STAT 605 GROUP 5

PROJECT REPORT

How Does Smoking Habit and Age Affect Your Insurance Rate in Different State?

Wenyi Wang
wwang584@wisc.edu
Yiran Wang
wang2559@wisc.edu
Yuxiao Li
li2268@wisc.edu
Zhihao Zhao
zzhao@wisc.edu
Zhiyu Ji
zji29@wisc.edu

December 10, 2020

1 Introduction

We want to find the number of distinct benefit count in each states and how does smoking habit affect the individual premium cost for insurance plan in different states. The data we use contain two parts: the benefit data set and rate data set. For benefit data set which includes benefits' name and corresponding state, we calculate the distinct benefit count of each state and draw them in a map. For rate data set which includes the premium for each individual, we run a big job in HPC to do simple linear regression for premium, StateCode, Age and Tobacco.
Finally, we find that for all states, the larger one's age is, the higher his cost of insurance premium will be. In addition, in most states tobacco use will increase the premium charged. An interesting fact is that in WY, WV, TX and SD, the estimated coefficient difference for tobacco is negative, which means tobacco users may pay lower premium than non tobacco user.

2 Data Introduction

2.1 Data Description

The data set employed in our project comes from *The Health Insurance Marketplace Public Use*, which contains data about health insurance plans targeting on individuals and small businesses through the US Health Insurance Marketplace. Ranging from 2014 to 2021 (updated on 18^{th}, Nov), covering states all over the United States, the whole data set is about 10.2G, divided by business year. The original data set can be found from this URL: https://www.cms.gov/CCIIO/Resources/Data-Resources/marketplace-puf.

The '$Benefits\_Cost\_Sharing\_PUF.csv$' data set for 2021 contains 433,042 rows, each representing a distinct benefit name and its corresponding information. When it comes to the '$Rate\_PUF.csv$' data set for 2021 that contains 2,136,450 rows, each row refers to the cost of premium for specific insurance plan divided by the rating area code within a state.

Table 1: Dependent Variables and Explanatory Variables

Variables	Type	Meaning
IndividualRate	Dependent Variable	Dollar value for the insurance premium cost applicable to a non-tobacco user for the insurance plan in a rating area
IndividualTobaccoRate	Dependent Variable	Dollar value for the insurance premium cost applicable to a tobacco user for the insurance plan in a rating area
StateCode	Parallel Job Divider	Two-character state abbreviation indicating the state where the plan is offered
Tobacco	Explanatory Variable	Categorical indicator of whether a subscriber's tobacco use is used to determine rate eligibility for the insurance plan
Age	Explanatory Variable	Categorical indicator of whether a subscriber's age is used to determine rate eligibility for the insurance plan
BenefitName	-	Name assigned to benefit

2.2 Visualization and Cleaning

2.2.1 Data Visualization

In order to vividly illustrate the distribution of different benefits, word cloud plot and map plot are utilized. The word plot in figure 1(a) displays the frequency of various benefit, where the size of each benefit name is determined by the count of their appearance in different insurance plan. In regard to the map plot, the numbers in figure 1(b) refers to distinct benefits counts within each state.

The impact of age and tobacco usage on the cost of premium draw our attention as well. There are 1658 different insurance plans with no preference on tobacco usage, but 3227 insurance plans that emphasize the smoking habit. In the following map plots and bar chart, mean costs of premium within each state are compared for non-smokers and smokers respectively. The side by side bar chart visualize the fact from another angle.

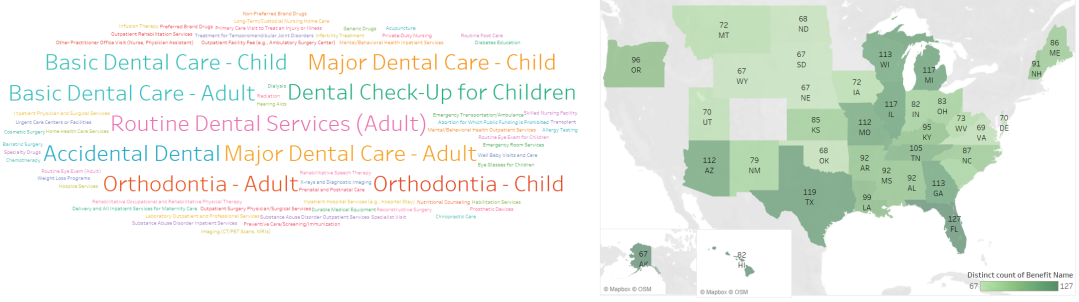
Interesting patterns are shown from these plots:

- The hottest five benefits covered in different insurance plans are all related to dental treatment, with the *Routine Dental Services(Adult)* reaching the top. The *Orthodontia* benefit for child and adult follows afterwards. On top of dental benefits, insurance plans are likely to cover benefit as *X-rays and Diagnostic Imaging*, *Eye Glasses for Children* and *Mental/Behavioral Health Inpatient Services*.

- In general, all states offer at least 67 distinct insurance plans. Specifically speaking, FL(127), TX(119), MI(113) and IL(113) contains the most diverse types of insurance plans. Our state Wisconsin also does a great job.

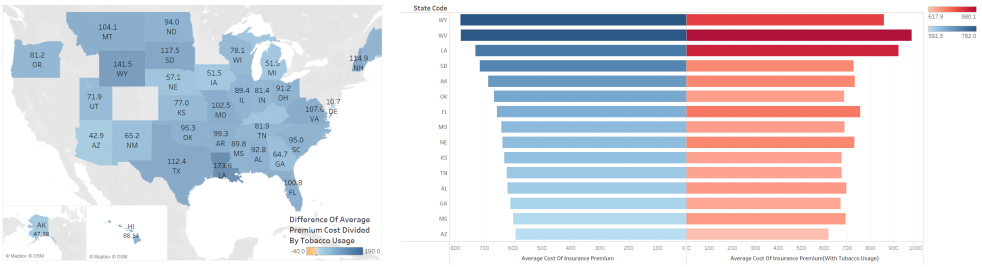
- The smoking habit increases the average cost of insurance plan in all states, which is not surprising. The penalty of tobacco usage for average premium cost reaches up to 173.6 dollar in LA, while only 42.9 dollar in AZ.

- The average cost of premium for tobacco user is 1.64 times of that for non-tobacco user in DE, the minimum increase also reaches 7.6% in MO. Except for that, another attracting fact is that the tobacco variable changes the order of states with highest cost of premium, from WY(782), WV(780.8), LA(729.8) to WV(980.1), LA(922.1), WY(859.6).



(a) Most Frequent Benefit in Different Plans (b) Distinct Benefit Count Within Each State

Figure 1: Distribution of Benefit (2021)



(a) Difference of Average Premium Cost By Tobacco(2021)

(b) Side-by-side Bar Chart for Cost of Premium (2021)

Figure 2: Map Plot and Bar plot of Insurance Premium Cost Divided By Tobacco Usage

2.2.2 Data Cleaning

Several steps are performed in the data cleaning part:

- Remove observations with unexpected values for dependent variables.
  Variable "Age" has three unusual categories 'Family Option', '0-14' and 'over 64'. 'Family Option' indicates that such person got a family insurance plan so his insurance premium has some other influential factors which is beyond the purpose of this project. So individuals with such tag are removed first. Then we replace '0-14' and '64 and over' with '14' and '70' in order to convert "Age" into a numerical variable. As for "Tobacco", values = NA is automatically removed since such value only occurs when "Age" = 'Family Option'.
- Combine dependent variables.
  The response variable is a combination of variables "IndividualRate" and "IndividualTobaccoRate" due to their unique structures. Suppose we have twins as our customers to buy a same type of insurance, since they are almost identical, they are very likely to be offered with same price. But in this case, one is an aficionado of tobacco while the other can never stand the smell of it. Thus the insurance company may offer a different price. So we merge variables "IndividualRate" and "IndividualTobaccoRate" into one variable and change categorical name of "Tobacco" accordingly. This process is illustrated in Table 2 and 3 for better comprehension.

