The Analysis about the Impact of Factors on the Health Insurance Charges

Croup19: Jie Yu, Qianwei Dong, Yi Jiang

Course Name: STAT 3340

December 11th, 2020

Abstract

This research paper focuses on the factors that influence the charges of health insurance and explores how they affect the medical insurance bills. According to the Insurance Dataset, the paper analyzed all the variables contained in it. It is concluded that the age of the insurer, BMI, number of children in the family, area of residence and smoking habits are important factors that influence how much the insurance company charges for individual health insurance. For an instance, age is an important basis for evaluation. The older an individual is, the more likely he or she is to require medical treatment, so the more insurance companies impose on the individual's health insurance. The more children there are in a family, the higher the premium for individual insurance charges. Above all, these factors do affect insurance charges and become important indexes for evaluation of the insurance fees.

Keywords: medical insurance, charges, age, BMI, children, region

Introduction

Nowadays, as people pay more attention to their health, medical insurance has long been the choice of most people. Especially, the charge of health insurance is one of the most important topics in this field. As health insurance is a part of living expenses, it is necessary to understand the charging system of medical insurance of health insurance companies. The average health care cost per family in Canada is 2000 Canadian dollars, and the private insurance premium is 4000 Canadian dollars (Amanda, 2020). Since medical insurance charges are assessed in many perspectives, we can start exploring some factors that may affect the costs of health insurance.

According to the Affordable Care Act (Obamacare) in America, Sterling Price (2020) states that the age of the policyholder is one of the few characteristics that insurers are permitted to use for determining health premiums. Larger applicants need more medical services and therefore need to pay higher health insurance charges. Secondly, medical insurance costs are even higher for smokers. The ACA allows employers who provide health insurance and public and private health insurers to charge smokers up to 50 percent higher rates (David, 2013). Besides, Japanese scholars Yoko Izumi, Ichiro Tsuji and others in 2001 published a paper "Impact of smoking habit on medical care use and its costs: a prospective observation of National Health Insurance beneficiaries in Japan", which shows that the medical expenses of male smokers are 11% more than that of "never smokers". Finally, the number of children in each family may also affect personal medical expenses. In 2018, the child insurance rate in the United States has reached 94.5% (Edward & Lryssa, 2019). Therefore, the more children there are, the greater the family burden will be.

In this project, including the above factors, we will explore whether some factors influence the charges of health insurance and their impacts on the charges by studying the 2017 medical expenses data set. We will mainly use the knowledge of statistics to analyze the relationship between them and draw final results.

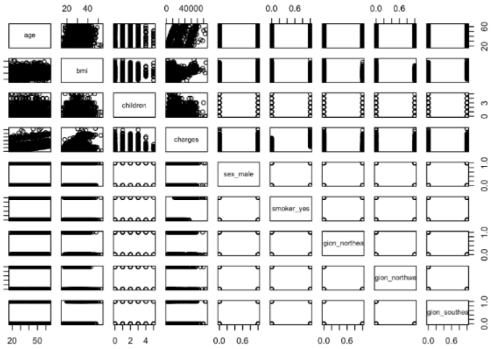
Data Description

The dataset used in this project is called insurance-group19 about the medical information and insurance charges. This dataset has multiple data relevant to various aspects of the health insurance costs, which is conducive to the analysis and research of the subject. This dataset contains a totally of 7 variables with 1338 observations and a new data point, mainly including age sex bmi children smoker region charges.

In this dataset, as a factor of interest, the variable “charges” means the individual medical cost bill of health insurance. The other quantitative variables are: “age” is the age of primary beneficiary; “bmi” is the Body mass index; “children” means the number of children covered by health insurance. As for the qualitative variables: “sex” is the gender of the insurer; “smoker” means whether the insurer smokes; “region” is the beneficiary's residential area in the US which are northeast, southeast, southwest, northwest. In our analysis, especially for the qualitative variables, we use the indicator variables to make them available in the model using R code. We set “sex_male” to represent the insurer is male if it is equal to 1. We set “smoker_yes” to represent the insurer smokes if it is equal to 1. We set “region_northeast”, “region_northwest” and “region_southeast” to represent the four regions, which represent southwest if all the three are 0. Also, we add a new additional data point into the insurance dataset. This data point introduces that a man lives in the southeast region, who is twenty years old with no children. And he does not smoke whose bmi is 28.7. The bill for his health insurance is 1784.826.

