Health insurance cost

Authored by Menna Mostafa

Introduction

In the health insurance industry, data analytic is an important process that helps insurers to identify and predict the cost of health insurance which prevents risk and cost for the companies. The Insurance data set contains 1338 observations and 7 variables which include age, sex, children, smoker, region, and charges.

Methods

We used multiple linear regression model which commonly used for cost estimation models, we use it to understand the relationship between variables and predict the value of one variable based on the others.

Data preparation phase

1- Understanding the data

- **Age**: insurance contractor age, years.
- **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, [yes, no].
- **Region**: the beneficiary's residential area in the US, [northeast, southeast, southwest, northwest].
- **Charges**: Individual medical costs billed by health insurance.

2- Structure of the data

```
> # display the structure of the data set
>
> str(insurance)
'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 ...
```

3- Summary of the data

```
C:/Users/moon/Desktop/excell/Data_Analysis/
> summary(insurance)
      age            sex                 bmi           children
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
 Median :39.00   Mode  :character   Median :30.40   Median :1.000
 Mean   :39.21                      Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                      Max.   :53.13   Max.   :5.000
    smoker             region            charges
 Length:1338        Length:1338        Min.   : 1122
 Class :character   Class :character   1st Qu.: 4740
 Mode  :character   Mode  :character   Median : 9382
                                       Mean   :13270
                                       3rd Qu.:16640
                                       Max.   :63770
```

4- Adding dummy columns

As sex, smoker and region are qualitative variables, so we need to convert them to binary variables,

sex -> sexmale and sexfemale
smoker -> smokeryes and smokerno
region -> regionnortheast, regionnorthwest, regionsoutheast and regionsouthwest

	age	sexfemale	sexmale	bmi	children	smokerno	smokeryes	regionnortheast	regionnorthwest	regionsoutheast	regionsouthwest	charges
1	19	1	0	27.9	0	0	1	0	0	0	1	16884.92
2	18	0	1	33.77	1	1	0	0	0	1	0	1725.552
3	28	0	1	33	3	1	0	0	0	1	0	4449.462
4	33	0	1	22.705	0	1	0	0	1	0	0	21984.47
5	32	0	1	28.88	0	1	0	0	1	0	0	3866.855
6	31	1	0	25.74	0	1	0	0	0	1	0	3756.622
7	46	1	0	33.44	1	1	0	0	0	1	0	8240.59
8	37	1	0	27.74	3	1	0	0	1	0	0	7281.506
9	37	0	1	29.83	2	1	0	1	0	0	0	6406.411

When using dummy variables, the number of variables must be one less than the number of categories, for example the sex variable has two categories (sexmale and sexfemale) so that when developing the model we must have only one dummy variable for sex and the same one dummy variable for smoker and three dummy variables for region because we have 4-categories for region variable.

Model planning phase

1- Checking if there is a missing variables in the data set or not

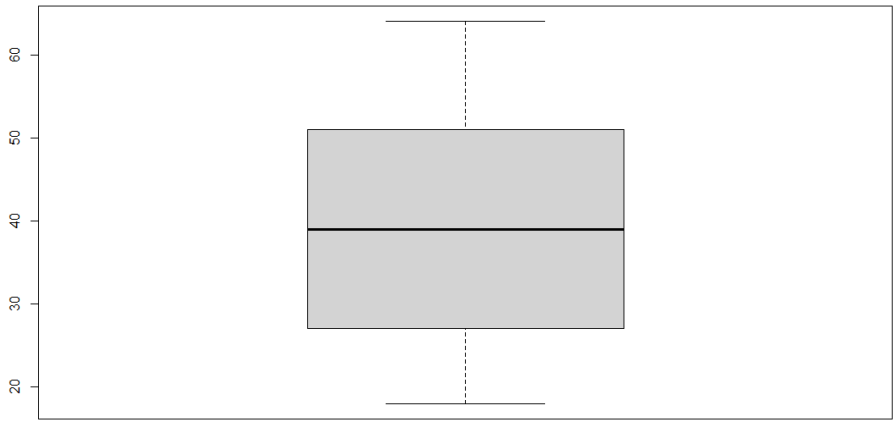
```
> any(is.na(insurance))
[1] FALSE
> |
```

As we see in the figure, no missing values in the dataset.

