Group 5
Mohammad Shaik & Jagadeesh Siripurapu

Price Prediction for Health care Insurance

1. Introduction

The most important thing for any nation world-wide is the health of the citizens. Hence, in some countries the governments provide free health care to the public and in some nations the governments provide some expense coverage. But this is only restricted to some countries as of now. With the development of the nations, the health care costs are also increasing rapidly in these countries. This is affecting the lives of the middle class and lower-class people in these countries. With the increasing health care costs, these people are not able to get good medical attention to their health problems. Seeing this as a major issue, both governments and private firms started the concept of Insurance. With insurance, people can pay their monthly fixed charge or premium, so that the insurance company will take care about the health expenses for these people. This solution has been helping the affected people in a good way. Slowly, the Organizations also started to give insurances for their employees and their immediate families. But, with the increase in the demand for insurance, the companies are facing hard to figure out the requirements for handing out an insurance. Earlier, as per the people's present health condition the monthly premium amount is fixed. But some companies have faced losses by taking inaccurate health reports into consideration and deciding the low monthly premium and ending up paying for heavy amounts for the insurance coverage of the people.

Here comes the solution of predicting the insurance expense by using various factors of an individual, like Age, Gender, Body Mass Index, Number of children, Smoking habits, and the geographical region, etc., By using these factors, we can program or develop systems that can decide the monthly fixed premium the individual has to pay for the insurance company. This kind of systems helps to reduce a lot of manual work for the insurance organizations, such as calculating the premium amounts. With this system, the organization can cut off the excess number of manual employees required to do these works, thus resulting in the organization profits.

These systems can be implemented with various technologies and solutions, like Bigdata, Machine Learning, Regression models, etc.., The advantage of using ML based prediction system is that the systems can do self-learning and improve the good predicting percent. The machine learning systems takes years of data in and store them in the database and develop schemas for processing new requests with precision and accuracy. But the model that is demonstrated in this paper is using different regression models to predict good premium amounts for the customers, which will reduce the expenses for the people. We are using two econometric modelling techniques in this paper, Simple regression model and multi regression model, using the factors Age, Gender, Body Mass Index, Number of children, Smoking habits, and the geographical region and charges. Out of all the above factors, charges is the only dependent variable. By using this factor, in simple regression factor, the relation between charges and one independent variable. Whereas with multi regression model, the charges factor is compared with many individual variables. For the same we are using the datasets from the Kaggle site.

2. Literature Review

There have been numerous studies on the insurance cost prediction in the health-related fields. These studies include various types of predictions, like predicting the insurance cost based on health factors, Machine learning based schemas, predicting the insurance based on last year's insurance expenses, etc.., The machine learning prediction depends on the selection of the good suitable schema for a particular prediction. Developing this schema from the past companies' data and training it with some random test data and then the same schema will be deployed for usage.

Sudhir [1] "they have developed a model to predict the monthly fixed premium using the regression models that include simple linear, multi linear , Ridge regression, Lasso regression and polynomial regression, by using health factors as basis."

Lahiri et al. [2] "discussed a model that predicts, if any person's health care amount will go up or down the upcoming year. This prediction is done by using an algorithm that takes previous year's health care data of a person and using a machine learning schema or algorithm to project the next upcoming costs and thus resulting in finding the next years costs."

In [3] [4] "They made a model by using hierarchical decision trees and machine learning models, using these models the insurance prediction will be done to filter out the risks and difficulties while finalizing cost."

In [5], "the authors have discussed a model that predict the upcoming expenses of an individual using his/her old health expenses and insurance claim details. With the use of the old health data, they project upcoming quarterly costs, by using the simple linear regression model and the random forest classifier."

3. Methodology

The data regression models used are statistical ways for creating an association among the target variable charges, dependent attribute and the other group of independent attributes. In this paper we have used the below two types of regression models.

1. Simple Linear Regression Model:

    This model will be showing a straight or linear connection between the target variable (which is charges) (Q) is dependent on an individual independent attribute(P).
    This model is tend to set the regressor line in the middle of the above discussed P and Q attributes.

$$Q = x + yP$$

The variable x and y are the model parameters known as regression coefficients. When P is equal to zero, the line's Q intercept is equal to "x," and the slope that represents the modification in Q with a modification in P is equal to "y." more of a minor modification in P results in a large modification in Q, and vice versa, if "y" is present. The Ordinary Least Squares method can be used to determine the values of "x" and "y."

2. Multi Linear Regression Model:

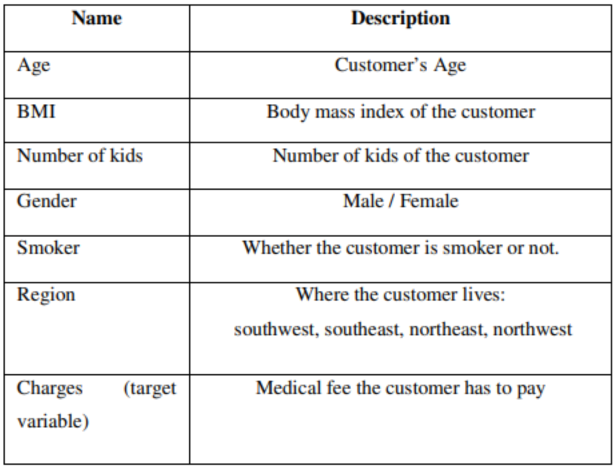
This model is same as the simple linear regression model, but the connection between one target attribute to the other group of independent attributes will be found out. Hence, the value of the target variable(charges)(Q) will be found out using the group of these independent attributes. We often consider that the dependency among the independent attributes in null. Assume that the regressor sets the regression line in an N-dimensional-space if the goal value depends on "n" independent attributes. The formulaic view of calculating the target variable is shown below.

