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

November 30, 2020

1 Introduction

We want to find the number of distinct benefit count in each states and how does smoking habit and age affect the individual premium cost for insurance plan in different states. The data we use contains two part: the benefit data set and rate data set. For benefit data set which includes the copay for each individual, 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 do parallel job in CHTC for each state to do generalized linear regression for premium, area, age and tobacco. Finally, we find that for all states, the larger age is, the higher premium will be charged. In addition, in most states tobacco use will increase the premium charged. But it is interesting that in some states tobacco user will 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 18th, 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
RatingAreaId	Explanatory Variable	Identifier for the geographic rating area within a state
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.

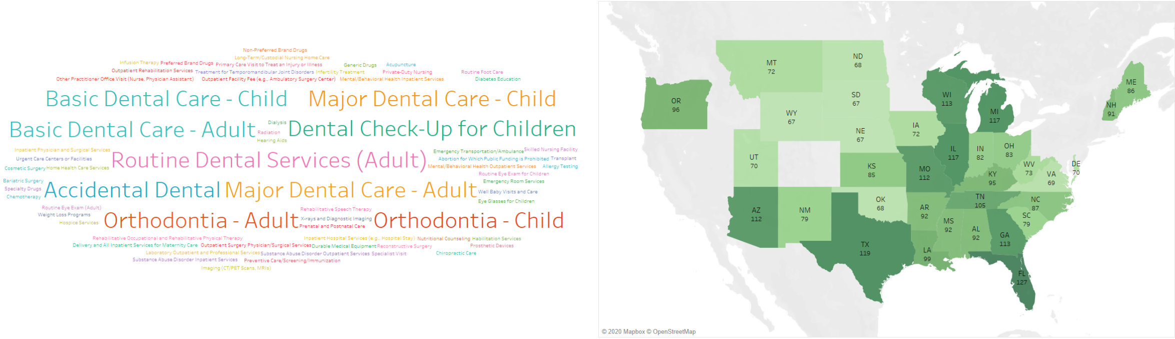
The impact of age and tobacco usage on the cost of premium draw our attention as well. There are 1658 different insurance plans with no preference on tobacco usage, but 3227 insurance plans that emphasize the smoking habit. In the following map plots and bar chart, mean costs of premium within each state 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.

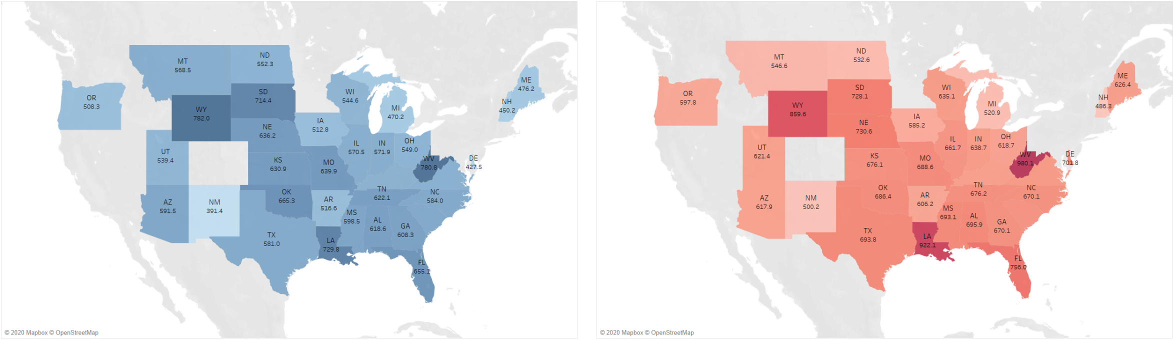
- The smoking habit increases the average cost of insurance plan in all states, which is not surprising. The average cost of premium for tobacco user is 1.64 times of that for non-tobacco user in DE, the minimum increase also reaches 7.6% in MO. Except for that, another attracting fact is that the tobacco 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) Non-Tobacco User

