Using Big Data To Solve Economic and Social Problems

Professor Raj Chetty

Head Section Leader Rebecca Toseland

Photo Credit: Florida Atlantic University

Improving Health Outcomes

- Research in economics typically focuses on earnings or wealth as key outcomes of interest

- But most people view health and life expectancy as among the most important aspects of well-being

- What interventions are most effective in improving health (holding fixed current frontier of medical technology)?

    - Research on these issues spans multiple fields, from epidemiology and public health to economics

Epidemiology and Public Health

- One common approach: randomized trials
  - Ex.: vary exercise regimes and examine impacts on short-term health outcomes
- Difficult to implement especially when studying long-term effects
  - Use observational data to estimate correlations, but many pitfalls
  - Ex: People who report dieting in a phone survey weighed more on average → dieting counterproductive?

Health Economics: Markets for Healthcare

- Health is a very complex market with many non-standard features

    1. Patients have private information → asymmetric information

    2. Hard for patients to judge quality and decide what to buy

    3. Third-party payers (insurance companies) → moral hazard

- Escalating costs of healthcare in America (now 17% of GDP)

  – Particularly timely topic in the context of political debate on Affordable Care Act (Obamacare vs. Trumpcare)

Improving Health Outcomes: Overview

- This lecture illustrates how big data is helping us learn how to improve health, in three segments:

1. Descriptive analysis of health outcomes in U.S. population [method: survival analysis]

   Chetty, Stepner, Abraham, Lin, Scuderi, Bergeron, Cutler. “The Association Between Income and Life Expectancy in the United States” *JAMA* 2016.

Improving Health Outcomes: Overview

- This lecture illustrates how big data is helping us learn how to improve health, in three segments:

1. Descriptive analysis of health outcomes in U.S. population
   [method: survival analysis]

2. Epidemiology application: using big data to forecast pandemics
   [method: predictive modeling]

   Ginsberg, Mohebbi, Patel, Brammer, Smolinski, Brilliant. “Detecting Influenza Epidemics Using Search Engine Query Data.” *Nature* 2009.

   Lazer, Kennedy, King, Vespignani. “The Parable of Google Flu: Traps in Big Data Analysis.” *Science* 2014.

Improving Health Outcomes: Overview

- This lecture illustrates how big data is helping us learn how to improve health, in three segments:

1. Descriptive analysis of health outcomes in U.S. population
   [method: survival analysis]

2. Epidemiology application: using big data to forecast pandemics
   [method: predictive modeling]

3. Economics applications: impacts of health insurance coverage
   [method: regression discontinuities]

   Wherry, Miller, Kaestner, Meyer. “Childhood Medicaid Coverage and Later Life Health Care Utilization” *REStat* 2017.

Income and Life Expectancy

- Most common measure of health: mortality rates
  - Crude but well measured in population data
- Begin with basic descriptive facts about life expectancy in America
- Chetty et al. (2016) examine relationship between life expectancy and income
  - Use data on entire U.S. population from 1999-2013 (1.4 billion observations)

Estimating Life Expectancy: Data

- Mortality measured using Social Security death records
- Income measured at household level using tax returns
- Focus on percentile ranks in income distribution
  - Rank individuals in national income distribution within birth cohort, gender, and tax year

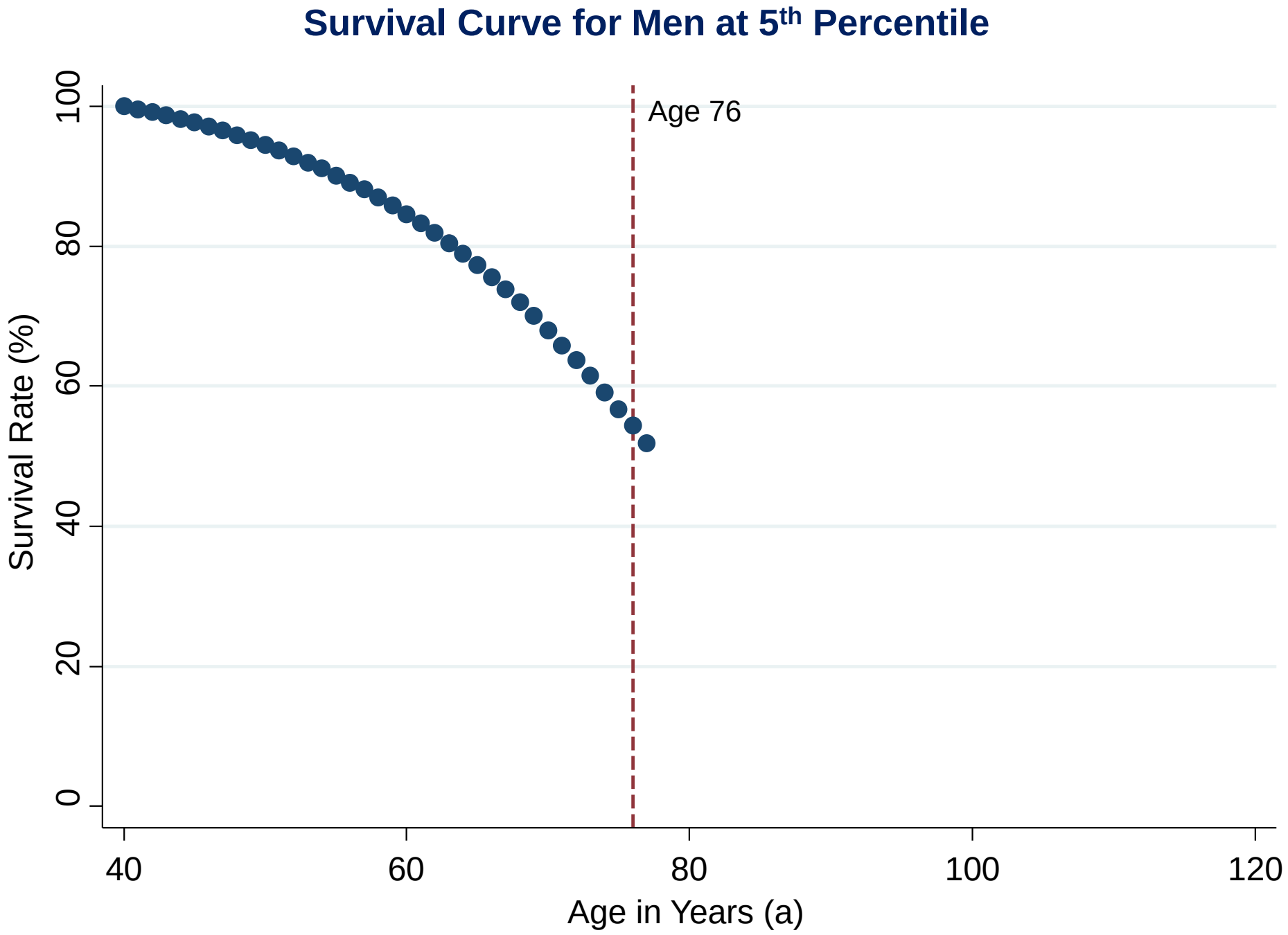
Methodology to Estimate Life Expectancy

- Goal: estimate expected age of death conditional on an individual's income at age 40, controlling for differences in race and ethnicity

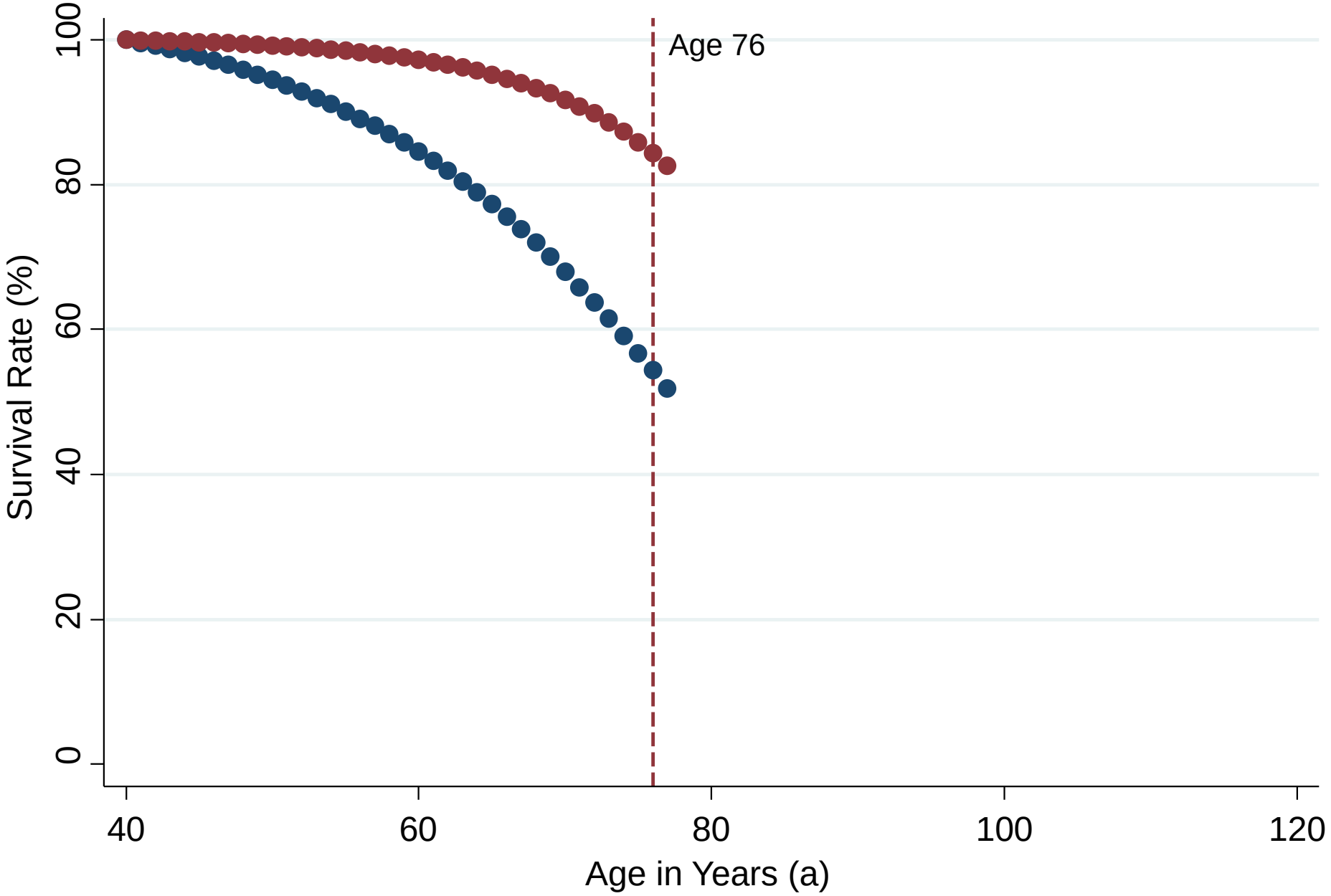
    - *Period* life expectancy: life expectancy for a hypothetical individual who experiences mortality rates at each age observed in a given year

