22/04/2021

Prediction of Insurance Premium Charges

Athulya Shanty
2048030
2MDS
CHRIST UNIVERSITY

TITLE: PREDICTION OF INSURANCE PREMIUM CHARGES.

DOMAIN:

The chosen domain is Insurance. Insurance is a contract between two parties such as the insurance company and the person seeking insurance, where the insurer agrees to hedge the risk of the insured against some specified future events or losses, in return for a regular payment from the insured as premium.

OBJECTIVE:

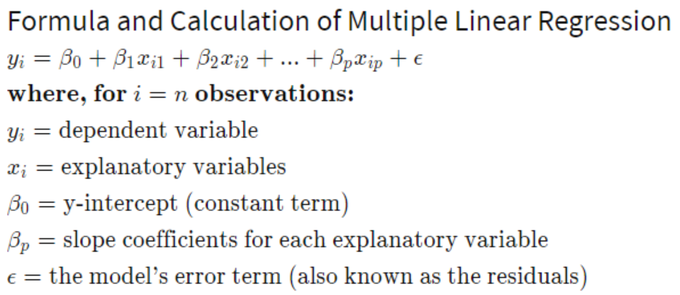
To predict the insurance premium charges charged by an insurance company using R programming.

SOLUTION:

Build a model with at most accuracy to predict the premium charges.

METHOD/TECHNIQUE

The method used is Multiple Linear regression Model. Linear regression follows the linear mathematical model for determining the value of one dependent variable from value of one/more given independent variable(s).



Formula and Calculation of Multiple Linear Regression

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$

where, for $i = n$ observations:

$y_i =$ dependent variable

$x_i =$ explanatory variables

$\beta_0 =$ y-intercept (constant term)

$\beta_p =$ slope coefficients for each explanatory variable

$\epsilon =$ the model's error term (also known as the residuals)

PROJECT DESCRIPTION

The project aims at building a suitable model for predicting the premium charges charged by the company using R Programming Language. The dataset for this project has been downloaded from Kaggle website.

The initial part of the project aims at the Exploratory Data Analysis (EDA). The dataset was free of null/empty values. EDA includes finding the structure of the dataset which in-return showed the number of numerical and categorical variables. Histograms, density plots, Boxplots were plotted for the numerical variables. The number of unique values of the categorical variable has been found. The categorical variables were factorized for ease of further analysis. The plotted Correlation matrix gave the relationship between the variables.

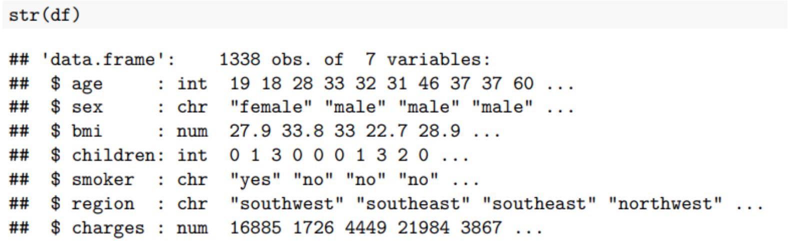
The second half of the project aims at the modelling and prediction. The dataset was divided into training and testing data for modelling. The relevant variables were selected by Random Forest Classifier for the linear regression model. All the assumptions of the Linear Regression was satisfied. The data was then modelled and that model was used to predict the charges.

DATASET

The chosen dataset was downloaded from the website https://www.kaggle.com/teertha/ushealthinsurancedataset. The dataset contains 1338 rows of data, where the Insurance charges are given against the Age, Sex, BMI, Number of Children, Smoker and Region. The variables are a mix of numeric and categorical variables. There are no missing or empty values in the dataset.

