Prediction of Medical Insurance Cost

ACTU 5840 Predictive Modeling

Professor Lina Xu

Po-An Chen (pc2950)

Zhongwei Wang (zw2731)

Yibing Chen (yc3934)

I. Introduction

As students in actuarial science, we have gained a profound understanding of how risk could bring actual loss to others. For many people, their life can easily be destroyed by a natural disaster or unexpected accident. The value of insurance is to minimize the damage and bounce back. Large Medicare expenses are often a major source of these huge losses. In this report, we are going to quantify the Medicare cost using a linear regression model. The predicted value of the medical insurance cost has a comprehensive application in the field of insurance. For the insured, the predicted medical expenses can help them choose the most cost-effective insurance plan. For insurance companies, expected medical expenses are an important reference for pricing and setting underwriting standards, which motivates our group to choose this topic.

II. Data

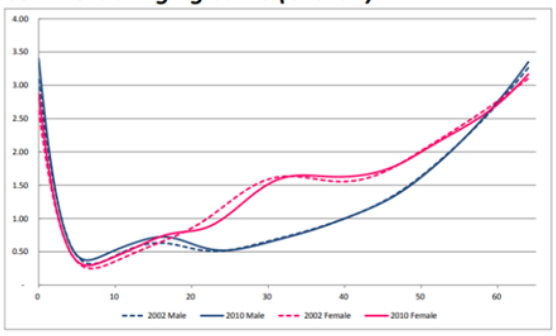
Overview

Our group selected a dataset from Kaggle that consists of 7 variables to study using linear models. Among them, the Medicare cost is a response variable, and the other six variables are explanatory variables, they are age, sex, body mass index, smoking, region, and a number of children. Before proceeding with the processing of the data, we first describe these variables and their relationship to medical expenditures.

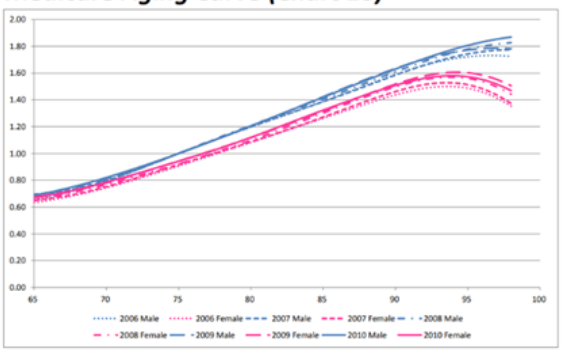
In the real world, age and sex are frequently considered as key factors that affect the expected healthcare cost. We use a graph from the SOA report for illustration. In general, the analysis shows that health care costs increase by age with the exception of the very youngest ages. Costs, on

average, are very high in the first year or two of birth and drop significantly by age five. At that point, costs increased modestly through the teen years. Female costs then begin to accelerate more quickly during child-bearing ages and flatten out in the 40s before increasing again. Male costs are relatively flat in the 20s and begin to accelerate after age 30, but remain lower on a per person basis than females in the same age group. The “cross-over age” occurs in the early 60s, when per capita spending for males exceeds that for females. Medicare costs for beneficiaries age 65 and older continue to increase with age. Males continue to have higher costs than females for whom per person costs start to decline around age 90.

Commercial Aging Curve (Chart 1)



Medicare Aging Curve (Chart 10)



