Analysis of factors related to health insurance premiums by linear regression

by

M.S., Wenxiao Zhou, Haoxi Ma

University of Connecticut, 2020

Project Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Statistics

in the

Department of Statistics

Advisor: Haiying Wang

Spring 2020

ABSTRACT

In today's society, purchasing insurance has become a mainstream trend, how well insurance companies can help their customers with the best plans to ensure a win-win is practical. In this report, we aim to build a linear model between the medical insurance charges for customers and their history information (Age, Sex, BMI, Children, Smoker and Region). Based on such statistical model, the company can fee insureds appropriate premiums and give them fitted insurance plans.

First, we draw scatterplot to give an intuitive mode of the influences and use stepwise model selection, AIC model selection to determine the “best” MLR model. Then, we do diagnose and make some remedies towards original model. What’s more, we also check the validation of the model and give some comments.

Besides, we extend to Random forest and Neural Networks to fit model and find that the Neural Networks method has the highest R^2 among three models.

Key words:

Multiple linear regression, Model selection, Diagnose and remedies in MLR, Random forest, Neural Networks

BACKGROUND

With the development of society, people's risk awareness is increasing. Under this condition, insurance has become more and more popular. In order to make profit and achieve win-win, it is important for insurance industry to determine insurance premium for different people.

After collecting and summarizing lots of historical user’s information, insurance company can build the statistical model on personal charge (medical costs billed by health insurance) with a system of the customers’ information. After analyzing the behaviors as well as their purchase ability, according to the possible future payment, the company can fee insureds appropriate premiums and give them fitted insurance plans.

This determination of premiums helps insurance companies in enhanced pricing, underwriting and risk selection. Besides, it is also helpful in making better decisions, understanding customer needs and be fair to the customers. As time goes by, statistic technology will be widely applied in insurance field.

MOTIVATION

After searching the relevant studies and researches regarding the insurance charges analyses, we find out that most of the studies prefer to just fit multiple linear regression model with several variables, however, without further determination on the model selection and diagnose. Such predictions are not persuasive and lack accuracy in prediction, which can easily lead to errors. Thus, we will focus more on model selection and validation with comparison of different statistical index such as $R^2, AIC, MSPR$ and so on. What's more, we will use some other algorithms to fit models and compare them with MLR.

Main goals:

(i) Determine which factor influences the premium significantly.

(ii) Select and construct the "best" MLR model

(iii) Do some diagnose and remedies (if needed)

(iv) Use Neural Network algorithm and Random Forest algorithm to fit regression models and make comparisons.

DATA DESCRIPTION

The dataset is available online through the GitHub repository called Machine Learning with R and the link is:

https://github.com/stedy/Machine-Learning-with-Rdatasets/blob/master/insurance.csv

This Dataset has1338 rows and 7 columns corresponding to 7 variables:

Age - age of primary beneficiary: continuous variable

Sex - gender of beneficiary: two categories—Female, Male

BMI - body mass index, defined as $\frac{kg}{m^2}$ (ideally 18.5 to 24.9): continuous variable

Children - number of children covered by health insurance: six categories

Smoker - whether beneficiary smoke: two categories—Yes, No

Region -the residential area in the US: four categories—northeast, southeast, southwest, northwest

Charges - medical costs by health insurance: continuous response variable.

Data summary:

Table 1 – Data summary

	Data type	Missing value
Age	Integer	None
Sex	Factor	None
BMI	Numeric	None
Children	Integer	None
Smoker	Factor	None
Region	Factor	None
Charges	Numeric	None

MODELS

1. Multiple Linear Regression model

The formula is:

$$\boldsymbol{Y} = \beta \boldsymbol{X} + \alpha \boldsymbol{D} + \gamma \boldsymbol{D}\boldsymbol{X} + \varepsilon$$

Where

$\boldsymbol{X}\ represents\ continous\ covariate$

$\boldsymbol{D}\ represents\ categorical\ covariate$

$\boldsymbol{DX}\ is\ their\ interaction\ term$

Assumptions for model:

(i) Linearity

(ii) Constant Variance

(iii) Uncorrelated error terms

(iv) Normality of error terms

Note: the last assumption is constructed for statistical inference (such as CI, PI) and parameter estimation. This means if normal assumption is satisfied, Least Square method can obtain asymptotic minimum variance or we should use Maximum Likelihood method. But, under large sample size, the last assumption is not as critical as the others.