$$Q = x + y1P1 + y2P2 + y3P3 + \dots \quad \dots + ynPn$$

4. Data and Sample

The dataset for this model development has been taken from the Kaggle site [6] and this dataset consists of 7 attributes and 1338 records. The seven attributes are as follows Age, Gender, Body Mass Index, Number of children, Smoking habits, and the geographical region and charges. From this list of 7 attributes, Gender, Smoking habits, and the geographical region are categorical data and the remaining i.e., Age, Body Mass Index, Number of children and charges are numerical values. Now from all the seven factors given, there are some null values in the data in age and body mass index columns, and the same are imputed by calculating the mean of the total values. And from the given 7 attributes, the charges is the target value, cause that’s what the model needs to find out using the other 6 attributes. And then the data set has been divided into two lists, one is for the purpose of training the models and the other list will be used for the purpose of testing the trained model.

Now, we have developed three models for calculating the charges. For each of these models, the insurance data has been properly fed, both the test and train data. As explained, that 3 out of these 7 attributes are category related values, but the regression models only take numbers as input and gives just numbers as output. Therefore, these 3 attributes have been changed into number with the help of label encoding. Now after all these steps, the models have been trained with the training list, so that the model will be more accurate and precise, when fed with real world data to better calculate results.

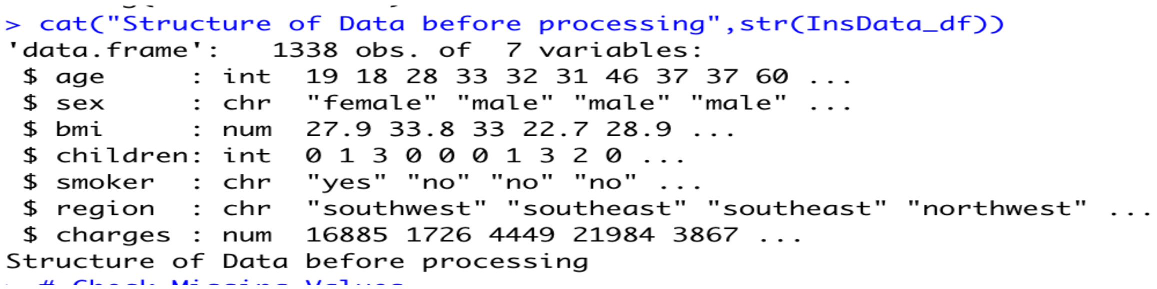


Name	Description
Age	Customer's Age
BMI	Body mass index of the customer
Number of kids	Number of kids of the customer
Gender	Male / Female
Smoker	Whether the customer is smoker or not.
Region	Where the customer lives: southwest, southeast, northeast, northwest
Charges (target variable)	Medical fee the customer has to pay

Dataset

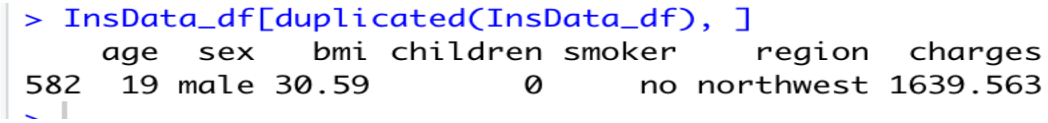
The loading of the insurance data and generating the summary of the same has been given below.

Data Processing:



```
> cat("Structure of Data before processing",str(InsData_df))
'data.frame':	1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ charges : num  16885 1726 4449 21984 3867 ...
Structure of Data before processing
```

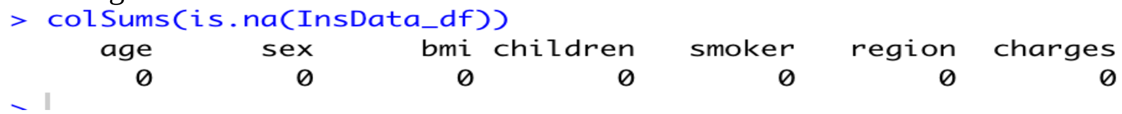
Checking duplicates in dataset



```
> InsData_df[duplicated(InsData_df), ]
    age  sex   bmi children smoker    region  charges
582  19 male 30.59        0     no northwest 1639.563
```

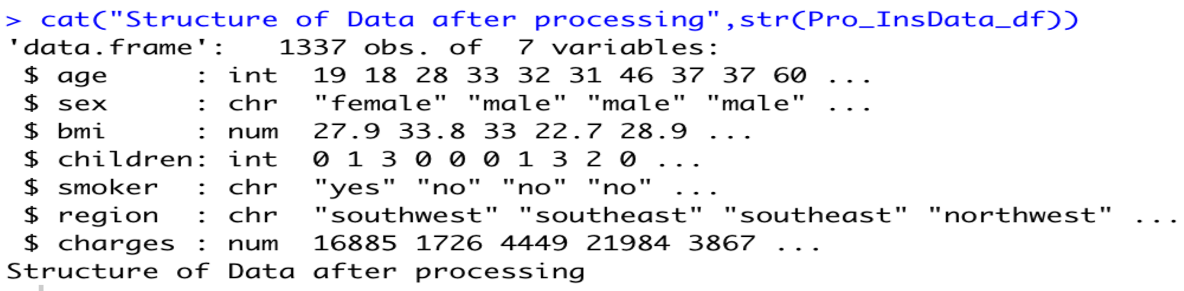
Here in the above snip, we can see that weFound 1 duplicate data in the dataset