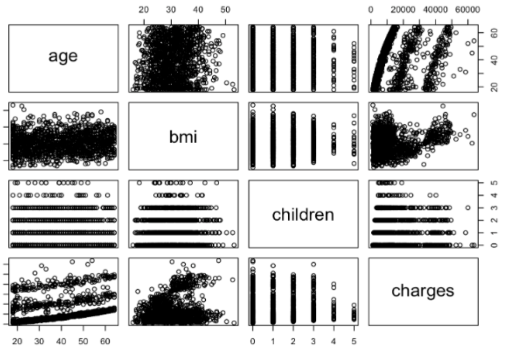
Then to make the data visualized, firstly, we make a matrix scatter plot (Figure 1) to preliminarily judge the relationship between the variables.

Figure1：

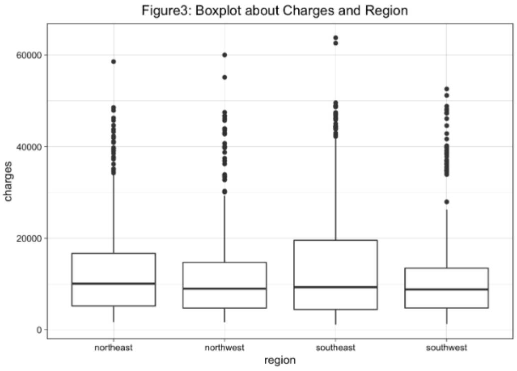


Through figure 1, we find that among all variables, only age, BMI, children and charges had an obvious correlation. Therefore, we remake the scatter diagram with only these four variables (Figure 2).

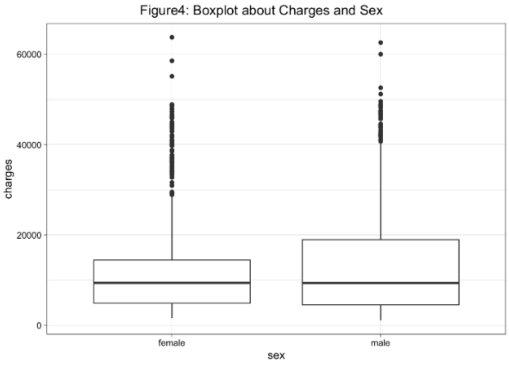
Figure2：



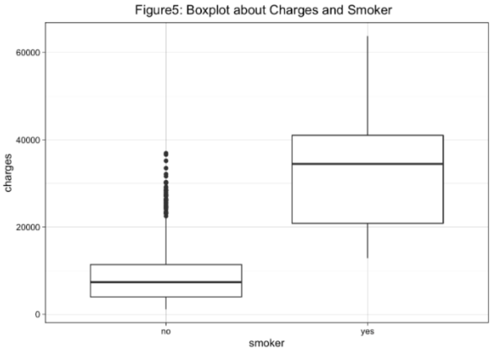
Through figure 2, we can see that age correlates with charges because their linear relationship is the most obvious. There is no obvious correlation between other variables and charges. To explore whether there is a correlation between the dependent variables and charges, we conduct three more ggplot code and get three boxplot pictures.



From Figure 3, we can see that there is no obvious correlation between charges and region.



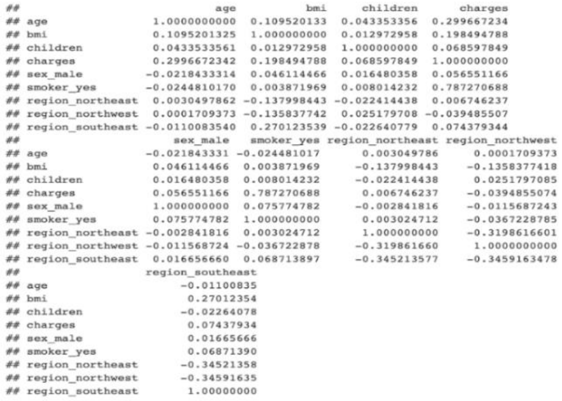
It can also be seen from Figure 4 that there is no obvious correlation between charges and sex.



