STAT350 Final Project

Student Name & ID:

Yunxiang Zhuang: 301301775

Xian He: 301323754

Course: STAT350

Instructor: Derek Bingham

Dataset: Medical insurance

Contents

Abstract

In this paper, we mainly investigate what factors influenced the charges of medical insurance. We analyze the charges distribution and influencing factors of medical insurance by applying the indicator variable, model selection, cross-validation, robust regression methods and data transformation. The final results indicate that the exponential variable smoker, age, body mass index (BMI), number of children and region has the mostly influence on the response variable charges. Meanwhile, we find that the factor with greatest impact on health insurance costs is ‘smoker or not’. In addition, we observe that at the same age, smoker needs to pay more for medical insurance.

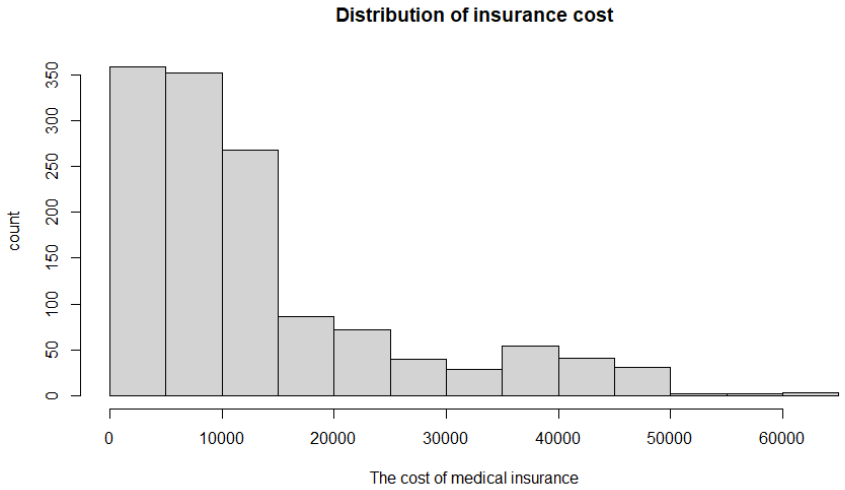
Introduction

With the advancement of medical technology, the range covered by medical insurance programs keeps expanding. Simultaneously, the price fluctuation interval of medical insurances is widening as well. In response to this phenomenon, we try to investigate which factors will influence the costs of health insurance mostly. Moreover, the relationship between different age groups and insurance costs whether depends on smoker is part of our analyze. We hope that this analysis will help those who want to estimate the cost of their health insurance.

Data Description

Response Variable:

Charges: The cost of medical insurance

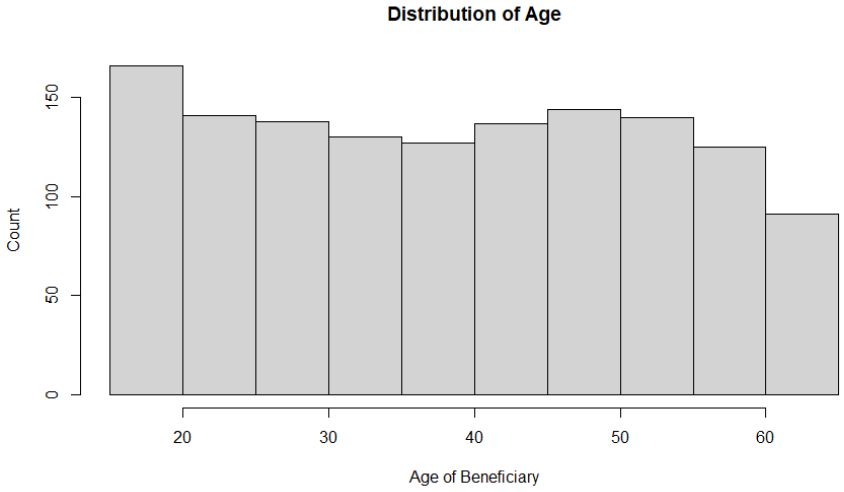


Min.	Median	Mean	Max.
1122	9386	13272	63770

***NOTE:** Charges does not follow normal distribution and we will make a data transformation in further steps.

Explanatory Variables:

