STAT 3340

Final Project Report

Regression Analysis of Medical Insurance Cost Dataset

Faculty of Statistics
Dalhousie University

Dec.5th, 2020

Group 8

Yuqiao Du	B00773343
Rao Huang	B00798538
Yurunyun Wang	B00776875
Ruoxin Xu	B00771053

Table of Contents

Section 1: Abstract

This report aims to provide readers with a systematic regression analysis of medical insurance cost with several potential predictors in the dataset. Our goal is to utilize multiple linear regression processes to fit those data points as close as possible. First,we use the “pairs()” function in R to get the whole possible relationship between numeric variables so that it can indicate the approximate relations from the pair graphs. Then we use ggplot to visualize categorical variables and use indicator variables to deal with categorical variables. Finally, we implement backward selection procedures to get the ‘best model’. After analyzing four graphs and VIF of the best model, it basically obeys the assumption, and both of the predictors are weak multicollinearity. Overall, the dataset basically fit our linear regression model, but it might fit the polynomial model better.

Section 2: Introduction

As a result, medical insurance costs have become a topic of great concern. Do you want to know how your health insurance costs will be affected? In this report, we apply a linear regression model to a set of personal medical costs datasets in order to predict future insurance costs and trends for individuals. Linear regression is one of the types of machine learning. It is the first machine learning algorithm based on "supervised learning". Through our analysis and study of this dataset, we will also give you a better understanding of the application of the model.

It’s well known that the charges reported by the insurance company vary much based on different physical conditions and other external factors. The primary purpose of our project is to find the best model to fit the given dataset. The other purpose is to figure out whether sex and smoker have a joint effects of insurance charges. These results can provide a suggestion when the customers are purchasing health insurance and estimate the premium.

The remainder of the report will be outlined as follows. In section 3, we will describe the source of data, columns of data, and the new added data point. In section 4, we will discuss the process of attaining the best model to represent the data. In section 5, we will describe the results of verify the performance of the models. The conclusion part can be found in section 6. In section 7, the data file and R markdown file will be attached in the appendix.

Section 3: Data Description

In this project, we are focusing on exploring the analysis of the Medical Insurance Costs dataset. This dataset is downloaded from an online community Kaggle and is originally inspired in the book *machine learning using R* by author Brett Lantz. It contains the medical information of observations and the premium charged by the health insurance company.

The original dataset contains 1338 rows of data, and the columns are age, gender, BMI, children, smoker, region, and insurance charges. For this dataset, we add a unique data point. The new data point records a 22-year-old female observation with a BMI of 23.71, who has no kids, doesn't smoke, comes from the northeast in the US, and pays $5503.7768 for the total insurance charge. The values for this unique data point come from real life. Each value is randomly selected from the real data provided by all four group members. As for the data structure of the columns, the age, BMI, children, and insurance charges are double-type data, while the data in the rest of columns are categorical data with String type.

Table 1 provides more details of columns of numeric data.

age	bmi	children	charges
Min. :18.00	Min. :15.96	Min. :0.000	Min. : 1122
1st Qu.:26.50	1st Qu.:26.25	1st Qu.:0.000	1st Qu.: 4742
Median :39.00	Median :30.40	Median :1.000	Median : 9378
Mean :39.19	Mean :30.66	Mean :1.094	Mean :13265
3rd Qu.:51.00	3rd Qu.:34.69	3rd Qu.:2.000	3rd Qu.:16622
Max. :64.00	Max. :53.13	Max. :5.000	Max. :63770

Table 1. Summary of numeric data

Table 2 provides more details of columns of categorical data.

Variable Name	sex	count	region	count	smoker	count
	female	663	northeast	325	no	1065
	male	676	northwest	325	yes	274
			southeast	364		
			southwest	325		

Table 2. Summary of categorical data

Figure 1 illustrates the relationships between numerical variables. In this scatterplot, there is a possible linear relationship between variable age and variable charges. For the rest pairs of variables, it is hard to say that there is any possible linear relationship between them based on the scatterplot.

