Identifying the factors that have effects on the Medical Insurance Costs

Student Name: Tianyu Yu, Leyao Lin, Ziying Wu, Langqi Liang

Instructor Name: Derek Bingham

Course: Stat350 (Fall 2020)

Content

Abstract

In this paper, we would investigate the factors which have effects on the Medical Insurance Cost in the US. For the data set given in the website, it contains 1338 data which have the response variable - insurance charges and the six explanatory variables. By analyzing the scatter plot and histogram, we will initially conclude that the main factors affecting insurance charges are age, BMI and smoker. After observing the initial analysis, we used the AIC and BIC model selection to determine the initial model from many possible models. Then we used residual plot to check and validate the model that the way could help us to confirm the final model. The final model determines and displays the factors which affect the insurance charges: age, the square of age, children, the interaction of age and region, the interaction of age and smoker, and the interaction of smoker and weight status.

Keywords:

Insurance charges, age, BMI, children, smoker, region, AIC, BIC, CP

Introduction

The medical service system occupies an important position in the economy and society in the US that doctors and medical institutions are the core supporting the entire system. Medical insurance payment is the main source of income for doctors and medical institutions and it is also the main expenditure of consumers. People focus on the medical insurance expense; hence it is important and useful to the research of influencing factors of medical insurance charges. In addition, the charges of medical insurance are increasing every year as time passes; however, the U.S. government is shrinking health insurance coverage against PPACA. This study will help people to determine which influence the costs of medical insurance -- the standard for measuring health will directly or indirectly affect the price of medical insurance.

1. Data set:

This data set was inspired by the book Machine Learning with R by Brett Lantz. The data contains medical information and costs billed by health insurance companies. It contains 1338 rows of data and the following columns: age, gender, BMI, children, smoker, region, insurance charges. (Miri Choi, 2020, Medical Cost Personal Datasets)

2. Response Variables and Explanatory Variables (Miri Choi, 2020)

Response Variable:

(1) Charges: Individual medical costs billed by health insurance

Explanatory variables:

(1) Age: age of primary beneficiary

(2) Sex: insurance contractor gender, female, male

(3) 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

(4) Children: Number of children covered by health insurance / Number of dependents

(5) Smoker: Smoking

(6) Region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

3. Hypothesis and analysis:

The insurance charges related to the health situations of the Insured, the explanatory variables the data given also related to the fitness factors which collected in four regions in the US.

(1) The relationship between age and the insurance charges may be the positive linear correlation because of age is closely related to health.

(2) For different genders, average life expectancy is different due to the morbidity and mortality which are also different when male or female are facing serious illness such as cancer.

(3) Body mass index (BMI) is an important index to measure human health; Different ranges of BMI may have different insurance charges.

(4) About the number of children, perhaps the health of females will be destroyed by the number of births and the cost of raising children. The insurance charges of females may be higher than male's.

(5) The weights of smoking or not may be the highest variable in insurance charges because smoking is the primary enemy of human health. In 2017, smoking was included in the list of carcinogens which may make our health destroyed.

(6) The beneficiary's residential area in the US is divided into four parts: northeast, southeast, southwest and northwest. We will analyze the balance of data collection about insurance charges that determines whether the data collected in four regions has the nearly number of samples.

4. To do List:

(1) By using box plots to analyze independent variables, to determine the important factors affecting dependent variables.
(2) Analyze the covariance between each variable to verify the importance of each as an influencing factor.
(3) Find the initial model using AIC, CP and BIC
(4) In order to display the model more intuitively and accurately, various methods will be used to optimize the model.

Methods and Result

1. Analyze variables and find the covariance

For determining the final model about the response variable insurance charges, we would analyze each explanatory variable, both categorical variables and numerical variables. In the data, the age, BMI, children and charges are the numerical variables and sex, smoker and region are the categorical variables. We would choose box plots to analyze each variable and observe the relationship between explanatory variables and response variable charges.

