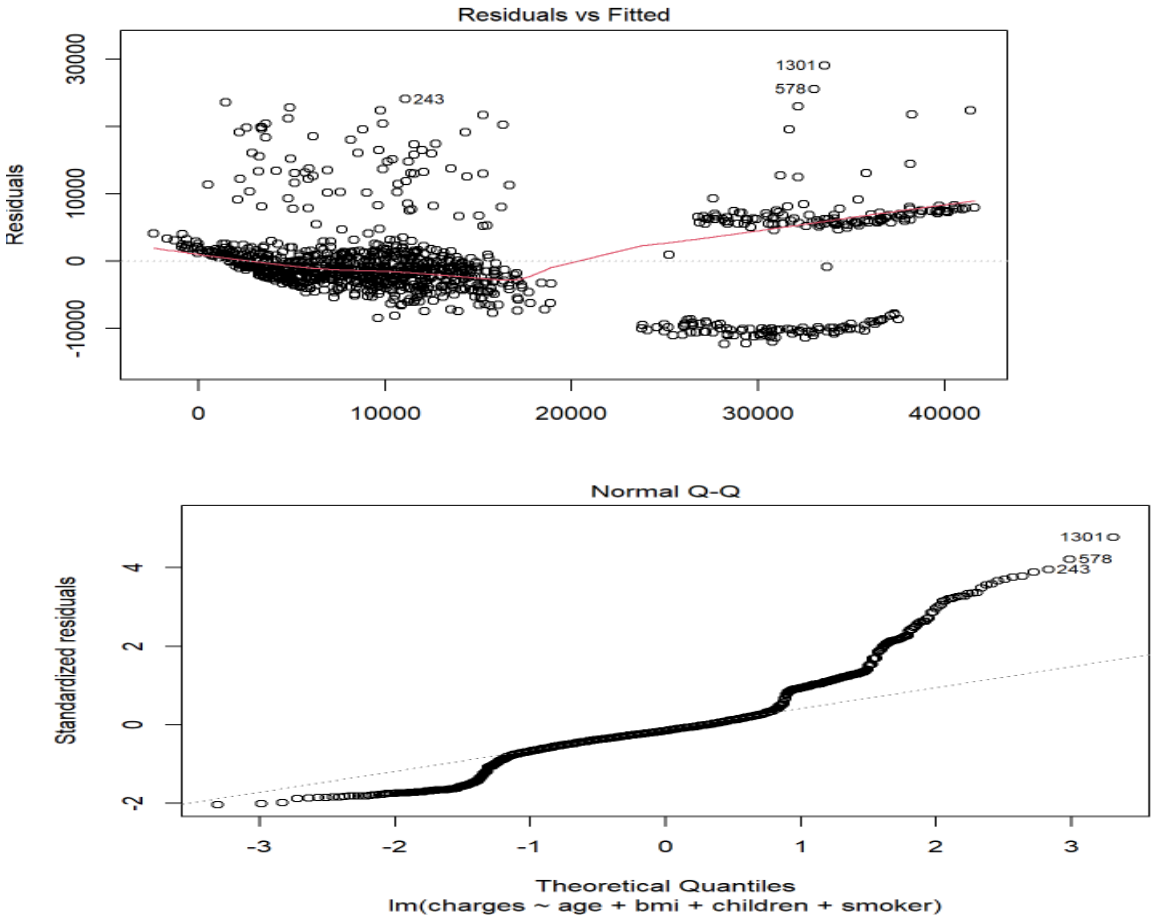
The result gives all significant factors which help to determine the medical insurance costs. The regression model has very low p-values. By applying the stepwise approach, we observe that age, BMI, children, and smoker contribute the most to our model as predictors. And in the coefficients from the regression model, we discover that smoker is the most significant factor within all these factors.