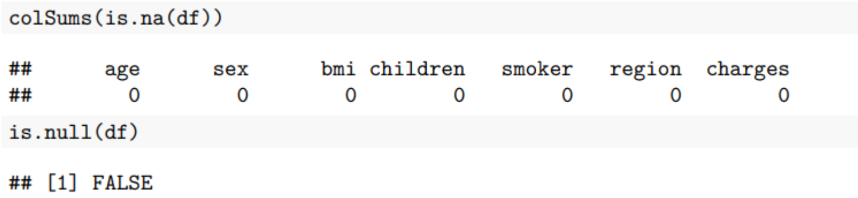
EXPLORATORY DATA ANALYSIS

- Structure of the dataset: The dataset contains 1338 observations of 7 variables. There were 3 categorical and 4 numerical variables.



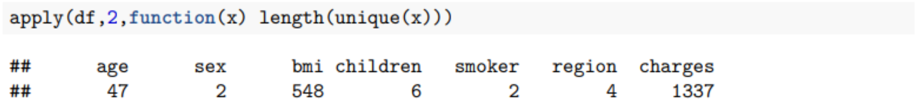
```
str(df)

## '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 ...
```

- Missing Values: The dataset was free of missing and empty values.



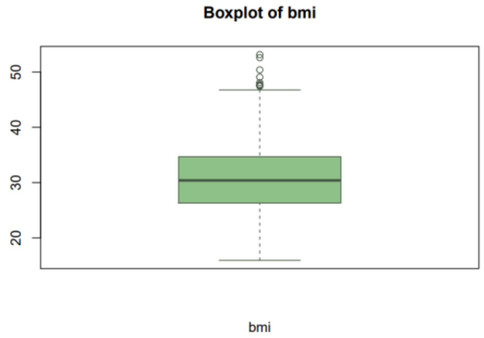
```
colSums(is.na(df))

##      age      sex      bmi children   smoker   region  charges 
##        0        0        0        0        0        0        0

is.null(df)

## [1] FALSE
```

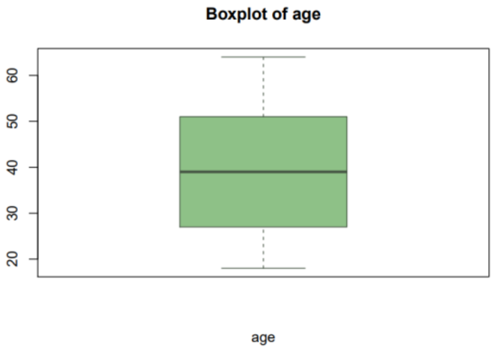
- Unique values: The unique values in each column has been displayed.

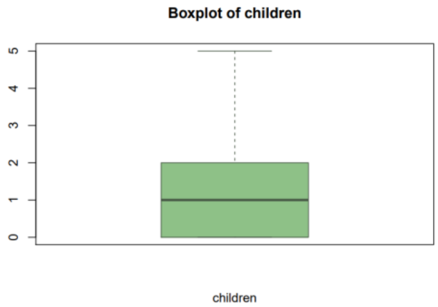


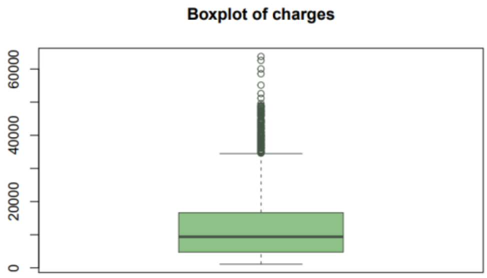
```
apply(df,2,function(x) length(unique(x)))

##      age      sex      bmi children   smoker   region  charges 
##       47        2      548        6        2        4     1337
```

- Handling Outliers: There were some outliers observed in ‘bmi’ variable. I am not removing the outliers here as there might be customers who have a bmi of above 45 and the charges may significantly affect on bmi of the person. There were outliers in ‘charges’ variable but those can’t be removed from the dataset as there will be customers who have to pay a huge amount. All other variables was free of outliers.