(1) Using box plots to analyze the categorical variables

First of all, we would check the data collection from the different regions in the US. From the box plot of the region, the amount of data from four regions in the US is similar, and then we initially conclude that the data collection is accurate.

Then, the numerical variable: children are also a categorical variable which displays the amount of the children. According to observing the box plot of children, there are no significant differences in different intervals of children. The mean of insurance charges in different intervals is similar and contrary to the initial assumption, the insured with five children have the lower insurance.

Next, for the BMI, we would divide the BMI into two parts: normal and overweight which display the weight status. We create a new variable ‘weight status’ to distinguish the BMI that when the BMI less than 30, which means ‘no obese’ otherwise means ‘obese’. The box plot of weight status display the different between obese and no obese, the obese have more insurance charges than the health. It means that the BMI has a strong relationship with charges and we need further research in building models.

The final categorical is smoker which displays the status of smoking that the box plot displays a huge difference between smokers and non-smokers. The difference is so huge that we could not ignore this variable that there is about a $20,000 insurance difference between smoker and non-smoker.

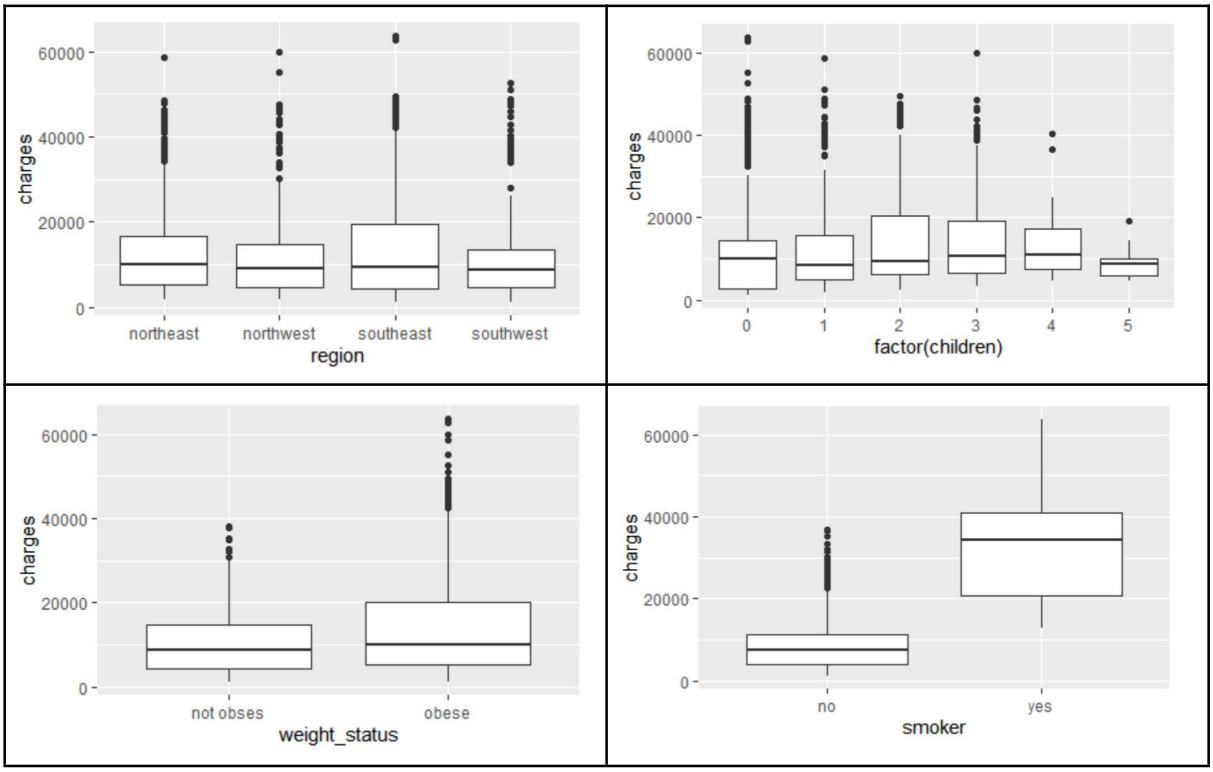

Figure1:the box plots of the categorical variables

(2) Observe the relationship between each variables

Using covariance function in R to analyze the covariance which is the measure of the joint variability of two random variables. About the other numerical variable, we would find the mathematical connection between each other. With observing the covariance matrix of the covariance, the covariance of relative strong relationships between age and BMI is 0.1092719, between age and charges is 0.29900819, and between BMI and charges is 0.19834097. Meanwhile, we build the scatter plots matrix to observe conveniently and intuitively the collection between each numerical variable. The initial analysis is not wrong that we would deeply discuss these connections in building models.

	age	BMI	children	charges
age	1.0	0.1092719	0.04246900	0.29900819
BMI	0.1092719	1.0	0.01275890	0.19834097
children	0.04246900	0.01275890	1.0	0.06799823
charges	0.29900819	0.19834097	0.06799823	1.0

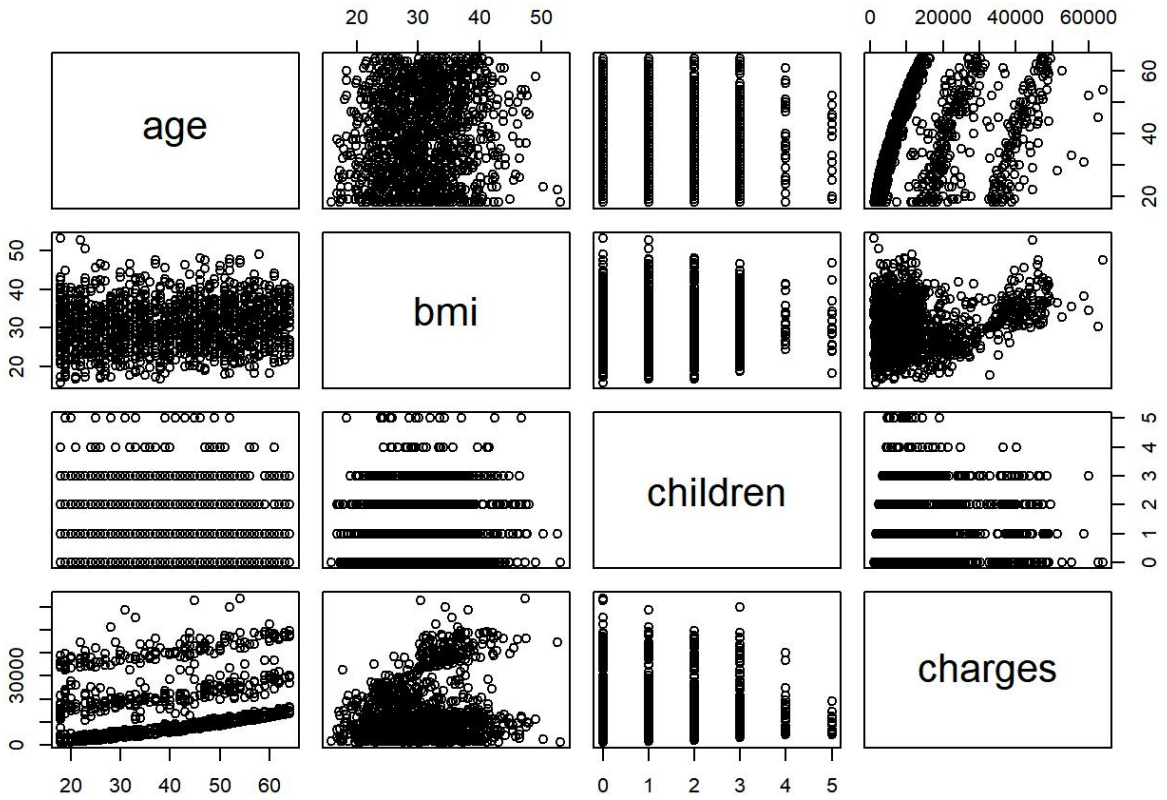

Figure 2: the scatter plot matrix of each numerical variables

2. Using BIC, CP, and AIC to find model

First of all, we generate a full model involving all of the variables, namely, age, sex, children, BMI, smoker, region, weight status (newly added) and their interaction terms with each other.