2. Random Forest model

We let Age, Sex, BMI, Children, Smoker and Region be the features of this dataset and build a random forest with 10000 trees.

Assumption: All features mutually independent

3. Neural Networks model

We build a two layers Neural Net with 8 nodes in deeper layer, 4 nodes in shallower layer and give one output.

Assumption: data is well arranged and has no missing value

EXPLORATORY DATA ANALYSIS

1. MLR

In this party, we will combine graphic and mathematic methods to identify a "best" MLR and check its validation, do some diagnose and remedies.

1.1 Find fundamental model

To reduce computation and improve efficiency, we should give an intuitive model based on scatterplots, which can help make intuitive judgements on the relationship between response and covariates.

First, we draw scatterplots for continuous covariates Age, BMI vs Charges respectively:

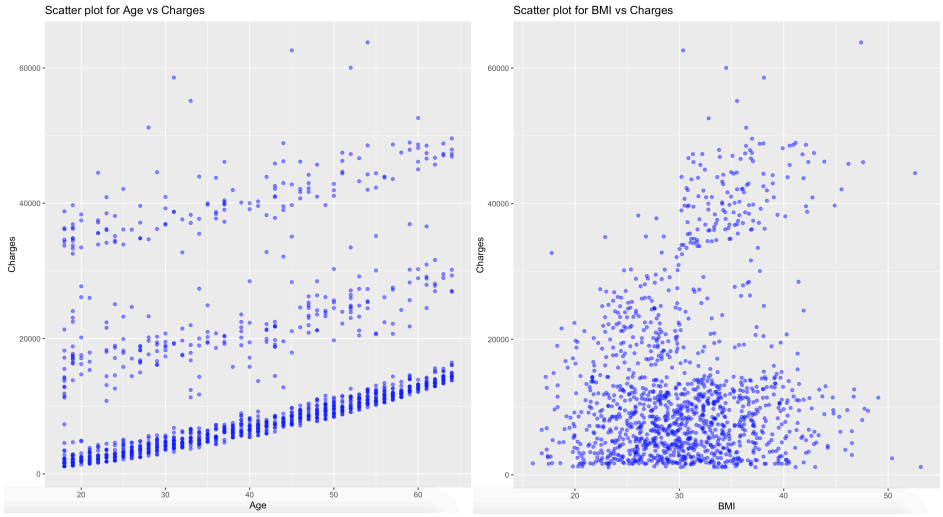
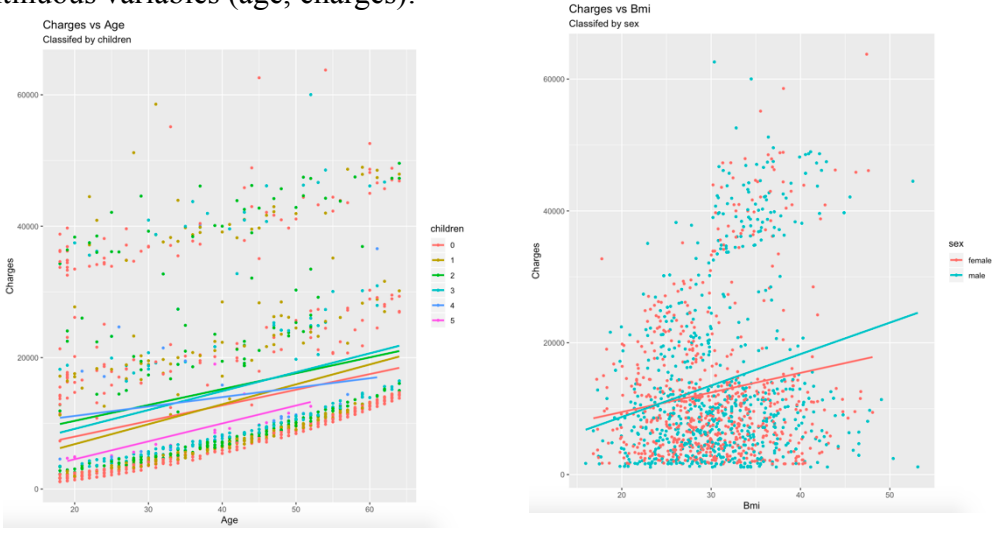