1. Age: Age of beneficiary

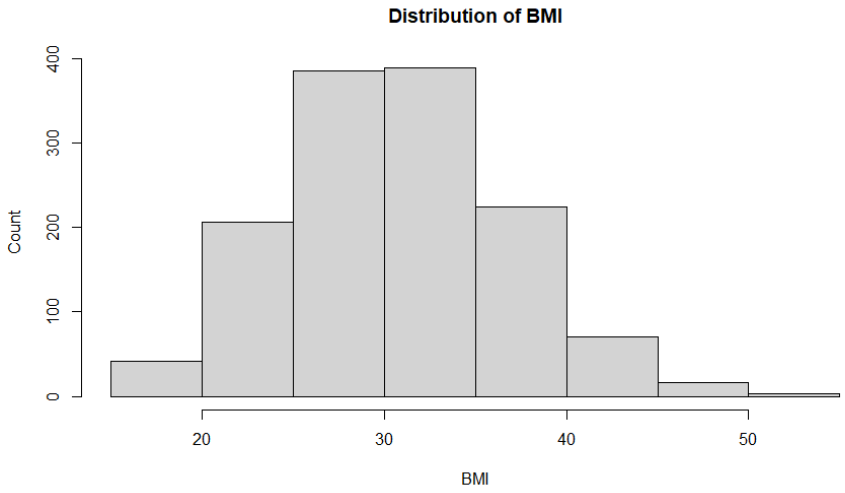


2. Sex: Gender of beneficiary

662 females, 677 males (female=1, male=2)

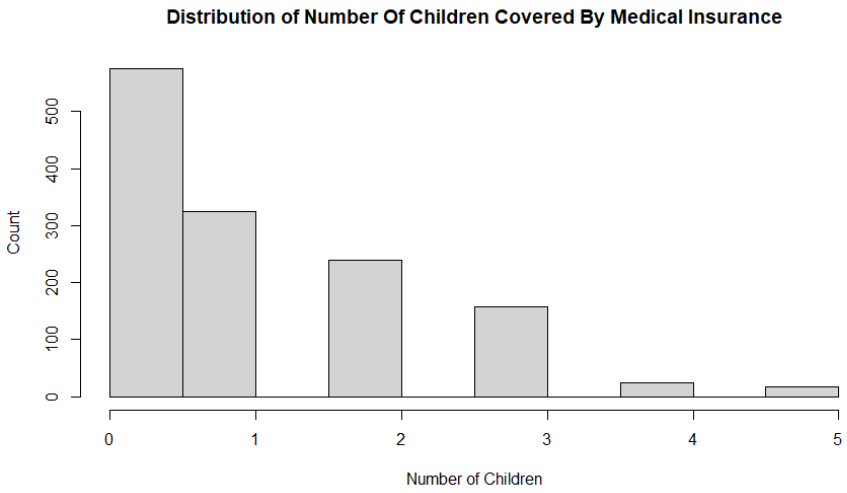
3. BMI: Body mass index.

Body mass index (**BMI**) is a value derived from the mass (weight) and height of a person. The BMI is defined as the body mass divided by the square of the body height and is universally expressed in units of kg/m^2, resulting from mass in kilograms and height in meters. [1] Commonly accepted BMI ranges are underweight (under 18.5 kg/m^2), normal weight (18.5 to 25), overweight (25 to 30), and obese (over 30). [2]