Table 2: Oringinal Structure

Tobacco	IndividualRate	Individual TobaccoRate
Tobacco User/ Non-Tobacco User	866.14	996.06
No Preference	482	NA

Table 3: Structure After Transformation

Tobacco	IndividualRate
Non-Tobacco User	866.14
Tobacco User	996.06
No Preference	482

3 Statistical Model

The core concern of this project is how do other factors affect the cost for insurance premium, thus variables "IndividualRate" and "IndividualTobaccoRate" are merged as our response variable as we illustrated before. Simple linear model is utilized to fit the statistical model and draw relevant inference.We denote "IndividualRate"as y, "StateCode"as u,"Age" as w,"Tobacco" as v."StateCode" and "Tobacco" are factor variables and "Age" is continuous. Thus, our model can be formulated as below, we also use an interaction

term for tobacco preference and state.

$$y = \sum_i a_i 1[u = x_i] + \sum_i b_i 1[v = z_i] + \sum_i c_i 1[u = x_i] \sum_j 1[v = z_j] + d * w + \epsilon, \epsilon \sim N(\mu, \sigma^2)$$

Since the whole data set is extremely large that cannot be calculated on our local computer to do regression, we plan to use HPC server to do regression for efficiency. We ran one big job that took 70 seconds, using 6GB of memory on the HPC cluster(after doing some clean-up in small jobs). We firstly obtain the coefficients of these parameters, among which we are most interested in the coefficients of interaction term for "StateCode" and "Tobacco". The coefficients of "Age" is positive in all states, meaning that the elder you are, the more insurance premium cost will be charged, which is consistent with our common sense. In most states, the coefficients for category "Non-Tobacco User" is positive, which indicates that non-tobacco users will pay more comparing with no preference people. For tobacco users in most states, their coefficients are even higher than non-tobacco users, so they will be charged more than non-tobacco users, and of course, much more than no preference people. The results are vividly illustrated in section 4.

We also do model diagnosis. We don't find high leverage point and the model is close to normal distribution.

4 Results and Insights

For the influence of Tobacco in different state, We draw the difference of the coefficients of tobacconon-tobacco term and interaction term in the map.

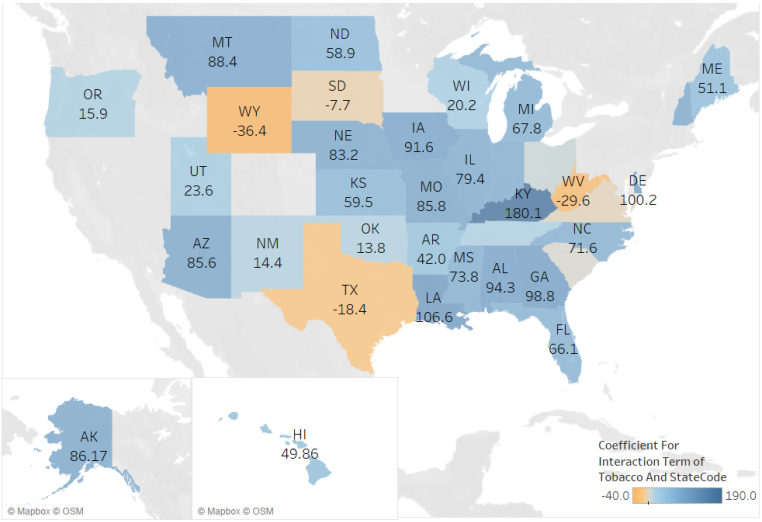


Figure 3: Coefficient For Interaction Term of Tobacco And StateCode (2021)

From figure 3, it is easy to find that when age is fixed, for most states, if a person change from non-tobacco user to tobacco user, the premium will increase. And the premium increase a lot in KY, LA and DE, which is greater than 100.
However, it is interesting that in some state such as WY, TX, WV and SD, when age are fixed, the estimated premium cost for tobacco user will be lower than the estimated premium of non-tobacco user.

5 Difficulty

For the former CHTC process, data for state Florida is too large, which contains 671,874 rows. Therefore, when we do regression on the server by using R script, the program will "hold" due to memory limitation. Later on, when using HPC to perform overall regression for all states, the scale of our data is no longer limited.

6 Conclusion

In this project, we show the distribution of different benefits count and tobacco influence among different states. In most states tobacco use will increase the premium of insurance. But there are 4 states(WY, TX, WV and SD) have the opposite result.