Logo, company name

Description automatically generated

**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**

1. Abstract
2. Introduction
3. Data description
4. Methods

4.1 Model selection procedure

4.2 Estimate of the chosen model

1. Result
2. Conclusion
3. Appendix

7.1 Data file

7.2 R markdown file

7.3 References

**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.

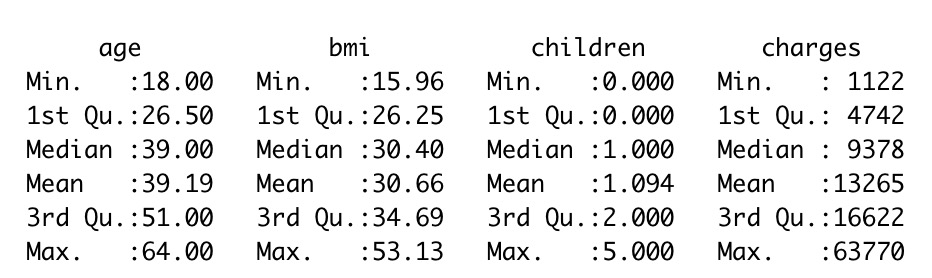
**Section 3: Data Description**

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.



***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.

Diagram, engineering drawing

Description automatically generated

***Figure 1. The correlation between numeric variables***

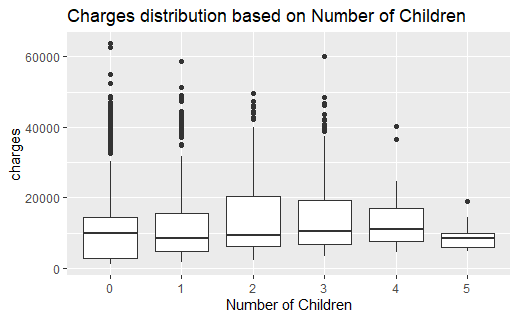
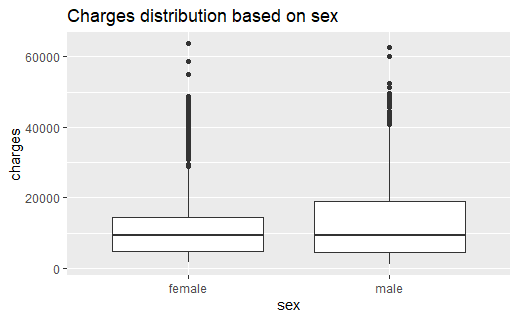
Figure 2 shows the correlation matrix between all numeric variables. The correlation matrix shows the same result as what we obtain in the scatterplot, which is to say, the correlation between age and charges is the largest among all pairs of numeric variables.

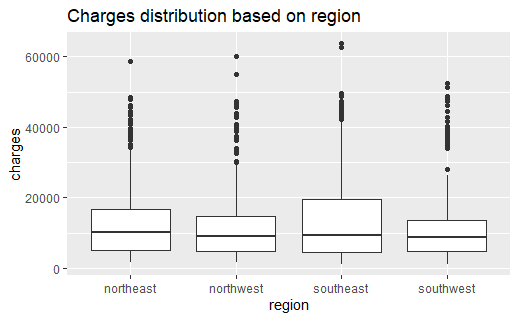
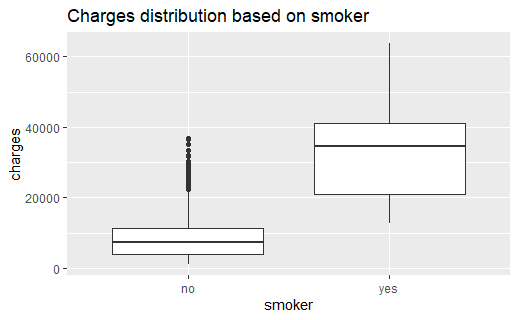
Text