Checking null values in data



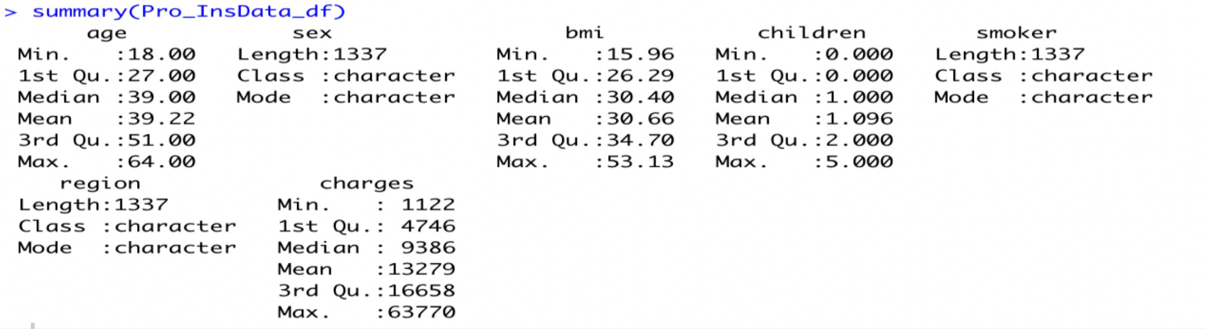
```
> colSums(is.na(InsData_df))
     age      sex      bmi children   smoker   region  charges 
       0        0        0        0        0        0        0 
```

As we can see that there are, No null values in dataset.



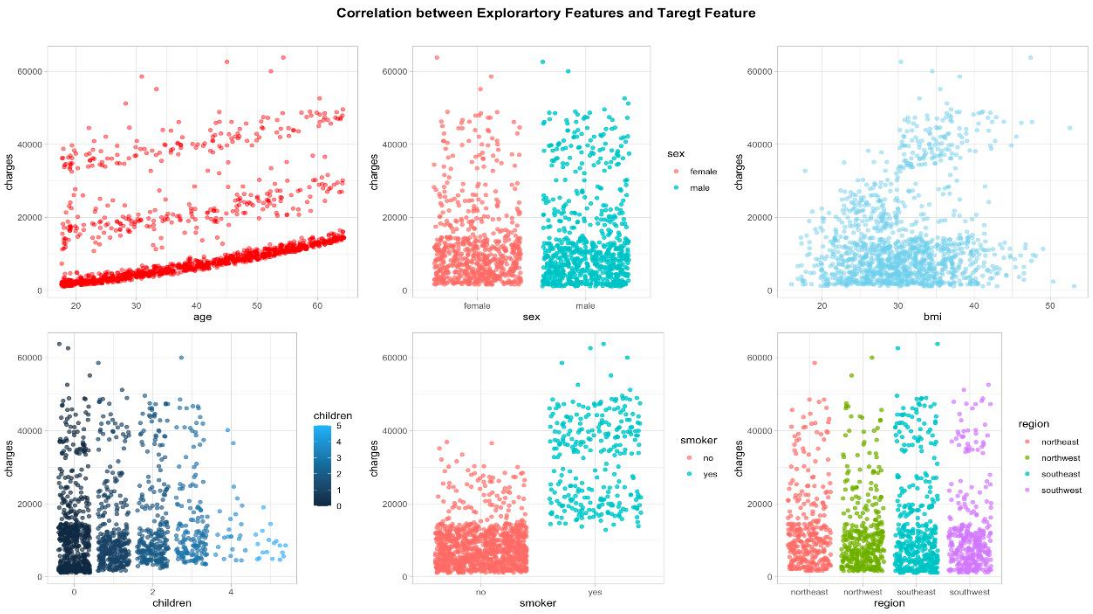
```
> cat("Structure of Data after processing",str(Pro_InsData_df))
'data.frame':	1337 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ charges : num  16885 1726 4449 21984 3867 ...
Structure of Data after processing
```