However, it can be seen from Figure 5 that there is an obvious correlation between charges and smokers. Therefore, we can draw a preliminary conclusion: smoker and age have an obvious correlation with charges, while other factors have little correlation with charges.

Then, to draw a more intuitive and accurate conclusion, we calculate the correlation coefficient matrix of all the data (Figure 6). If the absolute value of the correlation coefficient matrix approaches 0, the correlation does not exist or is weak; if the absolute value approaches 1, it indicates that there is a correlation between the two variables.

Figure6:



```
##                            age          bmi     children      charges
## age                1.0000000000  0.109520133  0.043353356  0.299667234
## bmi                0.1095201325  1.000000000  0.012972958  0.198494788
## children           0.0433533561  0.012972958  1.000000000  0.068597849
## charges            0.2996672342  0.198494788  0.068597849  1.000000000
## sex_male          -0.0218433314  0.046114466  0.016480358  0.056551166
## smoker_yes        -0.0244810170  0.003871969  0.008014232  0.787270688
## region_northeast   0.0030497862 -0.137998443 -0.022414438  0.006746237
## region_northwest   0.0001709373 -0.135837742  0.025179708 -0.039485507
## region_southeast  -0.0110083540  0.270123539 -0.022640779  0.074379344
##                       sex_male   smoker_yes region_northeast region_northwest
## age              -0.021843331 -0.024481017      0.003049786     0.0001709373
## bmi               0.046114466  0.003871969     -0.137998443    -0.1358377418
## children          0.016480358  0.008014232     -0.022414438     0.0251797085
## charges           0.056551166  0.787270688      0.006746237    -0.0394855074
## sex_male          1.000000000  0.075774782     -0.002841816    -0.0115687243
## smoker_yes        0.075774782  1.000000000      0.003024712    -0.0367228785
## region_northeast -0.002841816  0.003024712      1.000000000    -0.3198616601
## region_northwest -0.011568724 -0.036722878     -0.319861660     1.0000000000
## region_southeast  0.016656660  0.068713897     -0.345213577    -0.3459163478
##                  region_southeast
## age                   -0.01100835
## bmi                    0.27012354
## children              -0.02264078
## charges                0.07437934
## sex_male               0.01665666
## smoker_yes             0.06871390
## region_northeast      -0.34521358
## region_northwest      -0.34591635
## region_southeast       1.00000000
```

According to the data in Figure 6, the correlations between age, BMI, smoker_yes and charges are relatively large, which are 0.2996672342, 0.198494788 and 0.787270688, respectively. In contrast, the correlations between children, sex_male, region and charge are relatively not obvious. In addition, the correlations between several dependent variables are also obvious. For example, the data between age and BMI is 0.1095201325, and the three directions of region and BMI are closely related.

This is consistent with our previous conclusion through image analysis. Therefore, we predict that there will be at least three variables in the final model: age, BMI and smoker_yes.

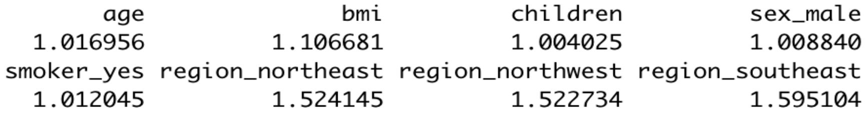
Methods

In this analysis, we are going to explore what factors affect the charges of the health insurance using R. This is done by considering linear regression models between charges and other regressors, which is a general tool for analysis.

First of all, we construct the full regression model with all independent variables including age sex_male bmi children smoker_yes region_northwest region_northeast region_southeast. Through the analysis of the significance of each P-value, we initially determine which variables may exist in the final model and which may not. Meanwhile, VIF is used to check whether there is multicollinearity between variables. Because of the defects of the regression model, we try to transform the variables to correct the model inadequacies. However, if we can't modify the model by transformation, some variables need to be deleted which are not important. Therefore, we first start to do the variable selection by performing all subsets models to find the best model. According to the adjusted R2 criterion, the model choice is the one with the largest adjusted R2. We then check to see if the model is satisfactory by doing a full residual analysis. Furthermore, if it is not feasible to fit all possible models, we use a stepwise procedure to select the best model and make a backward elimination to compare the best models they produce. In addition, because there may be some correlations between regressors, we add some reasonable interactions in the model to do the variable selection. Then, we compare and analyze all the selected models to get the final best model and make a thorough analysis of the best model. Finally, we summarize the advantages and disadvantages of the final model.