Description automatically generated

***Figure 2. The correlation matrix between numeric variables***

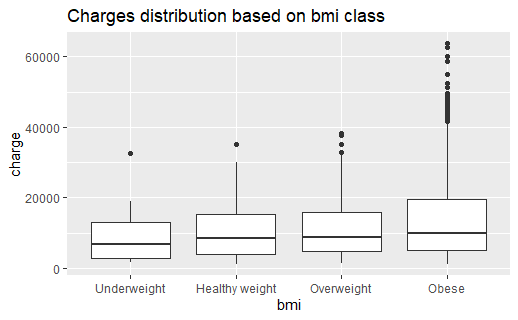
To have a clear overview of the relationships between every categorical variable and charges, we use the **ggplot2** package in R to show the scatterplots. Figure 3 shows the distribution of four categorical variables respectively.





***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:

: the coefficient of region southwest

: the coefficient of region southeast

: the coefficient of region northwest

: the coefficient of region northeast

: the coefficient of age

: the coefficient of bmi

: the coefficient of children

: the coefficient of sex male

: the coefficient of smoker yes

: the coefficient of sex female

: the coefficient of smoker no

: the coefficient of sex\_male \* smoker\_yes

: the coefficient of sex\_female \* smoker\_no

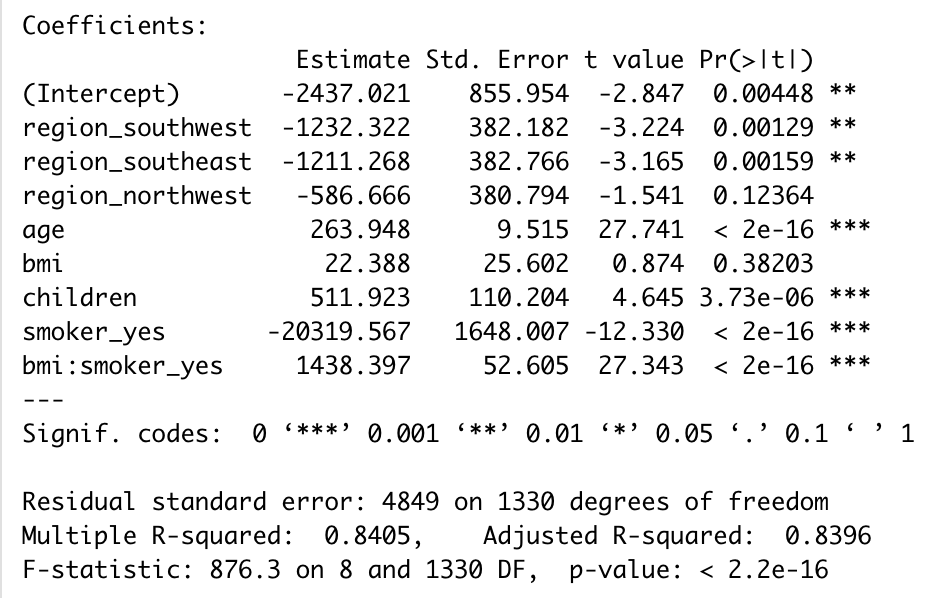
: the coefficient of sex\_female \* smoker\_no

: the coefficient of sex\_male \* smoker\_yes

The null hypothesis is and the alternative hypothesis is at least one of . 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

.

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.



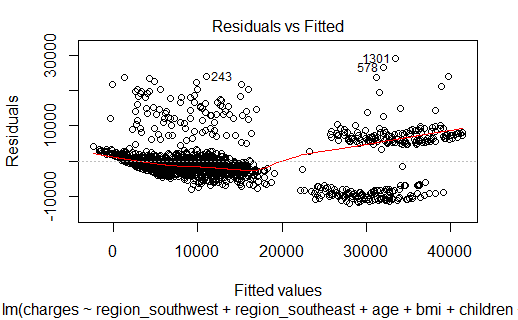
***Figure 5. The summary of the model obtained by backward selection.***