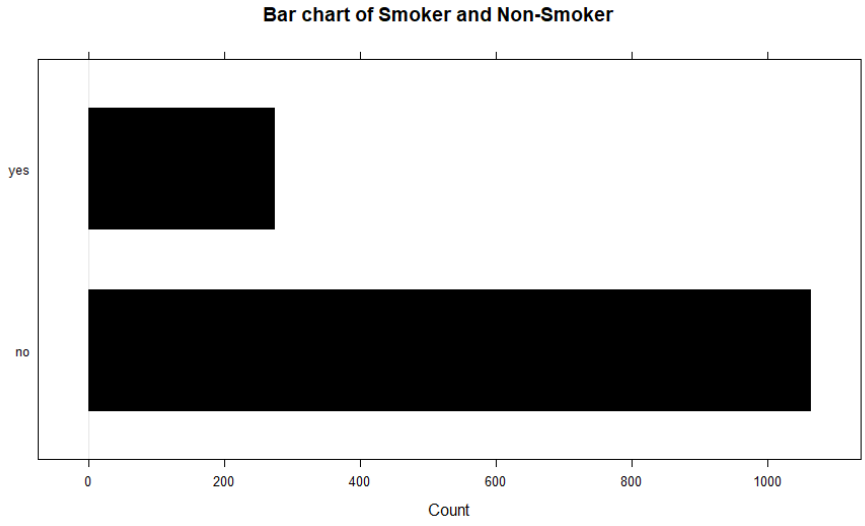


Min.	Median	Mean	Max.
15.96	**30.40**	**30.65**	**53.13**

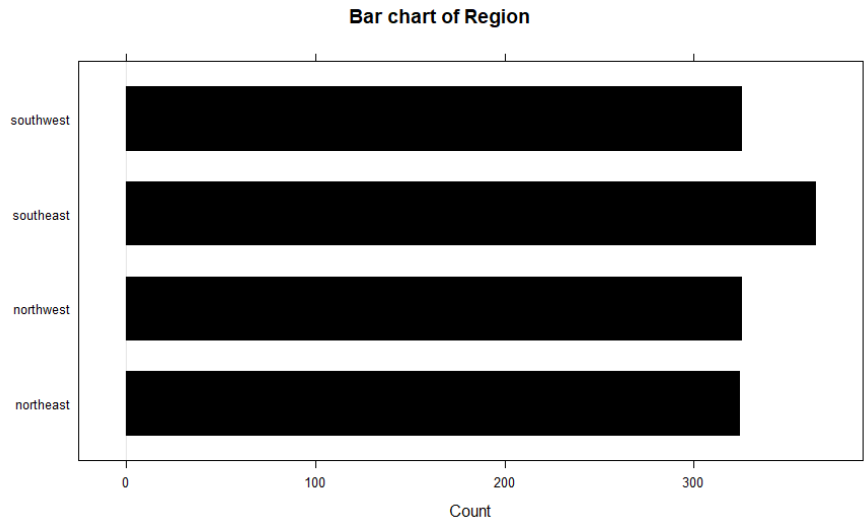
4. Children: Number of children of beneficiary covered by medical insurance



5. Smoker: Smoker and Non-Smoker (yes=1, no=0)



6. Region: The residential area of beneficiary



Additional Data Point

Age	Sex	Bmi	Children	Smoker	Region	Charges
22	male	17.32	0	Yes	Southwest	16003.29

We created a data point which is a young man with lower BMI and no children. We suspected that the man got lower BMI due to take drugs which lead his insurance cost higher.

Methods

➢ Check missing value in dataset

To ensure the accuracy of data analysis, we checked the missing value of dataset firstly. Fortunately, there were no missing values in this dataset.

➢ The use of Indicator Variables

In this paper, we want to investigate the relationship between age of the beneficiary and cost of medical insurance whether depends on smoker. We want to use indicator variable and the interaction term to answer this question.

From this diagram, we observed that smoker needed to pay more money for medical insurance than non-smoker at the same age.
Therefore, we used smoker as an indicator variable to investigate whether the relationship between age and charges depends on smoker.

Firstly, we build a model which only contained smoker and an interaction model:

$Y_1 = \beta_0 + \beta_1 X_1 + \varepsilon$

$Y_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) + \varepsilon$

X_1	X_2	$X_1 * X_2$
Smoker=1 Non-smoker=0	Age	Interactor

Then we set a hypothesis to test our analysis.

H_0: *Smoker does not matter*	**H_a**
$\beta_1=\beta_3=0$	$\beta_1 \neq 0$ or $\beta_3 \neq 0$

We will use F-test to confirm our conjecture in further steps.

➢ Variable Selection

People usually do not understand which factors influence their price of medical insurance. Thus, we wanted to investigate which predictors impact the charges mostly.
We used stepwise regression method and set Akaike's Information Criterion (AIC) as criteria in "backward" and "both" these two directions.

Full Model: contain ALL predictors

$Charges=\beta_0+\beta_1*age+\beta_2*sex+\beta_3*bmi+\beta_4*children+\beta_5*smoker+\beta_6*region+\varepsilon$

After stepwise model selection, we got a model as the result. But we needed to do cross validation to check in a further step to avoid overfitting in this model which was selected.