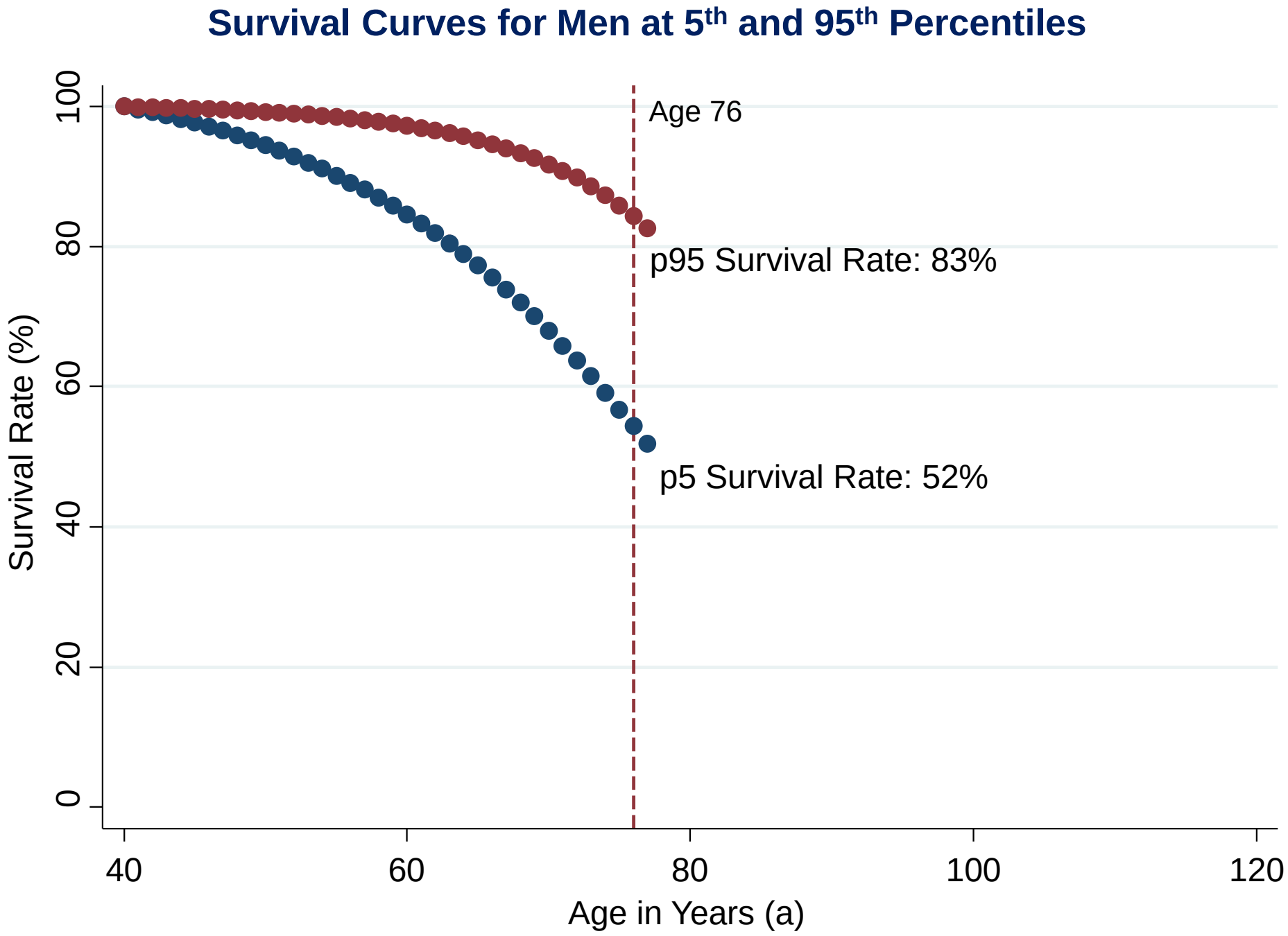
- Three steps:

    1. Calculate mortality rates by income rank and age for observed ages
    2. Estimate a survival model to extrapolate to older ages
    3. Adjust for racial differences in mortality rates


Survival Curve for Men at 5th Percentile
Age 76
Survival Rate (%)
0
20
40
60
80
100
40
60
80
100
120
Age in Years (a)

Survival Curves for Men at 5th and 95th Percentiles




Survival Curves for Men at 5th and 95th Percentiles
Age 76
p95 Survival Rate: 83%
p5 Survival Rate: 52%
Survival Rate (%)
0
20
40
60
80
100
40
60
80
100
120
Age in Years (a)

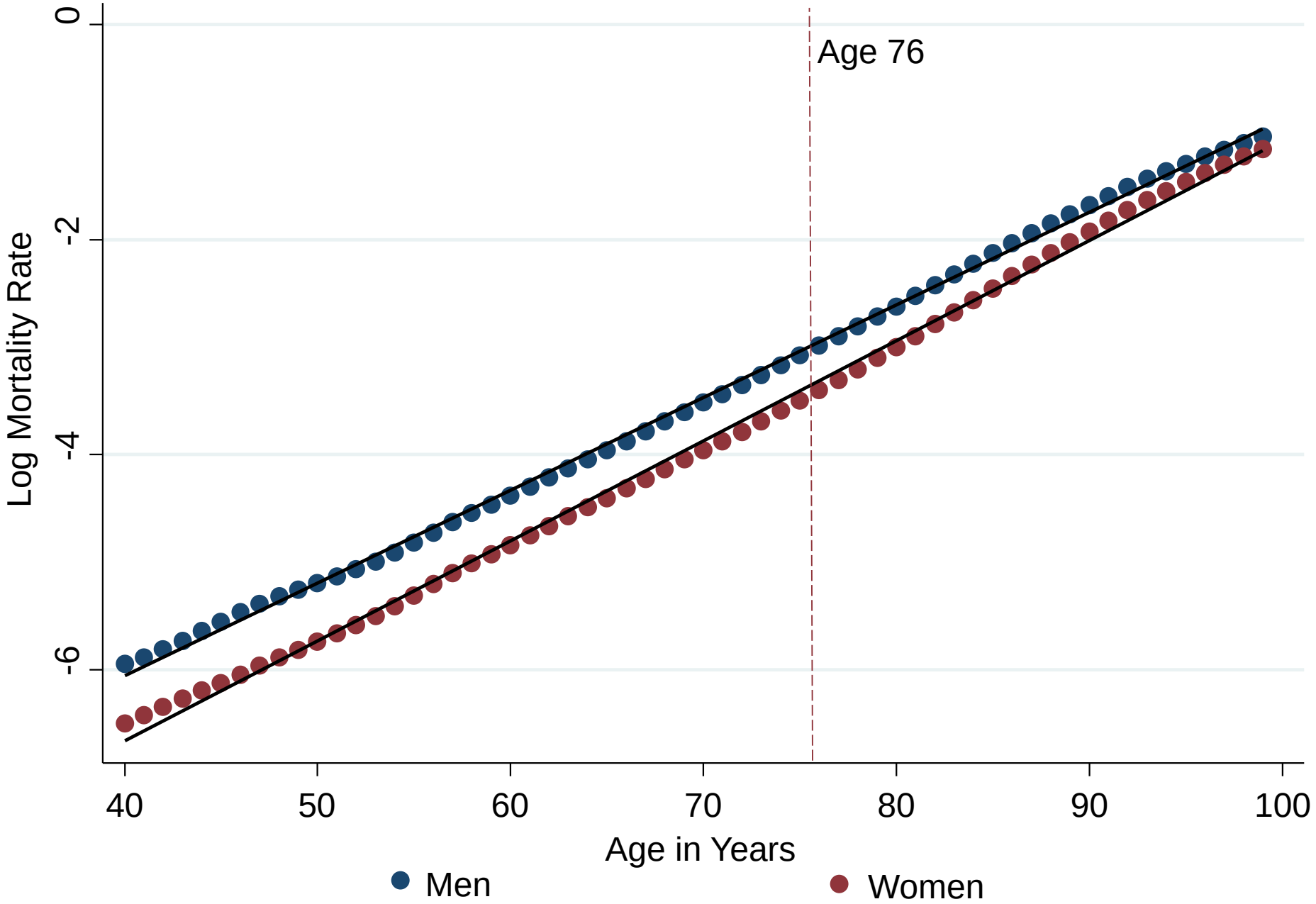
Step 2: Predicting Mortality Rates at Older Ages

