Medical Cost Prediction

Probability & Statistics -1 MATH01523
MSDS, Rowan University

1. Introduction

Health care is one of the most vital thing that person needs to have productive, secure and healthy living style. In US, residents get health coverage from private and public sectors, few of them have through their employers or direct purchase from private source or enrolled with medicate, medicaid or veterans' affairs programs. The insurance company calculates premiums based on the two major factors firstly, the cost of the premium, insurer predict under their policies and second, the cost of operating plans. There are several factors affect for the cost of medical expense such as policy holder's health status, employment status, wages, region, habits etc. In terms of machine learning, regression method is often used to predict the health care costs and calculate premiums. Regression analysis is used to identify the important factors that are affected to predict the insurance cost. Using proper classification techniques, it offers fair price to customer and maximize the company's profit.

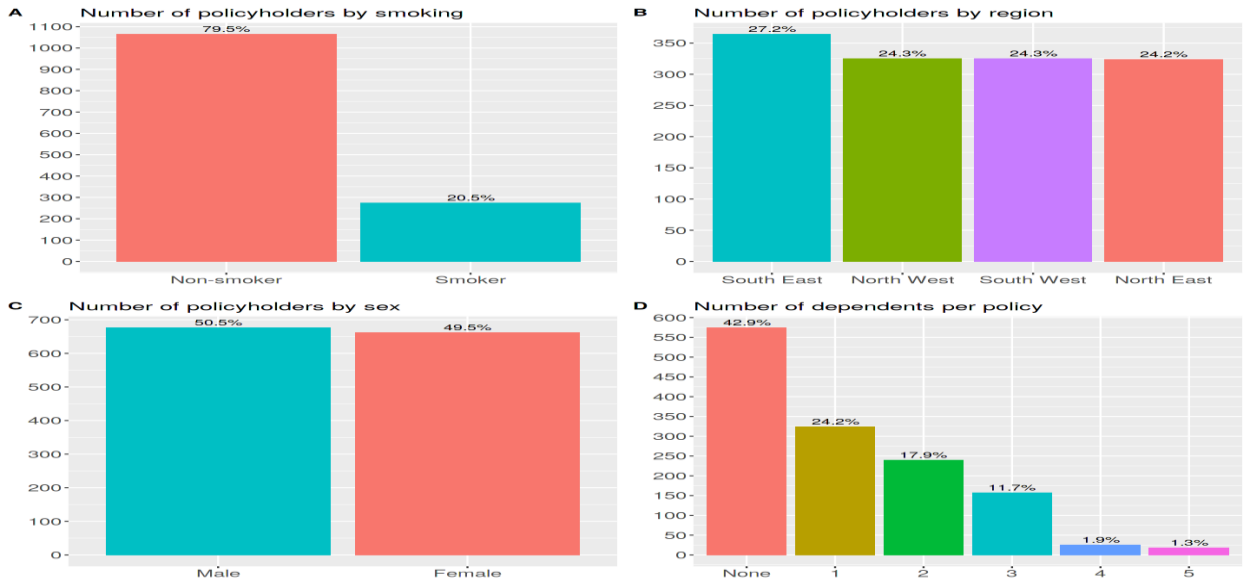
2. Dataset Description

The insurance dataset consists of 1338 policyholders with 7 attributes that describe their demographic region, southwest-southeast-northwest-northeast including age, gender, bmi and person's smoking habits, hospitalization charges and the no of dependents who are covered in the policy. The variable we would like to predict is the total claim amount that are billed to the insurance company. Using R we will perform EDA, we would analyze the distribution of various attributes and determine is there any relation between attributes and importance of attributes to calculate the medical cost with statistical testing and finally we fit the regression model to predict the medical cost.