Removed duplicate data when you check data before and after processing

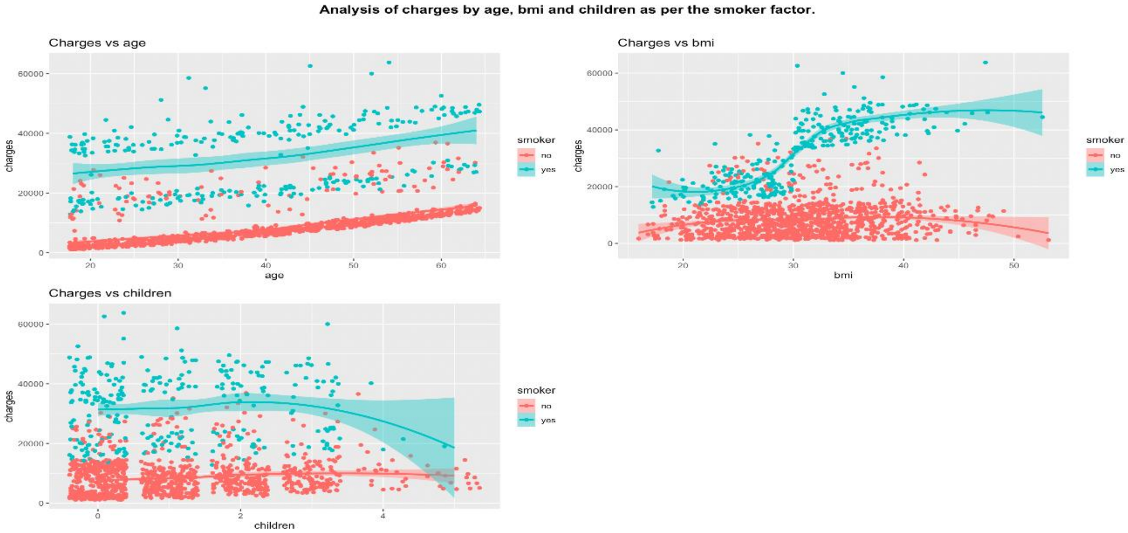


```
> summary(Pro_InsData_df)
      age            sex                 bmi           children        smoker         
 Min.   :18.00   Length:1337        Min.   :15.96   Min.   :0.000   Length:1337       
 1st Qu.:27.00   Class :character   1st Qu.:26.29   1st Qu.:0.000   Class :character  
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character  
 Mean   :39.22                      Mean   :30.66   Mean   :1.096                     
 3rd Qu.:51.00                      3rd Qu.:34.70   3rd Qu.:2.000                     
 Max.   :64.00                      Max.   :53.13   Max.   :5.000                     
    region             charges     
 Length:1337        Min.   : 1122  
 Class :character   1st Qu.: 4746  
 Mode  :character   Median : 9386  
                    Mean   :13279  
                    3rd Qu.:16658  
                    Max.   :63770  
```

Now as we have taken the dataset insurance and about to use this for our model, we need to analyze them and generate summary of their main characteristics and below are the sni9ps of the results we achieved after the execution.

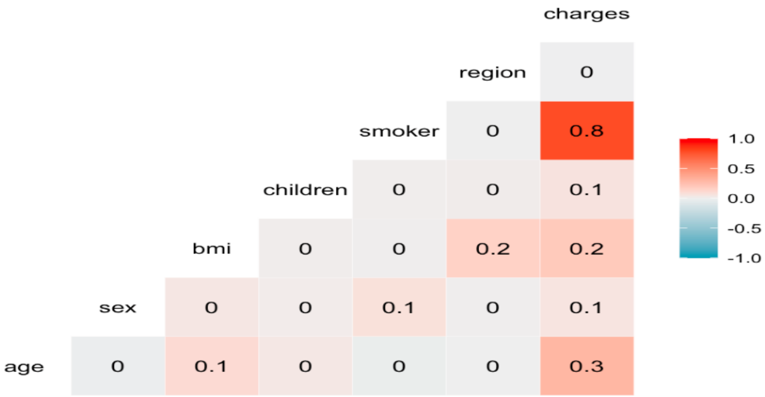


From the features age and bmi, as they increase the charges also increases. There is no apparent relationship between charges and for the features sex and region. As children count goes down charges decreases. Smoker has the high impact on charges as smokers have a high charge than a non-smokers.



Hence from the above plots Smoker have higher impact on Charges. Charges are directly to age and bmi and inversely proportional to children.

The correlation plots of the same dataset have been attached below from the execution



Correlation is a factor to find relationship between each attribute with the target feature. It ranges from 0-1. Higher the correlation factor higher the dependency. A good correlation ranges from 0.5 and vice-versa to the opposite. From correlation plot, you can see smoker is highly correlated to charges. Region, Sex and children doesn't have any impact with the charges.

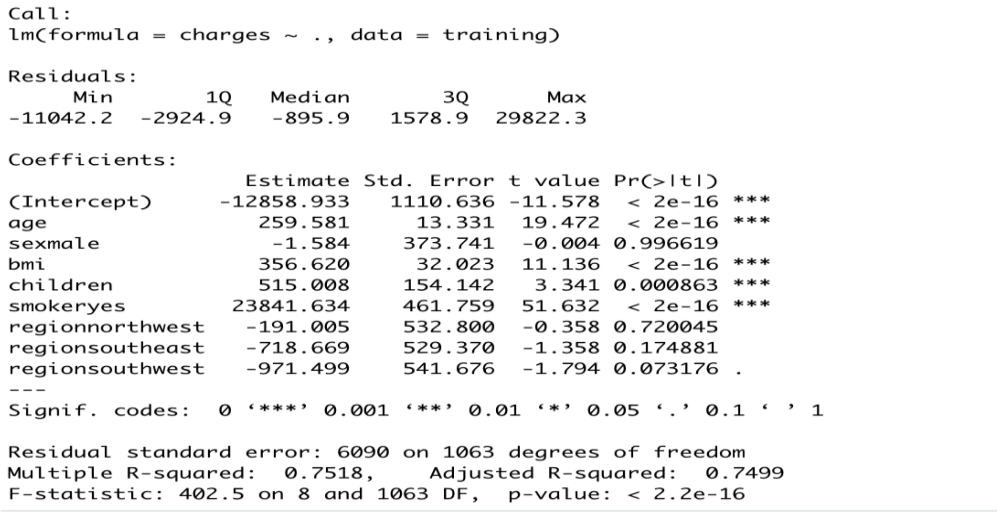
Data Modeling:

There have been three models developed using the multi regression models. The execution of each and every one is shown beow. Please the below step by step model.