- To calculate life expectancy, need estimates of mortality rates beyond age 76

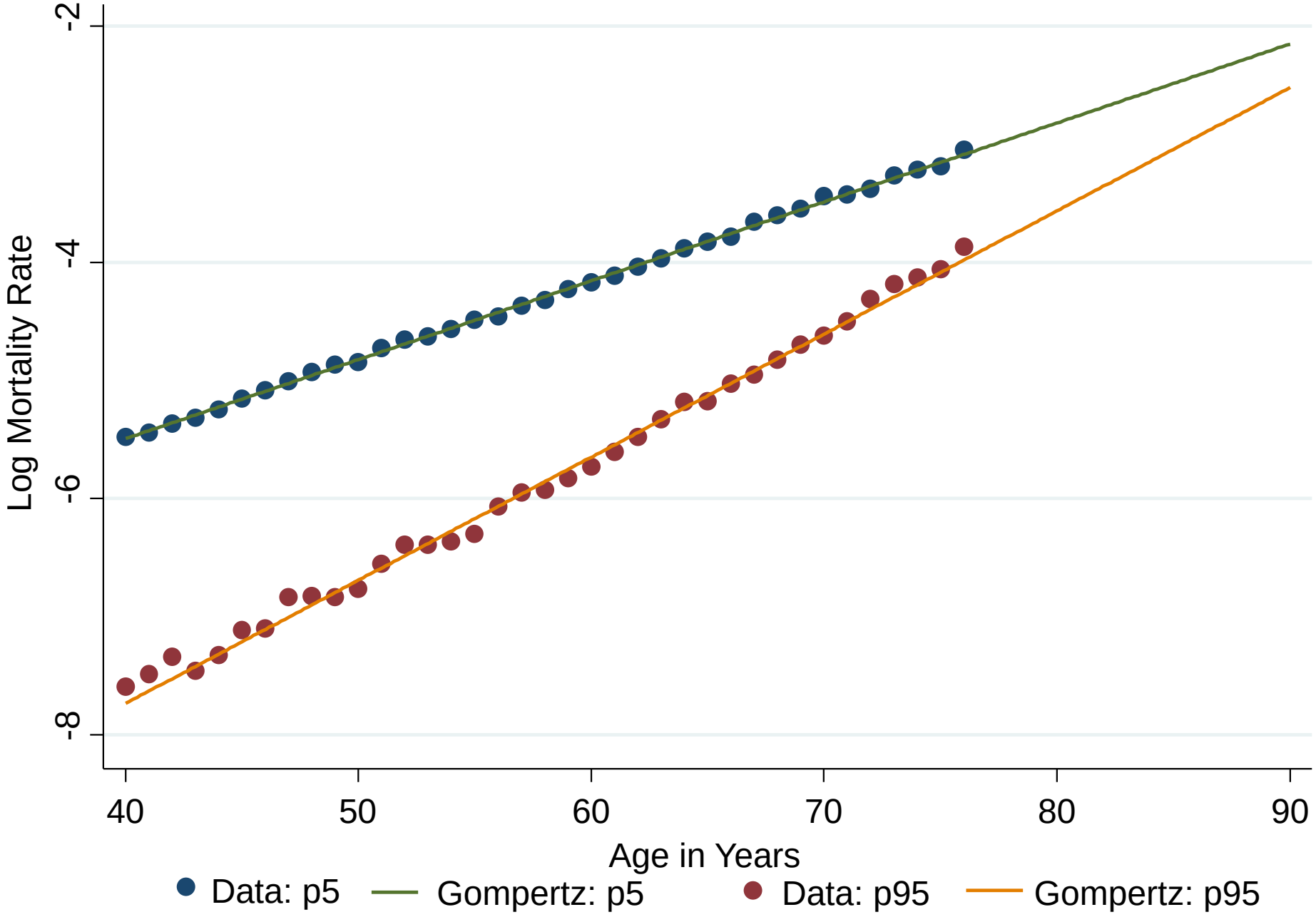
- Gompertz (1825) documented a robust empirical pattern: mortality rates grow exponentially with age

$$m(a) = k\, e^{\beta a}$$
$$\Rightarrow \log m(a) = \kappa + \beta a$$

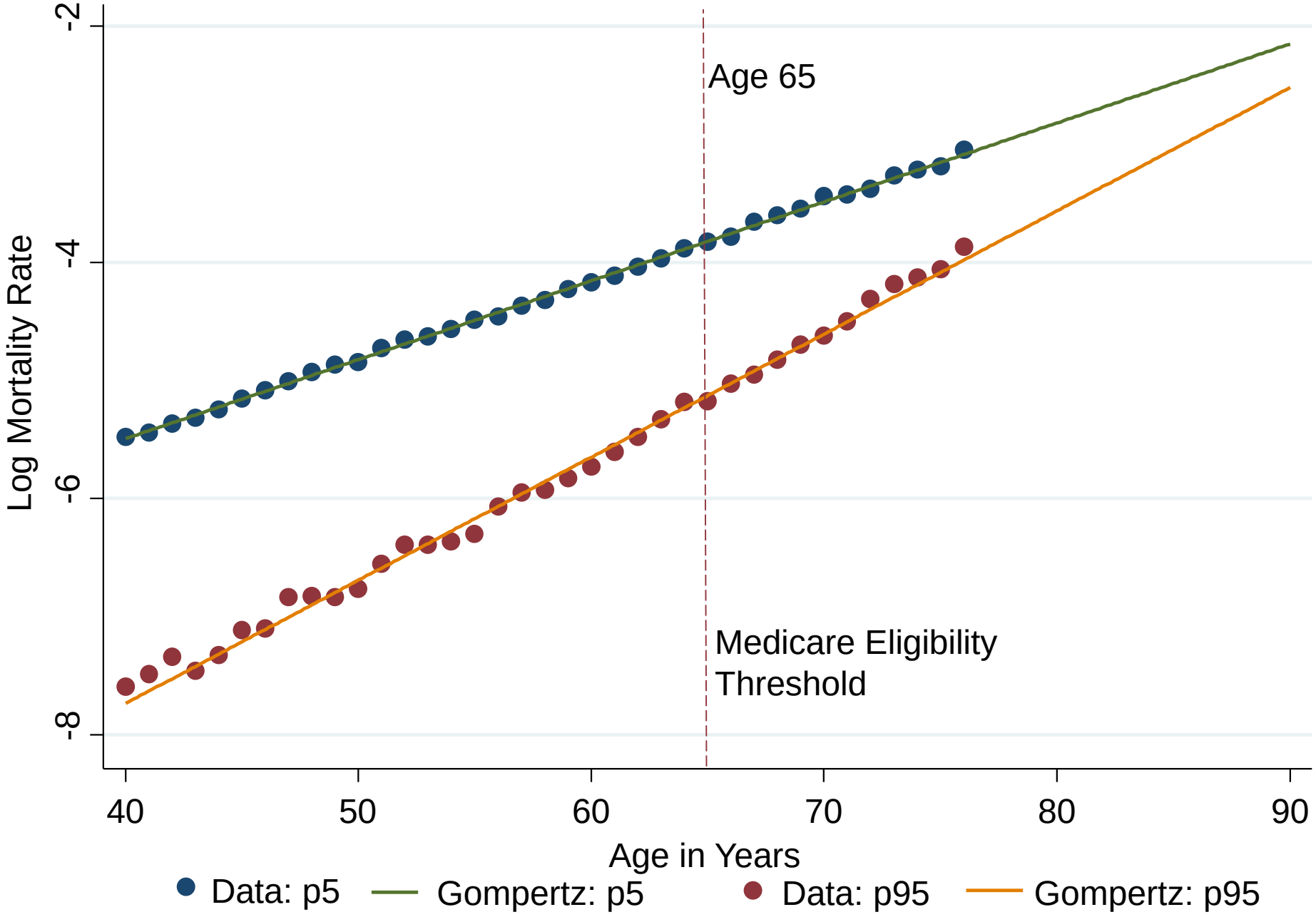
Mortality Rates by Gender in the United States in 2001: CDC Data