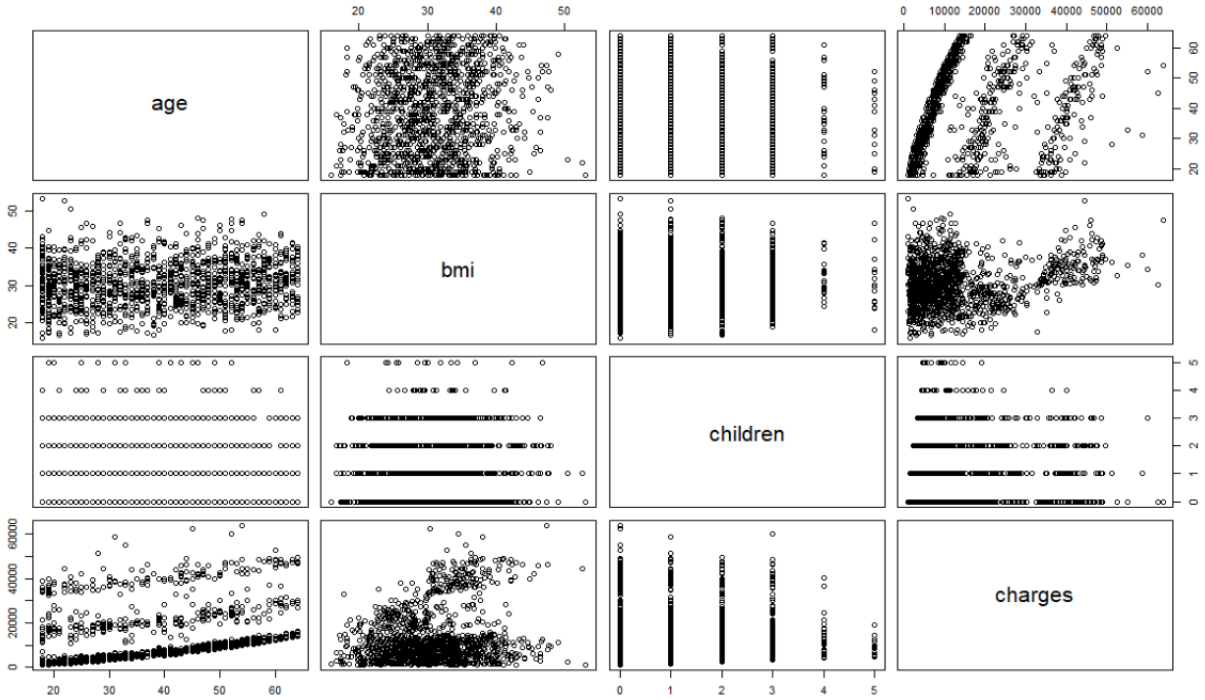


Figure 1. The correlation between numeric variables

Figure 2 shows the correlation matrix between all numeric variables. The correlation matrix shows the same result as what we obtain in the scatterplot, which is to say, the correlation between age and charges is the largest among all pairs of numeric variables.

	age	bmi	children	charges
age	1.00000000	0.11020026	0.04326278	0.29938143
bmi	0.11020026	1.00000000	0.01352219	0.19876041
children	0.04326278	0.01352219	1.00000000	0.06840199
charges	0.29938143	0.19876041	0.06840199	1.00000000

Figure 2. The correlation matrix between numeric variables

To have a clear overview of the relationships between every categorical variable and charges, we use the **ggplot2** package in R to show the scatterplots. Figure 3 shows the distribution of four categorical variables respectively.