Figure 1 – Scatter plots for Charges vs Age/BMI

From Figure 1, there is linear relationship between Charges and Age, BMI respectively. Then, we decide whether to add categorical variables and their interaction terms with continuous variables into original model. If the fitted lines in different categories are parallel, we consider there is no interaction between categorical variable with continuous variables (age, charges):



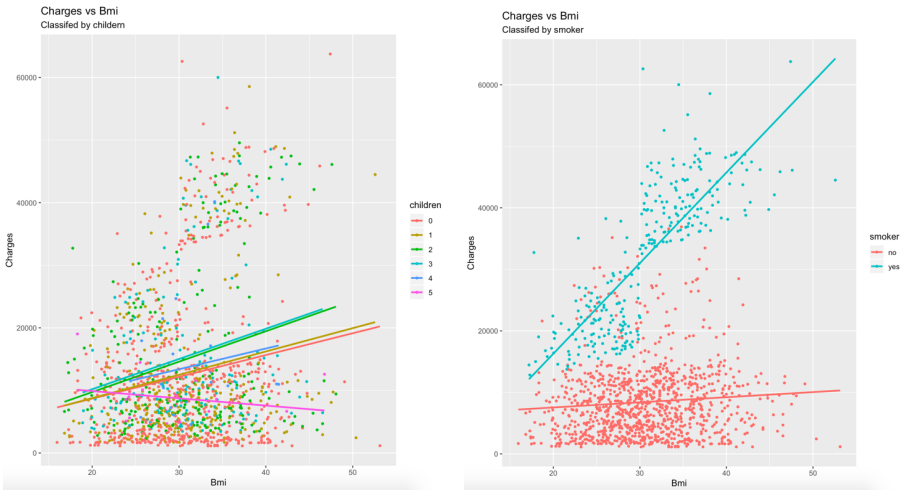


Figure 2 – Plots for Chagres vs Age/BMI group by categorical variables

We just show graphs indicating the interactions are significant above. You can find other graphs in Appendix-F-1. According to all the scatterplots, we will add interaction terms: $Age * Children, BMI * Sex, BMI * Children$ and $BMI * Smoker$ into original model.

After that, we also check the interaction between two continuous variables Age and BMI by dividing Age into three different levels: Young, Medium and Old.

Figure 3 – Plots for Chagres vs BMI group by three levels Age

Therefore, we ignore the interaction of these two variables.

All in all, we assume the fundamental linear model:

$$charges \sim age + sex + bmi + children + smoker + region + bmi * sex + age * children + bmi * children + bmi * smoker$$

Where

$A * B$ *represents interaction between* A *and* B

1.2 Model selection

First of all, we draw histogram plot for response variable Charges:

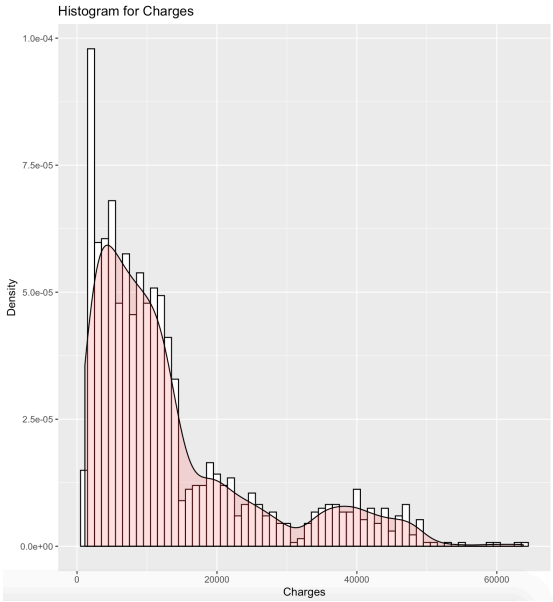


Figure 4 – Histogram for Charges

Figure 4 tells us Charges isn't corresponding to a normal distribution and since $Var(charges) \propto [E(charges)]^2$, perform a Log Transformation is a good choice which can help improve the model performance.

Before doing model selection, we can check multicollinearity by drawing a correlation plot and calculating VIF.