Model 1:

Now the model1 execution has been shown below. Now as you can see "*" values in the models, a three star model is the highest form of indication that this factor affects the total charges calculated by this model.

Lm(formula = charges - ., data = training)

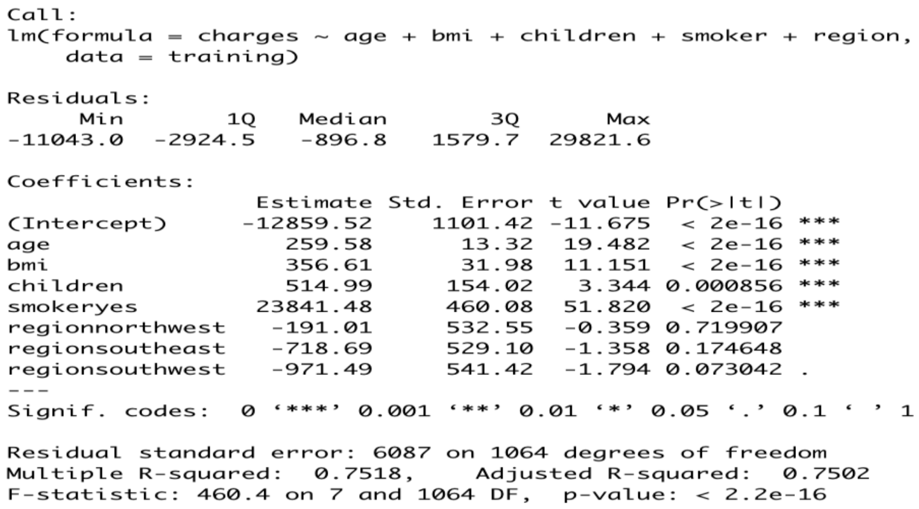


```
Call:
lm(formula = charges ~ ., data = training)

Residuals:
     Min       1Q   Median       3Q      Max 
-11042.2  -2924.9   -895.9   1578.9  29822.3 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12858.933   1110.636 -11.578  < 2e-16 ***
age                259.581     13.331  19.472  < 2e-16 ***
sexmale             -1.584    373.741  -0.004 0.996619    
bmi                356.620     32.023  11.136  < 2e-16 ***
children           515.008    154.142   3.341 0.000863 ***
smokeryes        23841.634    461.759  51.632  < 2e-16 ***
regionnorthwest   -191.005    532.800  -0.358 0.720045    
regionsoutheast   -718.669    529.370  -1.358 0.174881    
regionsouthwest   -971.499    541.676  -1.794 0.073176 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6090 on 1063 degrees of freedom
Multiple R-squared:  0.7518,    Adjusted R-squared:  0.7499 
F-statistic: 402.5 on 8 and 1063 DF,  p-value: < 2.2e-16
```

Model 2:

Now the model2 execution has been shown below. Now as you can see "*" values in the models, a three star model is the highest form of indication that this factor affects the total charges calculated by this model.

Lm(formula = charges _ age + bmi+ children+smoker+region, data=training)

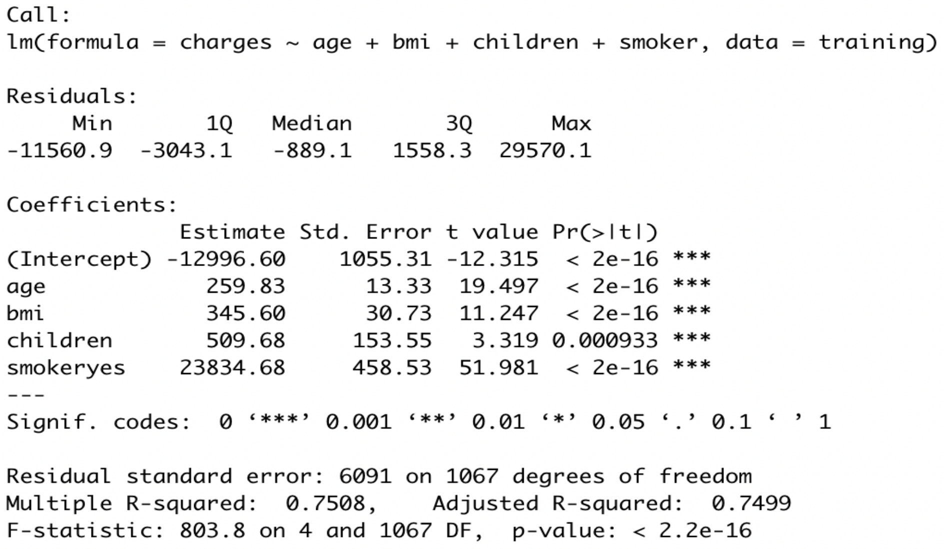


```
Call:
lm(formula = charges ~ age + bmi + children + smoker + region,
    data = training)

Residuals:
     Min       1Q   Median       3Q      Max 
-11043.0  -2924.5   -896.8   1579.7  29821.6 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -12859.52    1101.42 -11.675  < 2e-16 ***
age                259.58      13.32  19.482  < 2e-16 ***
bmi                356.61      31.98  11.151  < 2e-16 ***
children           514.99     154.02   3.344 0.000856 ***
smokeryes        23841.48     460.08  51.820  < 2e-16 ***
regionnorthwest   -191.01     532.55  -0.359 0.719907    
regionsoutheast   -718.69     529.10  -1.358 0.174648    
regionsouthwest   -971.49     541.42  -1.794 0.073042 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6087 on 1064 degrees of freedom
Multiple R-squared:  0.7518,    Adjusted R-squared:  0.7502 
F-statistic: 460.4 on 7 and 1064 DF,  p-value: < 2.2e-16
```