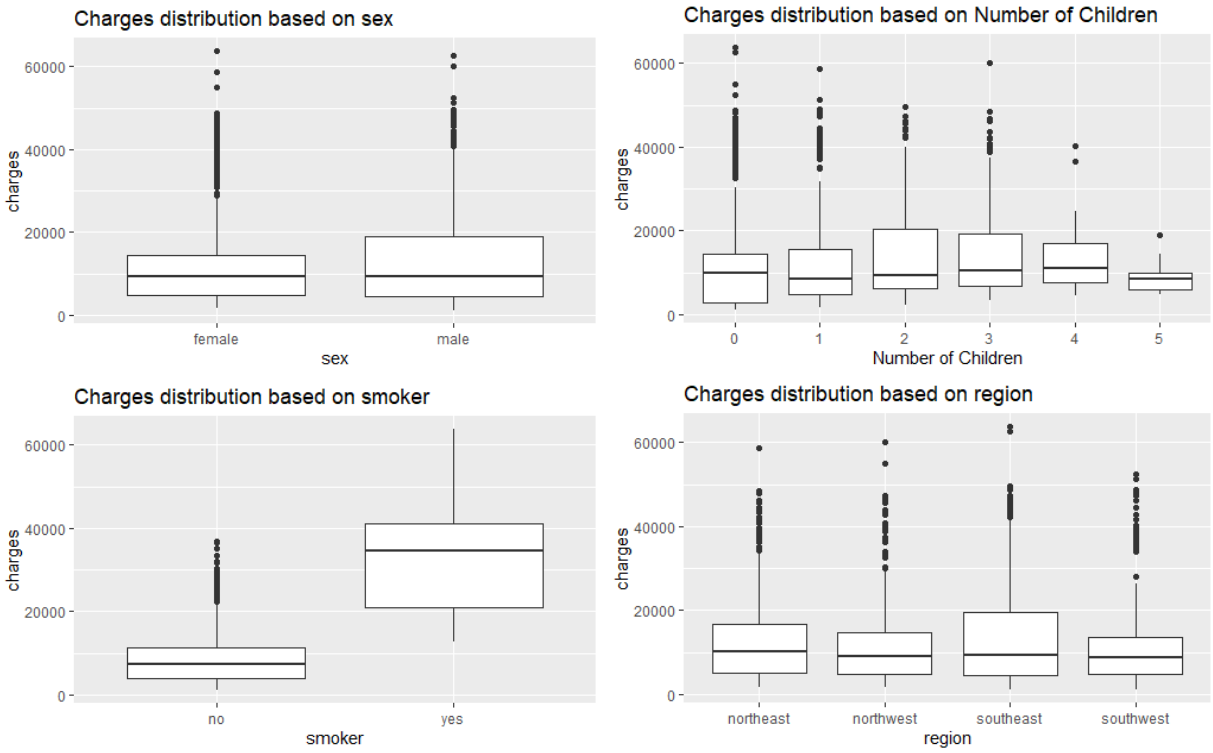


Figure 3. Ggplot of four categorical variables against charges.

The full name of BMI is Body Mass Index. It is a health indicator separating different weight categories. To help readers understand this variable, Figure 4 shows the ggplot of subcategories of BMI against charges. In the process of modeling, we still regard BMI as a numeric variable.

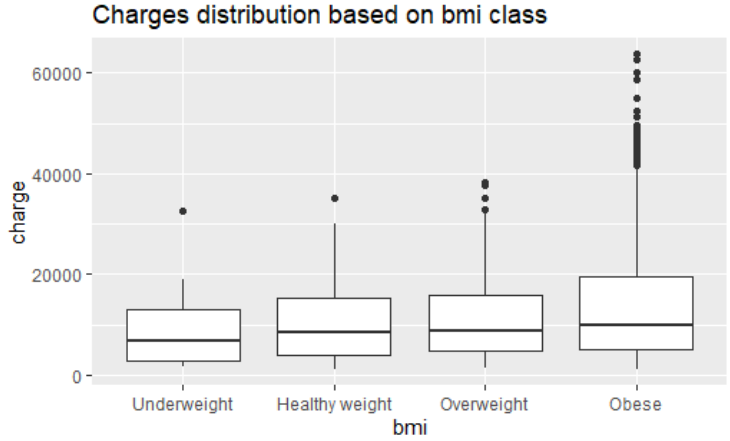


Figure 4. Ggplot of variable BMI and its subcategories.