- Charges = $\beta_0 + \beta_1$ *Age+ β_2* Sex + β_3*Children + β_4*BMI + β_5*Smoker + β_6*Region + β_7*Weight Status + β_8*(Age : Sex) + β_9*(Age : Children) + β_{10}*(Age : BMI) + β_{11}*(Age : Smoker) + β_{12}*(Age : Region) + β_{13}(Age : Weight Status) + β_{14}(Sex : Children) + β_{15}(Sex : BMI) + β_{16}(Sex : Smoker) + β_{17}(Sex : Region) + β_{18}(Sex : Weight Status) + β_{19}(Children : BMI) + β_{20}(Children : Smoker) + β_{21}(Children : Smoker) + β_{22} (Children : Region) + β_{23} (Children : Weight Status) + β_{24} (BMI : Smoker) + β_{24} (BMI : Region) + β_{25} (BMI :Weight Status) + β_{26} (Region : Weight Status)

- Where the variables Sex, Smoke, Region, and Weight Status are categorical variables.

By using regsubset() function, we are able to obtain different functions for regression subset selection with forward & backward methods, then look at the plot for BIC values of various number of variable, however, it would be not really obvious to find out the minimum BIC value at the plot.

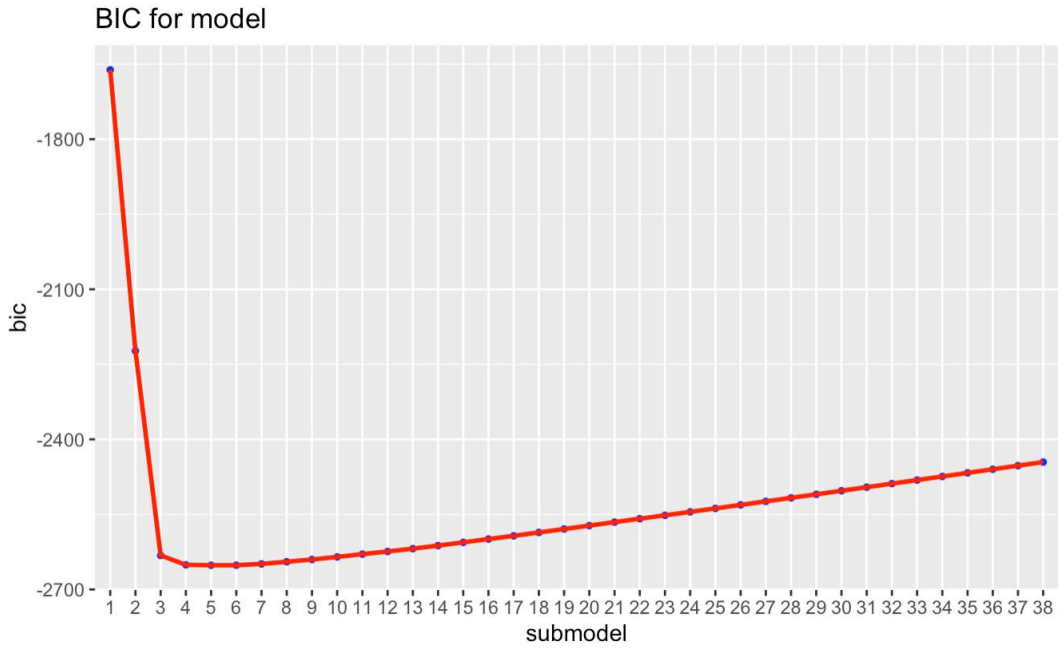

Figure 3: the BIC for model

Apart from the plot of BIC values, we also create an array about sub-models & BIC values in order to seek the minimum BIC value for optimizing the model. The model involving 5 variables would be the best choice based on BIC values, then makes use of forward and backward methods, respectively. Regardless of both of the methods used, the result is

- $Charges_{BIC} = \beta_0 + \beta_1$ *Age+ β_2* Children + β_3*(BMI : Region) + β_4*(BMI : Region) + β_5*(Smoker : Weight Status)
- Where the variables Sex, Smoke, Region, and Weight Status are categorical variables.

