Project Proposal

1 Group Members

Zhihao Zhao, zzhao357; Yuxiao Li, li2268; Yiran Wang, wang2559;
Wenyi Wang, wwang584; Zhiyu Ji, zji29

2 Description

The Health Insurance Marketplace Public Use Files contain data on health and dental plans offered to individuals and small businesses through the US Health Insurance Marketplace.
URL: https://www.cms.gov/CCIIO/Resources/Data-Resources/marketplace-puf

3 Questions

- Question1: What's the distribution pattern of different types of benefits divided by state or year?
- Question2: Are there any association between factors (age, smoking habit .etc) and amount of premium, copay and coinsurance for these benefits?
- Question3: Is it possible to build statistical model for prediction using these variables?

4 Code to read data

Use 2021 data as an example:

```
rate=read.csv("Rate_PUF.csv")
rate=rate[,c("BusinessYear","StateCode","Tobacco","Age",
"IndividualRate","IndividualTobaccoRate")]
rate$Age=as.numeric(rate$Age)
rate$IndividualTobaccoRate=as.numeric(rate$IndividualTobaccoRate)
head(rate)
summary(rate)
Benefit=read.csv("Benefits_Cost_Sharing_PUF.csv",nrows=100)
Benefit=Benefit[,c("BusinessYear","StateCode","BenefitName",
"CopayInnTier1","CopayInnTier2","CoinsInnTier1","CoinsInnTier2")]
head(Benefit)
summary(Benefit)
```

5 Variables Introduction

- `BusinessYear`
  Definition: Year for which plan provides coverage to enrollees
  Allowable Values: 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021

- `StateCode`
  Definition: Two-character state abbreviation indicating the state where the plan is offered
  Allowable Values: All 50 state abbreviations + 9 territory abbreviations

- `Age`
  Definition: Categorical indicator of whether a subscriber's age is used to determine rate eligibility for the insurance plan
  Allowable Values: 0-14, 15, 16. . . , 64 and over

- `BenefitName`
  Definition: Name assigned to benefit
  Allowable Values: Free text

- `Tobacco`
  Definition: Categorical indicator of whether a subscriber's tobacco use is used to determine rate eligibility for the insurance plan
  Allowable Values: Tobacco User/Non-Tobacco User

- . . .

6 Statistical Methods

We plan to use mean and variance statistics to analyze these variables we mentioned. Because of the large scale of the data, we plan to use parallel computing to obtain these values. Besides, we also plan to do some prediction, like do some linear regression or logistic regression so that we can understand the relation between these variables we interested.

7 Computational Tools and Computations

The original data is pretty large, so we decided to cut a small snippet to debug our R code and then run the code on Slurm. We can parallelize data cleaning procedure and it also works for some basic statistics like mean and variance. But since we expect to obstacle inner relationship between smoking habits and other factors by regression methods, it might not be appropriate to split it into parallel jobs.