Model 3:

Now the model3 execution has been shown below. Now as you can see “*” values in the models, a three star model is the highest form of indication that this factor affects the total charges calculated by this model.

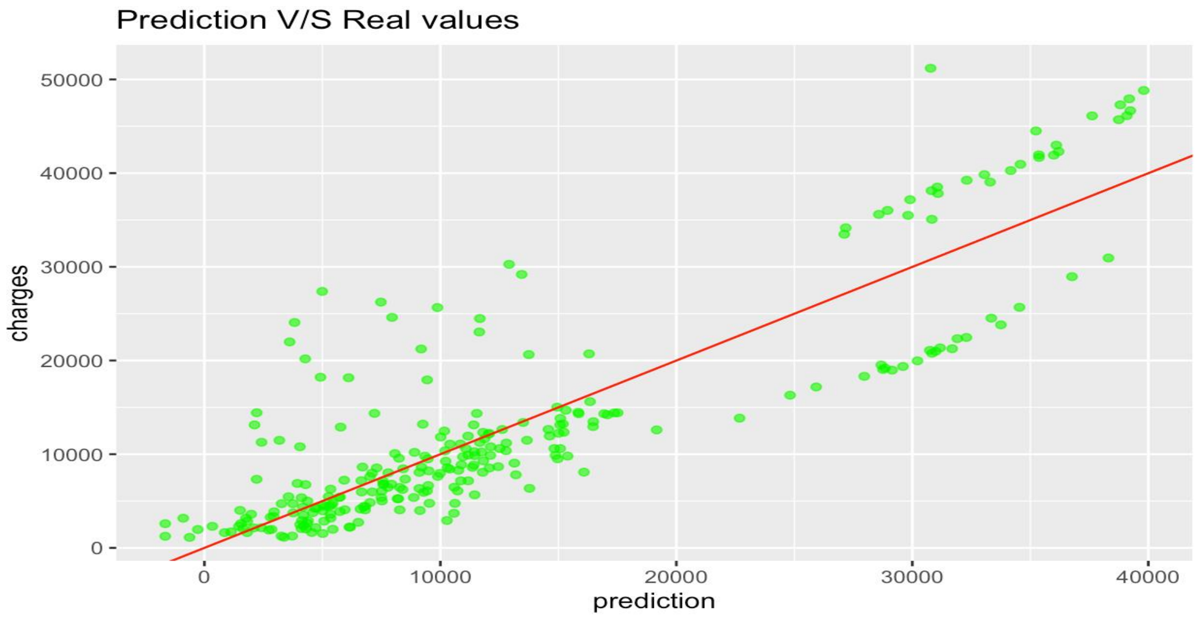
Lm(formula) = charges – age + bmi + children+ smoker, data=traning)