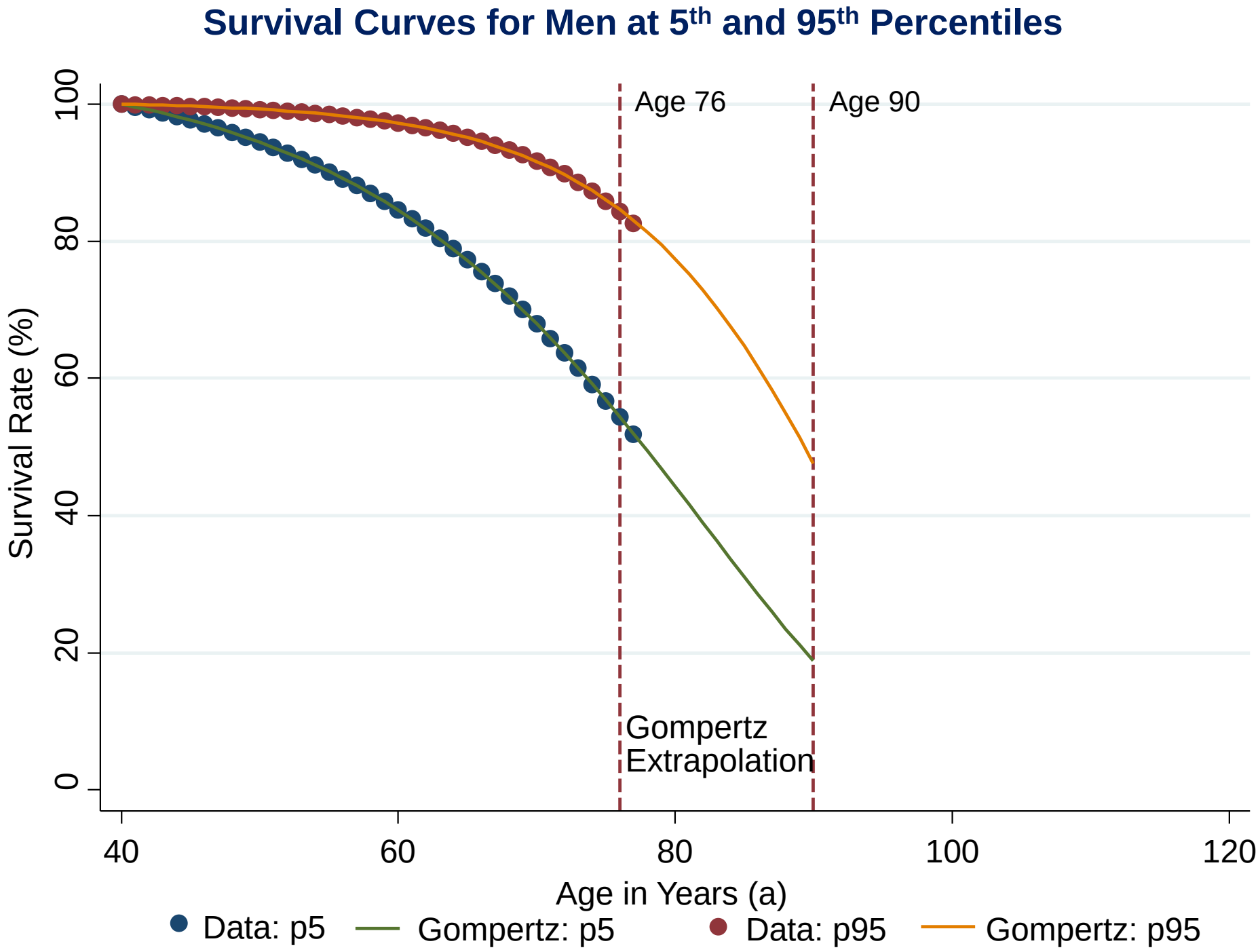


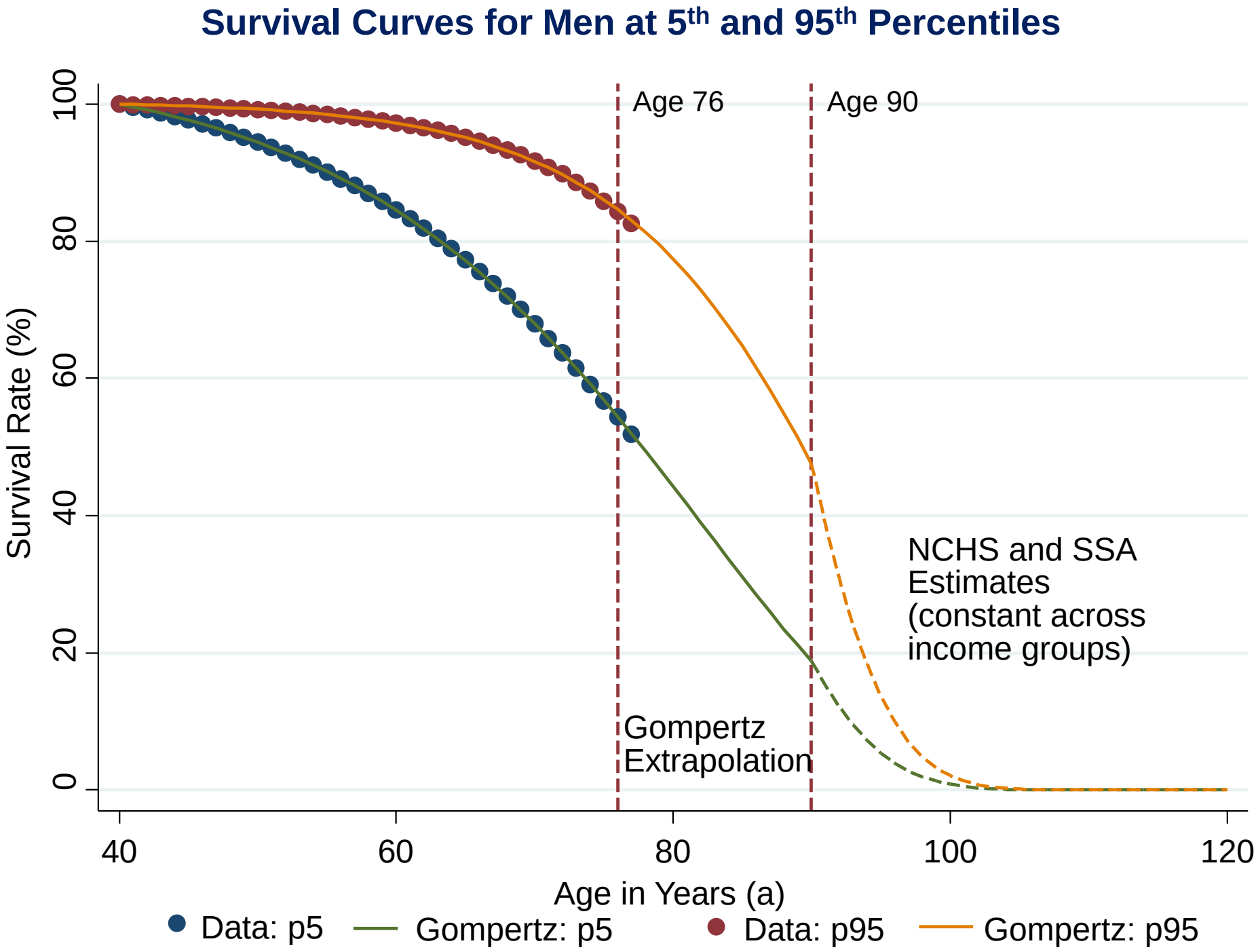
Log Mortality Rates for Men at 5th and 95th Percentiles



Log Mortality Rates for Men at 5th and 95th Percentiles




Survival Curves for Men at 5th and 95th Percentiles
Survival Rate (%)
0
20
40
60
80
100
Age 76
Age 90
Gompertz
Extrapolation
40
60
80
100
120
Age in Years (a)
Data: p5
Gompertz: p5
Data: p95
Gompertz: p95


Survival Curves for Men at 5th and 95th Percentiles
Survival Rate (%)
0
20
40
60
80
100
Age 76
Age 90
NCHS and SSA Estimates (constant across income groups)
Gompertz Extrapolation
40
60
80
100
120
Age in Years (a)
Data: p5
Gompertz: p5
Data: p95
Gompertz: p95

Race and Ethnicity Adjustment

- CDC data: for males, life expectancy of whites is 3.8 years higher than blacks and 2.7 years lower than Hispanics

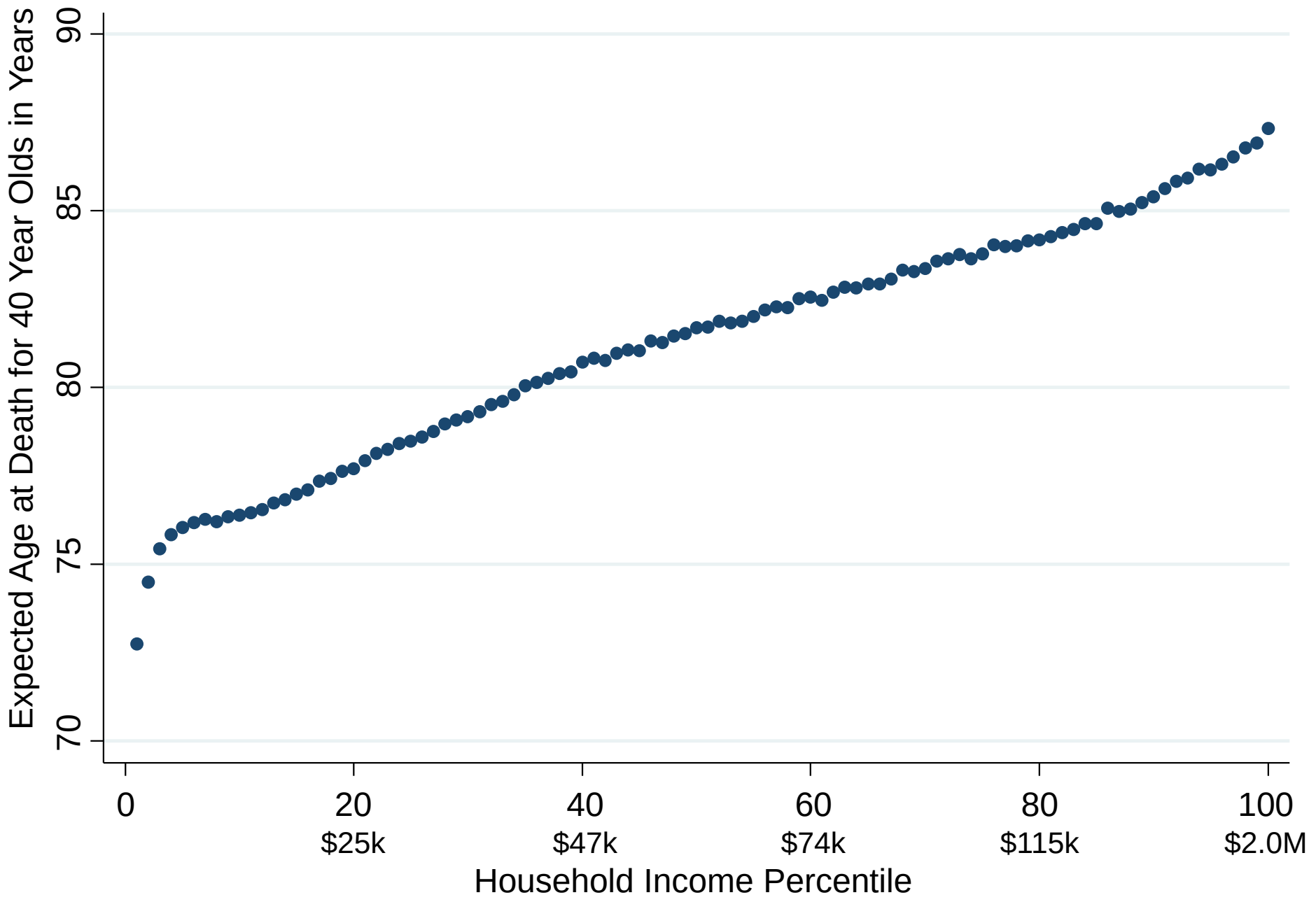
- Adjust for such racial and ethnic differences as follows:

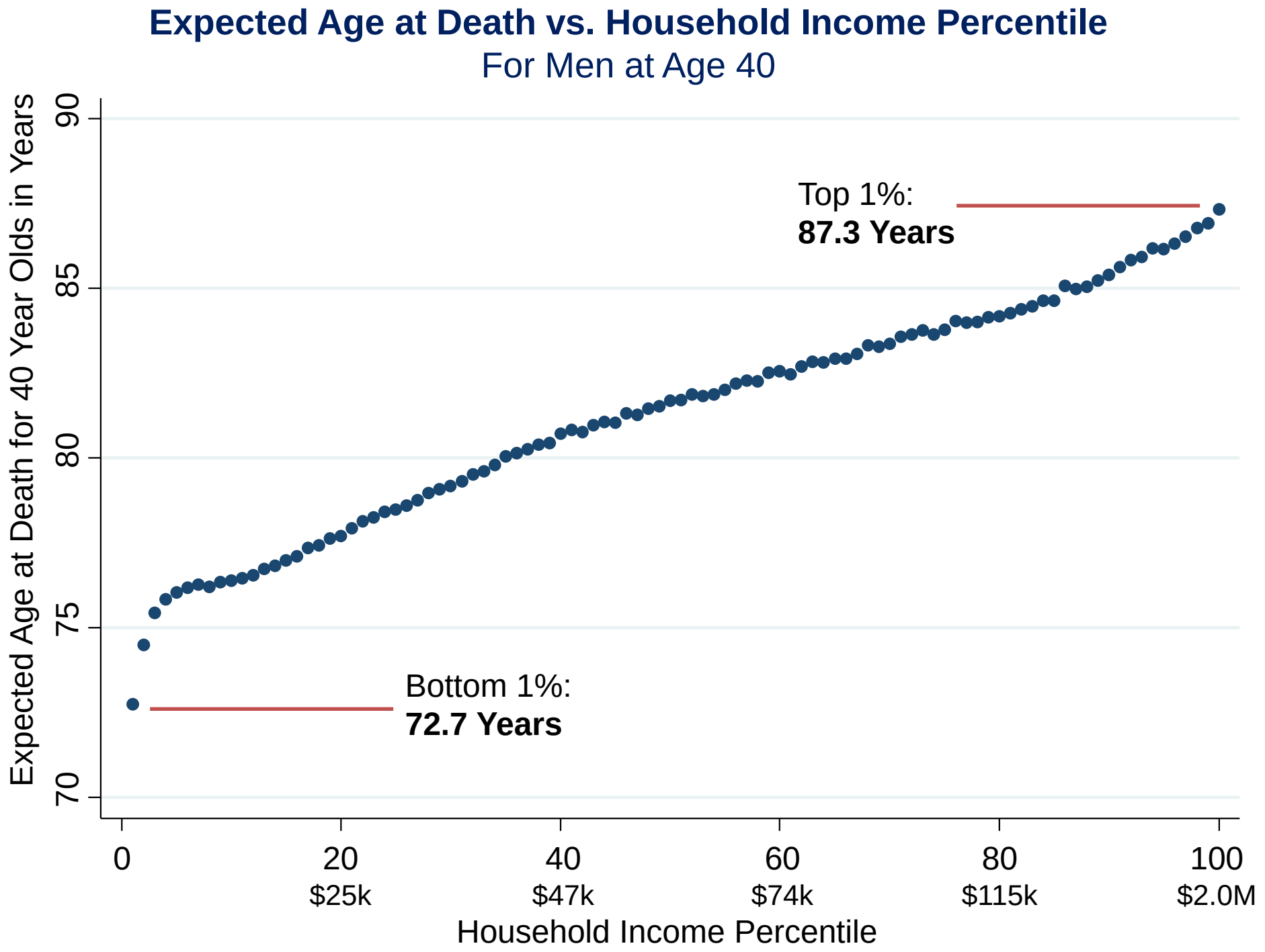
  - Use National Longitudinal Mortality Study to estimate racial and ethnic differences in mortality rates by age, controlling for income

  - Use Census data to construct race- and ethnicity-adjusted estimates of life expectancy, to answer the question:

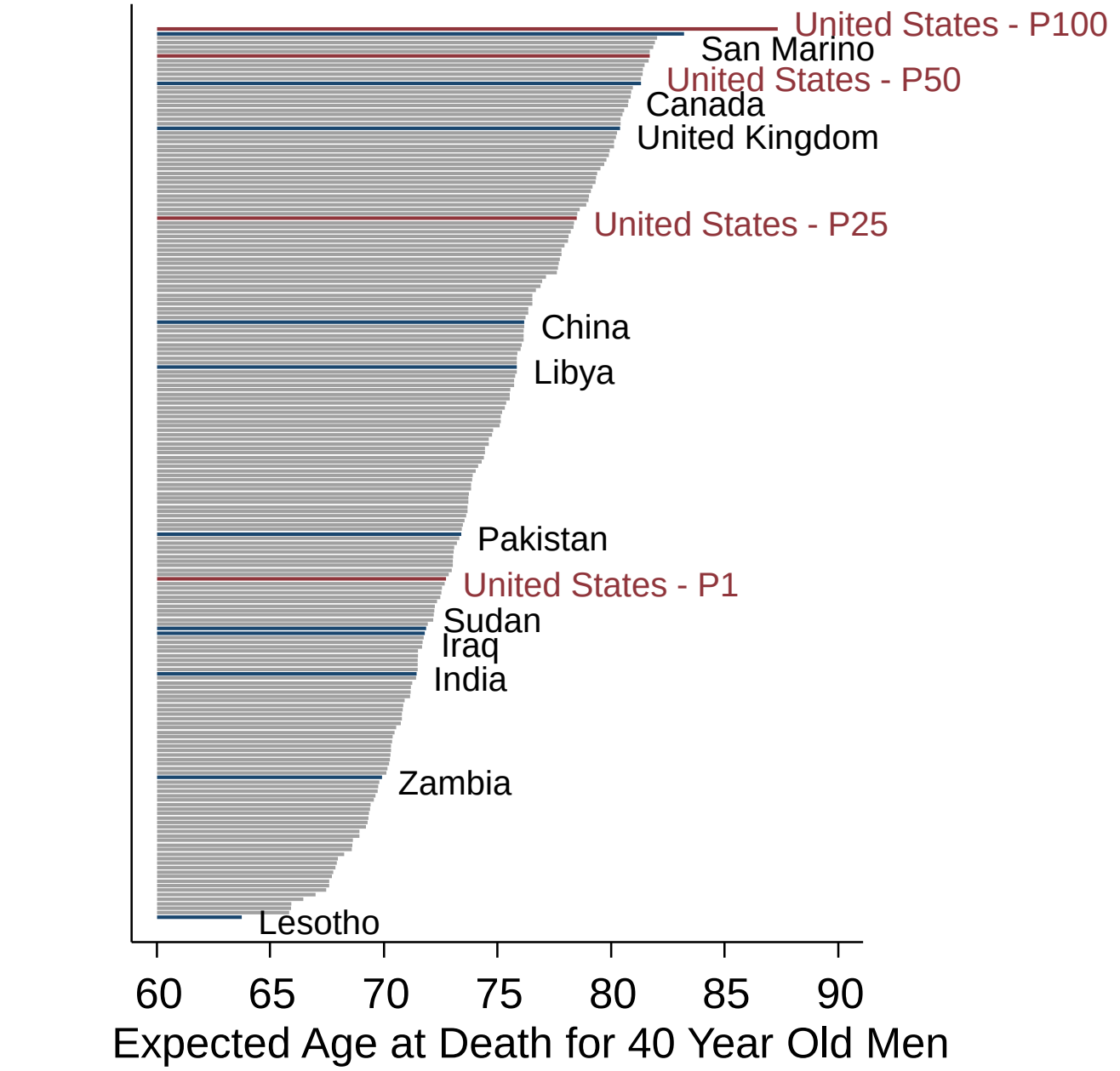
*"What would life expectancy be if each **income group** and **area** had the same black, Hispanic and Asian shares as the U.S. population as a whole at age 40?"*

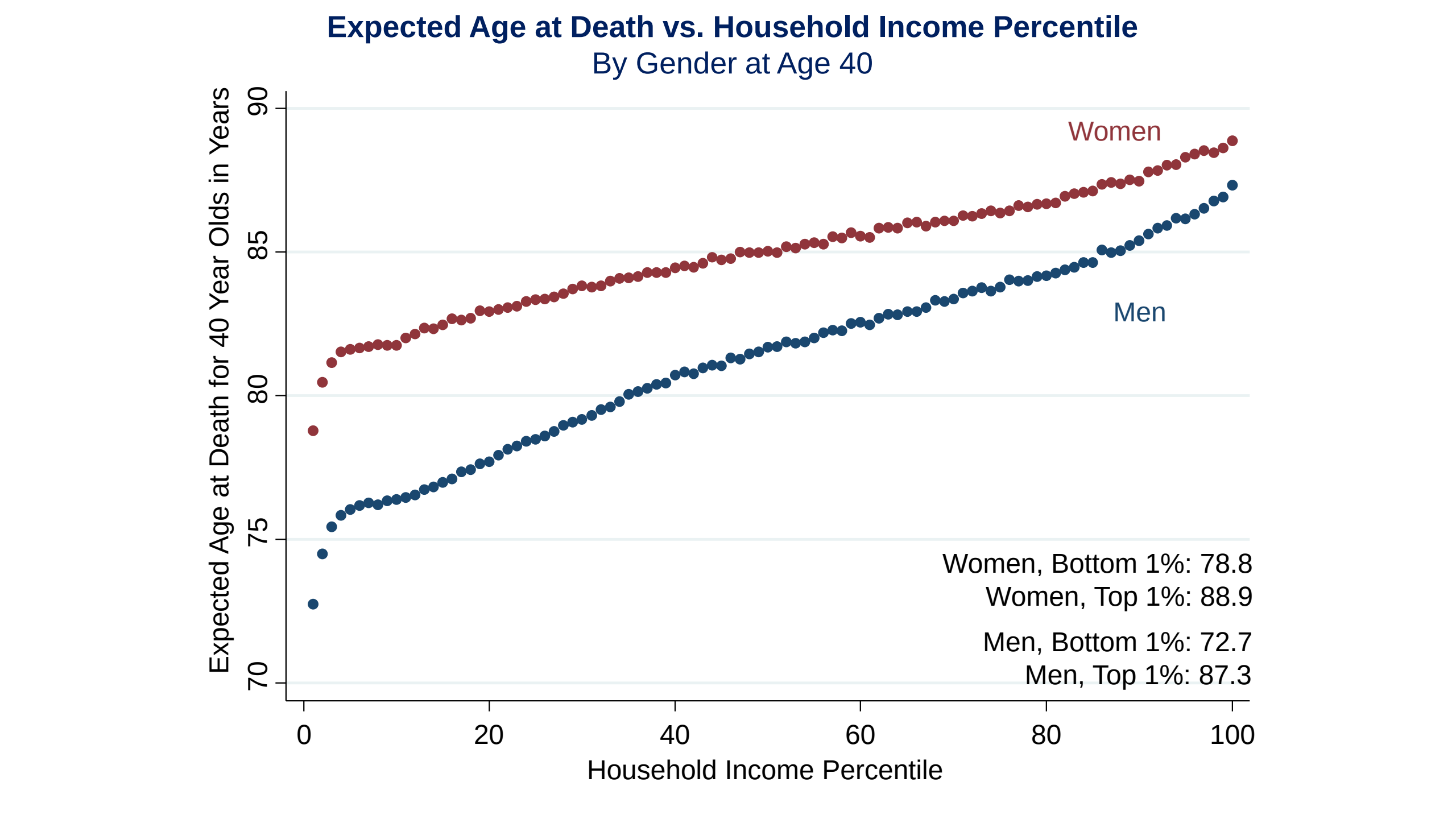
National Statistics on Income and Life Expectancy