Results

After regression of all variables, we found that P values of sex_ male，region_ northwest and region_ Southeast are very large, which are 0.692927, 0.202879 and 0.873607 (all greater than 0.1). This shows that sex and region have little effect on health insurance charges, and they may need to be excluded from the equation. At the same time, we use VIF to test whether there is multicollinearity between variables.



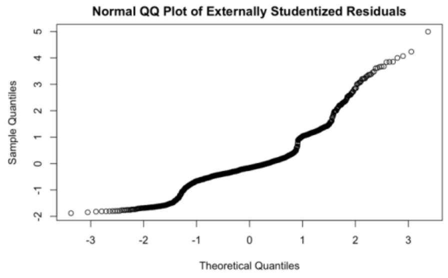
```
            age              bmi         children         sex_male 
       1.016956         1.106681         1.004025         1.008840 
     smoker_yes region_northeast region_northwest region_southeast 
       1.012045         1.524145         1.522734         1.595104 
```

It turns out that, the numbers of region_northeast, region_northwest and region_southeast are far away from 1, which indicates that there are collinearity problems between them that need to be adjusted. Its normal qq plot shows that residuals do not follow a normal distribution but it looks better when we transform the variable age to 1/age.

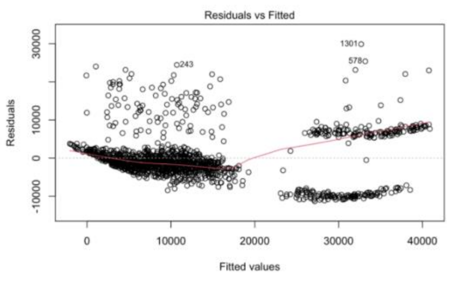
However, the remaining residual plots such as Residual vs Fitted and Scale-location all have strange trends and residuals are not evenly distributed in the graph. Another problem is that we found P values of sex_ male，region_ northwest and region_ Southeast are still large after we transform the model from changing the age to 1/age. The other one is adjusted R-squared changes from 0.7496 to 0.6619. The smaller the value, the worse the equation. Therefore, we start to select independent variables to get the best model.

To find the best model, we first using R code library leaps to perform all possible subset regressions. There are a total of 67 possible models and the corresponding adjusted R^2. According to the adjusted R^2 criterion, the best model should be the one with the largest adjusted R^2. Among them, we find the largest adjusted R^2 is 0.7499250429 which is the 49th model in the list. This model explains 74.99% of charges. It removes variables sex_ male and region_ Southeast and remaining regressors are significant with small p-value except for the variable region_ northwest; The p-value of region_ northwest reduces to 0.121171 which is less than p-value from the full model. Thus, the best model under adjusted R^2 criterion is: charges = 257.04*age+337.89*bmi+475.26*children+23832.15*smoker_yes+996.89*region_northeast + 644.54*region_northwest.

Then we plot the best model to check if it is satisfactory. Whether all possible regressions determined from evaluation of adjusted R^2 is feasible based on Gauss-Markov assumptions, especially, the normal qq plot of standardized residuals and externally studentized residuals vs fitted values.



From the figure, we can conclude that the error term of this model is not following the normal distribution and there are at least three possible outliers.



This output gives the plot of externally studentized residuals against fitted values to test the assumption that the expected value of the error term equals 0. From the plot, most residuals gather around 0. However, the red line starts to increase as fitted values become larger. Also, there are three outliers in the figure which are data#243, data#1301, and data#578 may result in an increasing trend of the red line. Above all, there are some inadequacies in this model. Therefore, it is reluctant to conclude the best model is feasible from the R^2_{Adj} criterion.