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:

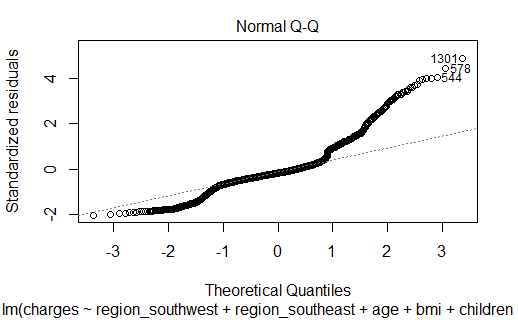
The AIC of this new ‘best model’ is 23325.66.

4.2 Estimation of the chosen model

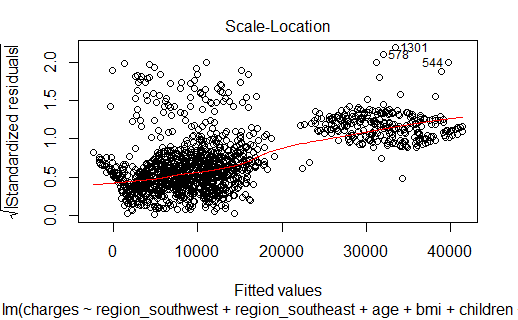
Looking at the Residuals vs Fitted plot, we can see that the red line is approximately flat, which means that the mean of error is approximately equal to 0.



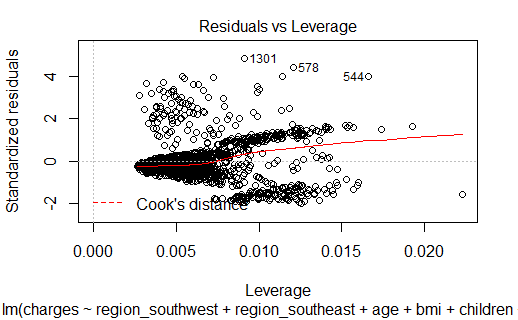
Looking at the Normal Q-Q plot, the residuals are approximately matched to the diagonal line, which means that the residuals are roughly normally distributed. And we observe both the upper and lower tail are ‘heavier’ (have larger values) than what we would expect under the gauss-markov assumptions.



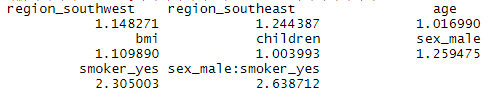
Looking at the Scale-Location plot, we can see that the red line is approximately flat, which means that there is no indication of having non-constant variance.



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

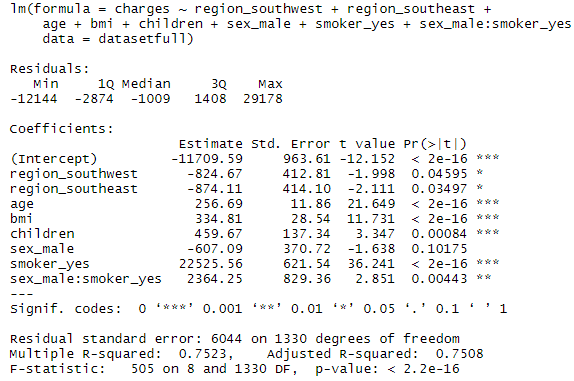


VIF value of each predictor:



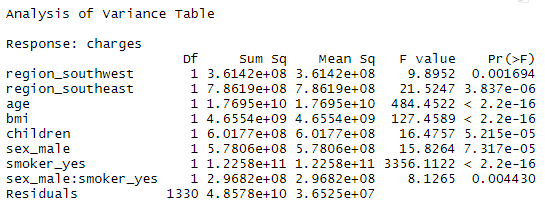
From the VIF values of each predictor, we can see that all of the variables in the model have weak multicollinearity.

**Section 5: Result**



***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.



***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>