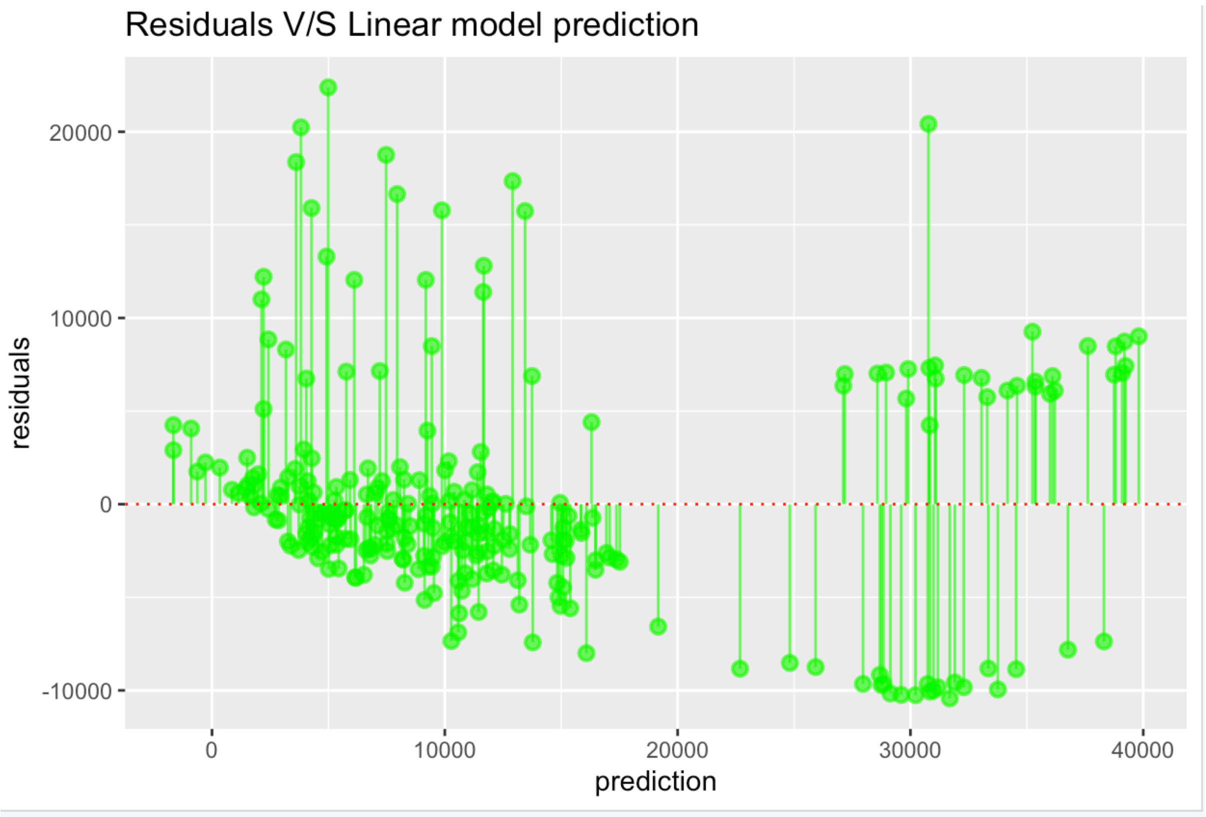


```
Call:
lm(formula = charges ~ age + bmi + children + smoker, data = training)

Residuals:
     Min       1Q   Median       3Q      Max 
-11560.9  -3043.1   -889.1   1558.3  29570.1 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12996.60    1055.31 -12.315  < 2e-16 ***
age            259.83      13.33  19.497  < 2e-16 ***
bmi            345.60      30.73  11.247  < 2e-16 ***
children       509.68     153.55   3.319 0.000933 ***
smokeryes    23834.68     458.53  51.981  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6091 on 1067 degrees of freedom
Multiple R-squared:  0.7508,    Adjusted R-squared:  0.7499 
F-statistic: 803.8 on 4 and 1067 DF,  p-value: < 2.2e-16
```

Model Performance:

The performance of the finalized model have been shown below. Please check the graphs.


Prediction V/S Real values
charges
50000
40000
30000
20000
10000
0
0
10000
20000
30000
40000
prediction


Residuals V/S Linear model prediction
residuals
20000
10000
0
-10000
0
10000
20000
30000
40000
prediction

5. Results and Discussions

In total, three models have been implemented and Model 2 shown above is the chosen model from model1 and model3, considering the below factors and reasons:

1. Residual Standard Error (RSE):

   Usually as low is the RSE, the accuracy will be as high. As you can see from the below image of the comparison of three models, this value is low for model2 than that of model1 and model3.
2. R-Squared:

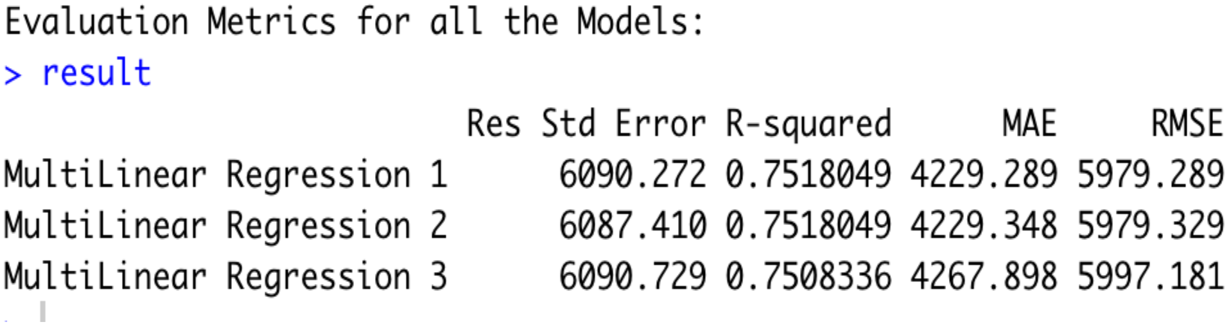
   Basically, the R-squared value comes in between 0 and 1 and R-squared value is directly proportional to accuracy.

   $$\text{R-squared} + (\text{SSE}/\text{SST}) = 1$$

   Now as we can see from the below results snip, the results for model1 and model2 are same i.e 0.7518049.

3. MAE, RMSE:

   This calculation is used for finding out model's prediction error. The root mean square error is usually inversely proportional to accuracy and if the RMSE value is zero then the model is said to be the perfect model. As per the below results snip, the RMSE value for the model 1 is less, which is not so less than the model 2.



