ILLINOIS INSTITUTE OF TECHNOLOGY

Medical Insurance Payout Prediction

A20516879 - Zainab Hasnain - zhasnain1@hawk.iit.edu
A20501739 - Santosh Reddy Edulapalle - sedulapalle@hawk.iit.edu

MATH-564

12-02-2022

Percentage contribution:
Zainab Hasnain: 50%
Santosh Reddy Edulapalle: 50%

Table of Contents

Abstract	**3**
Introduction	**3**
Problem Statement	**3**
Dataset Used	**4**
Descriptive Analysis and Statistical Summary	**5**
Methodology	**6**
Correlation Between Numerical Variables	**6**
Dummy Variables for Categorical Variables	**6**
Plotting the Target Variable	**7**
Models, Results and Analysis	**8**
1. Linear Regression Model	**8**
2. Generalized Linear Model	**9**
Conclusion	**10**
Appendix	**11**
References	**17**

Abstract

This project is about predicting how much insurance charges a person pays on average. The goal of this project is to which predictor variables have the most impact on the target variable, medical insurance charges. The model takes several information about the person as the input, such as age, sex, BMI, number of children, whether the person smokes, and the region a person belongs to, and outputs the amount of charges paid by that person. We perform exploratory data analysis first, preprocess the data, and then finally split the data for training and testing. In the process, we used multiple models such as linear regression, and generalized linear models to find out the significance of the independent variables. Finally, we compare different metrics to conclude the most important predictor variable.

Introduction

The healthcare industry in the USA is massive and is continually growing with the increasing demands of the population. Many companies and healthcare providers aim to maximize the satisfaction of their patients by providing better quality care. The marketplace has begun to address the patient's full health journey, leading to improved affordability, quality, access, and experience. Healthcare providers and payers alike are looking for not just better quality of care, but also focusing on business decisions to increase profitability. In this project, we will try to find out what are the factors which influence an individual's insurance payout. We will analyze the personal information of a person, interpret the relationship between the features, and find out which independent variable best predicts the target variable, insurance charges.

The motivation to do this project is to find out which independent variable has the most impact on the regression problem where the input is a combination of numerical and categorical variables. The output is a continuous variable. For this purpose, we are using the medical insurance dataset. To achieve our goals, we begin by understanding the data. We explore the data type of the variables and perform exploratory data analysis. We do correlation analysis on the numerical variables, convert the categorical variables to dummy variables, and apply two different models to study the coefficients and the significance of the predictors.

Problem Statement

This project is about Medical Insurance Payout Prediction. The goal of this project is to train a model to determine the annual medical charges incurred by an individual. The model takes several information about the person as the input, and outputs the amount of charges paid by that person. We want to determine which predictor variables are the most significant to predict the charges incurred by over 1300 customers, given their age, sex, BMI and several other attributes.

Dataset Used

The dataset used here is a Kaggle dataset, which is saved in the form of a CSV file. It contains a total of 7 features. Out of those 7 features, there are 3 categorical variables.

Input Variables:

1. Age (integer)
2. Sex (factor)
3. BMI (decimal number)
4. Number of children (integer)
5. Smoker
6. Region
7. Charges

Categorical variables:

1. Sex (male or female)
2. Smoker (yes or no)
3. Region (northwest, southwest, northeast, or southeast)

Target variable: Charges

Since our target variable is a numerical output, this is essentially a regression problem. However, we need to deal with the categorical variables as well, which we will discuss later in this report.

This is what the sample data looks like:

	age <int>	sex <fctr>	bmi <dbl>	children <int>	smoker <fctr>	region <fctr>	charges <dbl>
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
6	31	female	25.740	0	no	southeast	3756.622

This model is a regression problem where the inputs contain the person's personal data, such as age, sex, Body Mass Index, number of children, whether the person smokes, and the region the person is from. The class label or the dependent variable in this case is the medical charges paid by the person. The dataset contains 1338 records, and 7 columns. Therefore, the dimension of the dataset is 1338 by 7.

Descriptive Analysis and Statistical Summary

```
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 1 1 2 1 ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ region  : Factor w/ 4 levels "northeast","northwest",..: 4 3 3 2 2 3 3 2 1 2 ...
 $ charges : num  16885 1726 4449 21984 3867 ...
      age            sex           bmi           children     smoker           region
 Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064   northeast:324
 1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274   northwest:325
 Median :39.00                Median :30.40   Median :1.000              southeast:364
 Mean   :39.21                Mean   :30.66   Mean   :1.095              southwest:325
 3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                Max.   :53.13   Max.   :5.000
    charges
 Min.   : 1122
 1st Qu.: 4740
 Median : 9382
 Mean   :13270
 3rd Qu.:16640
 Max.   :63770
```

Let us first take a look at the statistics summary of the numerical variables.

