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.

After data wrangling, we save the tables in csv files table\_1.csv ~ table\_14.csv.

The python code of above jobs is in file Milestone-1.ipynb

1. exploration and initial findings

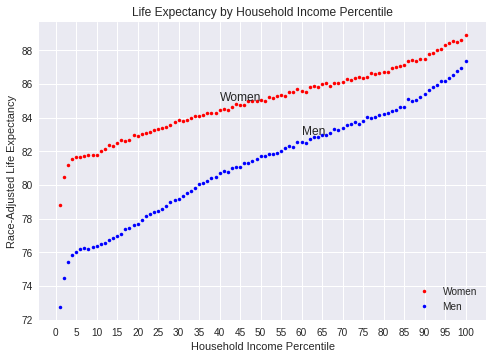
Next we will do exploratory data analysis. There are 3 steps analysis to answer the first three question:

1. **National Statistics on Income and Life Expectancy**

In this step, we’ll answer the question: What is the shape of the income–life expectancy gradient?

We use the dataset in table 1, which include National life expectancy estimates (pooling 2001-14) for men and women, by income percentile.

First we get the plot of Life Expectancy by Household Income Percentile of Men and Women:



From the plot, we can see that Race-adjusted life expectancy increases as Household income percentile increases, for both men and women. At the lowest 5 percentile, life expectancy increases very fast, then it turns slow and linear.

With the same household income percentile, women’s life expectancy is higher than that of the men, while the gap between top 1% and bottom 1% is less than that of the men:

Women, Bottom 1%: 78.8

Women, Top 1%: 88.9

Women, Life expectancy gap: 10.1

Men, Bottom 1%: 72.7

Men, Top 1%: 87.3

Men, Life expectancy gap: 14.6

But the gender gap decrease as income percentile increase:

Gender gap, Bottom 1%: 6.0

Gender gap, Top 1%: 1.5

When use linear regression, we can get the income–life expectancy gradient:

Women, Slope of linear regression: 0.07

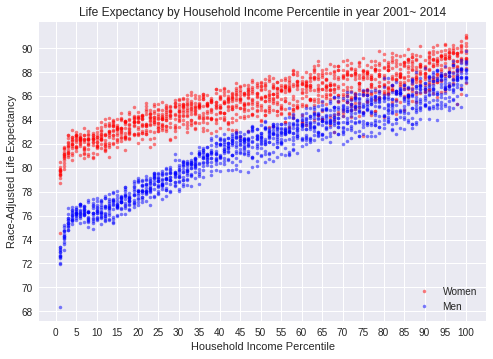
Men, Slope of linear regression: 0.11

**2. Trends in Life expectancy by year 2001~2014**

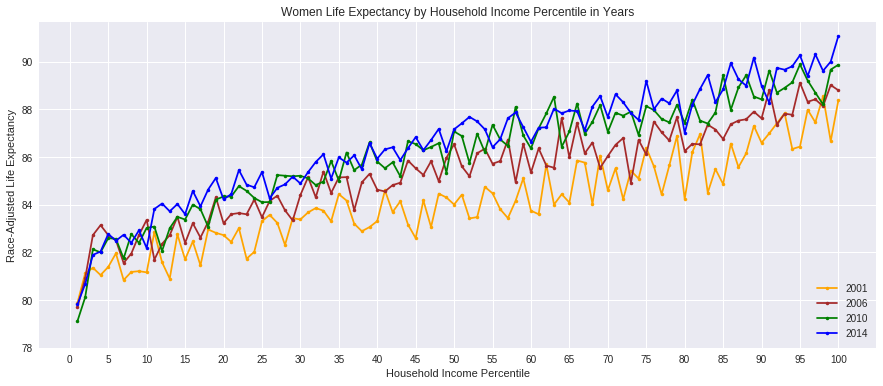
In this step, we’ll answer the question: How are gaps in life expectancy changing over time?

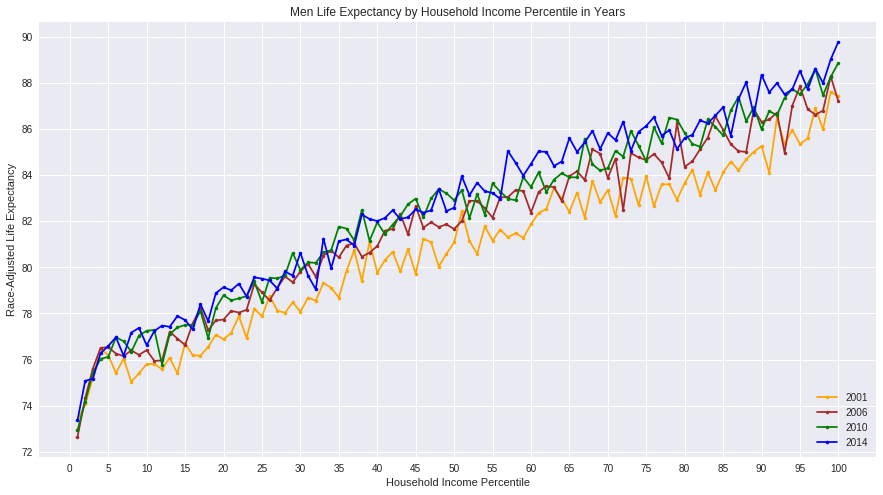
We use the dataset in table 2, which include National by-year life expectancy estimates for men and women, by income percentile.