2- Checking the outliers and removing it if exist

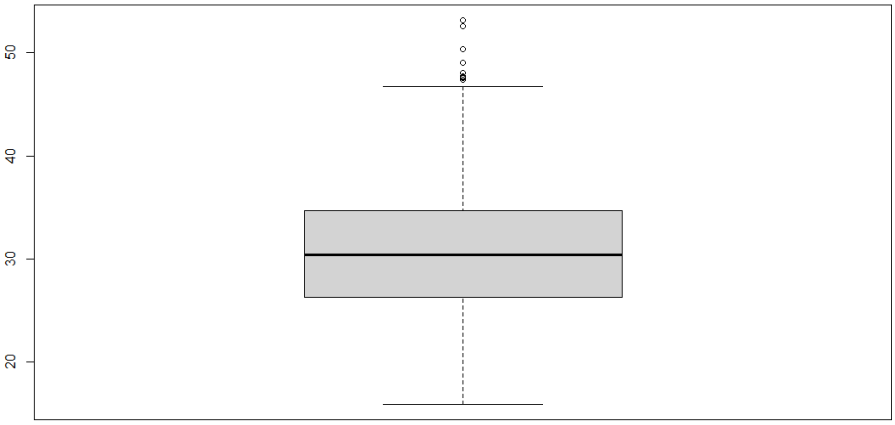
- Checking the boxplot of age and bmi to check the outliers.
- Removing the outliers if exists.

The boxplot graph of age



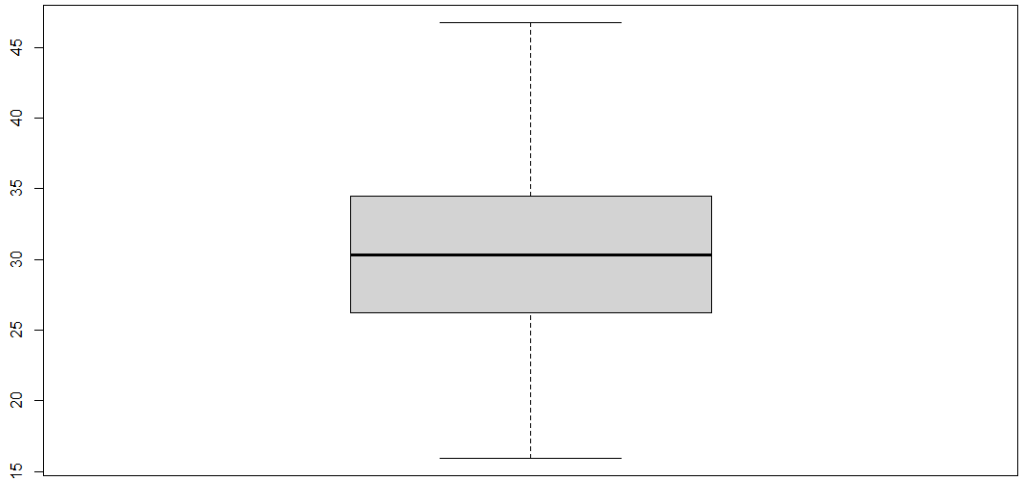
The graph indicates that there is no outliers in age.