- There are a total of 1338 records and 7 variables.
- The datatype of the variable ‘age’ is integer. The age of the pupils under consideration ranges from 18 to 64, with a mean of 39.2.
- The variable ‘sex’ is a factor with two levels, male and female and is relatively equally distributed between both the levels.
- The variable ‘bmi’ is a decimal number and it ranges from almost 16 to 53, with an average of 30.7.
- The variable ‘children’ is an integer value and it defines the number of children a person has. The individuals in this study either don’t have any children, or have up to 5 children. On average, they have 1 child.
- The variable ‘smoker’ is a factor with two levels, ‘yes’ or ‘no’. Out of 1388, there are 274 smokers and 1064 non-smokers.
- The variable ‘region’ contains 4 levels and is equally dispersed among all categories.
- Finally, the charges, the target variable, is in thousands, as expected from this study. The minimum insurance payout is about 1122 and the maximum is 63770. The average charges are 13270.

Methodology

To perform the task of predicting which variable is the most significant for predicting the insurance charges paid by a person, we follow the below mentioned steps:

1. First, use all the numerical variables and find the correlation coefficient between them to see if multicollinearity exists.
2. Next, use the most highly correlated variable with the target variable, and plot the graph.
3. Then, color code the graph by using the categorical variable as the color map. This will help us identify the pattern in data between different groups.
4. Now, we will use sample.split() function from the caTools library to split our data. We use 805 data (80%) for training the model and 20% for testing.
5. Then, after splitting the data, perform a linear regression model and a generalized linear model and observe the metrics to see if the models perform well.
6. Finally, summarize the findings and conclude which factor is most significant in predicting insurance charges.

Correlation Between Numerical Variables

	age	bmi	children	charges
age	1.0000000	0.1092719	0.04246900	0.29900819
bmi	0.1092719	1.0000000	0.01275890	0.19834097
children	0.0424690	0.0127589	1.00000000	0.06799823
charges	0.2990082	0.1983410	0.06799823	1.00000000

The correlation coefficient between the independent variables must be very small, to avoid multicollinearity. Furthermore, the higher the correlation coefficient between a predictor variable and the target variable, the more significant the predictor variable is to predict the output. Here, we can see that age is relatively most highly correlated with charges. However, the correlation is a weak one. Therefore, we must include the categorical variables and the model must be a combination of these predictors as just any one variable is not strongly correlated with the output variable.

Dummy Variables for Categorical Variables

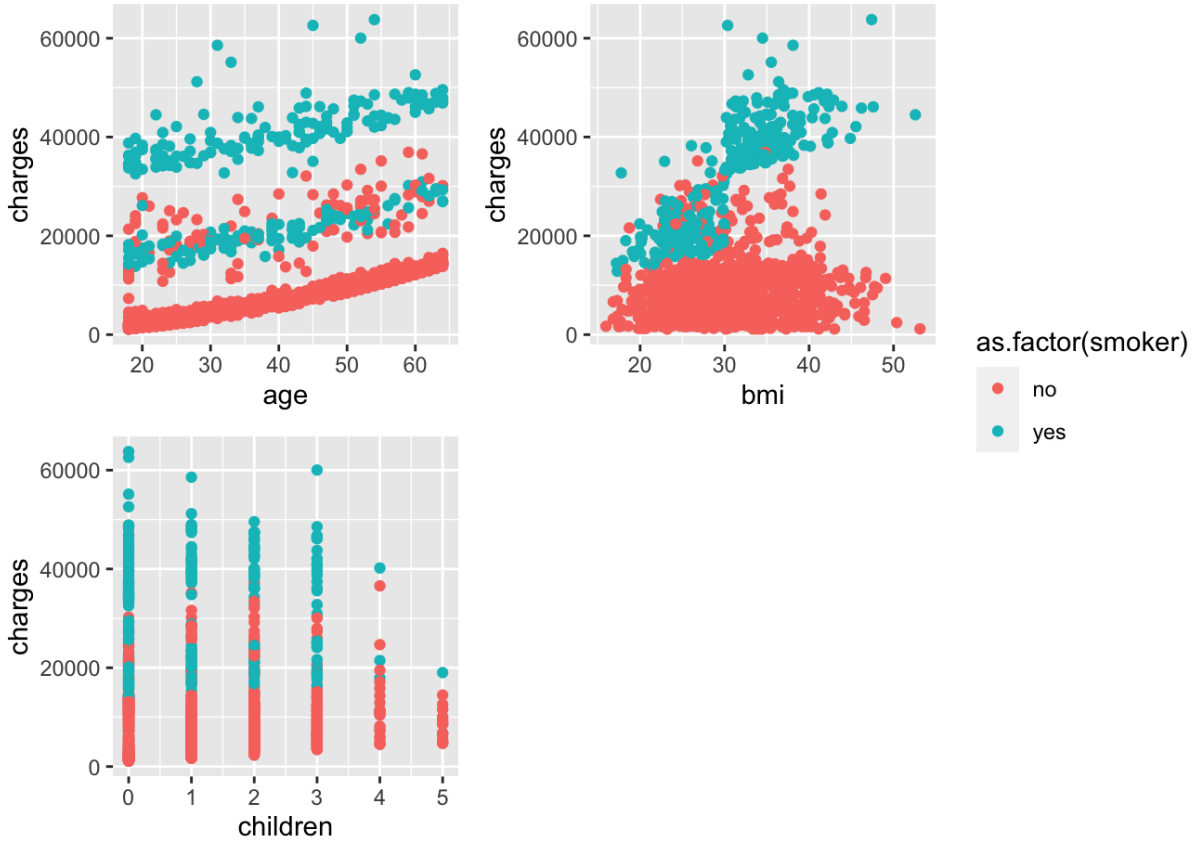
In order to do numerical analysis and to include the categorical variables in our model, we introduce dummy variables. If a variable has *n* levels, we will use *n-1* dummy variables.