Furthermore, similarly, based on the plot of C_p values and output of the minimum value, 9 variables would be our choice. However, forward method & backward method have different regression choices

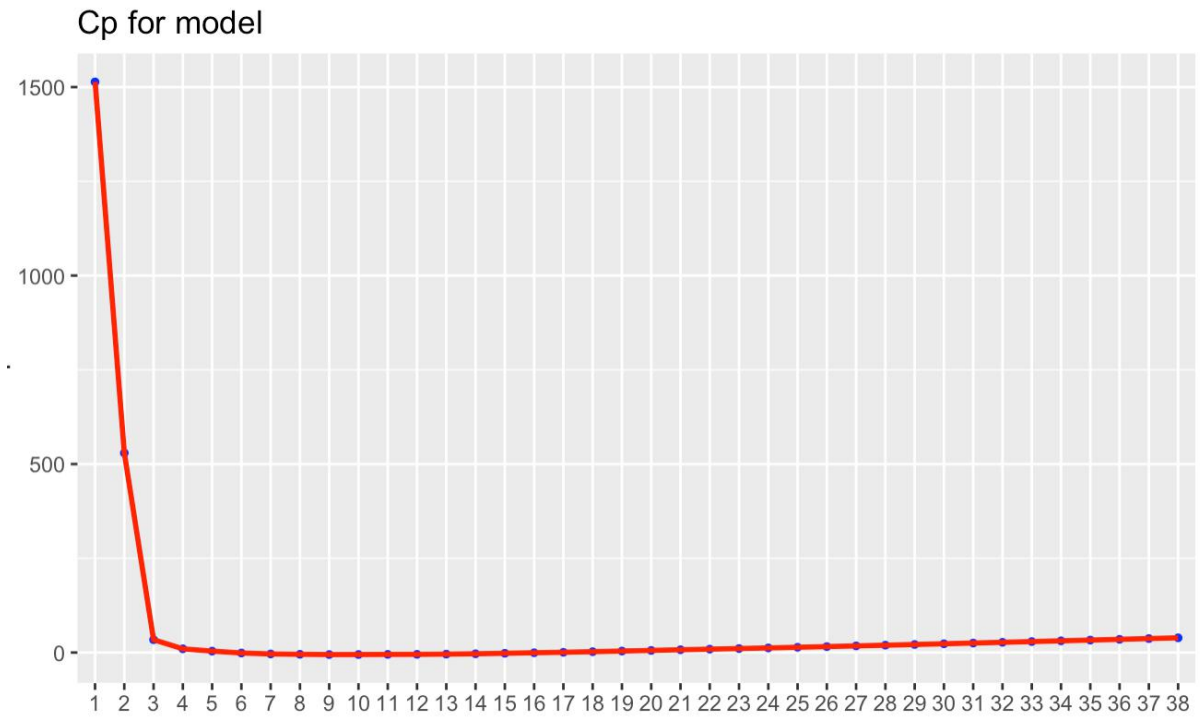

Figure 4: the CP for model

- $Charges_{CP:forward} = \beta_0 + \beta_1 * Age + \beta_2 * Sex + \beta_3 * Children + \beta_4 * Region + \beta_5 * (Age : Region) + \beta_6 * (Age : BMI) + \beta_7 * (BMI : Smoker) + \beta_8 * (BMI : Region) + \beta_9 * (Smoker : Weight\ Status)$
- *Where the variables Sex, Smoke, Region, and Weight Status are categorical variables.*

- $Charges_{CP:backward} = \beta_0 + \beta_1 * Age + \beta_2 * Sex + \beta_3 * Children + \beta_4 * (Age : BMI) + \beta_5 * (Age: Region) + \beta_6 * (BMI : Smoker) + \beta_7 * (Smoker: Weight\ Status)$
- *Where the variables Sex, Smoke, Region, and Weight Status are categorical variables.*

After that, using step() function choose a model by AIC in a stepwise algorithm from three directions of forward, backward and both, then select the model having the lowest AIC value. The model we selected appeared twice in the output; therefore, we confirm that the model would be the best choice in this case.