The boxplot graph of bmi



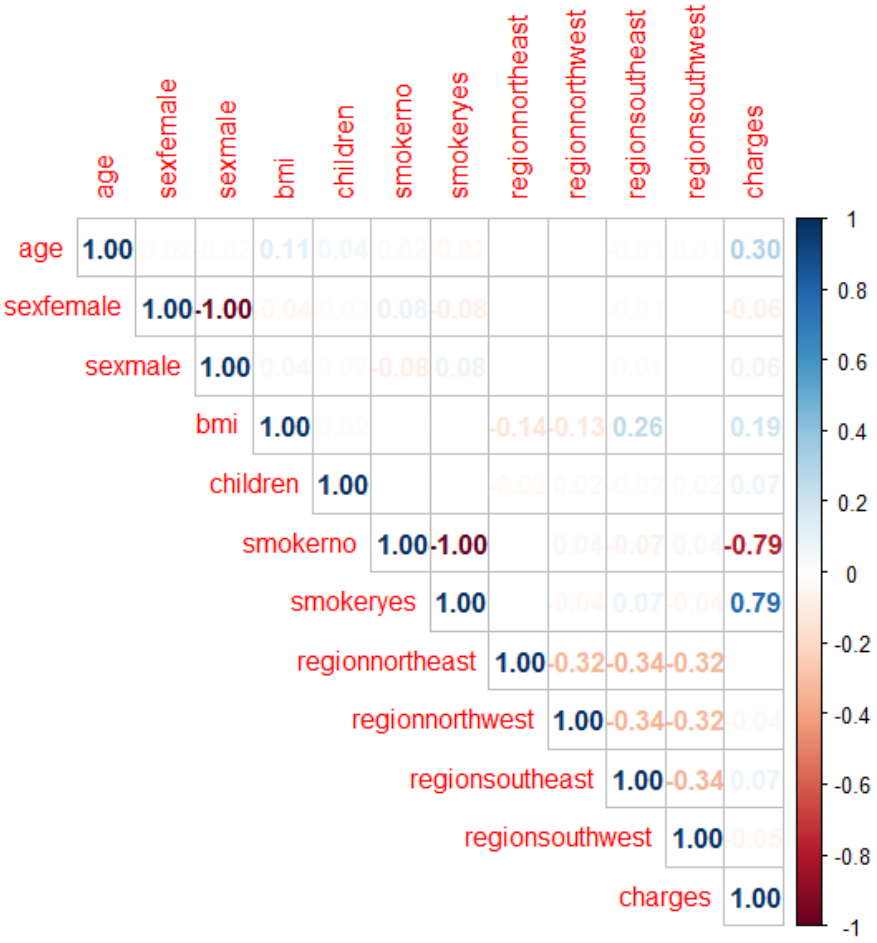
The graph indicates that there is outliers that need to be removed.

The boxplot graph of bmi after removing the outliers



3- Checking the correlation

When an independent variable is correlated with one other independent variable, the variable are said to be collinear. If an independent variable is correlated with a combination of other independent variables, the condition of multicolinearity exists. This can create problems in interpreting the coefficients of the variables as several variables are provided duplicate information. For example, if the information provided in one independent variable is also provided in the other independent variable, then several sets of regression coefficients for these two variables will yield exactly the same results. Thus, individual interpretation for these variables would be questionable, although the model itself is still well for prediction purposes.



As we see in the figure, no strong correlation between independent variables and there is a strong relationship between smoker and charges.

4- Selecting the model

- Multiple linear regression model is selected to research the data set as it is the best choice when there is more than one independent variable used for a prediction of a response variables.

- In this model we used age, bmi and smoker as independents variables and charges as dependent variable.

- The measured used to test the accuracy of the model are Adjusted R square, Statistic F test (P value) and testing the assumptions of the model (Residual plot, Plotting the model and Normality of errors plot).

Model building phase

First, splitting the data into train and test, then building the regression model using the train data set and finally, predicting the output from regression model using test data set.

We building two regression models, first one we used the charges as dependent variable and age, sexmale, bmi, smokeryes, children, regionnortheast, regionnorthwest, regionsoutheast as independent variables, then we building the second regression model with charges as dependent variable and age, bmi and smoker as independent variables.