In this case, we will have the following variables:

1. Sex: 2 levels, 1 dummy variable
2. Smoker: 2 levels, 1 dummy variable
3. Region: 4 levels, 3 dummy variables

In total, we will have 5 dummy variables and 3 numerical variables. Therefore, in our models, we will have a total of 8 predictor variables.

Plotting the Target Variable



By plotting these graphs and color-coding with the categorical variable smoker, we realize how significant the smoker variable is. When charges are plotted against any of the numerical independent variables, smokers (blue dots) are consistently above the non-smoker (pink dots).

Furthermore, there are some key observations which can be drawn from the above plots:

1. The charges increase with age and it shows a linear relationship. There are multiple lines showing that there are other variables which affect the charges.
2. The BMI plot against charges is divided into clusters. There is an obvious pink cluster and two blue clusters. Again, this shows there is a certain pattern in data and any 1 variable alone cannot predict charges adequately.
3. It is interesting to note that the insurance charges a person pays decreases as the number of children increase. Another thing to note here is that as the number of children

increases, the ratio of the smoker population decreases. There are very few blue dots when children are 4 or 5.

Finally, we can see from fig 4 and 5 in the Appendix section that sex and region are not as significant as the smoker variable. The colored dots corresponding to different categories of sex/region are scattered and intermixed when the charges are plotted.

Models, Results and Analysis

1. Linear Regression Model

We fit a linear regression model by taking into account all the variables, regardless of its correlation or significance to the target variable. The purpose of implementing this model is not just to accurately predict the charges, but to find out which variable has the highest influence on the target variable. For that purpose, we will look at the coefficients (beta values), p-values, and the t-values for each of the variables.

```
Residuals:
   Min     1Q Median     3Q    Max
-11112  -2830  -1014   1183  25515

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11426.76    1113.94 -10.258   <2e-16 ***
age                251.80      13.59  18.530   <2e-16 ***
sexmale            -96.41     377.77  -0.255   0.7986
bmi                332.15      32.74  10.144   <2e-16 ***
children           453.79     156.90   2.892   0.0039 **
smokeryes        23569.67     465.99  50.579   <2e-16 ***
regionnorthwest   -325.77     540.72  -0.602   0.5470
regionsoutheast  -1279.95     548.19  -2.335   0.0197 *
regionsouthwest  -1092.25     539.96  -2.023   0.0433 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6137 on 1061 degrees of freedom
Multiple R-squared:  0.7423,    Adjusted R-squared:  0.7404
F-statistic:   382 on 8 and 1061 DF,  p-value: < 2.2e-16
```

As we can see in the model summary, the magnitudes of the coefficient of the smoker (yes) variables is the highest. However, just the coefficient is not enough to conclude that the variable is the most significant, due to the scale of each variable being different. Therefore, we will look at the t-values, which is a ratio of the coefficient and its standard error, and the magnitude of the p-values. The greater the t-value, the more significant the predictor is to determine the charges. The smaller the p-value, the more significant the variable is to predict the target variable.

Here are some key observations to note from the above LM results:

I. Smoker (yes) dummy variable is the most significant factor to determine the charges paid by an individual. It has the highest t-value. It is significant if we take alpha in the significance level to be 0.05.
II. Age and bmi are also important factors, as their t-value is large and the p-value is the smallest. It is significant if we take alpha in the significance level to be 0.05.
III. Children is also a statistically significant variable if we take alpha to be 0.05. However, it is not the most significant variable.
IV. Overall, smokers have the highest coefficient, the beta value shows that the difference in charges paid by a smoker and non-smoker is estimated to be 23,570.

2. Generalized Linear Model

Similar to the linear regression model, we fit a generalized linear regression model by taking into account all the variables, regardless of its correlation or significance to the target variable. The purpose of implementing this model is not just to accurately predict the charges, but to find out which variable has the highest influence on the target variable. For that purpose, we will look at the coefficients (beta values), p-values, and the t-values for each of the variables.