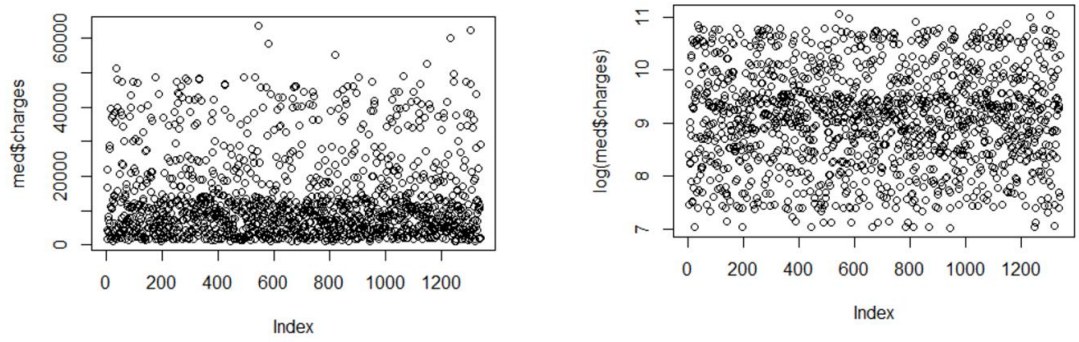
The second factor that might play a significant impact on healthcare spending is BMI, which is the short name of Body mass index. BMI is calculated as weight divided by height squared, where weight is in kilograms and height is in meters. People with a BMI between 18 to 24 are considered normal. People with a BMI over 30 are generally considered obese, and they have a higher chance of developing diabetes, high blood pressure, and fatty liver than normal people, leading to higher healthcare costs. And a BMI that's too low can also cost people's health care, only the difference is they're more likely to suffer from other types of diseases, such as anemia and osteoporosis.

Smoking is also a very important factor affecting medical spending. Various research suggests that a long-term smoker has a 30% probability of developing cancer than non-smokers, especially in oral cancer and lung cancer, which will significantly increase the Medicare cost. In addition, the number of children in coverage is also a factor that affects health care spending. Generally speaking, more children means more medical expenses. The last explanatory variable is region. Different regions may lead to differences in medical expenditure due to factors such as income and medical level. In this dataset, the entire United States is divided into four regions: northeast, southeast, northwest, and southwest.

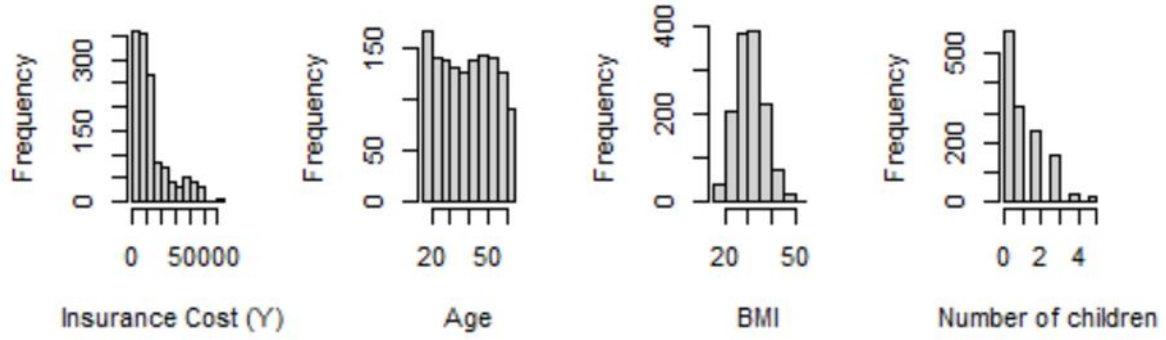
Summary of Data

A general look at all predictors and response variables. From summary statistics, age, BMI, children, and charges are continuous variables. Sex, smoker, and region are categorical variables. According to the summary, several conjectures are made before data visualization. For variable age, the minimum is 18 the maximum is 64, but the first quartile is 27 and the third quartile is 51, so the data are equally distributed. The minimum of BIM is around 16, the maximum is 53 and the median is 30.4, so apparently, BIM is normally distributed. For the number of children, the minimum and first quartile are 0, but the maximum is 5, so this looks more like an exponential distribution. As for the response variable charges, the minimum is 1,122 and the maximum is 63,770, but the median is only 9,382, so the data are highly skewed. As for the categorical variables, sex and region are pretty balanced but the variable smoker is unbalanced.

III. Data Visualization



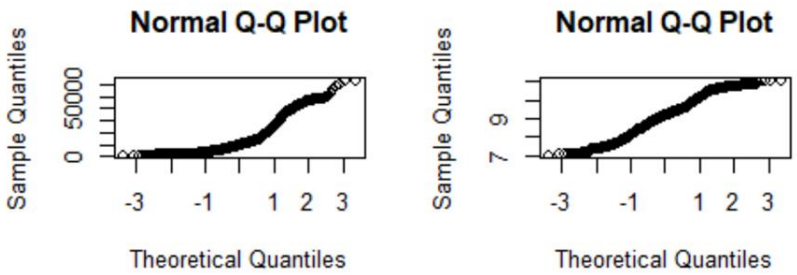
To test the conjecture, the following is the data visualization of the dataset. Here is a scatter plot of variable charges. Based on previous summary statistics and scatter plots variable charges are highly skewed and right-tailed. Thus, log transformation is applied to the variable charge. After applying the log transformation, the data are less skewed and become more valid.



