Factors associated with differences in Life Expectancy across the United States

A Capstone project:

Tonia Chu

Under the mentorship:

Dr. Vikas Ramachandra (Data Scientist at the Stanford Graduate School of Business, CA, United States)

For the course:

Data Science Career Track (Springboard)

1. INTRODUCTION to the problem

Addressing socioeconomic disparities in health is a major policy goal. Yet what is the magnitude of socioeconomic gaps in life expectancy? How these gaps are changing over time? And want are their determinants? Answers to these questions are not clear.

In this project, we use new data from 1.4 billion anonymous earnings and mortality records to construct more precise estimates of the relationship between income and life expectancy at the national level. We then construct new local area (county and metro area) estimates of life expectancy by income group and identify factors that are associated with higher levels of life expectancy for low-income individuals.

The purpose of this project is to characterize life expectancy by income, over time, and across areas. We will use de-identified data from tax records covering the US population from 2001-2014 to characterize income-mortality gradients. We will also characterize correlates of the spatial variation and construct publicly available statistics. We will build a model to predict the life expectancy of individuals by their age, income, living area and other aspects.

The analysis try to provide information for government agencies and health care companies to improve their services and environmental factors as well as help individuals to change their behaviors to get long life expectancy.

**In this study, I want to solve following problems:**

1. What is the shape of the income–life expectancy gradient?

2. How are gaps in life expectancy changing over time?

3. How do the gaps vary across local areas?

4. What are the factors associated with the longevity gap?

1. deeper dive into the data set

The dataset include 14 csv files of data tables. After loading these tables into python notebook, we can have a whole picture of the dataset. It provide life expectancy of people with different gender, household income, in different states, commuting zones, county during year 2001 to 2014. It also provide informations about fraction current smokers, fraction obese, percent uninsured, 30-day hospital mortality rate, percent of Medicare enrollees, percent religious, percent black, unemployment rate, labor force participation, population density and so on in commuting zones and county level. All these informations are important for the research.

The limitation of the dataset is that it doesn’t provide all the informations in each year during 2001 and 2014. So we can’t answer questions related with factors of life expectancy changing over time.

In data wrangling step, we first remove the unadjusted and Standard Error columns in the tables, then fill missing values in table 10 and table 12. There are 3 steps to fill missing values:

* A column is removed if there are more than 10% missing value.
* A commuting zone or county is removed if all the values of a column in that area are missing.
* Fill missing values with the mean value of that commuting zone or county.

1. exploration and initial findings

The datasets provide 14 formatted data tables in Excel format, along with documentation for each table. These tables contain life expectancy estimates by income and sex at various levels of geographic and temporal aggregation. I downloaded the data in Excel files.

1. APPROACH
2. Data preparation
3. Data Wrangling
4. Descriptive analysis
5. Exploratory Data Analysis
6. Data visualization
7. Modeling and Machine learning
8. Data Story
9. DELIVERABLES

Code, paper, slides