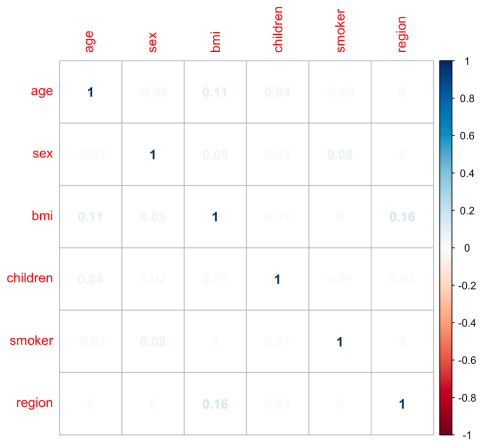


Figure 5 – Correlation plot for six main covariates

From Figure 5 and VIF (see in Appendix-F-2), multicollinearity doesn't exist.

Then we divide dataset into two part – train dataset (75% observations from original dataset) and test dataset (25% observations from original dataset). And use train dataset to do model selection.

First, we use Stepwise method with each 0.05 significant level joining and dropping elements. Following is a table of procedure.

Table 2 – Table for Stepwise model selection

	Join elements	Drop elements
Step 1	Smoker	None
Step 2	Age	None
Step 3	Children	None
Step 4	Age*Children	None
Step 5	BMI	None
Step 6	BMI*Smoker	None
Step 7	Region	None
Step 8	Sex	None

What's more, we also do AIC model selection and the output is attached in Appendix-F-3. These two methods come to a same "best" model:

$$charges \sim age + sex + bmi + children + smoker + region + age * children + bmi * smoker$$

Using ANOVA Type III error to recheck this model (see in Appendix-F-4), all factors are significant besides BMI and Smoker. However, since the interaction of $bmi * smoker$ is significant, in order to preserve hierarchy of the model, we add these two non-significant single variables.

Above all, we get the "best" model and it is:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 S_1 + \beta_4 C_1 + \beta_5 C_2 + \beta_6 C_3 + \beta_7 C_4 + \beta_8 C_5 + \beta_9 M_1 + \beta_{10} R_1 + \beta_{11} R_2 + \beta_{12} R_3 + \beta_{13} C_1 X_1 + \beta_{14} C_2 X_1 + \beta_{15} C_3 X_1 + \beta_{16} C_4 X_1 + \beta_{17} C_5 X_1 + \beta_{18} M_1 X_2$$

Where

$Y\ is\ Log\ transformed\ Charges$

$X_1\ is\ Age$

$X_2\ is\ BMI$

$Dimmy\ variables\ are\ defined\ in\ Appendix - F - 5$

1.3 Diagnose and remedy

We check the normality of the residuals first using QQ-norm plot and Shapiro-Wilk test. Following is the QQ-plot which seems affliction of normality assumption.

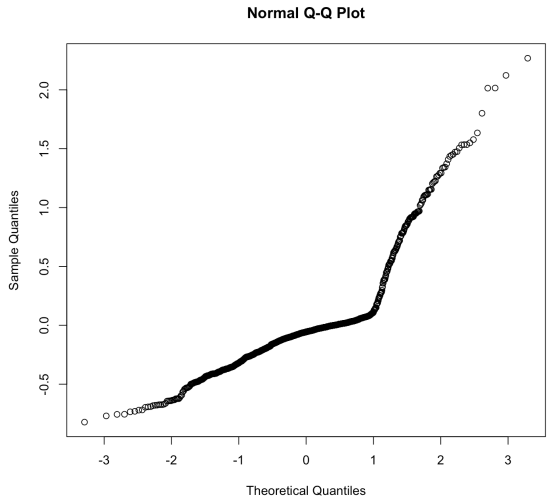


Figure 6 – QQplot for residuals

Since $P - value$ of Shapiro-Wilk test is less than 0.05 (see in Appendix-F-6), we conclude that normality assumption is not satisfied. Under this condition, in order to improve precision of estimator (mainly to reduce the variance), we will do Alternative Box-Cox Transformation on Charges, the response variable. The Power we select is showed in Appendix-F-7.

After transformation, we draw QQ-plot again. Unfortunately, the non-normal property still exists. We do lots of thing to fix it such as adding new predictors, interaction terms and quadratic terms and they all fail. But thanks to the large sample size, normality assumption is not as critical as the other assumption.

Then, because errors don't follow a normal distribution, we choose Brown-Forsythe Test to check constant variance assumption and get the result:

Table 3 – BrownForsythe Test result

t^*_{BF}	$P-value$
-0.4027716	0.6872022

$P-value$ is larger than 0.05, so we conclude constant variance assumption is satisfied.