There is another criterion for model selections named *Akaike Information Criterion* (AIC). According to the page 336 from textbook, it is shown that $AIC = nln(SS_{RES}/n)+2p$ in the case of ordinary least squares regression(Montgomery, 2012). It is much more practical to use AIC for testing complicated models than $R^2_{Adj.}$

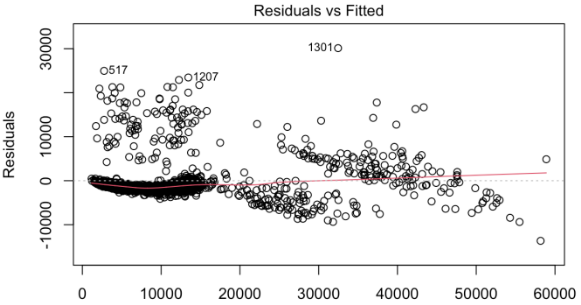
Firstly, we generate fullmodel1 that contains all regressors and fullmodel2 that contains all possible predictors and reasonable interactions. We believe that smoker_yes has interactions with age and bmi, age has interactions with bmi and children, and both bmi and children may interact with different types of regions. Secondly, we use AIC value and forward selection method to add regressors sequentially starting from nullmodel(*lm(charges ~ 1))* to fullmodel1(*lm(charges ~ age + bmi + children + sex_male + smoker_yes + region_northeast + region_northwest + region_southeast))* and the best model is charges ~ smoker_yes + age + bmi + children + region_northeast + region_northwest. The second best model is selected by using backward elimination and comparing the minimum value of AIC. That is, charges = smoker_yes + age + bmi + children + region_northeast + region_northwest + region_southeast + smoker_yes:bmi + bmi:region_southeast +

children:region_northwest. The next step is to decide whether we need to keep the first best model from the forward selection method. ANOVA test (anova(step1,step2)) allows us to find the lack of fit SSE between the first best model and the second-best model which is 191.59. Also, since the p-value of the F test is less than 2.2e-16, we can conclude the F test rejects H0 which is the first best model that does not fit the data. Thus, we remove the first best model and keep the second-best model from backward elimination. That is, charges = smoker_yes + age + bmi + children + region_northeast + region_northwest + region_southeast + smoker_yes:bmi + bmi:region_southeast + children:region_northwest. We call the model as lmbest2, and lmbest2 indicates all regressors and interactions are significant except the variable: region_northwest which is 0.715712(extremely large than 0.1). The adjusted R^2 of lmbest2 is 0.840955 and its p-value is less than 2.2e-16, which means the variables of lmbes2 explain 84.0955% of charges and prove lmbest2 exiting.

Moreover, the results could be inconsistencies between backward elimination and full stepwise method. Stepwise method selects model starting from nullmodel to fullmodel2 with adding and removing aggressors by checking F-statistics and AIC value. Our best model named lmbest1 is: charges ~ smoker_yes + age + bmi + children + region_northeast + region_northwest + smoker_yes:bmi + bmi:region_northeast + bmi:region_northwest + children:region_northwest.

In the lmbest1 from stepwise, except that the variables bmi and region_northeast are not significant, the other regressors and interactions are all significant. The adjusted R^2 is 0.8407513, which is larger than it in the best model selected from the leaps method with only the main variables in the model. For the statistic of the full model, its p-value is very small. This means that the lmbest1 exits.

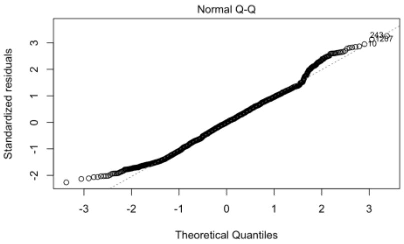
It is difficult for us to simply decide our best model through normal qq plot because of indicator variables. Then, we choose to use data splitting to measure the model performance precisely. Basically, the 10-fold cross-validation method splits the dataset into 10 folds that each fold has almost the same sample size and test those folds by 10 times. Then, the 10th subset is defined as a test set whereas the remaining 9 folds are fitted in the model. Next, we select our four potential best models from transformation, leaps and step-wise procedure method. Model1 has the highest adjusted R^2 during the leaps method test, and model4's normal qq-plot shows its residuals follow the normal distribution. Model2 and Model3 are selected from the step-wise procedure, which both have similar adjusted R^2 around 0.841. In the 10-fold cross-validation, there are two criteria to help decide which one works best: RMSE and R^2. RMSE (Root Mean Squared Error) assesses the accuracy of the model by measuring the standard deviation of average squared differences between predicted values and observed values. Thus, the less the RMSE, the better the model performance. R^2 measures how much of the variability in new observations the model might be expected to explain (Montgomery, 2012), so the best model should have the highest value of R^2. RMSE of model1 is around 6000 and the R-squared value is about 0.75. Model4's RMSE and R-squared are about 7000 and 0.66. Model2 and Model3 have similar RMSE and R-squared which all-around 4800 and 0.84. Thus, 10-fold cross-validation effectively indicates both model2 and model3 are the most accurate regression model.