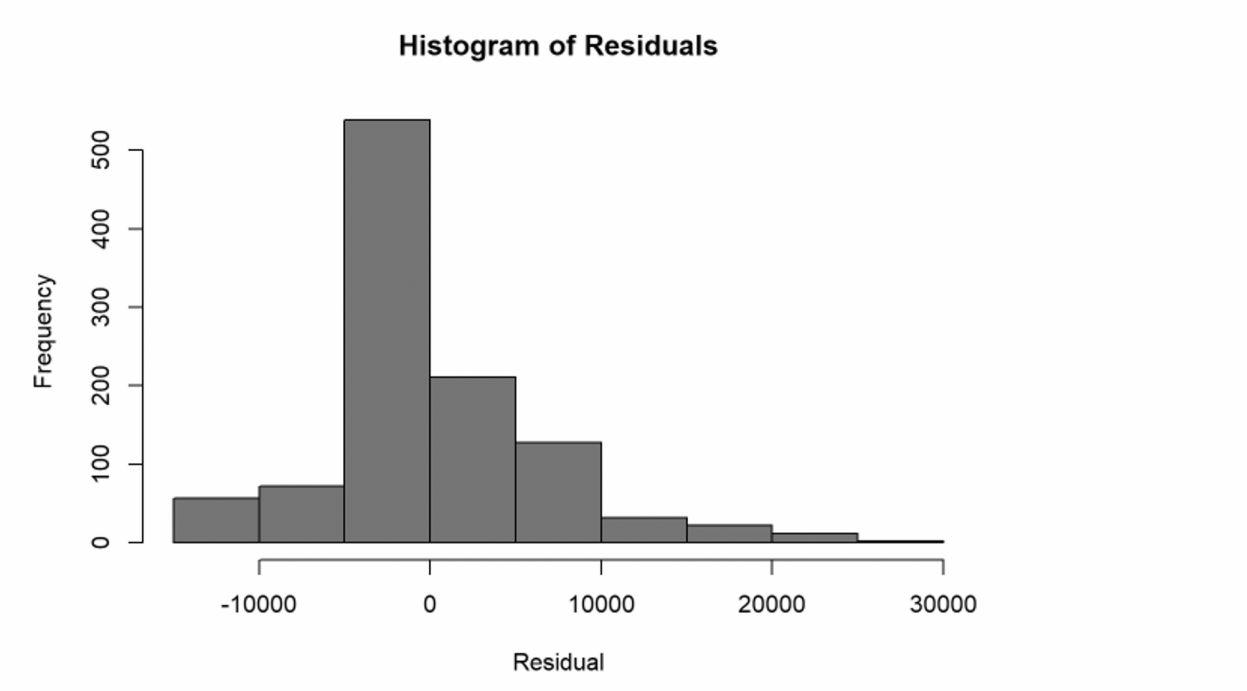
Check Model performance

The Adjusted R-Squared result is not ideal. If we do data cleaning to reduce the noise at the begging, for instance, remove missing values from medical insurance dataset.

```
##        mae       rsq
## 1 4172.758 0.7345454
```

Now let's perform the assumption checking to our linear model.





Histogram of residual is uniformly distributed as figure below. There is no obvious pattern between Residual and fitted value. It is random from the QQ plot.

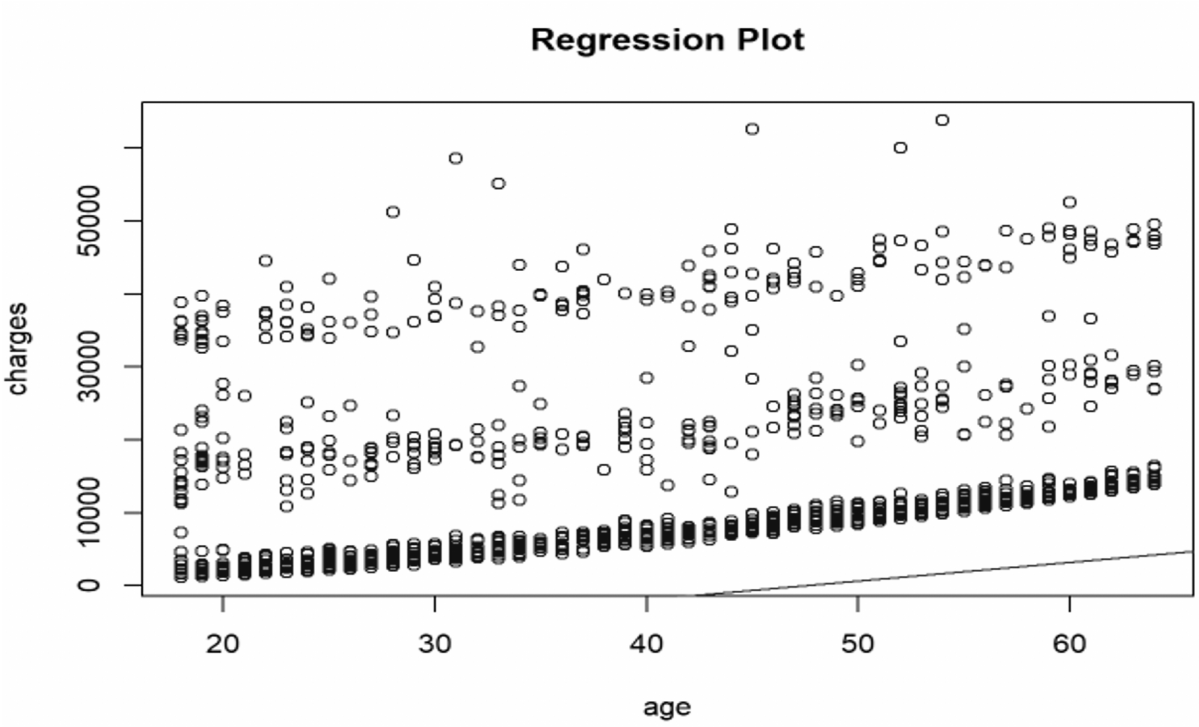
Multicollinearity check

According to the VIF score on our predictor, since the score is small. Therefore, we can conclude that there is no multicollinearity problem for our dataset.

```
car::vif(model_sp)
```

```
##      age      bmi children   smoker
## 1.010525 1.007849 1.002018 1.002557
```

Below plot shows the charges and age scatter plot along with regression line.



Conclusion:

Here, we built a linear regression model and used it to predict medical insurance costs. To do this, we can do the following:

- clean our data to reduce noise
- plot diagrams to Visualize
- Build Multiple Linear regression model
- Obtain information from model results and make some improvements based on the results.

Appendix:

Shown in GitHub .rmd file