At last, we are going to check whether Autocorrelation exists by Durbin-Watson Test. Result is showing below (also attached in Appendix-F-8):

Table 4 – DurbinWatson Test result

$D-W$ Statistic	$P-value$
2.007969	0.92

Because $P-value$ is larger than 0.05, we conclude that Autocorrelation doesn't exist which indicates uncorrelated error assumption achieves.

1.4 Check validation

As mentioned before, we have divided the original dataset into train and test datasets. We compute

$$R^2(1) = r^2(Y_1, \hat{Y}_1)$$

for the train group, where $r(Y_1, \hat{Y}_1)$ is the Pearson product-moment correlation between the observed Y_{1i}'s and fitted $\hat{Y}_{1i}$'s.

Then, use the fitted model from the training group to compute predictions $\hat{Y}_2^*$ for the test group corresponding to observed Y_2. Then compute their squared Pearson product moment correlation

$$R^{2*}(2) = r^2(Y_2, \hat{Y}_2^*)$$

After that, we pay attention to the Shrinkage statistic:

Table 5 – Shrinkage on cross validation

$R^2(1)$	$R^{2*}(2)$	$R^2(1) - R^{2*}(2)$
0.8120804	0.8031044	0.008975989

Because $R^2(1) - R^{2*}(2) < 0.1$, we can conclude that the fitted model is reliable.

What's more, we also need to measure model's Predictive Ability by calculate $MSPR$ and compare it with model's MSE.

Table 6 – model's MSE and MSPR

MSE	$MSPR$
0.1667906	0.1798896

According to Table 6, $MSPR$ is fairly close to MSE based on the regression fir to the model-building dataset, then the MSE for the selected regression model gives an appropriate indication of the predictive ability of the model. It's obviously that $MSE = 0.1667906$ is very small. Thus, the model preforms fairly well on prediction.

1.5 Summary

Finally, after all procedures, we get our "best" model:

$$\begin{aligned}\hat{Y} = 1.269 &+ 0.034X_1 + 0.004X_2 - 0.082S_1 + 0.181C_1 + 0.574C_2 + 0.493C_3 \\ &+ 1.22C_4 + 0.375C_5 - 0.166M_1 - 0.096R_1 - 0.13R_2 - 0.127R_3 \\ &- 0.002C_1X_1 - 0.009C_2X_1 - 0.007C_3X_1 - 0.017C_4X_1 - 0.001C_5X_1 \\ &+ 0.06M_1X_2\end{aligned}$$

Where

Y is Log and Box-Cox transformed Charges

Others are same as mentioned before

Interpretations:

(i) Coefficients for continuous variable term (we take $b_1 = 0.034$ as an example):

When all other factors fixed, transformed fitted Charges increases 0.034 as per unit increasing in Age which means older are the insureds more premiums they should pay.

(ii) Coefficients for categorical variable term (we take $b_3 = -0.082$ as an example):

When all other factors are same, intercept of models in Male and Female are -0.082 difference. This means male will pay less than female when all other identities are same.

(iii) Coefficients for interaction term (we take $b_{18} = 0.06$ as an example):

when all other factors fixed, slope of models in Smoke and Nonsmoker are 0.06 difference. This means if insureds smoke, they will pay more premiums per unit increasing in BMI compared to who don't smoke.

2. Random Forest

Because of independent features assumption and according to the interpretation in Figure 5, we only include six factors (Age, BMI, Sex, Children, Smoker and Region)

in model:

$$charges \sim age + sex + bmi + children + smoker + region$$

Then we build a Random Forest with 10000 trees (output seen in Appendix-F-9) and draw a Variable Importance Plot:

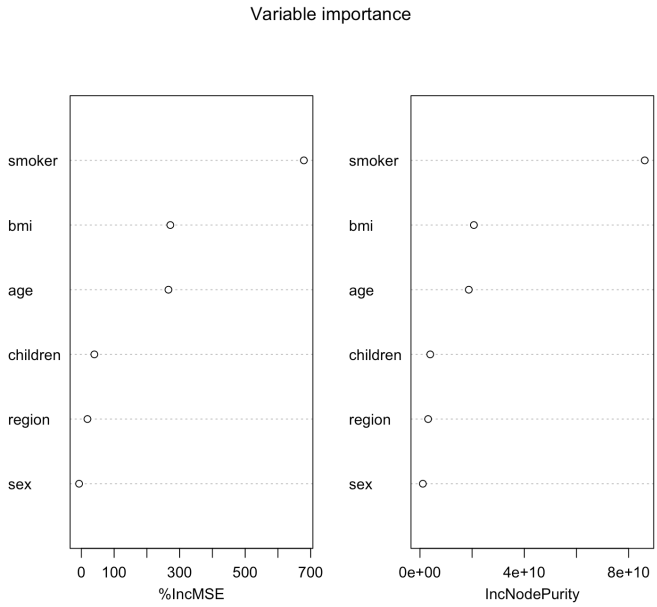


Figure 6 – Variable Importance Plot

According to Figure 6, it's obvious that Smoker influence Charges most significantly. Thus, smoking less helps save money!

3. Neural Networks

We first scale all continuous variables and check missing values, then build a two-layers Neural Net with 8 nodes in deeper layer, 4 nodes in shallower layer and give one output. Following is plot for Neural Networks:

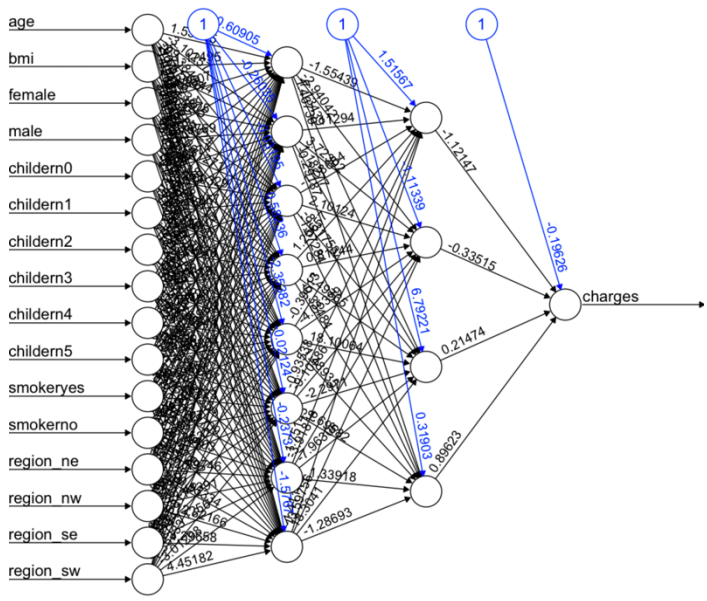


Figure 7 – Neural Networks for linear regression

DISCUSSION

Above all, we use three models to do linear regression and we are going to make some comparisons among them in this part.

We summarize R^2 in following table:

Table 7 – R^2 for three models

MLR	Random Forest	Neural Networks
0.8120804	84	0.9009527

From Table 7, we know that Neural Networks does best in linear regression of this dataset. We also using graphic way to show this:

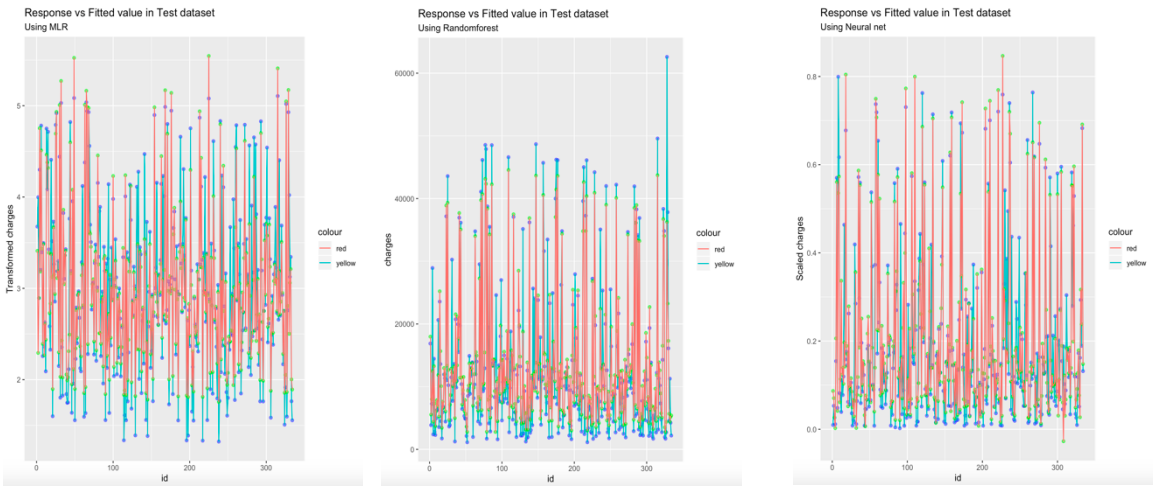


Figure 8 – Prediction vs observed value in three models

Figure above comes to the same conclusion as judging by R^2.

At last, we want to argue that there is an insufficiency left. That is, although having tried lots of methods to fix the non-normality property, we still fail to refine it. Hope we can solve it in the future as leaning deeper.

ACKNOWLEDGEMENTS

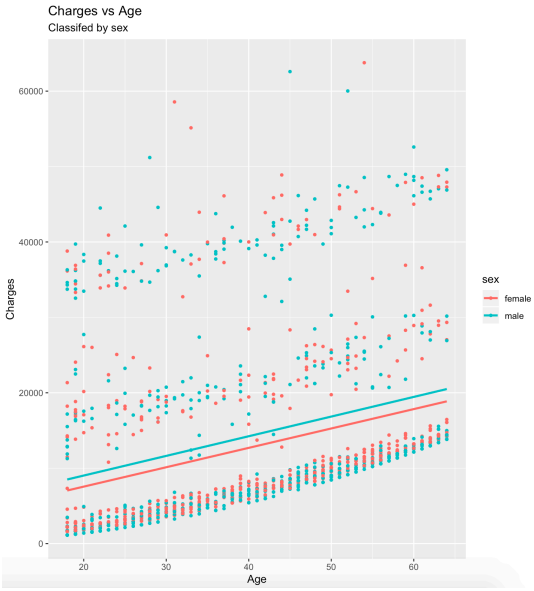
We would like to acknowledge professor Haiying Wang. This project would not have been possible without his constant support and guidance from the beginning till the end. We are highly indebted to him.

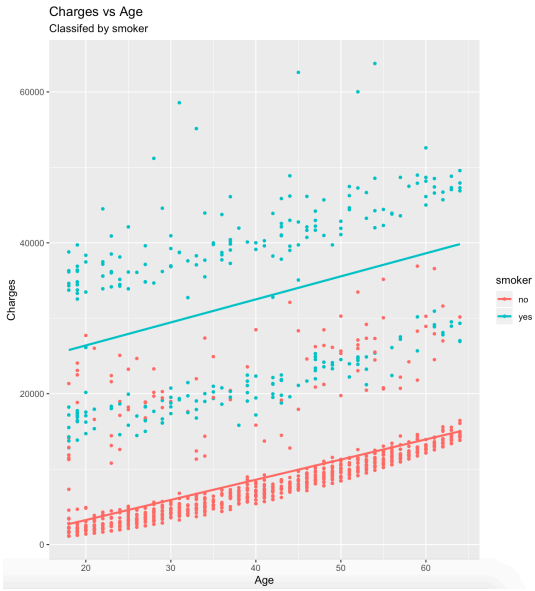
We would also like to extend our gratitude to our family, friends and peers. Additionally, we would also like to thank the graduate department faculty, other professors whose classes we have taken a and learned a lot from them and staff for making our time at University of Connecticut a wonderful and an enriching experience.

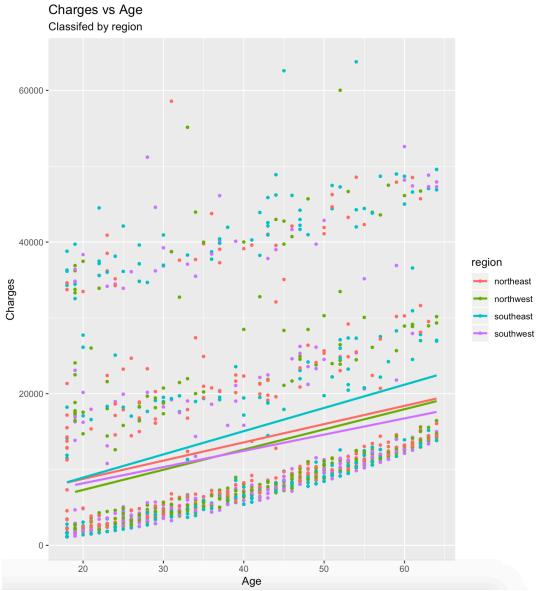
We have learned a lot during my master's journey at University of Connecticut and look forward to applying the knowledge that we have gained from my time here.