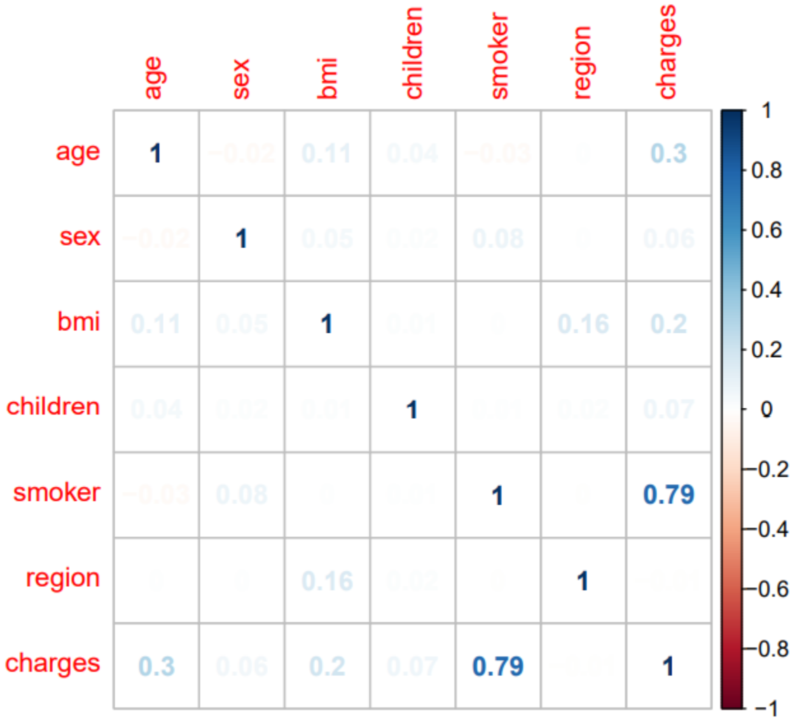






- <u>Correlation :</u> The correlation between the variables have been found.

```
df$sex=as.numeric(df$sex)
df$smoker=as.numeric(df$smoker)
df$region=as.numeric(df$region)
corrplot(cor(df),method = 'number')
```



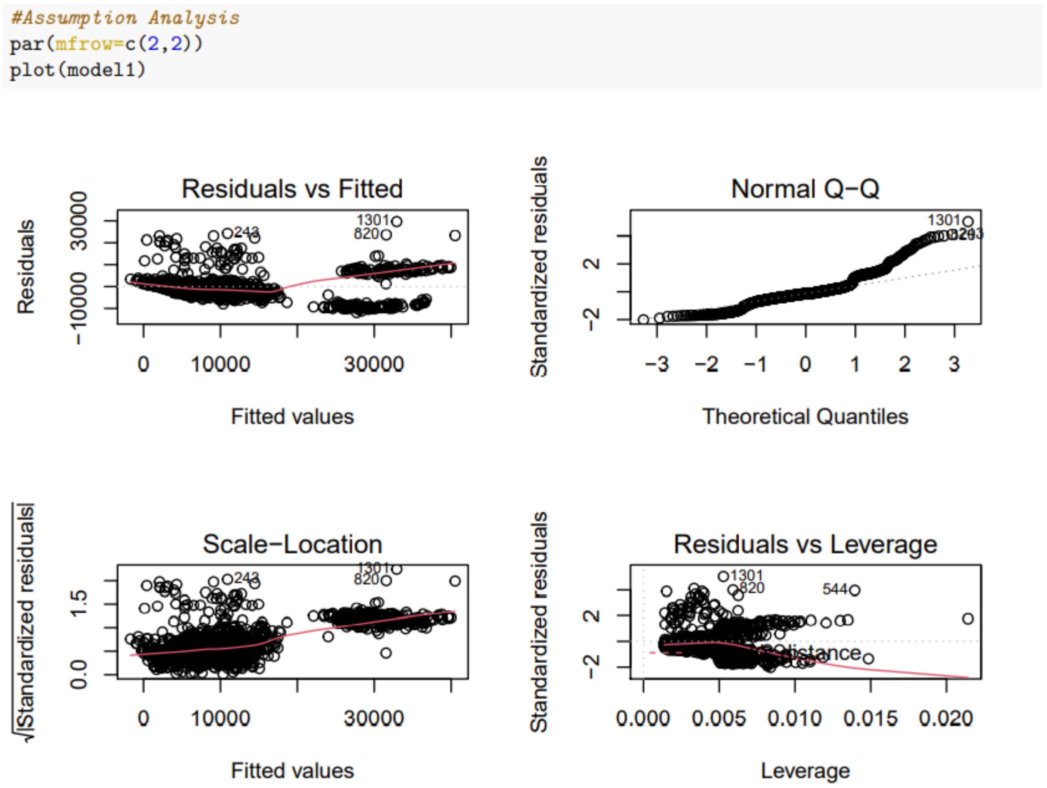
- <u>Assumptions:</u> The assumptions for multiple regression analysis were checked and all the assumptions were satisfied.