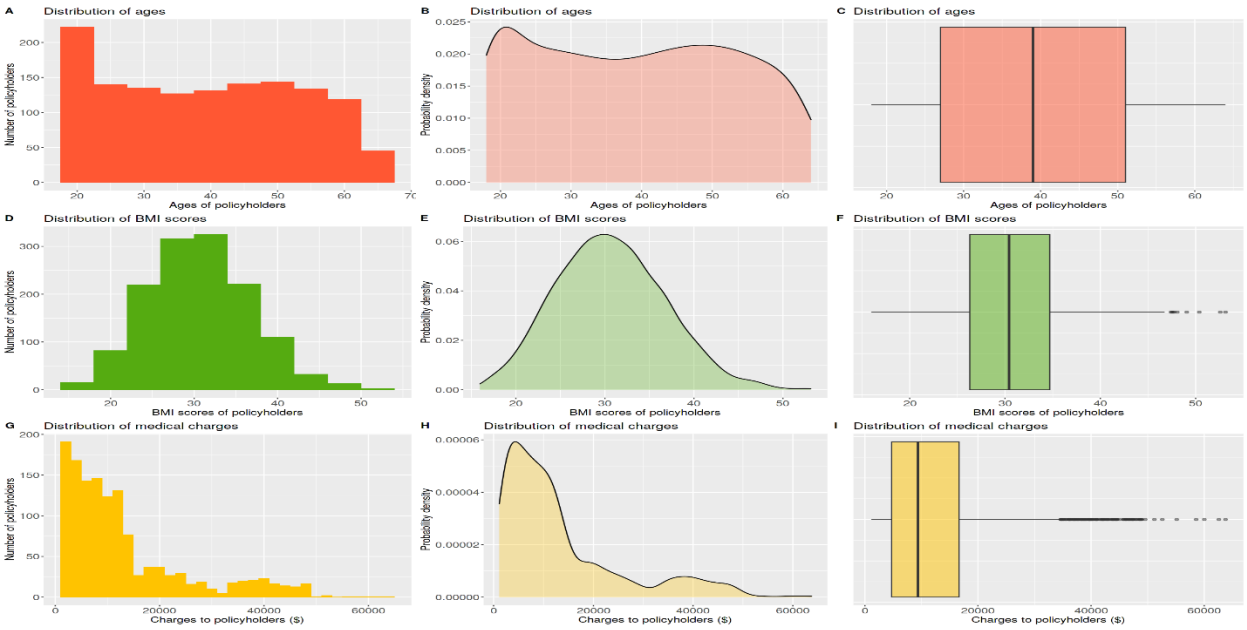
3. Exploratory Data Analysis

In any project, the data is most important starting point. After loading the data in R, we check for null values and data types. There are no missing values, and all the attributes are assigned with correct data types. There are three numerical continuous variables age, bmi, charges and four categorial variables sex, smoke, dependents and region. We perform the five-point summary and notice that mean