➢ Cross-Validation

We got a model which selected by model selection, we called this model M_1.
(Predictor ***Sex*** has been removed)
$M_1=\beta_0+\beta_1*age+\beta_2*bmi+\beta_3*children+\beta_4*smoker+\beta_5*region$

During this cross-validation, we split the total observations(n=1339) into training and testing datasets.

Training set	**Testing set**
Used to build the model	Used to test the model by prediction error

We used 70% of observations as training and 30% as testing.
To ensure M_1 model was not overfitted, we needed to build other training reduce models to check.
Training Model:

Remove region	$M_2=\beta_0+\beta_1*age+\beta_2*bmi+\beta_3*children+\beta_4*smoker+\varepsilon$
Remove smoker	$M_3=\beta_0+\beta_1*age+\beta_2*bmi+\beta_3*children+\beta_4*region+\varepsilon$
Remove children	$M_4=\beta_0+\beta_1*age+\beta_2*bmi+\beta_3*region+\beta_4*smoker+\varepsilon$

Remove bmi	$M_5=\beta_0+\beta_1*age+\beta_2*region+\beta_3*children+\beta_4*smoker+\varepsilon$
Remove age	$M_6=\beta_0+\beta_1*region+\beta_2*bmi+\beta_3*children+\beta_4*smoker+\varepsilon$

Then we made predictions and computed the R-square, RMSPE and MAPE for each training model. In order to reduce the error caused by random sampling, we repeated above process 10 times and used average of each R-square, RMSPE and MAPE as the final result.
Finally, we compared each R-square, RMSPE and MAPE of six models and selected a model which had the largest R-square and smallest RMSPE/MAPE as the final model.

➢ Robust Regression

The final model we selected was not satisfied with the multiple linear regression assumptions after we tested. Therefore, we were trying to use robust regression with Huber's t function to overcome this situation. We hoped to reduce the weight of outliers by robust regression, so as decreasing the influence of outliers on the model which we selected. After fitted a robust regression model, we plotted residuals versus fitted, Normal Q-Q Plot and residuals versus leverage to check whether we met the assumptions.

➢ Data Transformation

We used data transformation to satisfy the assumptions of multiple linear regression.

Predictors	Origin	Transformed
Charges	Charges	Log(charges)
Age	Age	1/Age

After transformed, we plotted distribution of residuals, residuals versus fitted and residuals versus leverage again. The results will present on further steps.

Results

➢ Indicator Variable

We interested in the relationship between age and charges whether depends on smoker.
Therefore, we used smoker as an indicator variable and set a hypothesis for analysis.

H_0: *Smoker does not matter*	**H_a**
$\beta_1=\beta_3=0$	$\beta_1\neq0$ or $\beta_3\neq0$

We used F test for this hypothesis, the result came with ***Table 1*.**

Model	**F**	**P-value**
Y_2	245.66	2.2e-16 ***

Table 1

From ***Table 1*** we observed that P-value closed to zero. Therefore, we had significant evidence to reject null hypothesis, which means smoker matters.

➢ Variable Selection

We used “backward” and “both” directions to select the final model. We set AIC as criteria to select.

Table 2: Backward

Variables Name	**AIC**
<None>	23329
Removed Region	23333
Removed Children	23339
Removed BMI	23466
Removed Age	23732
Removed Smoker	25014

Table 3: Both

Variables Name	**AIC**
<None>	23329
Added Sex	23331

Removed Region	23333
Removed Children	23339
Removed BMI	23466
Removed Age	23732
Removed Smoker	25014

Table 2 was the final step of "backward" method.
Table 3 was the final step of "Both" method.

From ***Table 2*** we noticed that when we removed any variable, AIC became larger. By contract, when we stopped to remove, AIC kept the smallest. If we removed smoker, AIC became largest. Therefore, we can assume smoker is the most influential predictor.
From ***Table 3*** we noticed that when we added sex into model, AIC became larger rather than did nothing. Thus, we removed predictor sex which means sex was the smallest influential predictor.