In the ANOVA test of lmbest2, it shows the p-value of region_southeast is extremely larger than 0.1 whereas all p-value of regressors in the ANOVA test of lmbest1 is less than 0.1. Compared to lmbest2, lmbest1 indicates all regressors are significant and have impacts on health insurance charges. However, the relative deficiency is the adjusted R^2 of lmbest1 is 0.8407513 which is smaller than the adjusted R^2 of lmbest2 (0.840955). According to the ANOVA test, the preferable model is lmbest1: formula = charges ~ smoker_yes + age + bmi + children + region_northeast + region_northwest + smoker_yes:bmi + bmi:region_northeast + bmi:region_northwest + children:region_northwest.

Compared to the plot of the leaps method, the resulting residual versus fitted plot of lmbest1 looks much better and most residuals are around 0 and evenly distributed in the graph. The problem of lmbest1 is the normal qq plot indicates residuals are normally distributed until the value of theoretical quantiles reaching 1. Then, we need to transform our best model. We start from the response transformation by using the Box-cox method. R helps us select the appropriate λ value(0.2626263) that maximizes the log-likelihood. We then fit our best model from step-wise method with the transformation, $y^{\lambda}-1/\lambda$, and 1/x applied to the response(charges): ((charges^0.2626263-1)/0.2626263) ~ smoker_yes + 1/age + log(bmi) + children +

region_northeast + region_northwest + smoker_yes:bmi + bmi:region_northeast + bmi:region_northwest + children:region_northwest



Even though our normal qq plot of the transformed model indicates residuals are normally distributed, the resulting residual versus fitted plot does not prove residuals are around 0 and randomly distributed in the graph. Besides, the adjusted R2 is 0.5844846 which means only 58.45% of predictors explain the insurance charges. We then decide to neglect the transformation model and keep lmbest1 as our best model: charges = smoker_yes + age + bmi + children + region_northeast + region_northwest + smoker_yes:bmi + bmi:region_northeast + bmi:region_northwest + children:region_northwest.

Conclusion

According to the final model which gets through leaps and stepwise, sex and region _ Southeast does not affect health insurance charges at all. This shows that in life, people's gender and living area have no direct impact on the insurance cost. In other words, this paper shows that the older people are, the worse their health is, the more children they have in their families, and the smoking addicts will greatly increase the cost of insurance. We believe that the result not only improves people's health awareness but also provides a theoretical basis for social management and legislation. In addition, we encountered a lot of problems in the derivation stage, and the final model is not perfect, so we hope this paper can provide some reference for future related research work.

Reference

Berchick, E., & Mykyta, L. (2019). *Children's Public Health Insurance Coverage Lower Than in 2017.* Retrieved December 3rd, 2020 from https://www.census.gov/library/stories/2019/09/uninsured-rate-for-children-in-2018.html

Frank, A. (2020). *How Much Do You Pay for Health Insurance?* Retrieved December 3rd, 2020 from https://www.monster.ca/career-advice/article/how-much-are-health-benefits-canada

Izumi, Y., Tsuji, I., Ohkubo, T., Kuwahara, A., Nishino, Y., & Hisamichi, S. (2001). Impact of smoking habit on medical care use and its costs: a prospective observation of National Health Insurance beneficiaries in Japan. *International Journal of Epidemiology,* Volume 30, Issue 3, June 2001, Pages 616–621.

Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to Linear Regression Analysis.* (5th Edition). VitalBook file:Wiley-Blackwell

Price, S. (2020). *How Age Affects Health Insurance Costs.* Retrieved December 3rd, 2020 from https://www.valuepenguin.com/how-age-affects-health-insurance-costs

Resnik, D. (2013). Charging Smokers Higher Health Insurance Rates: Is it Ethical? Retrieved December 3rd, 2020 from https://www.thehastingscenter.org/charging-smokers-higher-health-insurance-rates-is-it-ethical/