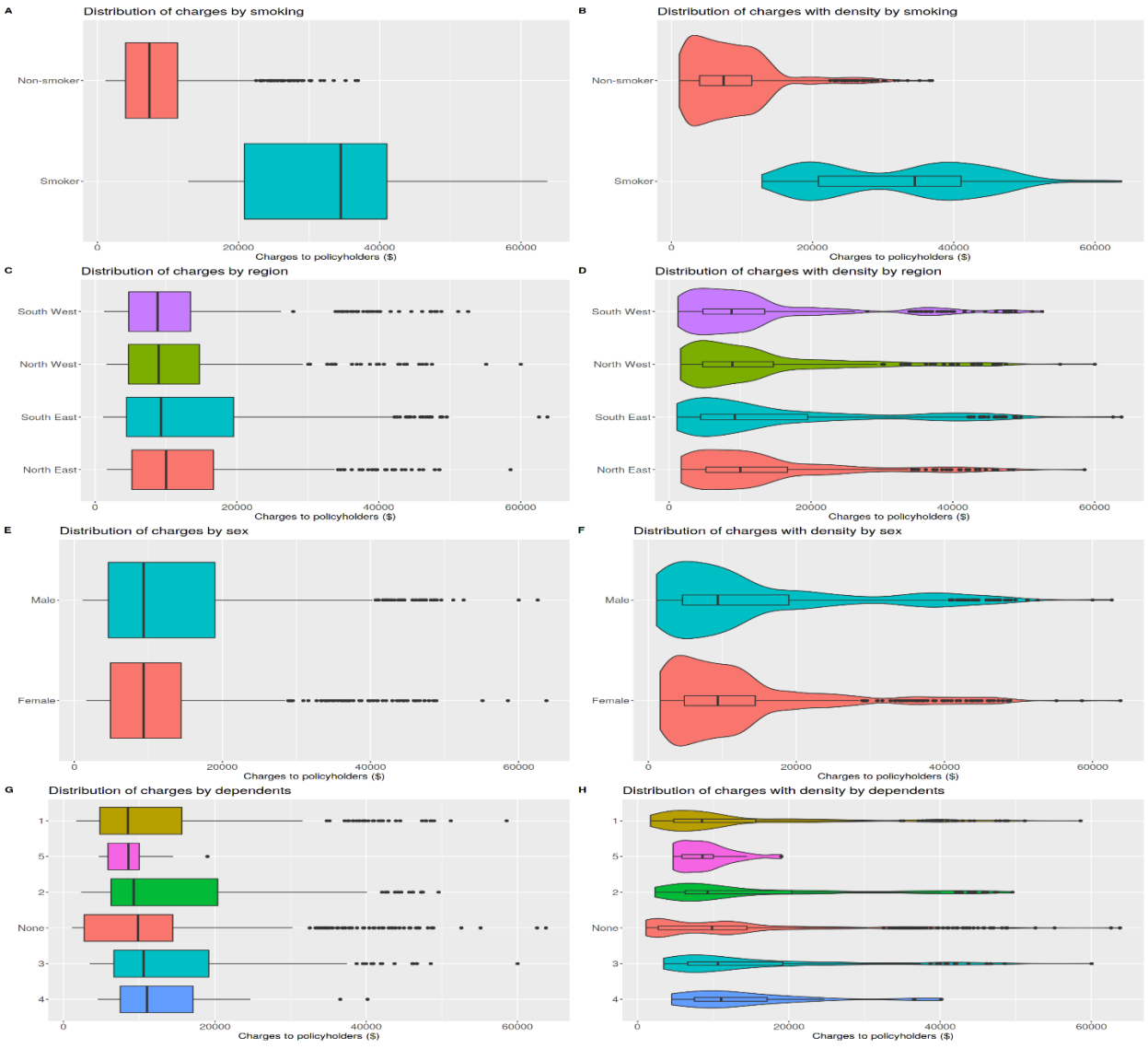
charges is 12k but median is 9k so there is a hint for outliers, mean bmi and age is 30 and 39 respectively whereas max no of children is 5.



By visualizing the bar chart for categorial variable, there are 80% nonsmoker than the 20% smoker. Policyholders are evenly distributed across the region; the most populated region is southeast. Data is also evenly distributed with gender, whereas almost 43% holders do not have children with only 1% have maximum no of children.



By visualizing the histogram for continuous variable, we can say that 18-23 years old are the most represented group whereas 60-65 years old are least represented. Bmi is normally distributed with identical mean and median having few outliers. Charges are highly right skewed with maximum outliers which means that most of the policyholders are with low charges compare to high charges.



If we compare the charges vs other factors by median, there is huge difference between smoker and nonsmoker, majority of the nonsmoker have low charges with many outliers (other factor affect for the high charges) and while it is fare that smoker have high charges due to the serious health risk. The southeast region has large IQR compared to fairly spread between other regions. Males and females have

almost equal value. The charges for the holders who do not have child are less likely then 2-3 dependents. If we analyze the medical charges vs bmi, we observed that person having high bmi has high charges means they are positively corelated and same correlation for age.

4. Statistical Test

When the sample are random, there are possibilities to make inference between sample and population. Hypothesis test is the way of the inference. The conclusion usually involves the estimation of population parameter, like mean or median value. We are trying to estimate population parameter using the sample. Assumption of the test are as follow. Dependent variable is continous, observations are independent of each other, variable doesn't contain any outlier and the shape of the distribution of the independent variable must be known. There are two competing hypothesis test: null (H0) and the alternative (H1). Null hypothesis is about no difference. We try to collect the evidence to support the alternative hypothesis during the process and reject the null hypothesis. We do test to check if we reject the test or fail to reject the test. Test will give p value, if p value less than rejection region, we reject H0, if p value is higher than the rejection region, we fail to reject H0. There another test, one way analysis of variance (ANOVA) is used to determine whether there is any statistically significant difference between the means of three or more independent group. We do have few hypothesis test as well as ANOVA which

\are as follow.

