PROJECT 1

Georgetown University

Graduate School of Arts & Science

Fengdi Li

Yifan Wu

Yarou Xu

**Data Science Problem**

In this project, our group are interested in the relationship between the health insurance coverage and the medical expenses.

Recent years, US medical expenses are becoming unfordable for more and more families. It is the third largest cause of bankruptcy for a family, and the first cause is the loss of job due to medical problems [1]. And, at the main time, healthcare cost increased tremendously. According to Brett O’Hara’s research, the total healthcare cost growth rate in 1996 was only 2%, while in 2001, it was already 10% [1]. Then from 2008 to 2010, the cost increased at a rate about three times inflation, reported by Charu Chandra, ect[2]. For individuals, Zeldin and Rukavina pointed out that in 2007, average indebted amounts of credit cards for non-medical expenses is $8333, and for medical expenses is $12515 [3]. Medical cost obviously is a huge burden, even for middle-class Americans [4]. In order to avoid going bankruptcy, besides improving physical quality, people can only turn to health insurance for help. And data showed that, health insurance can significantly minimize wealth depletion for older adults [5]. People without health insurance had 49% more debt than the average, however, those with Medicare or Medicaid had 29% and 13% less than the average, respectively.

From above research results, we are curious about that in order to reduce their bankruptcy probability, will more people buy health insurance if the medical costs are relatively higher in this area? So far, researches about health insurance and healthcare cost are most concentrate in how insurance reduces hospital expenses for individuals, nearly no people study the opposite relationship. However, if we found out the correlation of insurance occupation rate and hospital cost, maybe insurance companies can create more reasonable insurance products.

**Potential Analyzes that Can Be Conducted Using Collected Data**

In order to finish this project, we need to collect two datasets. The first one is about the insurance coverage in United State. And the second one is about the medical expenses in US. Then this two dataset need to be combined into one, so that we can conduct following analysis procedures. Therefore, we need to find two datasets contain a similar variable. Finally, we collect two datasets respectively contains following variables:

Dataset1:

|  |  |
| --- | --- |
| Variables |  |
| County\_State | County name and State name. |
| Number\_Insured | Number of people with health insurance. |
| NInsured\_Cl\_LowerBound | The lower bound of 90% confidence interval of the insured number. |
| NInsured\_Cl\_UpperBound | The upper bound of 90% confidence interval of the insured number. |
| Number\_Uninsured | Number of people without health insurance. |
| NUninsured\_Cl\_LowerBound | The lower bound of 90% confidence interval of the uninsured number. |
| NUninsured\_Cl-UpperBound | The upper bound of 90% confidence interval of the uninsured number. |
| time | Survey time. |
| State\_Code | FIPS code for state. |
| County\_Code | FIPS code for county. |

Why are these variables useful:

County\_State: We need to merge two data sets together using County and States information and also it is also helpful to keep a track of the hospital expenses in terms of geographical location.

Number\_Insured: We are here to investigate the relationship between people with health insurance and hospital expenses so the number of people with health insurance is critical.

NInsured\_CI\_Lowerbound & Ninsurance\_CI\_Upperbound: The 90% confidence interval is obtained here so we can get a whole picture of how the mean of the insured population of this county is distributed.

Number\_Uninsured: We are here to compare the rate of people with insurance, which is obtained using the formula: Number of people with insurance/ (Number of people with insurance + number of people without insurance). Numer\_Uninsured is here to obtain the number of total sample space.

NUninsured\_Lowerbound & NUinsured\_Upperbound: The 90% confidence interval is obtained here so we can get a whole picture of how the mean of the uninsured population of this county is distributed.

Time: It is critical to know that survey time as a future reference if we need to validate if result of our investigation is time-consistent.

State\_Code: It is easier to merge with future data set using FIPS code, which is the combination of State Code and County Code so state code is here to generate a new column of data indicating the unique county-specific geographical ID.

County\_Code: It is easier to merge with future data set using FIPS code, which is the combination of State Code and County Code so county code is here to generate a new column of data indicating the unique county-specific geographical ID.