Let's take a further look at the distribution of each variable. Here are the histograms of the continuous variables. The histogram of charges again reinforced the previous conjecture, which is positively skewed. BMI is a normal distribution because in reality most people are healthy and only a small portion of people are overweight and extremely thin. This also implies that the dataset

is randomly sampled. As for age, data are equally distributed and quite stable. The distribution of the number of children is exponential.

Here are the bar plots for categorical variables. Both sex and region are balanced but for variable smokers is unbalanced, which means more people are not smoking today.



To see whether the response variable and log response variable are normally distributed as the linear regression model assumed, a Q-Q-plot is applied to the data. From the first graph, the curve is an S shape crossing the straight line, so we could conclude that normal is not assumed. After applying log transformation, more points now lie on the straight line. Thus, the response variable is normally distributed now.

IV. Models

Model Selection

A full linear regression model is fitted. The residual standard error is 0.4443, and the adjusted r-squared is 0.7666, so the model is not bad. However, in order to improve this model by reducing unnecessarily predictors, an exhaustive search is applied to each subset of models with a maximum subset size of 8 and find the best model in each subset. The comparison of models is based on the MSE and other valuation methods. In the following model selection, three methods to measure MSE are applied. The first one is the full dataset and then is train test splitting methods and the last one is the cross-validation method:

Full dataset method:

Based on the valuation metrics, model 1 has a relatively low adjusted r square. Model 2 has relatively high cp and bic. Compared with other models, model 3 and model 4 are not very complicated and their adjusted r square is similar to the full model. Therefore, under the full dataset method, model 3 or model 4 are under consideration.

Train-test splitting method:

Under the train-test splitting method, error is pretty high for models 1 and 2, but starting from model 3 errors become close to each other. Thus, under the train test splitting methods, model 3 is the best selection.

Cross validation method:

A 10-fold cross-validation method is applied during the model selection. Compared with other models, model 1 has a large error, model 2 is moderate, and model 3 has less error and is not complicated. After all, there is a trade-off made between interpretability and accuracy. Model 3 is chosen since it is not complicated and the accuracy is relatively high. The factors considered in model 3 are age, children, and smokers. The residual standard error is 0.4535 and the adjusted r square is 0.7567, which is pretty close to the full model.

Ridge and Lasso

We also performed Ridge and Lasso regression analysis on the data to complement the results of the previous analysis. Ridge and Lasso both increase the penalty factor on the basis of SSE. The difference is that Ridge selects the square value, while Lasso selects the absolute value. From the corresponding picture display, some factors are positively correlated with medical expenses, and some are negatively correlated. In addition, from the MSE display, we believe that limiting the value of log(lambda) between -2.5 and -1.9 is a good result for the ridge. For Lasso, we think it is a good result to limit the value of log(lambda) between -7 and -3. When the value of log(lambda) exceeds this range, although the variance decreases slightly, the bias increases rapidly. Combining the two, there is a significant increase in MSE, which is something we don't want to see.