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-11112   -2830   -1014    1183   25515

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11426.76    1113.94 -10.258   <2e-16 ***
age                251.80      13.59  18.530   <2e-16 ***
sexmale            -96.41     377.77  -0.255   0.7986
bmi                332.15      32.74  10.144   <2e-16 ***
children           453.79     156.90   2.892   0.0039 **
smokeryes        23569.67     465.99  50.579   <2e-16 ***
regionnorthwest   -325.77     540.72  -0.602   0.5470
regionsoutheast  -1279.95     548.19  -2.335   0.0197 *
regionsouthwest  -1092.25     539.96  -2.023   0.0433 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 37662586)

    Null deviance: 1.5506e+11  on 1069  degrees of freedom
Residual deviance: 3.9960e+10  on 1061  degrees of freedom
AIC: 21713
```

As we can see in the model summary, the magnitudes of the coefficient of the smoker (yes) variables is the highest. However, just the coefficient is not enough to conclude that the variable is the most significant, due to the scale of each variable being different. Therefore, we will look at the t-values, which is a ratio of the coefficient and its standard error, and the magnitude of the p-values. The greater the t-value, the more significant the predictor is to determine the charges. The smaller the p-value, the more significant the variable is to predict the target variable.

Here are some key observations to note from the above GLM results:
I. Smoker (yes) dummy variable is the most significant factor to determine the charges paid by an individual. It has the highest t-value. It is significant if we take alpha in the significance level to be 0.05.
II. Age and bmi are also important factors, as their t-value is large and the p-value is the smallest. It is significant if we take alpha in the significance level to be 0.05.
III. Children is also a statistically significant variable if we take alpha to be 0.05. However, it is not the most significant variable.