(b) Tobacco User

Figure 2: Cost of Average Premium For Non-Tabacco User and Tabacco User (2021)

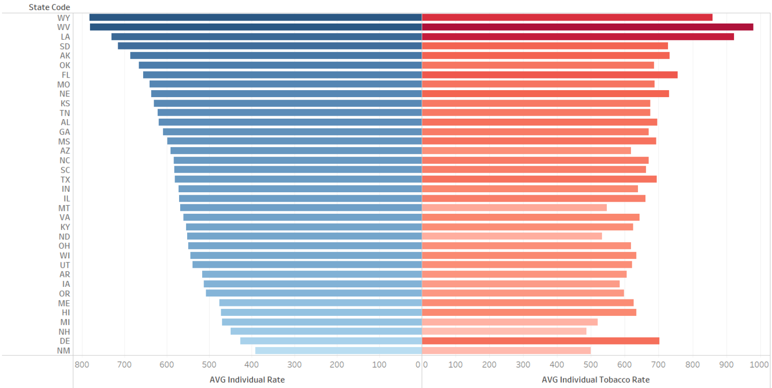


Figure 3: Side-by-side Bar Chart for Cost of Premium (2021)

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

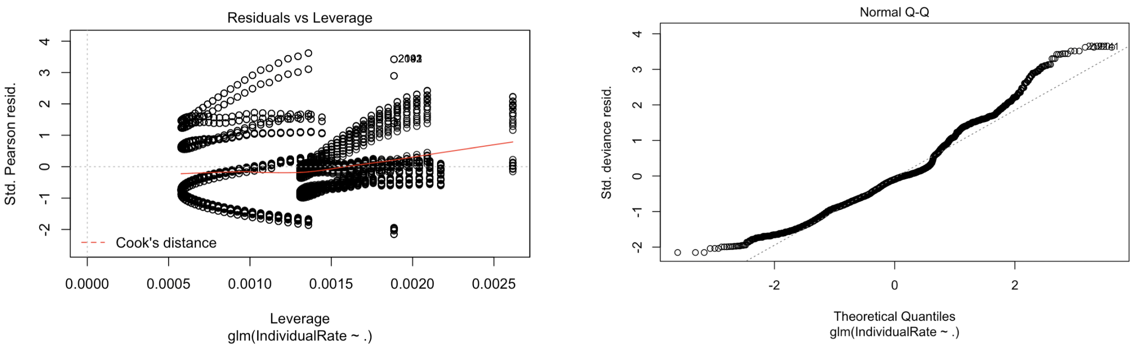
Tobacco	IndividualRate
Non-Tobacco User	866.14
Tobacco User	996.06
No Preference	482

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. General linear model is utilized to fit the statistical model and draw relevant inference.We denote "IndividualRate"as y, "Rating area id"as u,"Age" as w,"Tobacco" as v."Rating area id" and "Tobacco" are factor variables and "Age" is continous. Thus,our model can be formulated as below,

$$y = \sum_i a_i 1[u = x_i] + b * w + \sum_i c_i 1[v = z_i] + \epsilon, \epsilon \sim N(\mu, \sigma)$$

Since the whole data set is extremely large that cannot be calculated on our local computer nor WISC server, it is decomposed into 36 different set according to sample's state code (36 states in total). Besides, the decomposition enables parallel jobs on CHTC in a more efficient way. We firstly obtain the coefficients of these parameters, among which we are most interested in the coefficients of variable "Age" 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.



(a) Residuals.vs.Leverage

(b) Normal Q-Q

Figure 4: Model Diagnosis (Alaska)

We also do model diagnosis for the Alaska State. We don't find high leverage point from the plot and we think the model is close to normal distribution from plot above.

4 Results and Insights

For the influence of Tobacco and age in different state, We draw the coefficients of them in the map.