First, we get the plot of Life Expectancy by Household Income Percentile of Men and Women in year 2001~2014:

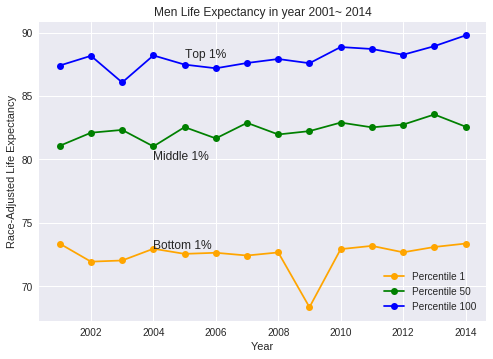
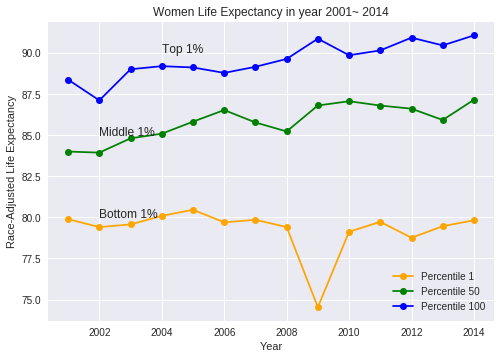
 From the plot, we can see that the life expectancy vary by year and trends are almost the same.

Second, we choose 4 years’ data to have a close look:

 Above is the women life expectancy by household income in years 2001, 2006, 2010 and 2014. With a close look, we can see that the life expectancy increases by year.

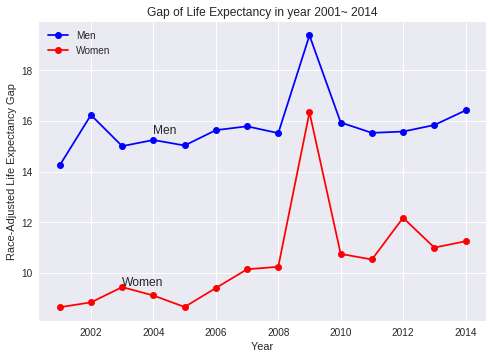
 From above figure, we can see that men have the same feature, but vary less than women by years.

Third, let’s have a look at the life expectancy trend of bottom, middle and top household income percentiles in years 2001~2014:



We can see that for people with bottom 1 percentile household income, their life expectancy didn’t increase by year, while for people with top 1 percentile household income, their life expectancy did increased over years.

Below is the Gap of life expectancy’s trend by years:

 From the figure, we can see that men’s gap is much higher than that of the women, and the gap increased by year.

### **3. Local Area Variation of Life expectancy gaps**

### In this step, we’ll answer the question: How do the gaps vary across local areas?

### We use the dataset in table 3, which include State-level life expectancy estimates for men and women, by income quartile.

### First, let’s have a look at the state level variation of life expectancy.

### The 5 states with the highest women life expectancy of bottom quartile income are: Maine, New York, Vermont, Massachusetts, and North Dakota. While the 5 states with the lowest women life expectancy of bottom quartile income are: Nevada, Oklahoma, Indiana, Hawaii and Michigan.

### The 5 states with the highest men life expectancy of bottom quartile income are: California, New York, Montana, Idaho, and Vermont. While the 5 states with the lowest men life expectancy of bottom quartile income are: Indiana, Oklahoma, Nevada, Alabama and Tennessee.

### Below are the figures of women and men life expectancy with quartile household income by states.

### From above figures, we can see that life expectancy of different states vary a lot.

### Then, let’s check life expectancy gap by states.

### The 5 states with the highest Men gap life expectancy are: District Of Columbia, Wyoming, Indiana, Ohio and Delaware. The 5 states with the highest women gap life expectancy are: Kansas, Iowa, Michigan, Indiana, and Oklahoma.

### The 5 states with the lowest Men gap life expectancy are: California, New York, New Jersey, Hawaii, and Illinois. The 5 states with the lowest women gap life expectancy are: California, Hawaii, New York, New Jersey, and Florida.

### Next, we’ll check CZ(Commuting Zone) level variation of life expectancy. This time we use the dataset in table 7, which include CZ-level life expectancy estimates for men and women, by income ventile.

### Following are CZs with the highest life expectancy gap:

### And Czs with the lowest life expectancy gap:

### From the results above, we choose commuting zones in California, New York, Indiana and Michigan for research.

### We can see that Men’s life expectancy gaps are higher than that of the women and the gaps vary in different Czs.

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