Section 4: Methods

From figure 3 shown in Section 3, we can know that there exist real patterns between variables, despite the great variability between them. In this project, we will fit the given data by using linear regression modeling. Following the statement of model selection procedure, we will estimate the chosen 'best model' obtained.

4.1 Model selection procedure

To find a model with better performance, we must take the interactions between different variables into consideration. Among all the interactions, we will focus on figuring out whether conclude the interaction between variable sex and variable smoker in the model.

First of all, we use a hypothesis test to help the decision of whether adding an interaction between BMI and smoker in the model. Variable sex is divided into variable sex_female and variable sex_male, while variable smoker is divided into smoker_yes and smoker_no as both of them are categorical variables. Then we can set the coefficients like this:

β_1: the coefficient of region southwest
β_2: the coefficient of region southeast
β_3: the coefficient of region northwest
β_4: the coefficient of region northeast
β_5: the coefficient of age
β_6: the coefficient of bmi
β_7: the coefficient of children
β_8: the coefficient of sex male
β_9: the coefficient of smoker yes
β_{10}: the coefficient of sex female
β_{11}: the coefficient of smoker no
β_{12}: the coefficient of sex_male * smoker_yes
β_{13}: the coefficient of sex_female * smoker_no
β_{14}: the coefficient of sex_female * smoker_no
β_{15}: the coefficient of sex_male * smoker_yes

The null hypothesis is $\beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 0$and the alternative hypothesis is at least one of $\beta_{12}, \beta_{13}, \beta_{14}, \beta_{15} \neq 0$. Then we begin modeling with given data by using linear regression. To identify the 'best model', we implement backward selection procedures to get the 'best model'. The 'best model' we obtain from backward selection is

$$charges = \beta_0 + \beta_1 region\ southwest + \beta_2 region\ southeast + \beta_5 age + \beta_6 bmi + \beta_7 children + \beta_{10} smoker\ yes.$$

What's more, the AIC of our 'best model' is 23329.98. Since the p-value of this model is far less than significance level 0.05, we can make the conclusion that we reject the null hypothesis, which means that we should conclude the interactions in our model. The p-value comes from the followed Figure 5.