1- This is the summary of the two models :

```
>
> summary(regression_model1)

Call:
lm(formula = charges ~ age + sexmale + bmi + children + smokeryes +
    regionnortheast + regionnorthwest + regionsoutheast, data = train_set)

Residuals:
   Min     1Q Median     3Q    Max
-10859  -3038  -1099   1266  30230

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -12767.82    1168.68 -10.925  < 2e-16 ***
age                252.51      13.69  18.439  < 2e-16 ***
sexmale           -269.47     380.98  -0.707  0.47953
bmi                343.05      33.12  10.357  < 2e-16 ***
children           456.67     157.09   2.907  0.00373 **
smokeryes        23648.04     480.08  49.259  < 2e-16 ***
regionnortheast   1057.21     545.11   1.939  0.05272 .
regionnorthwest   1097.86     542.94   2.022  0.04342 *
regionsoutheast    -25.90     537.46  -0.048  0.96158
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6168 on 1054 degrees of freedom
Multiple R-squared:  0.7325,    Adjusted R-squared:  0.7304
F-statistic: 360.7 on 8 and 1054 DF,  p-value: < 2.2e-16
```

```
> #Analyzing  the second model
>
> summary(regression_model2)

Call:
lm(formula = charges ~ age + bmi + smokeryes, data = train_set)

Residuals:
   Min     1Q Median     3Q    Max
-11676  -3124  -1106   1308  29062

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11348.70    1078.64  -10.52   <2e-16 ***
age            255.77      13.73   18.63   <2e-16 ***
bmi            321.63      31.91   10.08   <2e-16 ***
smokeryes    23604.89     478.48   49.33   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6201 on 1059 degrees of freedom
Multiple R-squared:  0.7283,    Adjusted R-squared:  0.7275
F-statistic: 946.1 on 3 and 1059 DF,  p-value: < 2.2e-16
```

The adjusted R square value provides a measure of how well the model explains the values of the dependent variable.

From the two figures we can conclude that :

1- Adjusted R square for first model = .7304
2- Adjusted R square for second model = .7275
3- P-vale of the two models is very small

As the performance between the two models is quite similar, we will keep the second model as it is simpler.

2- Model prediction

- From the summary of the second regression model, the model equation is: **Y = -11348.70 + 255.77 X1 + 321.63 X2 + 23604.89X3**
  Where X1 -> age, X2 -> bmi and X3 -> smokeryes.

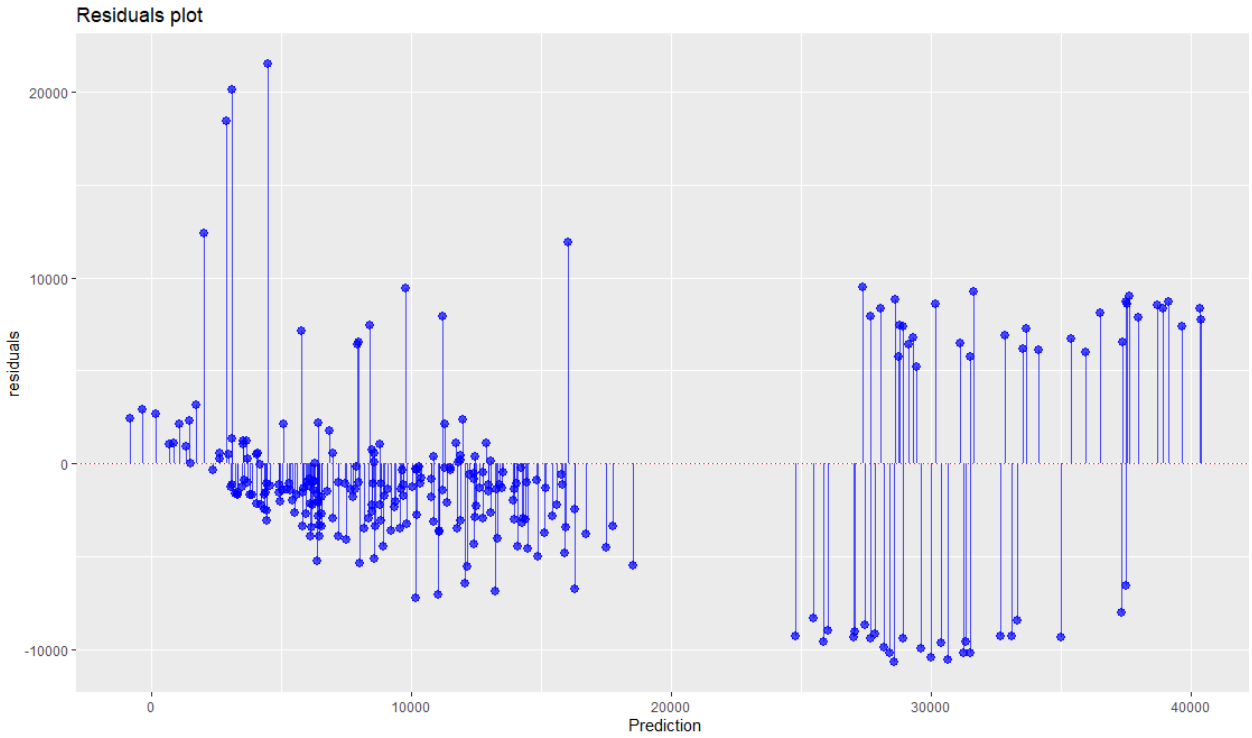
- The equation indicates that the person who smokes will cost **23604.89** more than about the person who does not smoke.