These two methods of model selection selected the same model M_1 as the final model.
$\mathbf{M_1}=\boldsymbol{\beta_0+\beta_1*age+\beta_2*bmi+\beta_3*children+\beta_4*smoker+\beta_5*region+\varepsilon}$

➢ Cross-Validation

In order to avoid overfitting in M_1, we used cross-validation to continue select predictors.

Training Model:

Remove region	$M_2=\beta_0+\beta_1*age+\beta_2*bmi+\beta_3*children+\beta_4*smoker$
Remove smoker	$M_3=\beta_0+\beta_1*age+\beta_2*bmi+\beta_3*children+\beta_4*region$
Remove children	$M_4=\beta_0+\beta_1*age+\beta_2*bmi+\beta_3*region+\beta_4*smoker$
Remove bmi	$M_5=\beta_0+\beta_1*age+\beta_2*region+\beta_3*children+\beta_4*smoker$
Remove age	$M_6=\beta_0+\beta_1*region+\beta_2*bmi+\beta_3*children+\beta_4*smoker$

***Note: Each R^2, RMSPE and MAPE were the averages of 10 times calculated.**

Training Model	R^2	RMSPE	MAPE
M_1: Final Model	0.7541	5993.825	4155.836
M_2: Remove region	0.7432	6194.076	4271.347
M_3: Remove smoker	0.1372	11434.38	9022.574
M_4: Remove children	0.7518	6080.456	4230.362
M_5: Remove bmi	0.7317	6268.288	4078.475
M_6: Remove age	0.6735	7054.189	5379.976

***Sorted** by RMSPE from smallest to largest.*

Training Model	R^2	RMSPE	MAPE
M_1: Final Model	0.7541	5993.825	4155.836
M_4: Remove children	0.7518	6080.456	4230.362
M_2: Remove region	0.7432	6194.076	4271.347
M_5: Remove bmi	0.7317	6268.288	4078.475
M_6: Remove age	0.6735	7054.189	5379.976
M_3: Remove smoker	0.1372	11434.38	9022.574

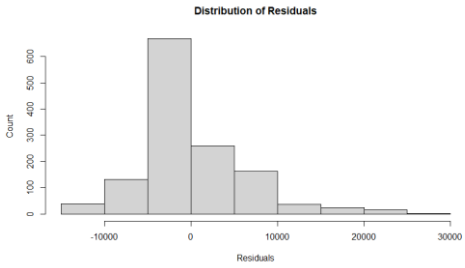
From this sorted table, we found M_1 had the largest R^2 and lowest RMSPE.
The result of cross-validation was same as we did in model selection. Meanwhile, we noticed when we removed smoker, the training model only explained 13.72% which means smoke is the most influential predictor.

➢ Check Assumptions

i. The relationship between the response y and the regressors is linear, at least approximately.
ii. The error term ε has zero mean.
iii. The error term ε has constant variance σ^2.
iv. The errors are uncorrelated.
v. The errors are normally distributed

Final model

$$M_1=\beta_0+\beta_1*age+\beta_2*bmi+\beta_3*children+\beta_4*smoker+\beta_5*region+\varepsilon$$



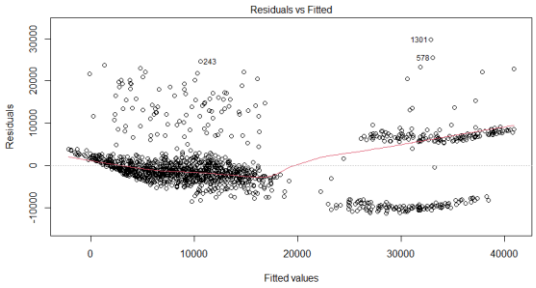
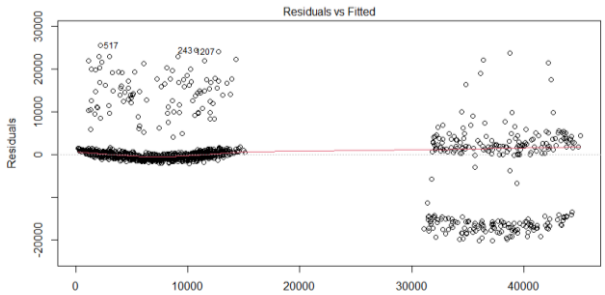