Expected Age at Death vs. Household Income Percentile
For Men at Age 40
Expected Age at Death for 40 Year Olds in Years
70
75
80
85
90
0
20
40
60
80
100
$25k
$47k
$74k
$115k
$2.0M
Household Income Percentile


Expected Age at Death vs. Household Income Percentile
For Men at Age 40
Expected Age at Death for 40 Year Olds in Years
70
75
80
85
90
Top 1%:
87.3 Years
Bottom 1%:
72.7 Years
0
20
$25k
40
$47k
60
$74k
80
$115k
100
$2.0M
Household Income Percentile

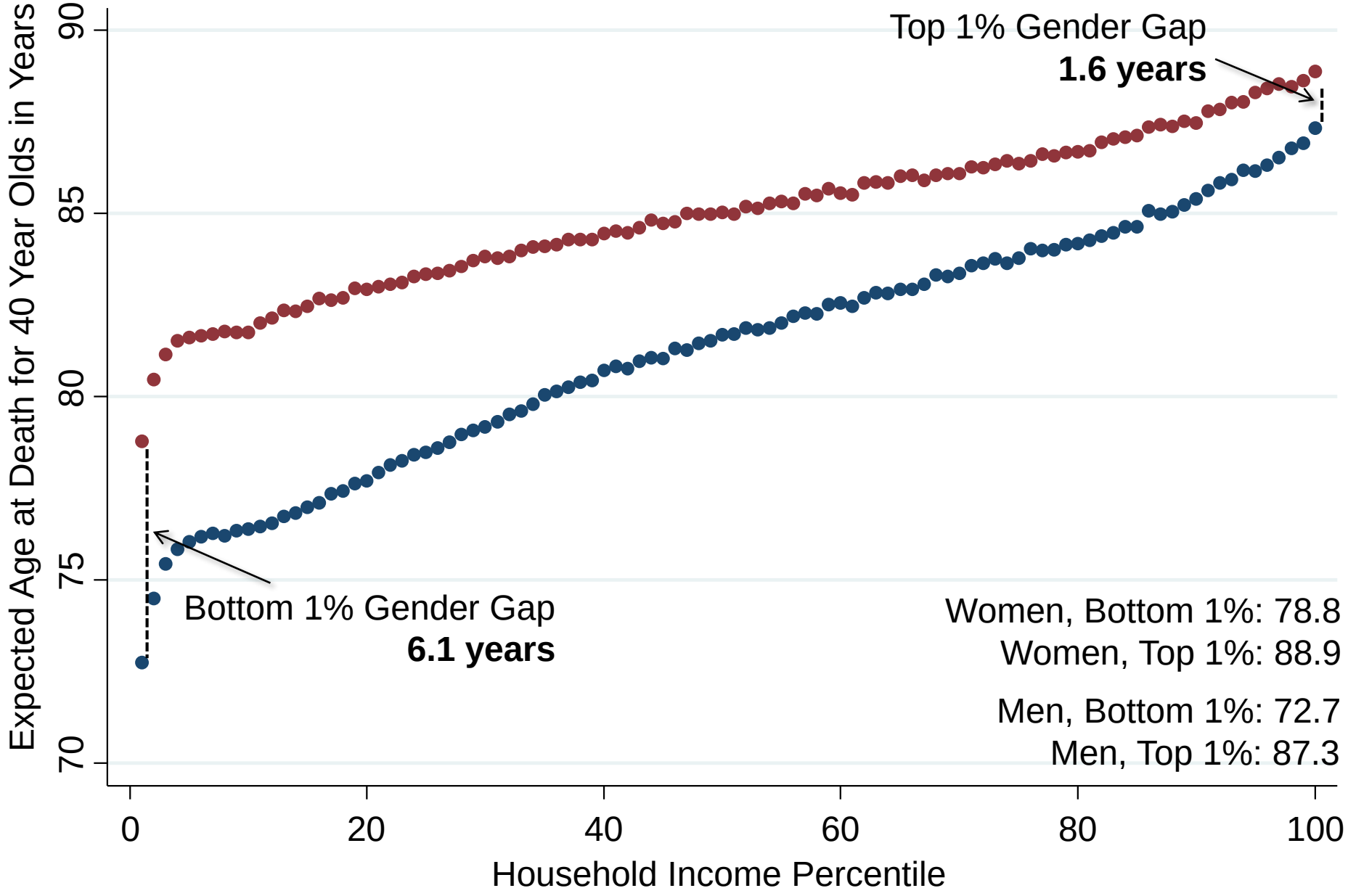
U.S. Life Expectancies by Percentile in Comparison to Mean Life Expectancies Across Countries




Expected Age at Death vs. Household Income Percentile
By Gender at Age 40
Expected Age at Death for 40 Year Olds in Years
70
75
80
85
90
0
20
40
60
80
100
Household Income Percentile
Women
Men
Women, Bottom 1%: 78.8
Women, Top 1%: 88.9
Men, Bottom 1%: 72.7
Men, Top 1%: 87.3