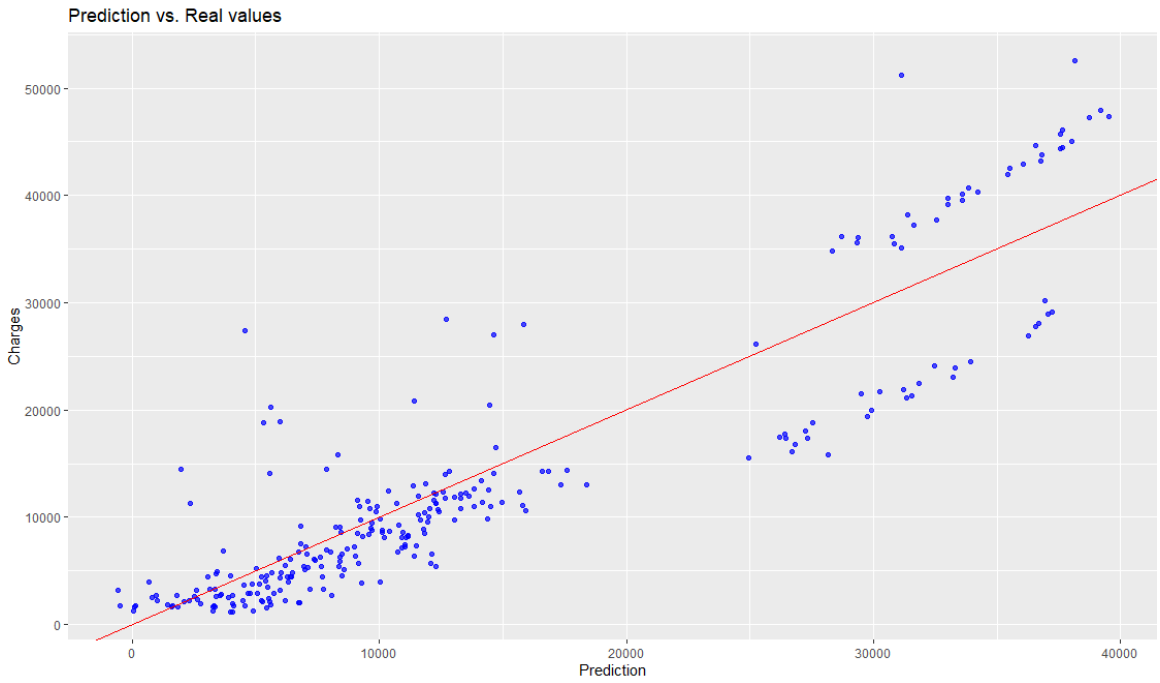
- The prediction can done using this equation by substituting of age, bmi and smoker in the equation, also the prediction can done by using predict function in R as follow :