IV. Overall, smokers have the highest coefficient, the beta value shows that the difference in charges paid by a smoker and non-smoker is estimated to be 23,570.

Conclusion

After analyzing the results of our linear and generalized linear model, we can conclude that both the models give very similar information about the significant variables. Smoker variable has the most significance in both the models.
Moreover, as we can see in the figure 6 and 7 in the appendix section, there is some trend in the data when we plot the predicted against the true values from the test dataset. This can be due to multiple factors, including that we have not taken into account the patient's past health history.
Finally, from our results from the t-values and p-values, we can conclude the smoker is the most significant categorical variable, and age is the most significant numerical variable. We can also infer that from our results of the model in the project rmd file where we find linear regression on filtered dataset with a particular group of a categorical variable. We analyzed that the model coefficients vary greatly between smoker and non-smoker groups, whereas the coefficients do not differ significantly between male, female, and different regional groups.

Appendix

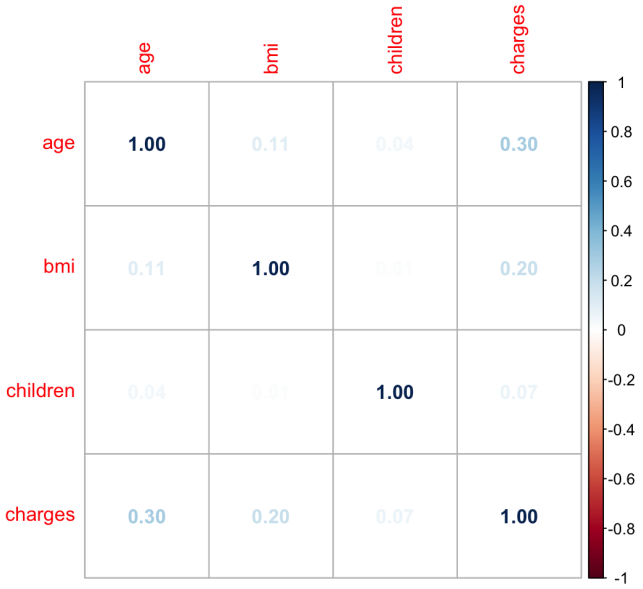
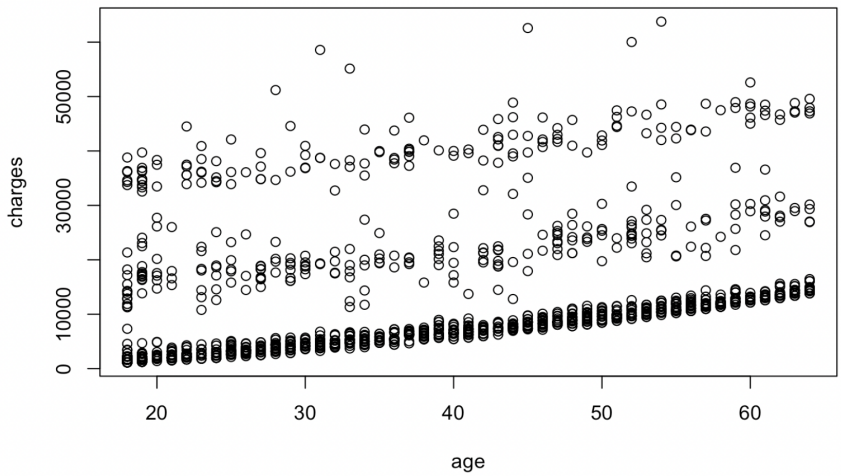
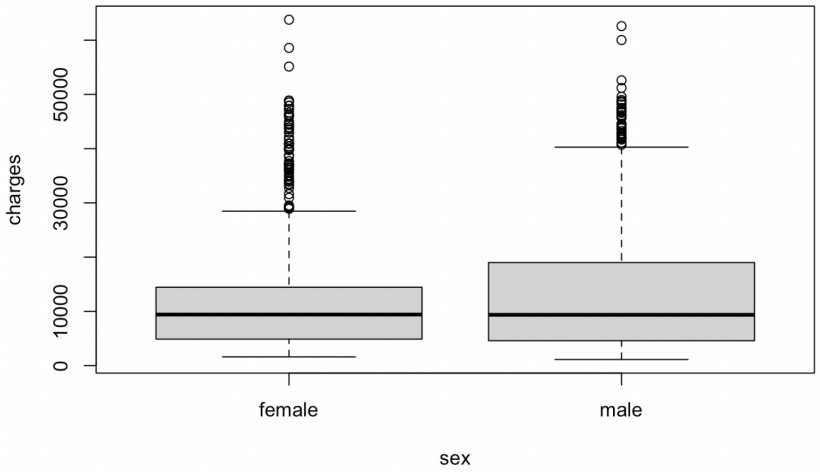
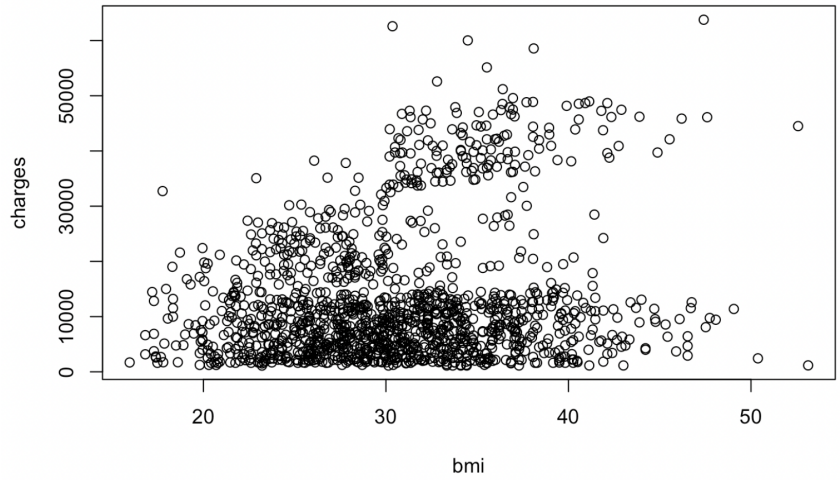
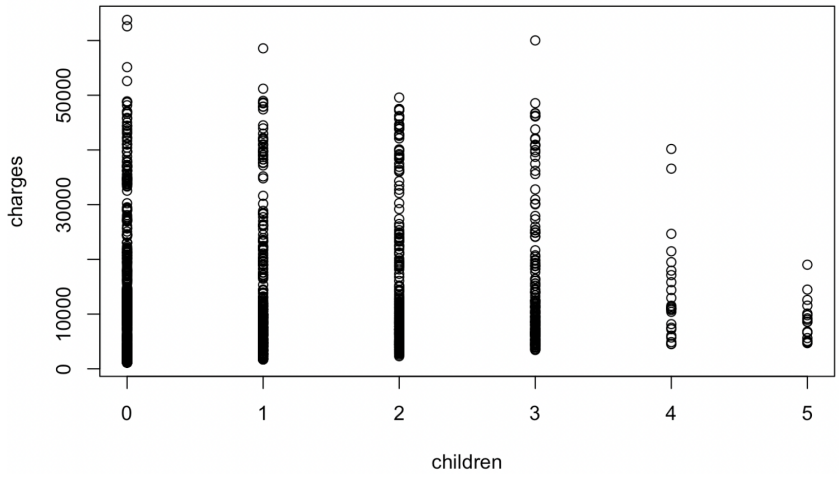


Fig1: Correlation plot between charges, age, bmi, children




charges
0
10000
30000
50000
female
male
sex


charges
0
10000
30000
50000
20
30
40
50
bmi


charges
0
10000
30000
50000
0
1
2
3
4
5
children

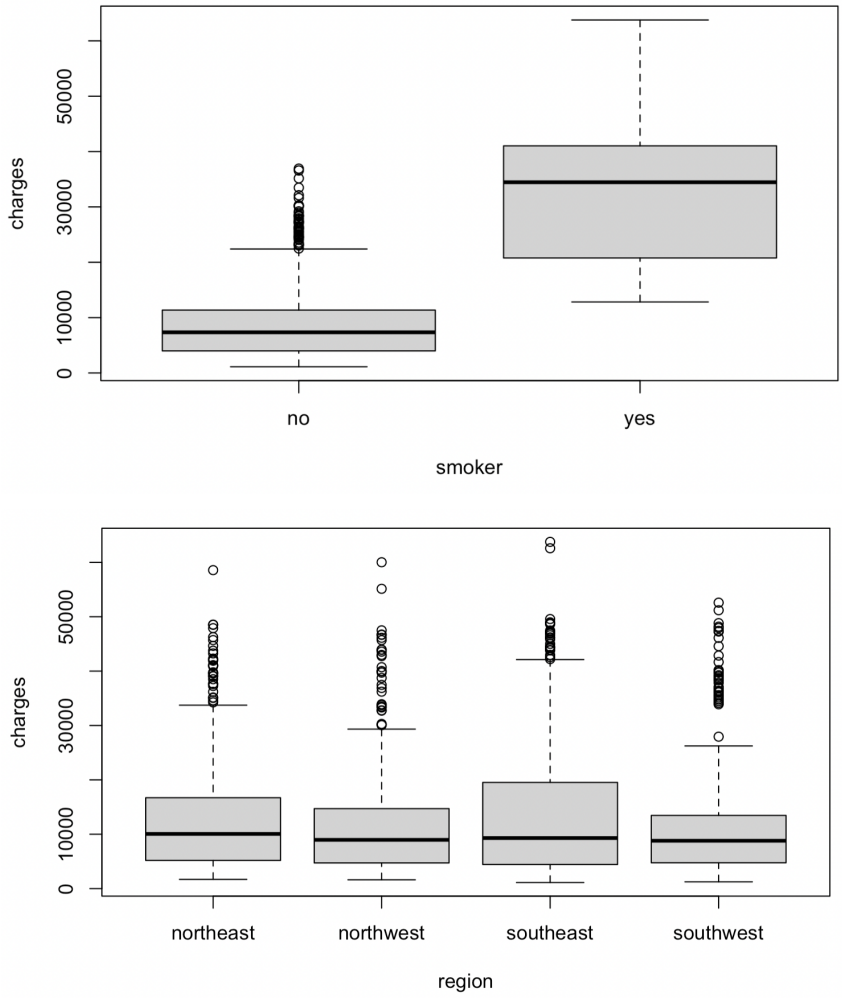


Fig2: Plot of charges vs other variables.

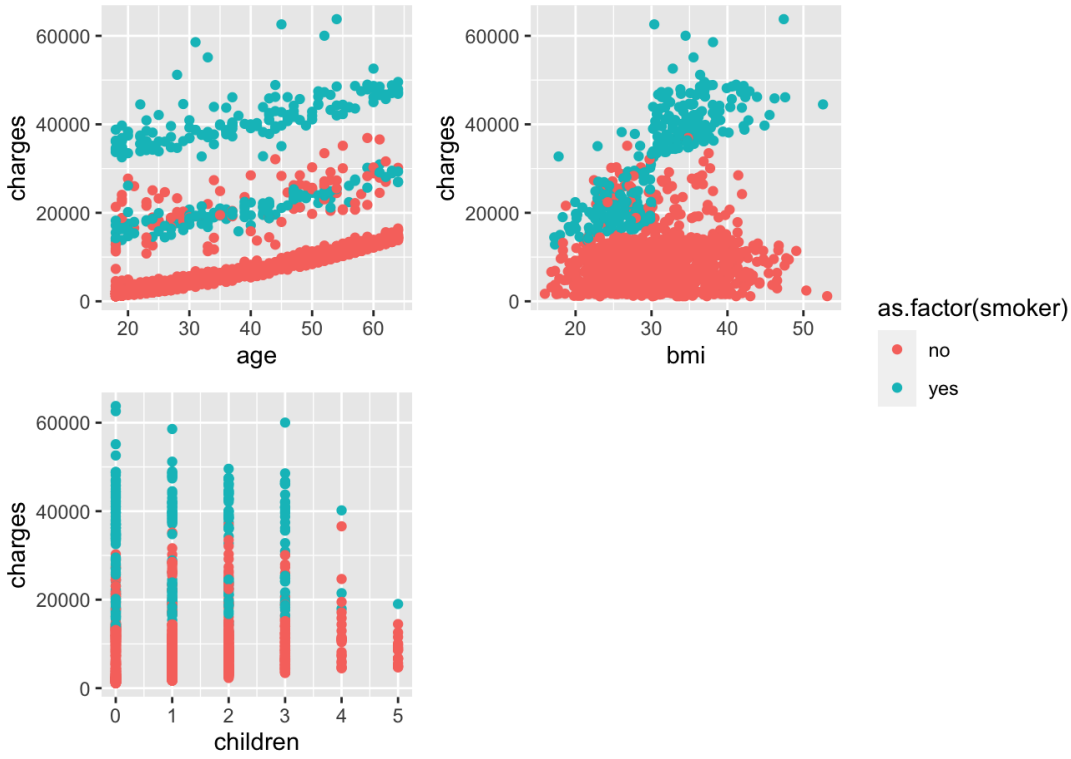


Fig3: charges vs age, bmi, children with color coded by smoker.

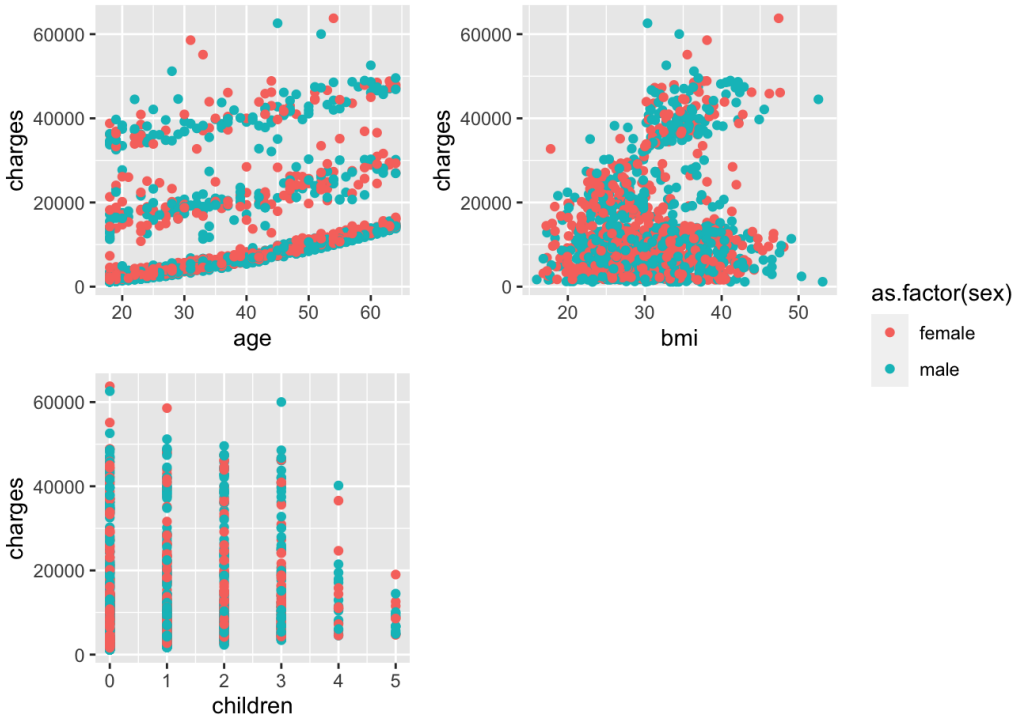


Fig4: charges vs age, bmi, children with color coded by sex.

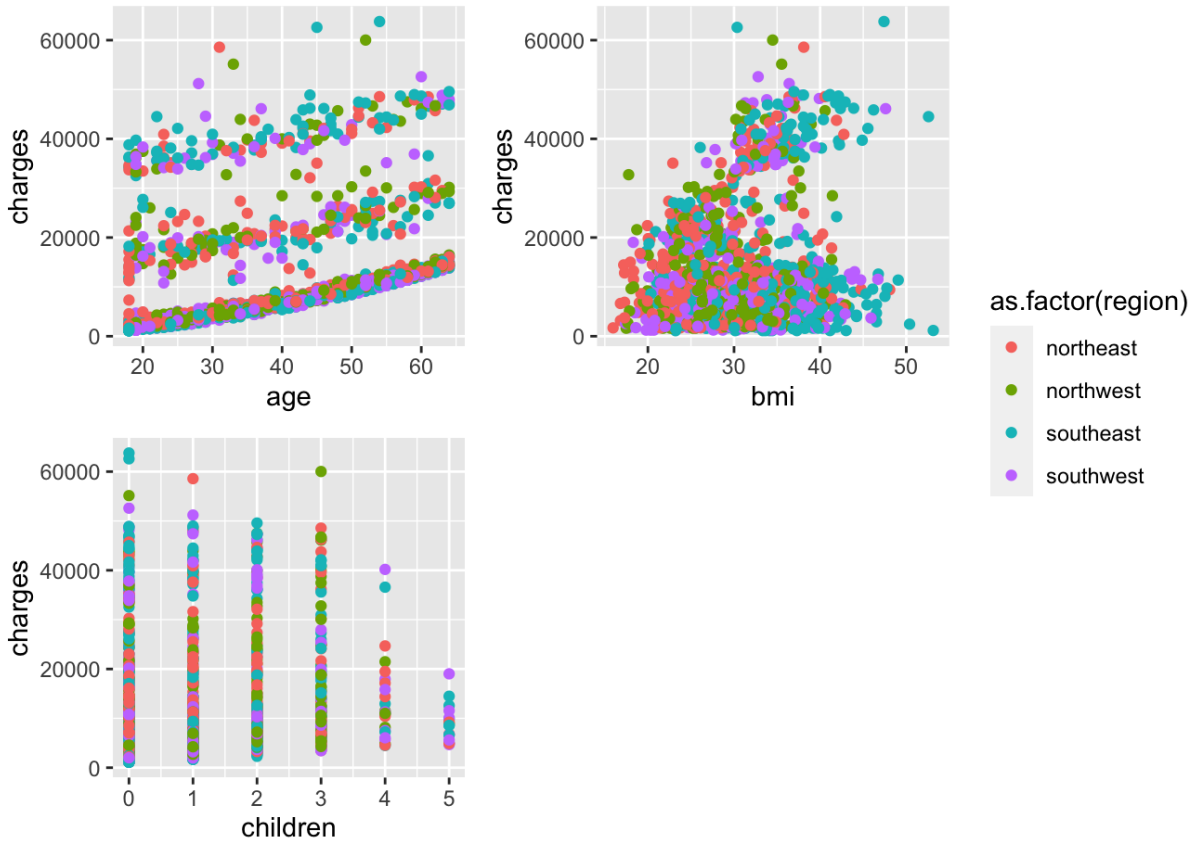