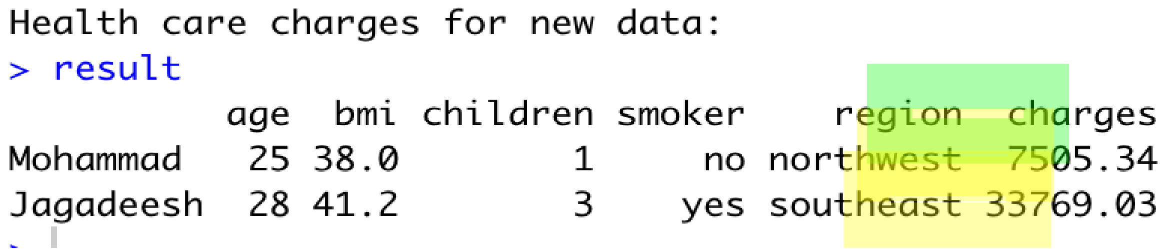
```
Evaluation Metrics for all the Models:
> result
                           Res Std Error R-squared      MAE     RMSE
MultiLinear Regression 1       6090.272 0.7518049 4229.289 5979.289
MultiLinear Regression 2       6087.410 0.7518049 4229.348 5979.329
MultiLinear Regression 3       6090.729 0.7508336 4267.898 5997.181
```

Comparison of the three implemented models

As per the results snip attached above, by comparing all three RSE, R-Squared and RMSE values, Only Model2 is a close to perfect model than Model1 or Model3. Hence, we are finalizing the use of this model in this paper.

Below you can find a sample test data and the prediction from the model2. We are giving two individual person's health details and the model will be generating the insurance charges for them, by going though the loads of data, which we already fed to the system for training.

1. Mohammad: 25 years old, with BMI 38.0, has 1 child, he doesn't smokes, from northwest region.
2. Jagadeesh: 28 years old, with BMI 41.2, has 3 children, he smokes, from southeast region.



```
Health care charges for new data:
> result
          age  bmi children smoker    region  charges
Mohammad   25 38.0        1     no northwest  7505.34
Jagadeesh  28 41.2        3    yes southeast 33769.03
```

Prediction of data using the developed model2

Now you can see that, as per the given factors for Jagadeesh and mohammad, the charges have been calculated and the charge for Jagadeesh is high due to the smoking habit, BMI and age.

6. Conclusion

The manual calculation of the monthly fixed premium cost is very difficult and time taking as there are too many things to be considered and due to too many factors in considerations, there can be error and incorrect amounts be released. The proposed idea in this paper takes simple and multi regression models to calculate the best possible premium, without any human interaction, thus saving time for the insurance company. A sample of set of outcomes are also attached above in the paper in the results section for reference. As per the current model, we are using regression model for our algo, in future, we can also use the hierarchical decision trees and vector machines for getting the most accurate results. Data sets that we used are playing a crucial role, because by using these we are training our schema. In future, we can add more records for our dataset and correct some existing data for better training of the schema, so that when it is deployed there are more accurate and precise results. There is even scope for deploying this program in cloud platform and outsource as Software-as-a-service(Saas) for the insurance companies, thus resulting in costs for insurance companies as well, without going for individual development and maintenance of such models.

References

1. Sudhir Panda, Biswajit Purkayastha, Dolly Das, Manomita Chakraborty, Saroj Kumar Biswas "Health Insurance Cost Prediction Using Regression Models", 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON), 26-27 May 2022

2. Lahiri B, Agarwal N. "Predicting healthcare expenditure increase for an individual from Medicare data". Proceedings of the ACM SIGKDD Workshop on Health Informatics, 2014.

3. B. Nithya and V. Ilango, "Predictive analytics in health care using machine learning tools and techniques," in *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 492–499, IEEE, Madurai, India, June 2017.
4. A. Tike and S. Tavarageri, "A medical price prediction system using hierarchical decision trees," in *Proceedings of the IEEE International Conference on Big Data (Big Data)*, pp. 3904–3913, IEEE, Boston, MA, USA, December 2017.
5. S. Sushmita, S. Newman, J. Marquardt, P. Ram, V. Prasad, M. D. Cock, A. Teredesai et al., "Population cost prediction on public healthcare datasets," in Proceedings of the 5th International Conference on Digital Health 2015. ACM, 2015, pp. 87–94. Association for Computing Machinery, New York, NY, USA, 87–94.
6. https://www.kaggle.com/datasets/mirichoi0218/insurance

Contributions

The research paper “Price Prediction for Health Care Insurance” has been contributed equally by both Mohammad and Jagadeesh by constantly working together attending zoom meetings in building the project from scratch. End to end orchestration has been done by both the team embers. Although, each of the member has his own individual contribution towards the project, but both of them work on their individual task and review each other task so that we are on the same page after every task assigned to each.

We would like to thank Prof. Rezwana Rafiq for teaching us the subject which laid great foundation for developing the project as well as providing the valuable insights for the outcome. Additionally, it’s a privilege to be a part of this class in enhancing the skillset towards role of R language in Data Science for getting insights from other teams. Also I would like to thank peer students which helped us their valuable feedback during the presentation in building better solutions.

Mohammad: Mostly took the ownership in finalizing the project by going through the research papers and deep research has been done on the datasets from the kaggle matching as per the requirement Major focus is on creating r file where he performed Data Processing(ETL), Exploratory Data Analysis and model building. Faced multiple challenges in building the model as well as visualization which required different libraries. Installed packages with latest versions and imported those libraries which helped in model building and plotting.

Jagadeesh: Mostly took ownership for documentation along with model building. Deep research has been done on the problem statement by reading multiple papers. . As per the Professor suggestions, both Mohammad and Jagadeesh look for the various datasets and finalized the dataset which is suitable for the research requirement. Involved in design and architecture of the research. Major focus on presentation and report Also worked on the r file for model building using simple regression and multi linear regression. Worked on the analysis, visualization and interpreting the by creating the sample data and by using the model, predicted the price for Health Care Insurance Sample data.