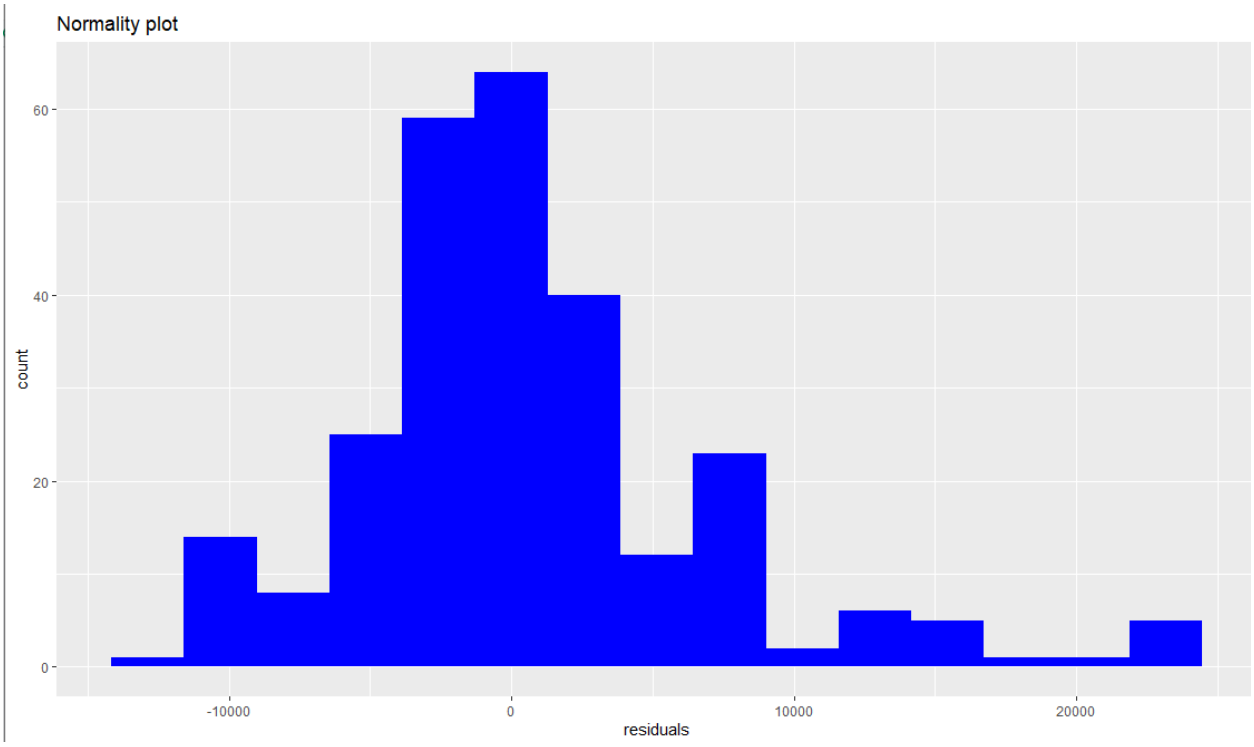
  Prediction <- predict(Our Regression Model , Test data)

3- Checking the performance of the second model

3.1- Checking the assumptions of the model




Prediction vs. Real values
Charges
Prediction
0
10000
20000
30000
40000
50000


Normality plot
count
residuals
0
20
40
60
-10000
0
10000
20000

- In residuals plot graph as we see in the figure, errors seem random, no discernible pattern is present and errors have a constant variance.

- In prediction vs. Real values graph as we see in the figure, the graph appears linear.

- In Normality plot as we see in the figure, the errors are normally distributed.

3.2- Discuss the accuracy of the model

- Adjusted R Square = 0.73 indicating that about 73% of the variability in charges can be explained by the regression model with age, bmi and smoker or at least one of them as the independent variables.

- The P-value of overall model is very small indicating that, the model is statistically significant. Thus, there is sufficient evidence in the data to conclude that the model is useful, and there is a relationship between charges and age, bmi and smoker or at least one of them as the independent variables.