- $Charges_{AIC} = \beta_0 + \beta_1 * Age + \beta_2 * Sex + \beta_3 * BMI + \beta_4 * Children + \beta_5 * Smoker + \beta_6 * Region + \beta_7 * Weight\ Status + \beta_8 * (BMI : Smoker) + \beta_9 * (BMI : Region)$
- *Where the variables Sex, Smoke, Region, and Weight Status are categorical variables.*

```
model_cp_backward  model_cp_forward      modelfrombic        model_step
         26277.43          26274.77          26272.62          26274.35
```

By comparing AIC values of all of the models above, we temporarily choose the model which is selected based on the BIC algorithm.

3. Model Optimization, Re-observation of variables

Moreover, we need to check if the model satisfies the assumptions associated with a linear regression model: linearity, normality, independence, homoscedasticity. However, after checking the residual plots, it may have some non-linear relationships between response variable and explanatory variables.

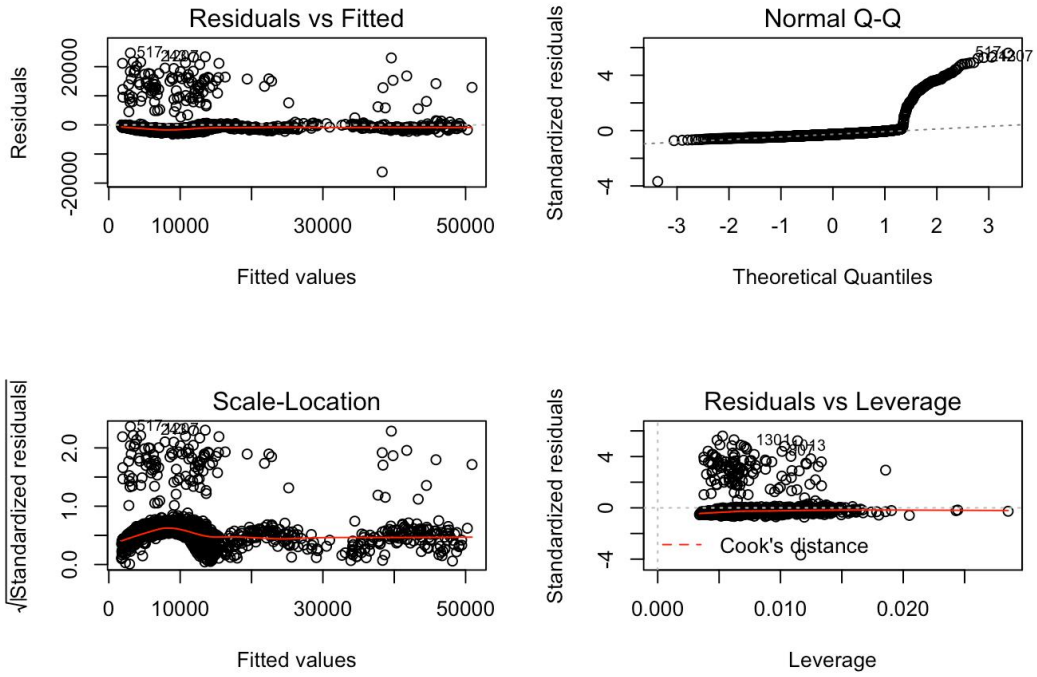

Figure 5: the residual plots to analyse error

Since there is a doubt about linearity, then we try to make the model as a polynomial. At the beginning, make age and children as the variables having the power of 2 and 3, respectively. Take advantage of summary(), then found that age may have a nonlinear relationship with charges. We determine the model will be $Charges = \beta_0 + \beta_1 * Age + \beta_2 * Age^2 + \beta_3 * Children + \beta_4 * (BMI: Region) + \beta_5 * (BMI: Region) + \beta_6 * (Smoker: Weight Status)$