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2437.021    855.954  -2.847  0.00448 **
region_southwest  -1232.322    382.182  -3.224  0.00129 **
region_southeast  -1211.268    382.766  -3.165  0.00159 **
region_northwest   -586.666    380.794  -1.541  0.12364
age                 263.948      9.515  27.741  < 2e-16 ***
bmi                  22.388     25.602   0.874  0.38203
children            511.923    110.204   4.645 3.73e-06 ***
smoker_yes        -20319.567   1648.007 -12.330  < 2e-16 ***
bmi:smoker_yes     1438.397     52.605  27.343  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4849 on 1330 degrees of freedom
Multiple R-squared:  0.8405,    Adjusted R-squared:  0.8396
F-statistic: 876.3 on 8 and 1330 DF,  p-value: < 2.2e-16
```

Figure 5. The summary of the model obtained by backward selection.

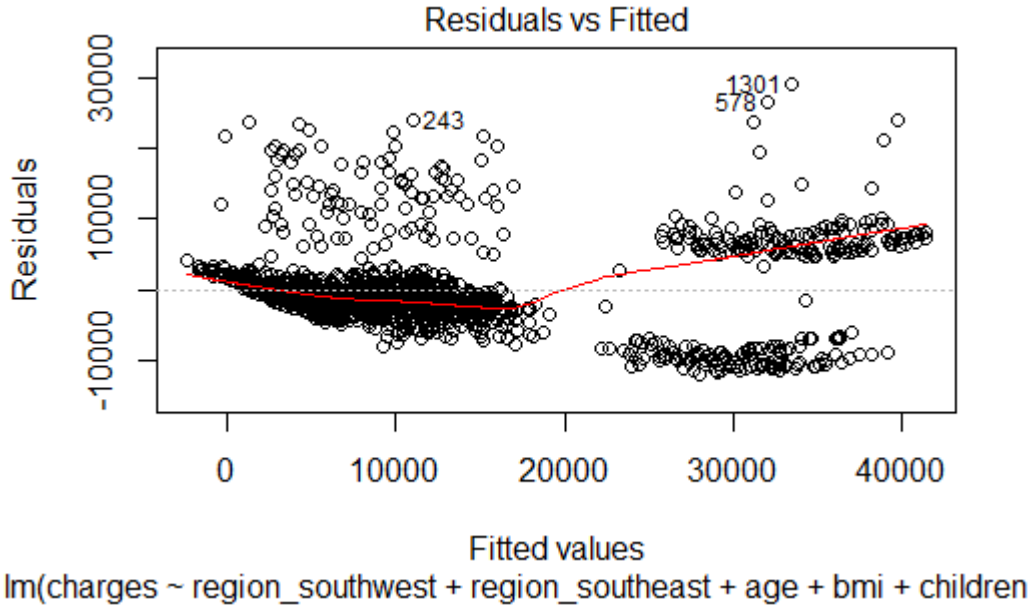
Refer to the conclusion drawn from the hypothesis test, the new model should conclude the interactions between variable BMI and variable smoker. Same as before, the new 'best model' obtained from backward selection is:

$$charges = \beta_0 + \beta_1 region\ southwest + \beta_2 region\ southeast + \beta_5 age + \beta_6 bmi + \beta_7 children + \beta_8 sex\ male + \beta_{10} smoker\ yes + \beta_{12} sex\ male * smoker\ yes$$

The AIC of this new 'best model' is 23325.66.

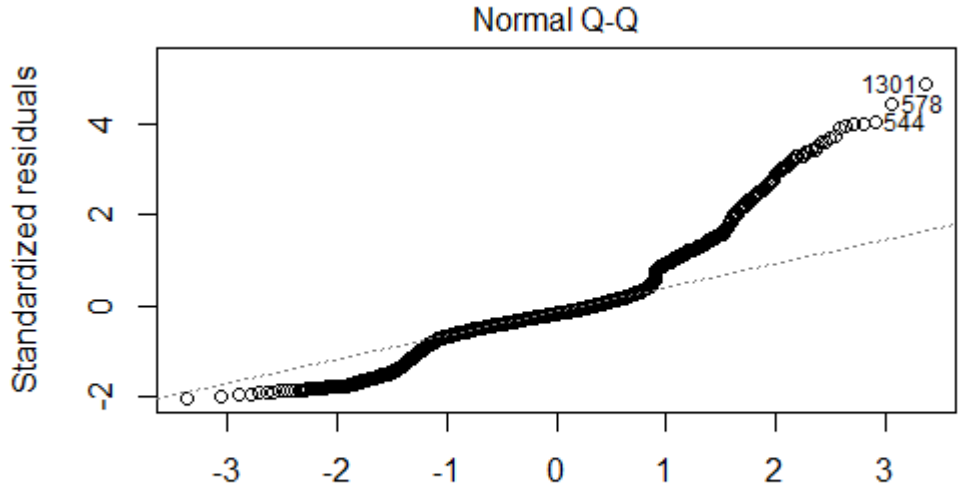
4.2 Estimation of the chosen model

Looking at the Residuals vs Fitted plot, we can see that the red line is approximately flat, which means that the mean of error is approximately equal to 0.

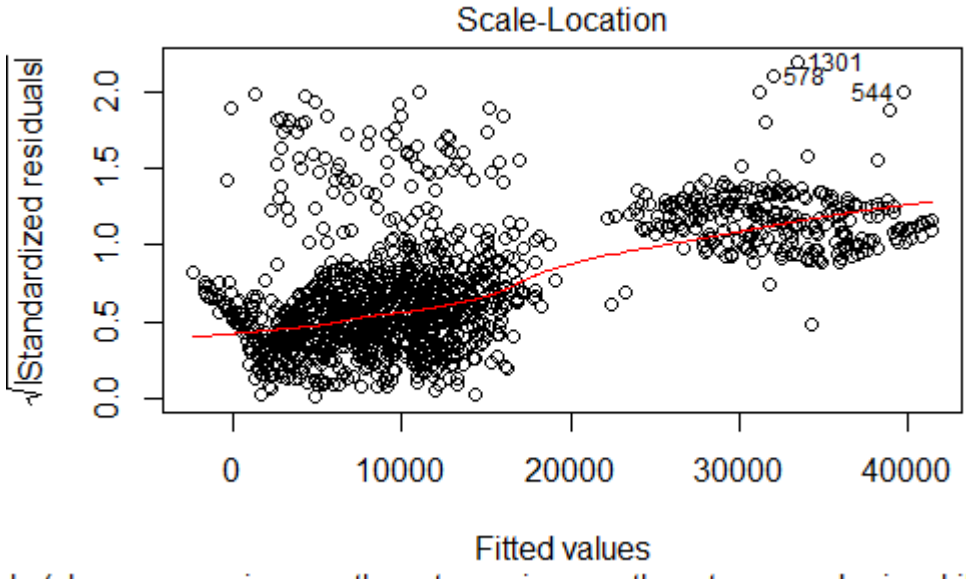


Looking at the Normal Q-Q plot, the residuals are approximately matched to the diagonal line, which means that the residuals are roughly normally distributed. And we observe both the upper

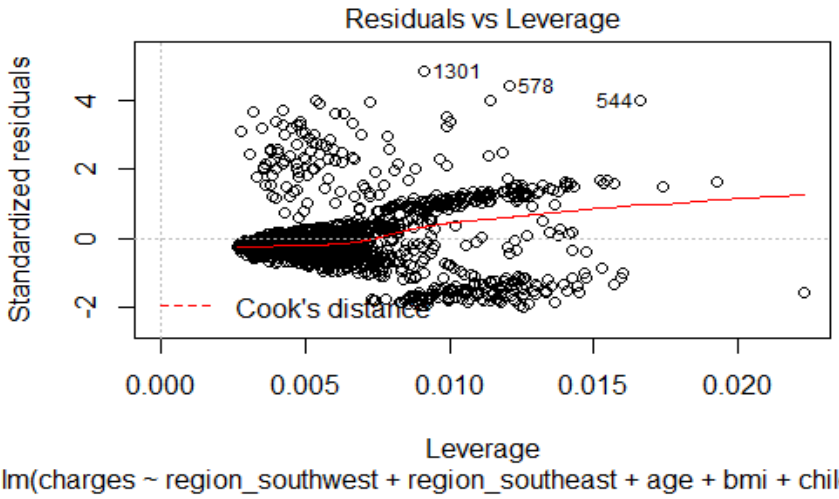
and lower tail are ‘heavier’ (have larger values) than what we would expect under the gauss-markov assumptions.



Looking at the Scale-Location plot, we can see that the red line is approximately flat, which means that there is no indication of having non-constant variance.



The Residuals vs Leverage plot indicates that there exist some outlier points: 1301st observation, 578th observation, 544th observation.



VIF value of each predictor:

```
region_southwest    region_southeast                age
        1.148271           1.244387           1.016990
             bmi           children           sex_male
        1.109890           1.003993           1.259475
      smoker_yes sex_male:smoker_yes
        2.305003           2.638712
```

From the VIF values of each predictor, we can see that all of the variables in the model have weak multicollinearity.

Section 5: Result

```
lm(formula = charges ~ region_southwest + region_southeast +
    age + bmi + children + sex_male + smoker_yes + sex_male:smoker_yes
    data = datasetfull)

Residuals:
   Min     1Q Median     3Q    Max
-12144  -2874  -1009   1408  29178

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -11709.59     963.61 -12.152  < 2e-16 ***
region_southwest     -824.67     412.81  -1.998  0.04595 *
region_southeast     -874.11     414.10  -2.111  0.03497 *
age                   256.69      11.86  21.649  < 2e-16 ***
bmi                   334.81      28.54  11.731  < 2e-16 ***
children              459.67     137.34   3.347  0.00084 ***
sex_male             -607.09     370.72  -1.638  0.10175
smoker_yes          22525.56     621.54  36.241  < 2e-16 ***
sex_male:smoker_yes  2364.25     829.36   2.851  0.00443 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6044 on 1330 degrees of freedom
Multiple R-squared:  0.7523,    Adjusted R-squared:  0.7508
F-statistic:   505 on 8 and 1330 DF,  p-value: < 2.2e-16
```

Figure 5. The summary information of the 'best model' with an interaction between sex male and smoker yes.

From the Figure 5 above, they illustrate the summary information of the 'best model' from our backward process. And we can observe that the estimated coefficient of smoker_yes is 22525.56, which indicates that there is a strong positive relationship between smoker_yes and charges. Although in this case we can get R-squared up to 0.7523 and adjusted R-squared up to 0.7508, from the normal QQ plot we know other regression might have better performance compared to linear regression, such as polynomial regression.

Analysis of Variance Table

Response: charges

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
region_southwest	1	3.6142e+08	3.6142e+08	9.8952	0.001694
region_southeast	1	7.8619e+08	7.8619e+08	21.5247	3.837e-06
age	1	1.7695e+10	1.7695e+10	484.4522	< 2.2e-16
bmi	1	4.6554e+09	4.6554e+09	127.4589	< 2.2e-16
children	1	6.0177e+08	6.0177e+08	16.4757	5.215e-05
sex_male	1	5.7806e+08	5.7806e+08	15.8264	7.317e-05
smoker_yes	1	1.2258e+11	1.2258e+11	3356.1122	< 2.2e-16
sex_male:smoker_yes	1	2.9682e+08	2.9682e+08	8.1265	0.004430
Residuals	1330	4.8578e+10	3.6525e+07		

Figure 6. ANOVA analysis of the 'best model' with an interaction bwteen sex male and smoker yes.

From Figure 6 above, it shows that the p-values of all variables are less than significance level (0.05), which means all variables in the current model are significant.

There are two potential reasons why the normal Q-Q plot doesn't strictly follow a normal distribution. The first reason is that we only take one interaction into consideration, and there may exist another interaction that could make the model better. So, if we try all the potential combinations, the model will become more explainable. Secondly, the data we get don't strictly fit the linear relationship. it may obey the rule of other distributions such as Poisson or polynomial distribution.

Section 6: Conclusion

From this project, we implemented a regression analysis based on the Medical insurance costs data from the real world, which is more complex than practice data in lectures. During this project, all group members reviewed the regression knowledge and R language, learned the working knowledge of GitHub, designed the regression analysis, and discussed the results obtained from the chosen 'best model'.

According to the 'best model' we got, it can be observed that there is a strong positive relationship between smoker_yes and charges, which means that smokers' medical insurance charges are more likely higher than the people who don't smoke. That is to say, if someone is a smoker, his or her health insurance costs will be more likely expensive than those who don't smoke.

Section 7: Appendix

7.1 Data file

Github link: https://github.com/huangrao1212/3340Project

7.2 R markdown file

Github link:https://github.com/huangrao1212/3340Project

7.3 References:

About Adult BMI. (2020, September 17). Centers for Disease Control and Prevention. Retrieved 10, December 2020, from: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Quick Guide: Interpreting Simple Linear Model Output in R. (2015). Feliperego. Retrieved 10 December 2020, from: https://feliperego.github.io/blog/2015/10/23/Interpreting-Model-Output-In-R

Understanding Q-Q Plots | University of Virginia Library Research Data Services + Sciences. (2015). Retrieved 10 December 2020, from: https://data.library.virginia.edu/understanding-q-q-plots/

Contributors, D. (2020). Data visualization with ggplot2. Retrieved 10 December 2020, from https://datacarpentry.org/R-ecology-lesson/04-visualization-ggplot2.html