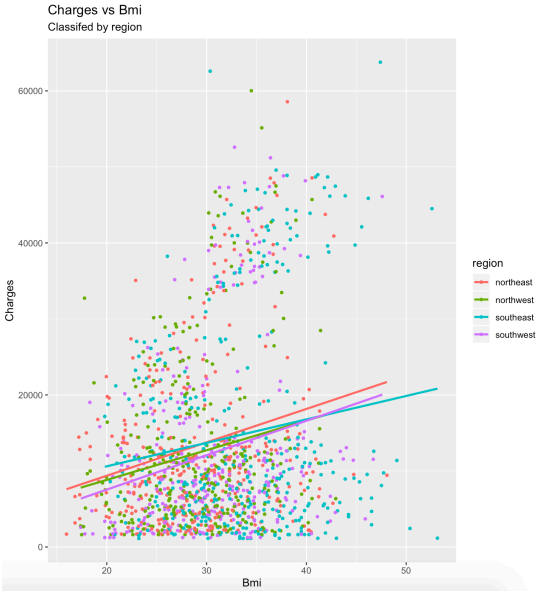
Appendix

F-1:









F-2:

	GVIF	Df	GVIF^(1/(2*Df))
age	1.020607	1	1.010251
sex	1.009334	1	1.004656
bmi	1.108836	1	1.053013
children	1.024871	5	1.002460
smoker	1.018024	1	1.008972
region	1.109919	3	1.017533

F-3:

```
Coefficients:
    (Intercept)              age          sexmale              bmi        children1        children2
       7.182173         0.038960        -0.088659         0.002357         0.286493         0.978903
       children3        children4        children5       smokeryes  regionnorthwest  regionsoutheast
       0.680409         1.396918         0.702599         0.067049        -0.068719        -0.164590
regionsouthwest    age:children1    age:children2    age:children3    age:children4    age:children5
      -0.133330        -0.003354        -0.016948        -0.010575        -0.019799        -0.006223
  bmi:smokeryes
       0.048529
```

F-4:

```
Response: charges
               Sum Sq  Df   F value    Pr(>F)
(Intercept)  1030.25   1 5814.8158 < 2.2e-16 ***
age           159.78   1  901.8193 < 2.2e-16 ***
sex             2.43   1   13.7083 0.0002254 ***
bmi             0.54   1    3.0727 0.0799293 .
children       11.05   5   12.4705 9.354e-12 ***
smoker          0.16   1    0.9049 0.3416989
region          3.25   3    6.1190 0.0004005 ***
age:children    4.92   5    5.5513 4.737e-05 ***
bmi:smoker     14.04   1   79.2239 < 2.2e-16 ***
Residuals     174.52 985
```

F-5:

Sex	S_1
Male	1
Female	0

Children	C_1	C_2	C_3	C_4	C_5
1	1	0	0	0	0
2	0	1	0	0	0
3	0	0	1	0	0
4	0	0	0	1	0
5	0	0	0	0	1
0	0	0	0	0	0

Smoker	M_1
Yes	1
No	0

Region	R_1	R_2	R_3
Northwest	1	0	0
Southeast	0	1	0
Southwest	0	0	1
Northeast	0	0	0

F-6:

```
        Shapiro-Wilk normality test

data:  fit1$residuals
W = 0.80365, p-value < 2.2e-16
```

F-7:

```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    3.1282         3.13       2.5881       3.6682
```

F-8:

```
 lag Autocorrelation D-W Statistic p-value
   1    -0.004163875      2.007969    0.92
 Alternative hypothesis: rho != 0
```

F-9:

```
Call:
 randomForest(formula = charges ~ age + sex + bmi + children +      smoker + region, data = train, importance = TRUE, ntree = 10000)
               Type of random forest: regression
                     Number of trees: 10000
No. of variables tried at each split: 2

          Mean of squared residuals: 23149730
                    % Var explained: 84
```