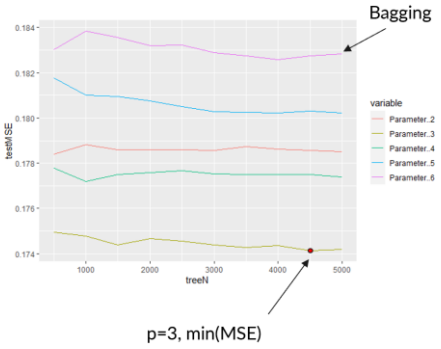
Decision Tree

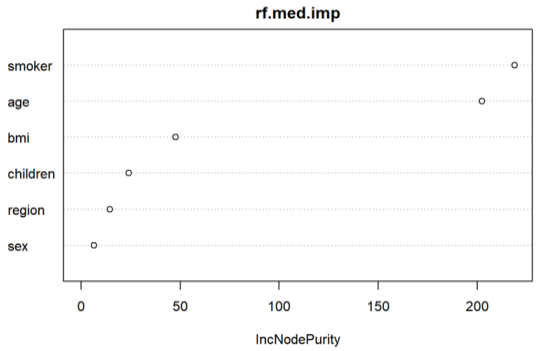
The model selection process points out which factors are important since it is easily implemented and the model is not a bad performance. However, in order to rank the importance of the factors, the decision tree method is applied in the following. Start with building the decision tree using the train test splitting method. Control the tree so that each node contains 2 or fewer data points. The test MSE shows that the error is 0.246. Even though the error is pretty small, the decision

tree seems to be too complicated. Thus, we decided to reduce the complexity of the decision tree by pruning the tree.

Pruned Tree

The cross-validation method is applied to decide the best size for the pruned tree. Since the minimum mean cross-validation error happens when branches are equal to 11, a training set is considered to fit the pruned tree. After fitting the model, the test MSE is down to 0.1890, which is lower than the MSE from the previous decision tree. Decision trees and pruned decision trees are defined as possessing the characteristic of high interpretability,but both of them tend to have high variances. Since decision trees generally don't have the same level of predictive accuracy as a regression model does, we apply random forest methods as a means to increase the level of accuracy.





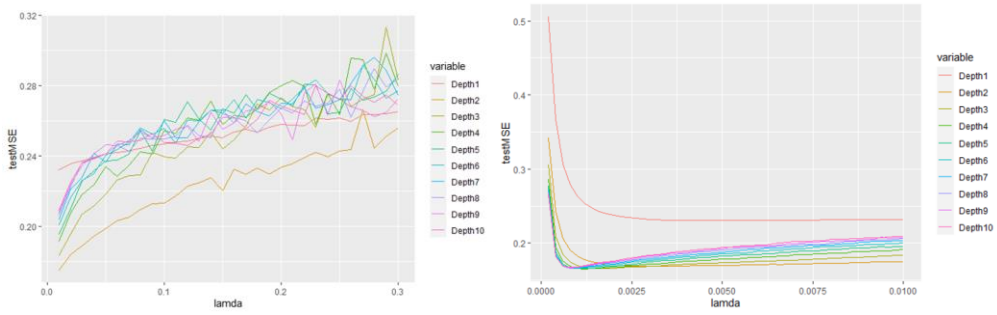
Random Forest

By aggregating the result of many decision trees, a bagging method is applied in order to reduce the variance. Since the model contains six different variables, we use all parameters with different tree numbers to draw the test MSE line as shown in the graph. It is noticeable that the

test MSE is around 0.183 and generally decreases as the number of trees increases. Nevertheless, the test MSE is close to the MSE which is under the pruned tree method. The reason is because of the similarity between bootstrap samples. Thus, we can conclude that averaging high correlated trees will not lead to a significant reduction of variance. A further step needs to be taken to reduce the MSE.

We decorrelate the decision trees from each sample set, and randomly select three predictors with the purpose of generating a bootstrap sample to build the random forest. After all, a relatively lower MSE is achieved when the tree equals 4500. The importance of variables is also evaluated as shown in the scatter plot. Apparently, smoker and age are the two most important variables under the random forest method.

Gradient Boosted Trees



With the aim of avoiding potentially overfitting, a boosting method is applied. Each of the trees is rather small and with just a few terminal nodes, which is determined by the parameter depth. From the graph, when the rate approaches 0.01, MSE has a relatively lower value. Thus, we take a closer look and zoom into the segment from 0 to 0.01. From the second graph, we can conclude that the lowest MSE is obtained when the learning rate is 0.00012 with a depth equal to 4.