Fig5: charges vs age, bmi, children with color coded by region.

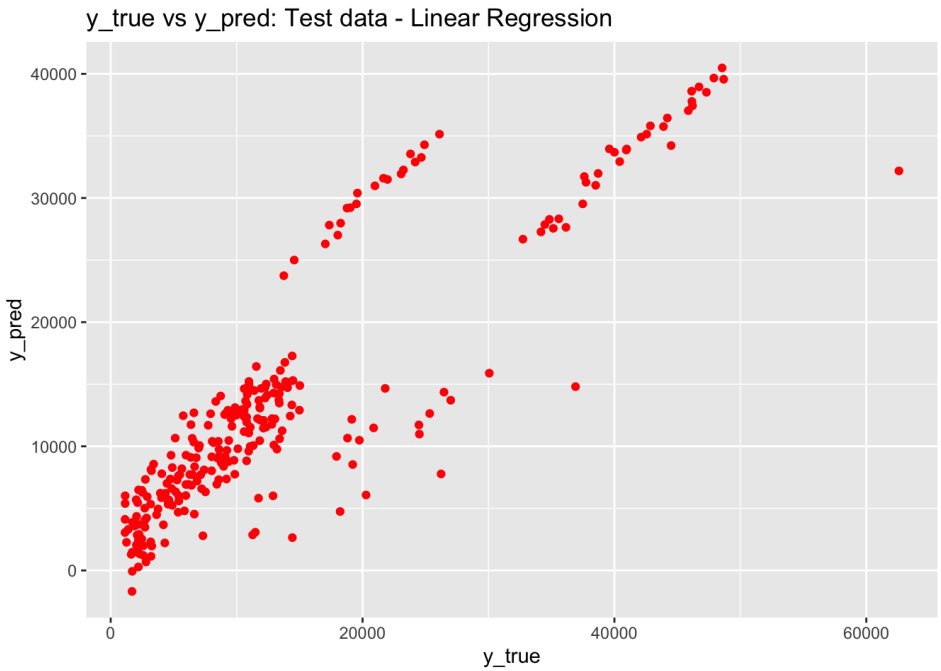


Fig6: True v/s Predicted values for Test data for Linear Regression model.

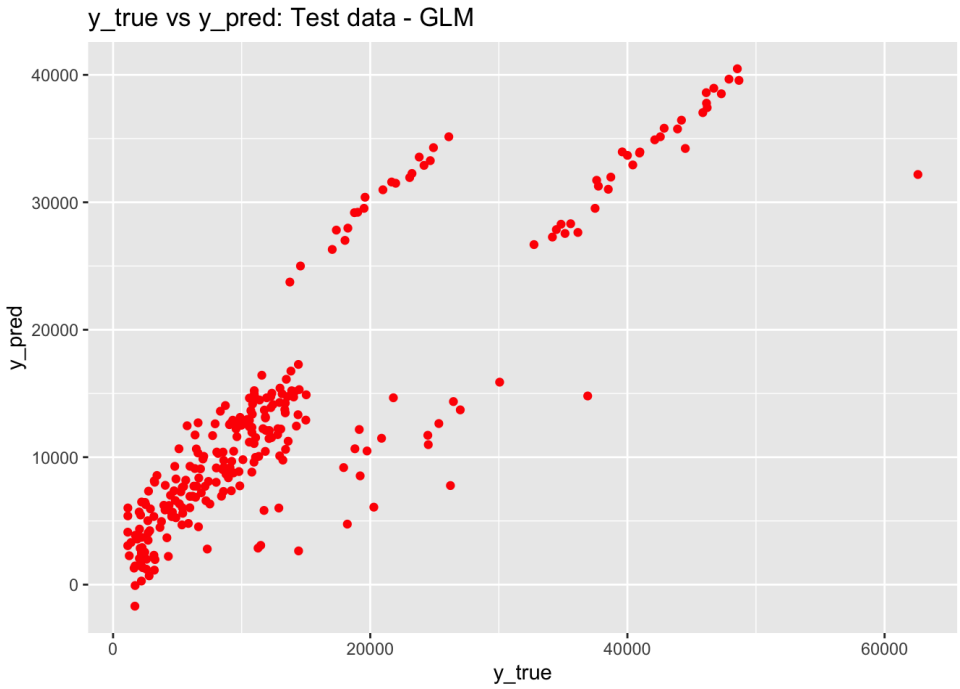


Fig7: True v/s Predicted values for Test data for GLM model

References

1. Singhal, S., & Patel, N. (2022, August 5). *The future of US healthcare: What's next for the Industry Post-covid-19*. McKinsey & Company. Retrieved December 2, 2022, from https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-future-of-us-healthcare-whats-next-for-the-industry-post-covid-19
2. Medical Insurance Payout Prediction : FLAML : RMSE: 2022. *https://www.kaggle.com/code/gauravduttakiit/medical-insurance-payout-prediction-flaml-rmse/data*. Accessed: 2022-12-02.
3. Machine learning in healthcare - benefits & use cases: 2022. *https://www.foreseemed.com/blog/machine-learning-in-healthcare#:~:text=Machine%20learning%20in%20healthcare%20can,that%20indicate%20a%20particular%20disease.* Accessed: 2022-12-02.