Dataset2:

|  |  |
| --- | --- |
| Variables |  |
| Provider\_id | The id that hospital registered in Medicare. |
| County | County name. |
| State | State name. |
| Lower\_Payment\_Est | The estimated lower bound for certain treatments in this hospital. |
| Ave\_Payment | Average payment for certain treatments in this hospital. |
| Higher\_Payment\_Est | The estimated upper bound for certain treatments in this hospital. |
| Start\_Date | The start date for the survey. |
| End\_Date | The end date for the survey. |

Why are these variables useful:

Provider\_id: This is the unique ID-like code for each hospital investigated in this dataset. It serves as an index for each of these hospitals and it makes us easier for us to locate each row of hospital information.

County: Used as one of the geographical components to associate the hospital information with the insurance information in the previous date set. Together with the county code, it is used to merge two data set together.

State: Used as one of the geographical components to associate the hospital information with the insurance information in the previous date set. Together with the state code, it is used to merge two data set together.

Lower\_Payment\_Est & Higher\_Payment\_Est: The range is obtained here so we can get a whole picture of how much money people spend for certain treatments in this hospital.

Ave\_Payment: This is the most representative measurement for how much people spend for certain treatments in this hospital.

Start\_Date & End\_Date: Besides to keep track of the time of this survey, this set of time variables also serves the role of validating if the two date sets are within the same time period.

We conjectured that with higher hospital cost, there will be more people tend to buy health insurance, for two reasons. First, the hospital may charge higher if the area is richer than average considering that the number of people who are affordable for health insurance is bigger. Second, if seeing a doctor is too expensive, it may force people who are hesitant in health insurance decisions to finally buy it.

**Data Issues**

In our datasets, there are following issues:

Dataset1:

|  |  |  |
| --- | --- | --- |
| Variables | Issues | |
| Missing Value | Noise Value |
| County\_State |  | missing state or county name |
| Number\_Insured | Yes |  |
| NInsured\_Cl\_LowerBound | Yes |  |
| NInsured\_Cl\_UpperBound | Yes |  |
| Number\_Uninsured | Yes |  |
| NUninsured\_Cl\_LowerBound | Yes |  |
| NUninsured\_Cl-UpperBound | Yes |  |
| time |  |  |
| State\_Code |  |  |
| County\_Code |  |  |

Dataset2:

|  |  |  |
| --- | --- | --- |
| Variables | Issues |  |
| Missing Value | Noise Value |
| Provider\_id |  |  |
| County | Yes |  |
| State |  |  |
| Lower\_Payment\_Est | Yes |  |
| Ave\_Payment | Yes |  |
| Higher\_Payment\_Est | Yes |  |
| Start\_Date |  |  |
| End\_Date |  |  |

**Data Cleaning Effectiveness of the Algorithms:**

Before we performed any data cleaning on the collected data set, the cleanness for the first data set is 99.99 and for the second data set is 93.56 on a scale of 0-100. (Our data is relatively clean from the beginning.) After we performed the cleaning algorithms, the score turned to 100 for all variables we performed on. Thus during the cleaning process, we removed all of the missing and unreasonable entries and corrected all of the mismatching data type at the same time. We also organized the variables into a better form for future analysis.

**GitHub link:**

<https://github.com/AnalyticsProject/Project-1>

**References:**

[1] O'Hara, Brett. "Do medical out-of-pocket expenses thrust families into poverty?" Journal of Health Care for the Poor and underserved 15.1 (2004): 63-75.

[2] Chandra, Charu, Sameer Kumar, and Neha S. Ghildayal. "Hospital cost structure in the USA: what's behind the costs? A business case." International journal of health care quality assurance 24.4 (2011): 314-328.

[3] Zeldin, C., and M. Rukavina. "Borrowing to stay healthy: how credit card debt is related to medical expenses. The Access Project and Demos, 2007." (2008).

[4] Himmelstein, David U., et al. "Medical bankruptcy in the United States, 2007: results of a national study." The American journal of medicine 122.8 (2009): 741-746.

[5] Kim, Hyungsoo, Wonah Yoon, and Karen A. Zurlo. "Health shocks, out‐of‐pocket medical expenses and consumer debt among middle‐aged and older Americans." Journal of Consumer Affairs46.3 (2012): 357-380.