  1.Residuals vs fitted graph is used to check the linear relationship assumptions. There are no distinct pattern, therefore, this is an indication for linear relationship

  2. Normal Q-Q plot is used to determine whether the residuals are normally distributed. We can see that the residuals points donot follow straight line. Therefore, the residuals are not normally distributed

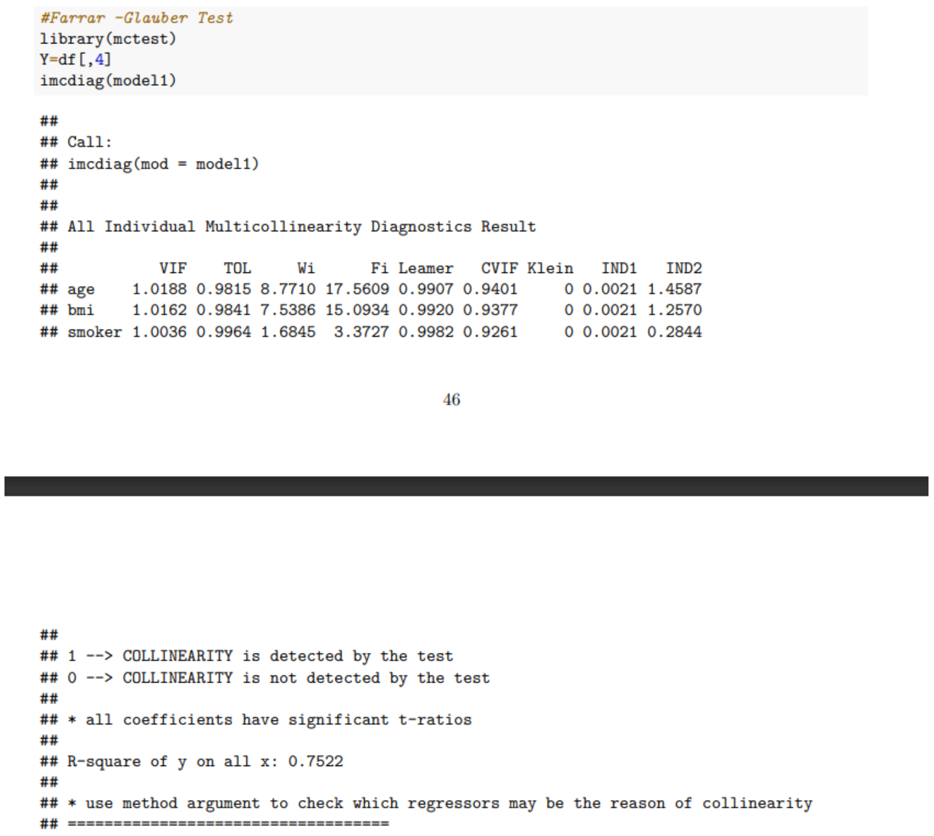
  3. Scale-location(or spread-location) is homogeneity of variance(homoscedasticity) determination