1. There is difference between charges in region. The assumption is H0: no difference between medians. H1: there is difference between the medians. P (0.19) value is higher than the significant level (0.05), we fail to reject the null hypothesis, which conclude that no difference is exists.
2. There is difference between mean of male and female. The assumption is H0: group means are equal. H1: there is difference between mean of the two group. The p value (0.1014) is greater than the significance value, we fail to reject and conclude that there is no variance of the two-sample group's mean value.
3. One-way Anova test to compare the mean of more than two independent variable, is the bmi equal among all region or not? The p value is less than the 5%, we reject the null hypothesis of equal mean among the regions. Some of them have different bmi. To observe the difference, we use the Tukey pairwise comparison.

5. Regression

To predict the medical cost, we perform the multiple linear regression model. The aim of the regression model is to find the equation between predictors and outcomes. If only one predictor, it is a simple linear regression. Other then it is multiple linear regression. The goal is to find best fitted regression line. For that we split the data for training (80%) and testing (20%).

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.40628 -0.09013 -0.02321  0.03314  0.93626

Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)      3.0308795  0.0342260  88.555 < 0.0000000000000002 ***
smokeryes        0.6760329  0.0144515  46.779 < 0.0000000000000002 ***
bmi              0.0058070  0.0009931   5.848  0.00000000663898698 ***
age              0.0153611  0.0004142  37.090 < 0.0000000000000002 ***
children1        0.0538452  0.0146927   3.665             0.000260 ***
children2        0.1286999  0.0161328   7.978  0.00000000000000385 ***
children3        0.1086741  0.0189414   5.737  0.00000001254227630 ***
children4        0.2109837  0.0411729   5.124  0.00000035470555674 ***
children5        0.1835554  0.0552900   3.320             0.000931 ***
sexmale         -0.0304837  0.0115905  -2.630             0.008661 **
regionnorthwest -0.0305449  0.0164321  -1.859             0.063325 .
regionsoutheast -0.0599089  0.0168307  -3.559             0.000388 ***
regionsouthwest -0.0562769  0.0165515  -3.400             0.000699 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1884 on 1059 degrees of freedom
Multiple R-squared:  0.7789,    Adjusted R-squared:  0.7764
F-statistic: 310.9 on 12 and 1059 DF,  p-value: < 0.00000000000000022
```

A significant equation is found with adjusted R-squared of 0.7764. Which means, The model explains the 77.6% of the total variance. The p value<0.001 of F-statics is very low, so there the predictors related to the outcomes. Almost all predictors are important according to the P-value except gender. Smoking habits and age plays vital role for predicting the outcomes as it has lowest p value.

To measure the robustness of regression, RMSE (test = 0.208) and RMSE (training = 0.187) is calculated, and it is indicating that model is not overfitted. After the back transformation, RMSE for test set is 9000, which means that the predictions are usually off by this amount.

6. Discussion & Conclusion

Smoking and age have the high potential to increase the medical cost as it quite expected. Smoking has high impact on cost, so we should create awareness to stop smoking. Normal bmi score not indication of ill health, only people with underweight and underweight has poor health outcomes. Middle age and elderly people face rapid decline in health compared to 18-22 years age, so we can promote healthy living at Yung age to avoid those charges. Medical cost will increase as no of dependent increase; however, 3 children cost is cheaper than having 2 children. Finally, R-square value of 77.9 % indicated total variance of the overall model. So, company can deal with better prediction, though some other factors also may affect.

7. Reference

1. https://www.aha.org/guidesreports/report-importance-health-coverage
2. https://rpubs.com/MasoKan/581892
3. https://www.kaggle.com/code/goldens/statistics-with-r/notebook#4--Regression
4. https://www.kaggle.com/code/mayank2896/insurance-eda-hypothesis-testing#Recommendations--