Expected Age at Death vs. Household Income Percentile

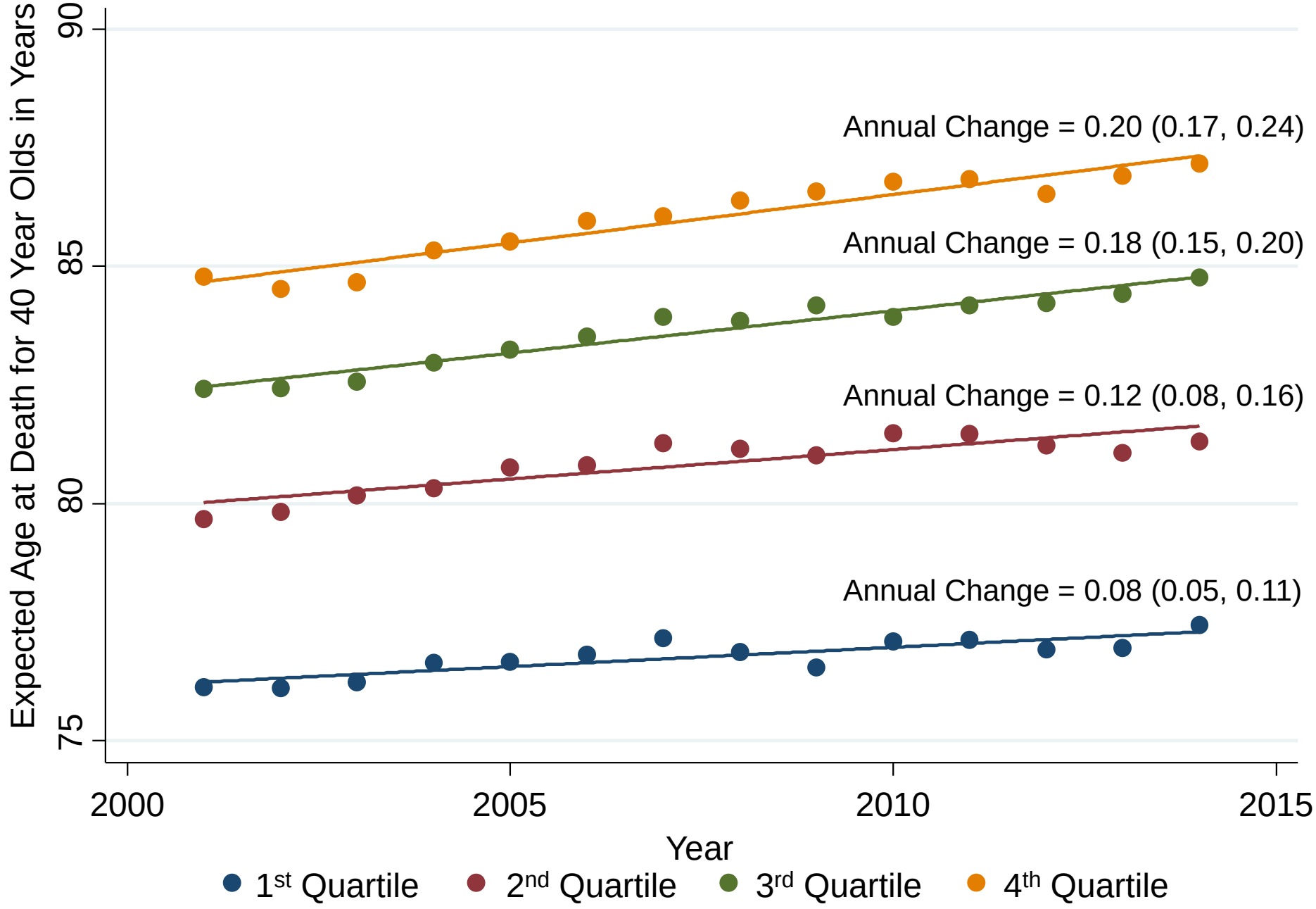
By Gender at Age 40



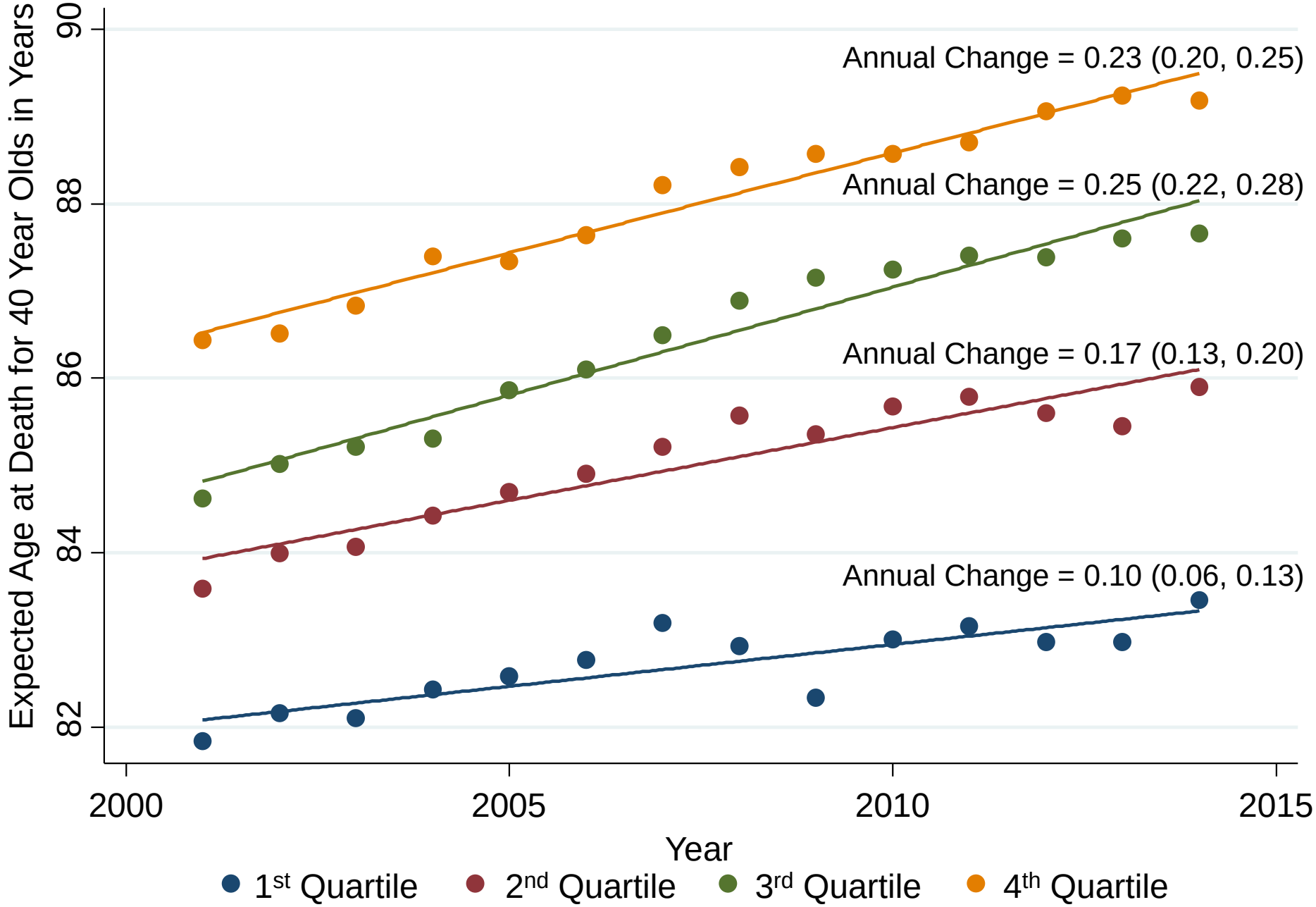
Time Trends

- How are gaps in life expectancy changing over time?

Trends in Expected Age at Death by Income Quartile in the US For Men Age 40, 2001-2014

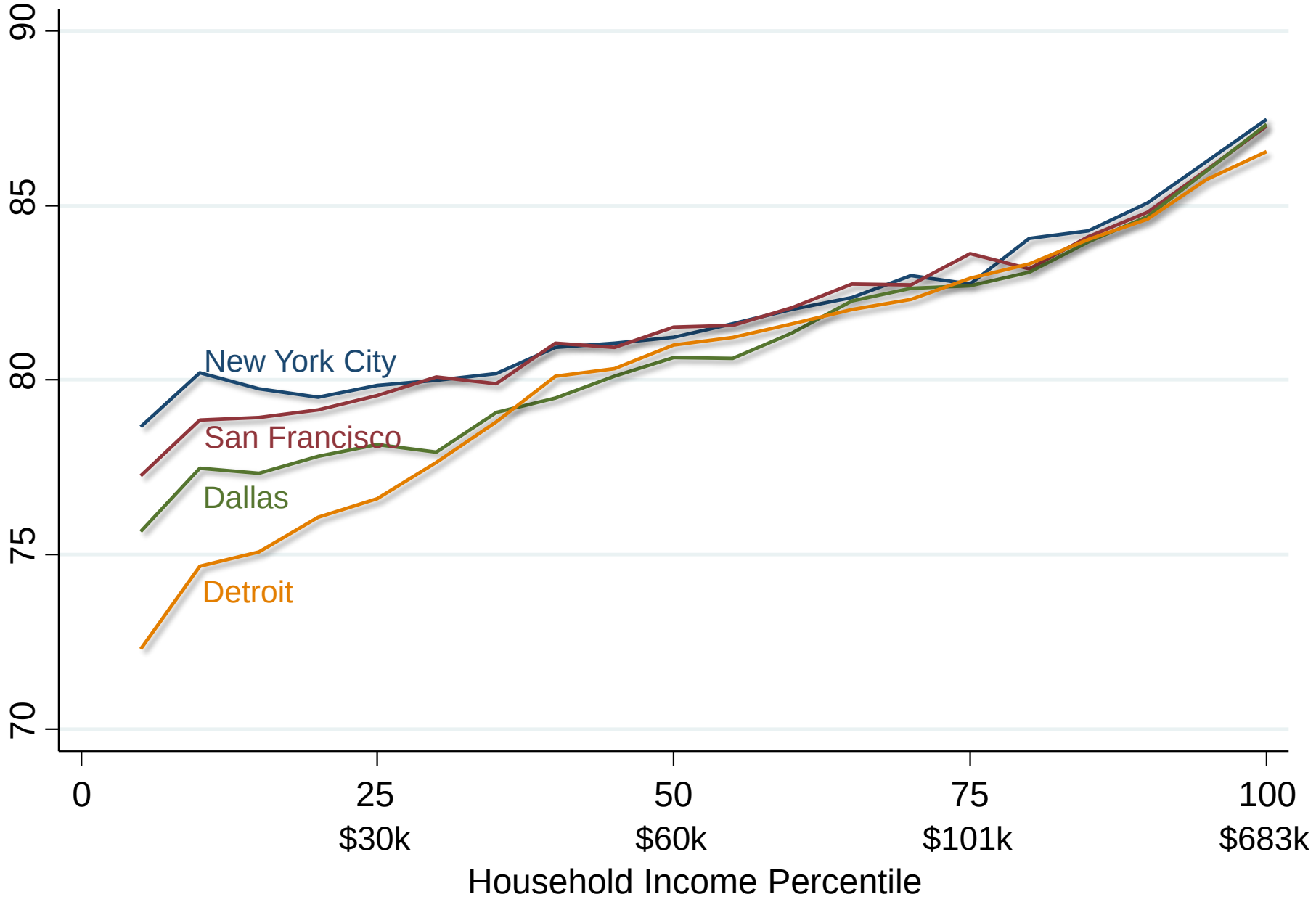


Trends in Expected Age at Death by Income Quartile in the US
For Women Age 40, 2001-2014