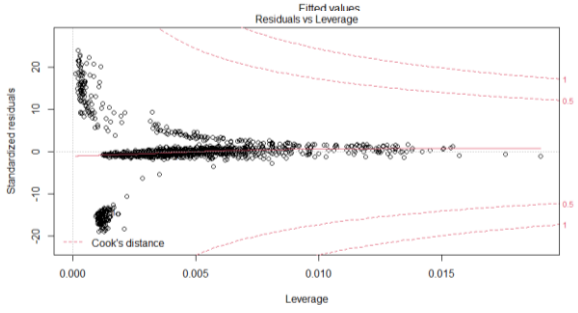
Figure 1

We used M_1 as the final model to plot diagrams of residuals. Unfortunately, from ***Figure1*** we can observe that distribution was not perfectly normal distributed (***Assumption Five***×). Meanwhile, on the right part of graph errors were not distributed among zero (***Assumption Two***×). Therefore, this model was not satisfied by the assumptions and it was a non-linear relationship between charges and predictors. Thus, we attempted to use robust regression reduced the impact of outliers and data transformation in further steps.

➢ Robust Regression

We constructed a robust regression model by M_1 with Huber's t function and plotted ***Figure2***.





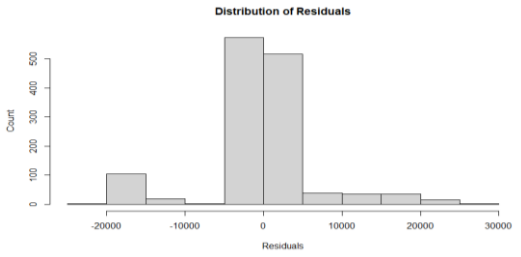
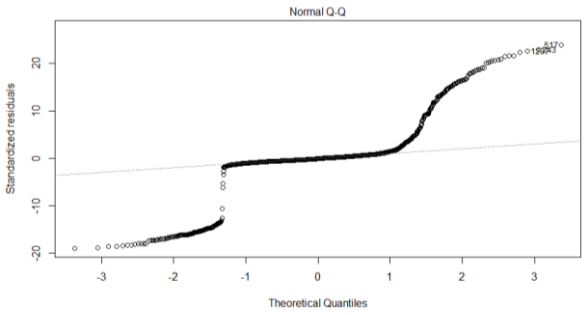


Figure2



From ***Figure2,*** we found distribution of residuals looked more normally distributed

than ***Figure1***, but there were two peaks. In residuals versus fitted' s plot, although red line going straighter than ***Figure1***, data points were widely dispersed. There was a perfect linear between -1 to 1 in Normal Q-Q plot, but it looked uncomfortable on the side part. Therefore, we still cannot use this model. Finally, we wanted to try the method of data transformation to check if we can get better plots and satisfy the assumptions.

➢ Data Transformation

Predictors	Origin	Transformed
Charges	Charges	Log(charges)
Age	Age	1/Age

Transformed Model in Robust Regression with function Huber's t:

$$Log(charges)= \beta_0+\beta_1*1/age+\beta_2*bmi+\beta_3*children+\beta_4*smoker+\beta_5*region+\varepsilon$$

When we transformed predictors, we got plots on ***Figure3***.