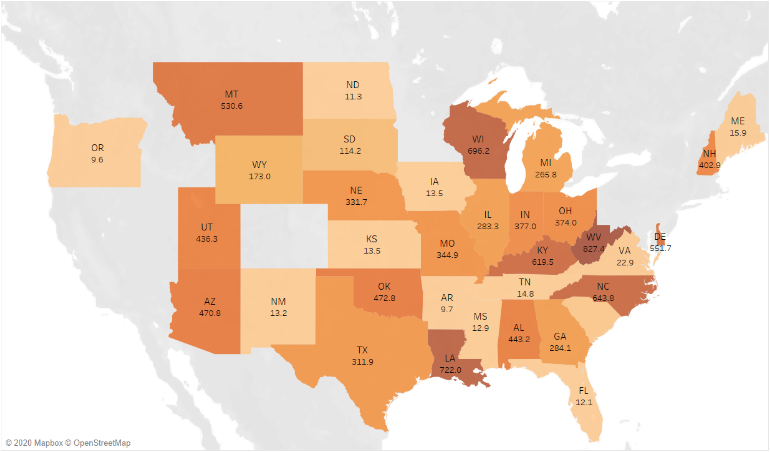
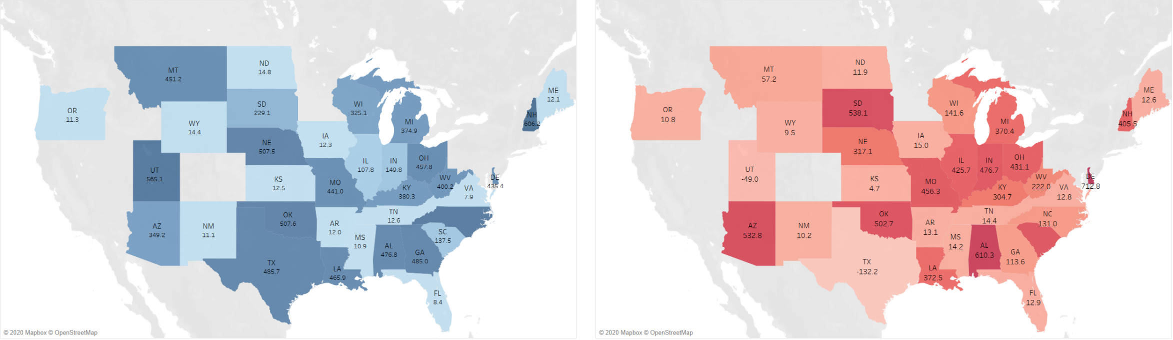


Figure 5: Age Coefficient (2021)

From Figure 5, we can see that if the tobacco use and area are fixed, the influence of age in WV, LA, WI, NC, KY are great. If the age of the individual increase by 1, the premuim of the insurance will increase by at least 600 in these states.



(a) Non-Tobacco User

(b) Tobacco User

Figure 6: Coefficient for Non-Tabacco User and Tabacco User (2021)

From figure 6, it is easy to find that when age and area of individual are fixed, if a person change from no preference to non-tobacco user, the premium will increase by at least 500 in UT, NE, OK and NH. If a person change from no preference to tobacco user, the premium of SD, AZ, OK, AL and DE will increase by at least 500. However, it is interesting that in some state such as OR, MT, ND, NE, KS, NM, TX and WI, when age and area are fixed, the premium of tobacco user are lower than the premium of non tobacco user and the premium of tobacco user are lower than the premium of no preference user.

5 Difficulty

When we do regression in parallel on the server, we got two problems.

- data in Florida is too large, it contains 671,874 rows. When we do regression on the server by using R script. The program will "hold". However, It only happens in Florida. The process can run successfully in other states.
- Some states only have one Rating Area ID. Since we consider Rating Area ID as a factor, It will raise an error when the state only contains one Rating Area ID. We encountered this problem in two states. To solve this problem, we just delete this variable and do regression then.

6 Conclusion

In this project, we show the distribution of different benefits count and age/tobacco influence among different states. In most states large age and tobacco use will increase the premium of insurance. In future work, we can try to find the relationship between premium and years.