4. Statistical model validation

Finally, using data splitting as model validation of the final model. At first, random selecting 75% data from original dataset as training data and the rest 25% data as test data. Second, training model based on training dataset. Then, calculating the predicted charges and finally calculate the r-squared (R2), root mean squared error (RMSE) and mean absolute error (MAE) by predicted charges and the function from the "caret" package. We do the 5 times of above steps, the table as below shows the result of 5 times R2, RMSE and MAE.

index	R2	RMSPE	MAPE	RMSPE_N
1	0.88298451	4035.57462	2273.31225	0.34573132
2	0.85073522	4783.48294	2451.67998	0.38761322
3	0.88007946	4510.85541	2373.38921	0.3469244
4	0.89927485	3806.26143	2130.26256	0.31845245
5	0.88746352	3965.42677	2147.62185	0.33688741

According to the result above, the final model is good to predict the insurance charges. The r-squared fluctuates from 85% to 89% which could explain most of the data in the dataset. Since it is not sure whether RMSPE is larger or smaller, a new variable RMSPE_N which is calculated by RMSPE divide test data standard deviation is used to test the model. The results show that RMSPE_N is in the range from 0.31 to 0.38 which is relatively close to zero. Therefore, the final model is determined.

Final model:

$$charges = \beta_0 + \beta_1\, age + \beta_2 age^2 + \beta_3 children + \beta_4\, bmi{:}region + \beta_5\, bmi{:}smoker + \beta_6 smoker{:}weight\_status$$

Conclusion and Discussion

According to the final model we determined, we could conclude that the insurance charges is linear related to:

(1) The age of insured (age)

(2) The square of insured's age (age^2)

(3) The number of children (children)

(4) The interaction of insured's age and their place of residence (BMI:region)

(5) The interaction of insured's age and the situation of smoking (age:smoker)

(6) The interaction of the situation of insured smoking and weight status (smoker: weight status)

According to the results of the summary table of the final model, there are five explanatory variables that are partly detected significant, which are age, age^2, children, the interaction terms BMI: smoker and smoker: weight status. For the significant terms, their estimated coefficients are:

Estimated coefficients of significant explanatory variables in final model	
(intercept)	7330.173
age	135792.113
age^2	22621.361
children^1	656.262
bmi:smoker_yes	524.935
smoker_yes:weight_status_obese	14560.002

Base on above table, we conclude these statement:

(1)According to the increasing age, the elder insured should pay more insurance than the young.

(2)The insured which has one child should pay less insurance than the ones who have more children.

(3)The smoker should pay more insurance than the non-smoker.

(4)The insured who is both smoker and overweight should pay more insurance than other insured.

(5)The insurance charges in different regions have no significant difference.

In order to reduce insurance expenses, people need to keep exercising to ensure BMI in a normal range because the overweight person should pay more insurance. Meanwhile, people need to quit smoking because of the difference in insurance charges between smoker and non-smoker over $20,000. However, about the final model, we build the residual plot to analyze the error. The plot of residual displays that most of the data is in the normal range, but there are some points that are not distributed evenly which means the model is flawed yet. Consequently, we need to investigate more deeply about reducing the error in the future study.

Reference

Medical Cost Personal Datasets Insurance Forecast by using Linear Regression. Retrieved from Miri Choi *https://www.kaggle.com/mirichoi0218/insurance*

Appendix

More detail on insurance.rmd on github link below.

These four links included the same data and code but belonged to different accounts.

https://github.com/LeyaoLin/2020Fall-STAT350-team-project
https://github.com/langqiliang/stat350
https://github.com/mila328/stat350teamproject
https://github.com/tianyu-yu-97/STAT350