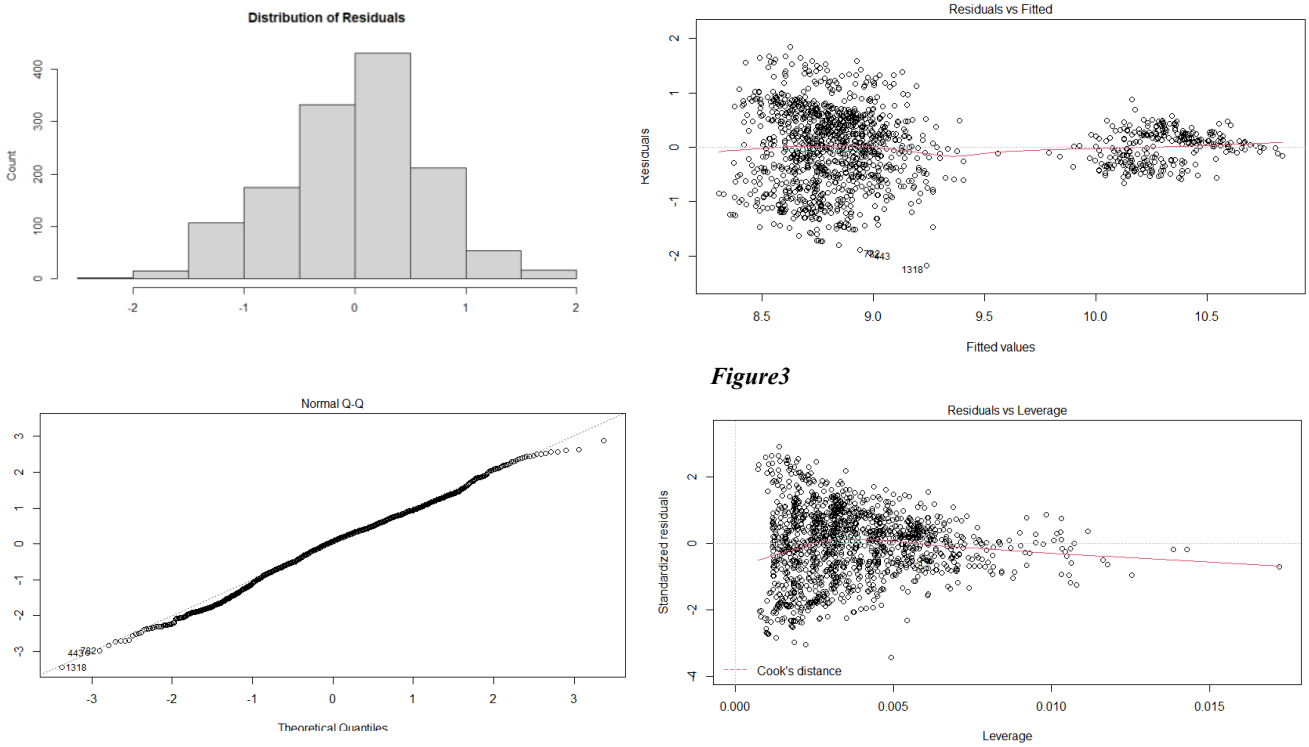


Figure3

From ***Figure3***, the distribution of residuals distributed almost perfectly in normal. As same as the Normal Q-Q plot, almost perfectly linear. In residuals versus fitted plot, we can observe red line almost keep on zero, which means satisfying the constant variance assumption for errors. Overall, the new transformed model satisfied the assumptions for regression analysis and the plots look pretty good. Finally, we did it!

➢ Check Multicollinearity

Moreover, we checked multicollinearity between predictors. Fortunately, there was no multicollinearity issue between explanatory variables, because of all VIF values closed to 1 and no more than 10.

Explanatory Variables	VIF
Smoker	1.000040
bmi	1.024758
Children	1.000427
Region	1.024802

Conclusion

We used a variety of analytical tools to draw conclusions about what factors affect the cost of health insurance. One of the key factors is whether the beneficiary is a smoker or not. We realized that smoking is not only hazardous to our health, but also burdensome to our insurance premiums! Moreover, we noted that both the mean and median of BMI reach around 30, which means that the majority of people in this dataset have health risks due to high BMI. We want people to live a high quality of life while maintaining a regular and healthy diet. A healthy body will not only keep you productive, but it will also reduce the amount of money you spend on medical care. Especially during this difficult phase of COVID-19, we need to pay more attention to our health status. In conclusion, stay safe, healthy and well, we navigate the pandemic together!

Appendix

In GitHub “Appendix” file: https://github.com/vincely-zz/STAT350-Final-Project/tree/main/Appendix

Dataset from:https://www.kaggle.com/mirichoi0218/insurance

Reference

[1] “Body Mass Index” *Wikipedia*, Wikimedia Foundation, https://en.wikipedia.org/wiki/Body_mass_index.

[2] *"WHO Mean Body Mass Index (BMI)". World Health Organization. Retrieved 5 February 2019.*

Cover Photo: *https://www.aarp.org/health/health-insurance/info-2019/health-care-cost-concerns-middle-aged-adults.html*