  4. Residuals vs Leverage is for identifying influential points, which are points that might impact regression results when included or excluded from the analysis



```
#Assumption Analysis
par(mfrow=c(2,2))
plot(model1)
```

-
- <u>Multicollinearity:</u> There is no multicollinearity while observing the correlation plot. Also, There is no multi-collinearity between the independent variables from Farrar -Glauber test.



```
#Farrar -Glauber Test
library(mctest)
Y=df[,4]
imcdiag(model1)
```

```
##
## Call:
## imcdiag(mod = model1)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##           VIF    TOL     Wi      Fi Leamer   CVIF Klein   IND1   IND2
## age    1.0188 0.9815 8.7710 17.5609 0.9907 0.9401     0 0.0021 1.4587
## bmi    1.0162 0.9841 7.5386 15.0934 0.9920 0.9377     0 0.0021 1.2570
## smoker 1.0036 0.9964 1.6845  3.3727 0.9982 0.9261     0 0.0021 0.2844
```

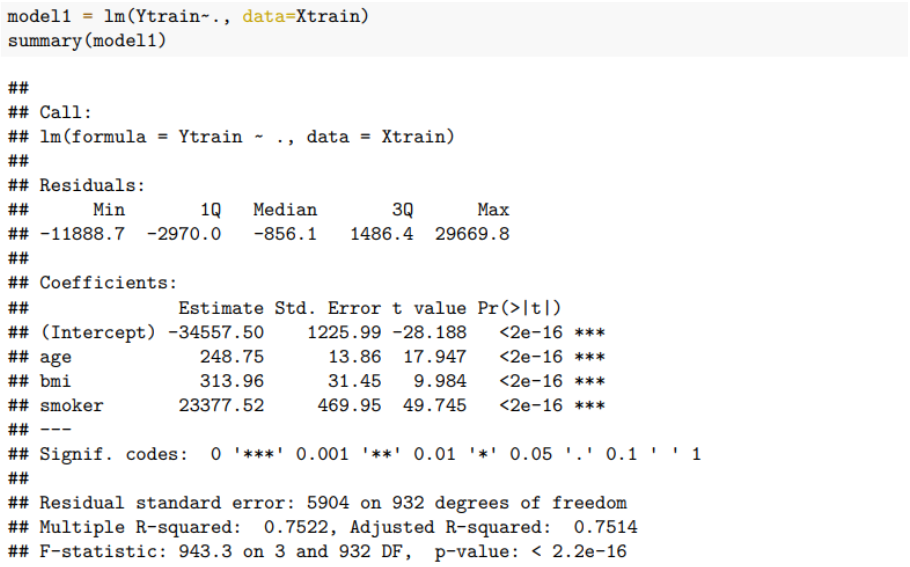
```
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.7522
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

ACCURACY OF THE MODEL

In order to increase the model accuracy, the relevant variables were chosen by Random Forest Classifier which indicated that ‘smoker’,’age’ and ‘bmi’ are the important factors. The model was modelled using these important variables and the model is charges=-34557.50+ 248.75(age)+ 313.96(bmi)+ 23377.53(smoker).

R square(Coefficient of determination) - measures the proportion of variance in the dependent variable that can be explained by the independent variables. The model gave R square value of 0.7522 which indicates that 75.22% of the variation in the dependent variable is explained by the independent variable. Adjusted R-square corrects the positive bias created by the sample, its value is 0.7514 about 75.14%.

F-ratio in the ANOVA table tests whether the overall regression model is a good fit for the data. We have 943.3.The p value is less than 0.05 which indicates that the model is good. We can conclude that all the independent variables statistically significant since every p for each independent variable is < 0.05.

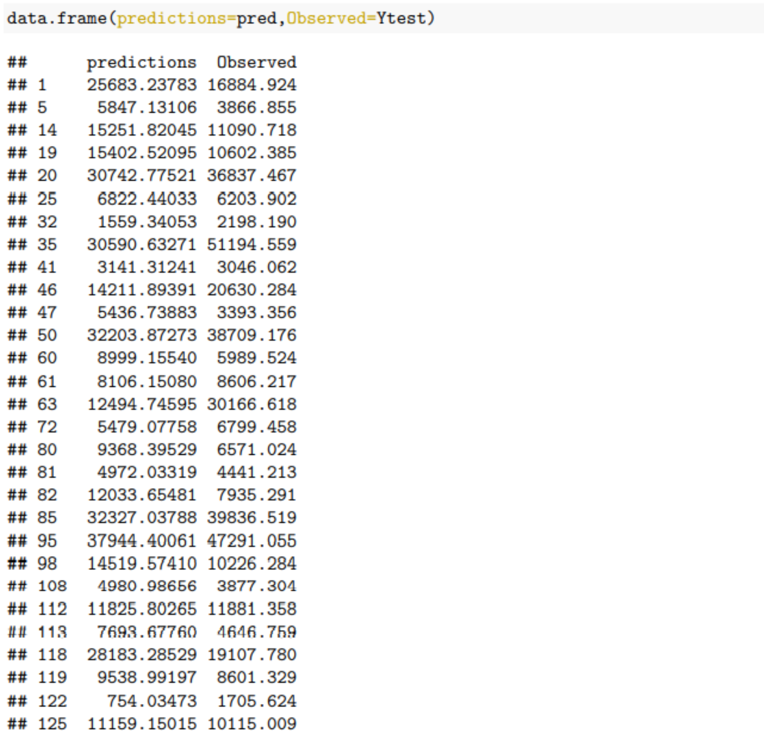