Models Comparison

model <int>	MSE.lm <dbl>	branchs <int>	MSE.tr <dbl>	treeN <dbl>	MSE.rf_p3 <dbl>	Depth <chr>	MSE.bo_l0.0012 <dbl>
1	0.4586110	5	0.2207118	1500	0.1743677	Depth3	0.1673605
2	0.2126700	6	0.2029762	2000	0.1746719	Depth4	0.1652118
3	0.1962058	7	0.2043064	2500	0.1745387	Depth5	0.1656875
4	0.1916911	8	0.1908290	3000	0.1743732	Depth6	0.1667413
5	0.1911752	9	0.1904854	3500	0.1742815	Depth7	0.1675478
6	0.1905866	10	0.1907478	4000	0.1743549	Depth8	0.1684804
7	0.1886521	11	0.1890266	4500	0.1741359	Depth9	0.1692687
8	0.1881355	12	0.1915374	5000	0.1741971	Depth10	0.1702054

Accuracy and interpretability are taken into consideration when comparing the models. The predictive models with the highest accuracy are circled in red. The predictive models in the same level with the higher interpretability are circled in blue. Compared with other methods, fandom forest and boost methods are more complicated to interpret, so they have relatively low interpretability.

V. Conclusion

From the analysis above, our group decided on the pruned tree with 6 branches to be our final model because it is easy to interpret and the accuracy is relatively high. And the three most important variables are smoker, age, and BMI. Then our conclusion is very close to the practice of the real insurance industry, For example. For a man who is a smoker and his BMI is higher than 30.1, on average his medical insurance cost would be exp(10.62) or $41,000.

Finally, we have integrated the data analysis results with the actual insurance field for further interpretation. For most healthcare insurance products, age and gender play a decisive role in premiums. For long-term insurance products, such as critical illness insurance, each age and gender are associated with a unique premium rate. For YRT medical insurance products, people are usually divided into various age groups. For example, people in the age range 20-24 years old

have the same rate, while people at 25-29 years share another rate. For such products, gender generally does not result in a difference in premium. This might be attributed to the difference between underwriting. For critical illness insurance, they will only underwrite once, which is when the insurance is purchased. In the future, the physical status will not be a reason for denial. For one-year short-term insurance, since the physical condition of the insured is reviewed every year, and people who are in poor health are rejected. The premium rate will ignore the difference within a certain range. BMI is also a very important indicator, especially when underwriting. This is an important criterion. People with a BMI that is too high or too low have a higher risk of being exposed to various diseases, such people need to pay more premiums to get promised benefits, or be denied coverage. Smoking is also a factor that affects premiums. In general, smokers pay 5%-30% more premiums than non-smokers, but smoking is usually not a reason for denial. Although the number of children will also have an impact on the total medical expenditure to a certain extent, as in the previous table, the medical expenditure of children in youth is relatively low, so the increase in the number of children caused by the increase in medical expenditure is not as much as other factors. The last factor is Region. Indeed, there will be differences in medical expenditures between regions. However, these differences are more reflected between different states, cities, and even various communities. These local differences are narrowed when the country is divided into only four great regions so that their impact on health care costs is less important than other variables.

The suggestion we would like to give from the project is that the best way to minimize the Medicare cost would be to keep BMI at a normal range and don’t smoke. Both of them are significant factors that affect Medicare costs. Age and sex are something that we can’t control and the number of children, and region is not as significant as other factors.

VI. Reference

Society of Actuaries. (2013, January). *Health Care Costs—From Birth to Death*. Dale H. Yamamoto. https://www.soa.org/globalassets/assets/files/research/projects/research-health-care-birth-death-report.pdf

Wikipedia contributors. (2001, December 11). *Body mass index*. Wikipedia. https://en.wikipedia.org/wiki/Body_mass_index

M.C.P.D. (2020, July 14). *Medical Insurance Cost with Linear Regression*. Kaggle. https://www.kaggle.com/code/mariapushkareva/medical-insurance-cost-with-linear-regression/notebook

Dobson, A. J., & Barnett, A. G. (2018). *An Introduction to Generalized Linear Models (Chapman & Hall/CRC Texts in Statistical Science)* (4th ed.). Chapman and Hall/CRC.

Frees, E. W. (2010). *Regression Modeling with Actuarial and Financial Applications (International Series on Actuarial Science)* (1st ed.). Cambridge University Press.

Cowpertwait, P. S. P., & Metcalfe, A. V. (2009). *Introductory Time Series with R*. Springer Publishing.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R (Springer Texts in Statistics)* (2nd ed. 2021 ed.). Springer.