```
model1 = lm(Ytrain~., data=Xtrain)
summary(model1)
```

```
##
## Call:
## lm(formula = Ytrain ~ ., data = Xtrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11888.7  -2970.0   -856.1   1486.4  29669.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34557.50    1225.99 -28.188   <2e-16 ***
## age            248.75      13.86  17.947   <2e-16 ***
## bmi            313.96      31.45   9.984   <2e-16 ***
## smoker       23377.52     469.95  49.745   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5904 on 932 degrees of freedom
## Multiple R-squared:  0.7522, Adjusted R-squared:  0.7514
## F-statistic: 943.3 on 3 and 932 DF,  p-value: < 2.2e-16
```

CONCLUSION

The summary of the model showed that the model is good. The charges are predicted for the training data.



```
data.frame(predictions=pred,Observed=Ytest)
```

```
##      predictions  Observed
## 1    25683.23783 16884.924
## 5     5847.13106  3866.855
## 14   15251.82045 11090.718
## 19   15402.52095 10602.385
## 20   30742.77521 36837.467
## 25    6822.44033  6203.902
## 32    1559.34053  2198.190
## 35   30590.63271 51194.559
## 41    3141.31241  3046.062
## 46   14211.89391 20630.284
## 47    5436.73883  3393.356
## 50   32203.87273 38709.176
## 60    8999.15540  5989.524
## 61    8106.15080  8606.217
## 63   12494.74595 30166.618
## 72    5479.07758  6799.458
## 80    9368.39529  6571.024
## 81    4972.03319  4441.213
## 82   12033.65481  7935.291
## 85   32327.03788 39836.519
## 95   37944.40061 47291.055
## 98   14519.57410 10226.284
## 108   4980.98656  3877.304
## 112  11825.80265 11881.358
## 113   7693.67760  4646.759
## 118  28183.28529 19107.780
## 119   9538.99197  8601.329
## 122    754.03473  1705.624
## 125  11159.15015 10115.009
```

REFERENCES

https://en.wikipedia.org/wiki/Insurance

https://www.kaggle.com/teertha/ ushealthinsurancedataset