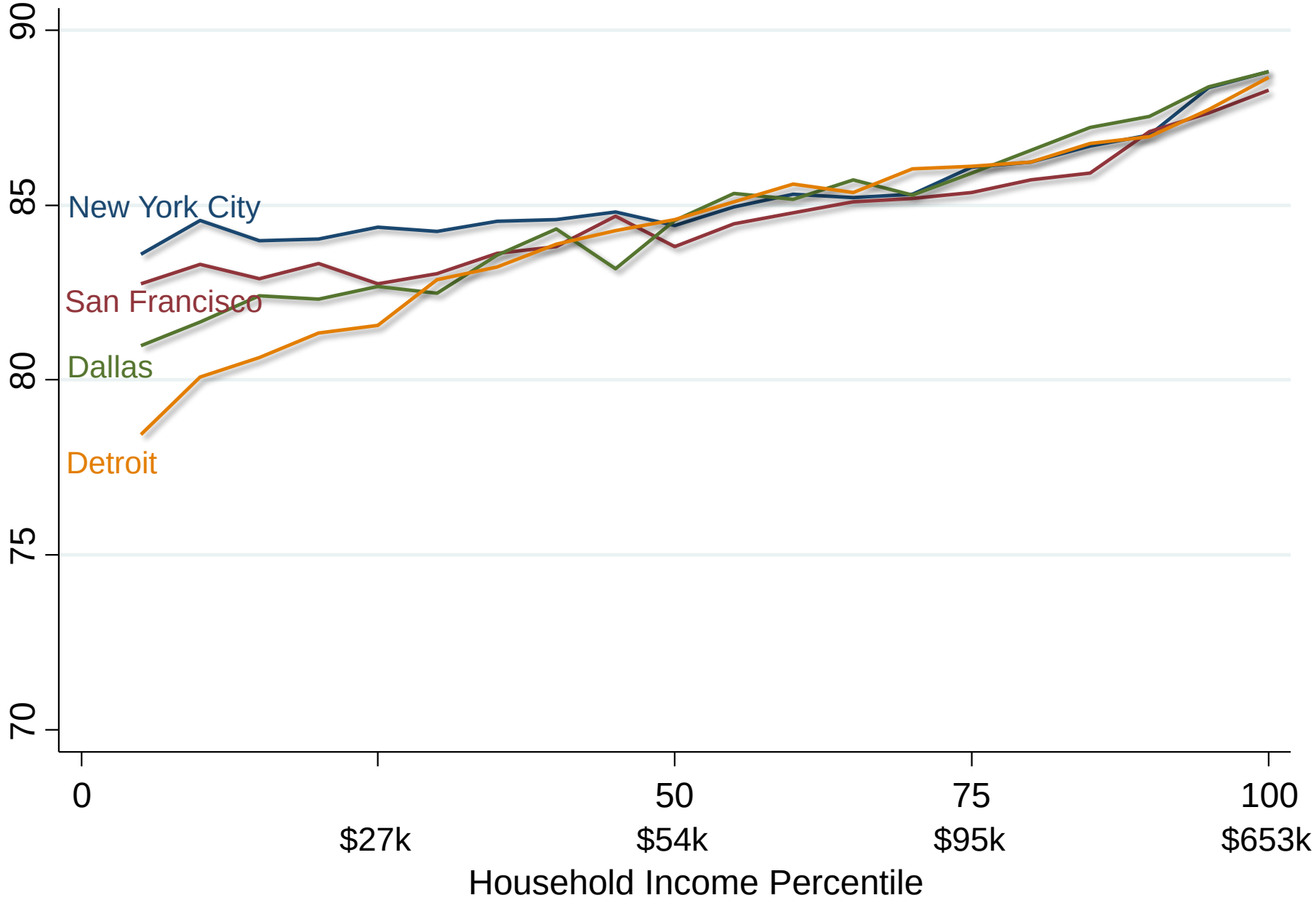


Local Area Variation in Life Expectancy by Income

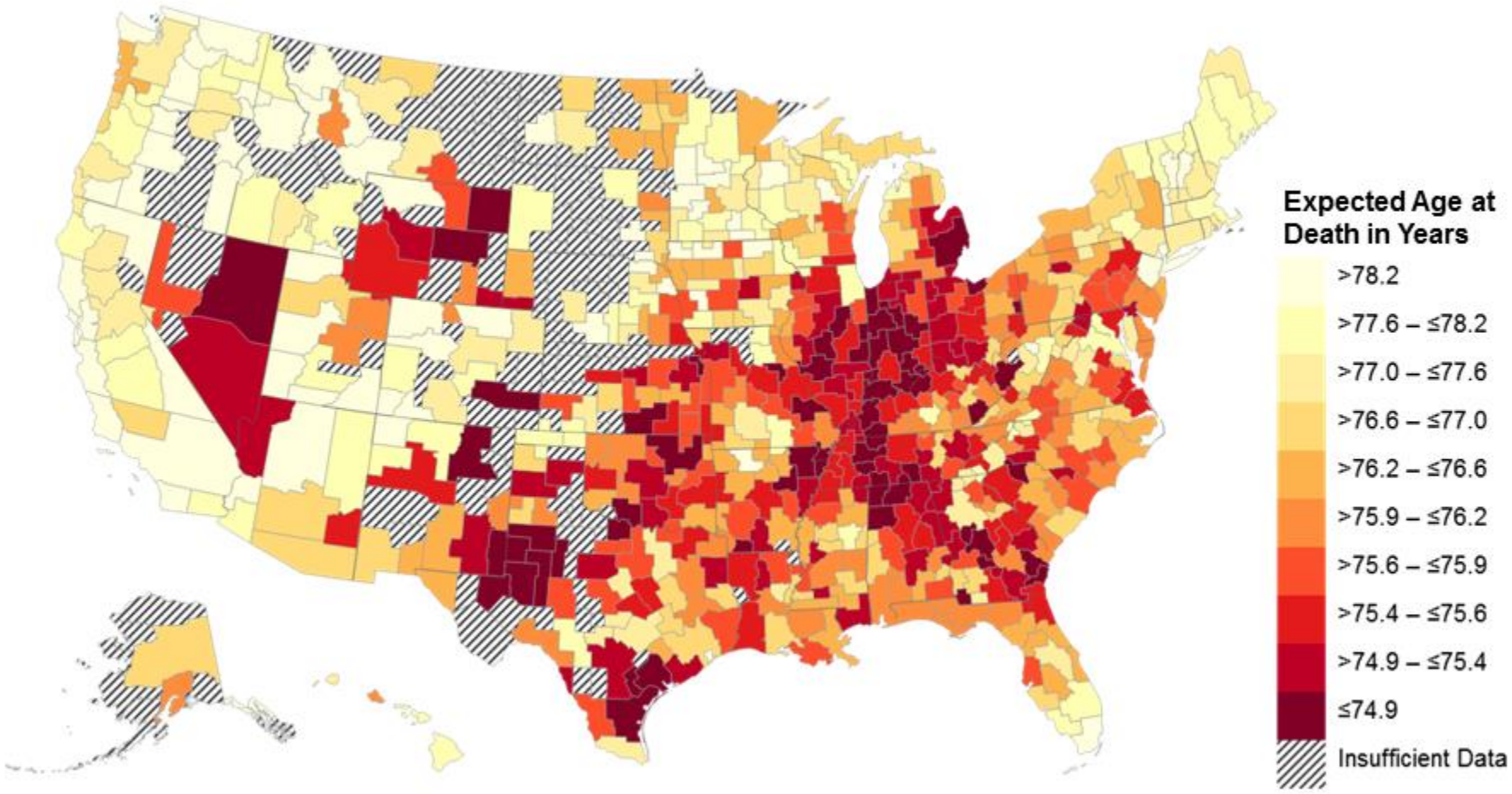
Expected Age at Death vs. Household Income for Men in Selected Cities



Expected Age at Death vs. Household Income for Women in Selected Cities

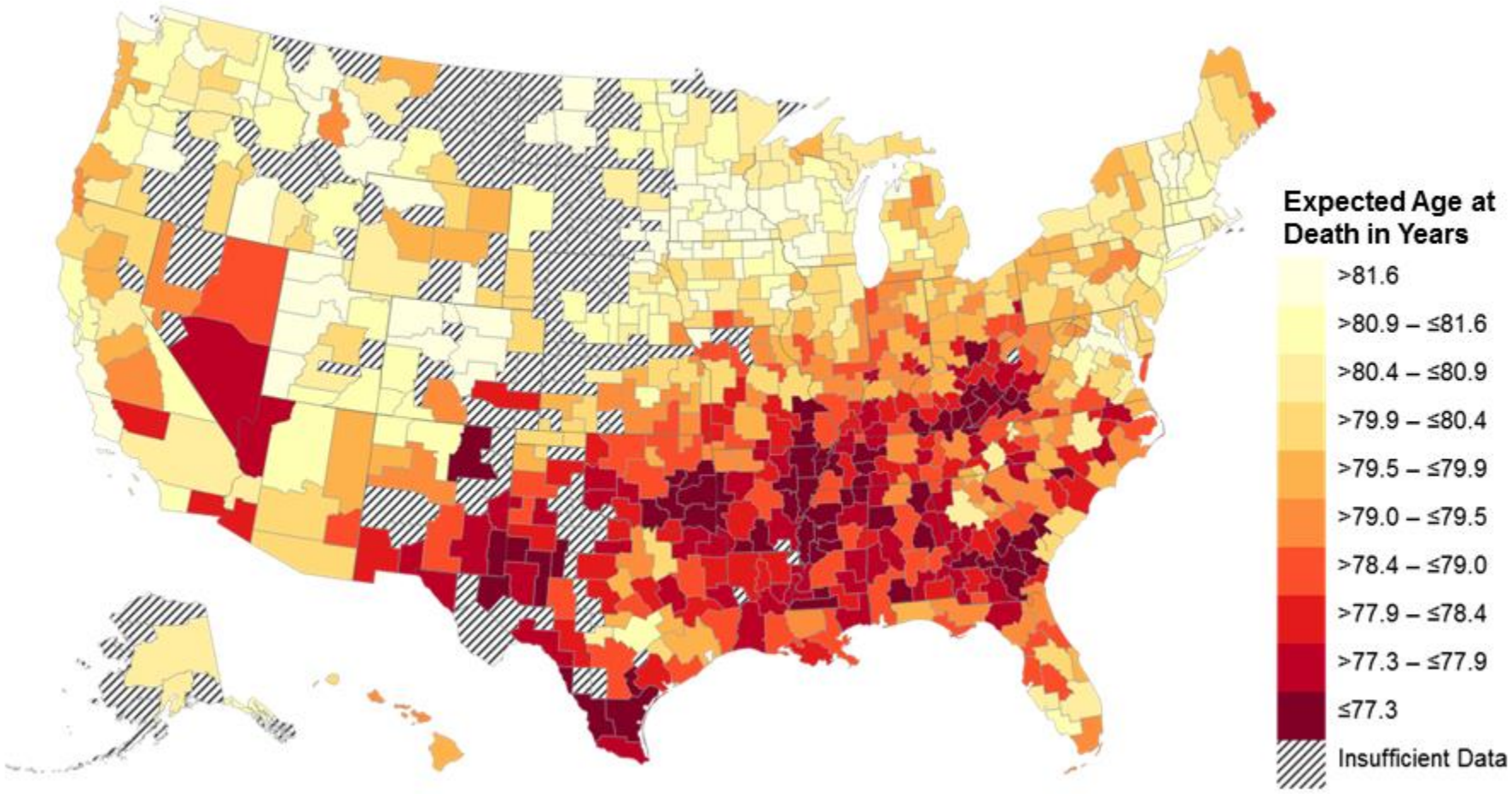


Expected Age at Death for 40 Year Old Men
Bottom Quartile of U.S. Income Distribution



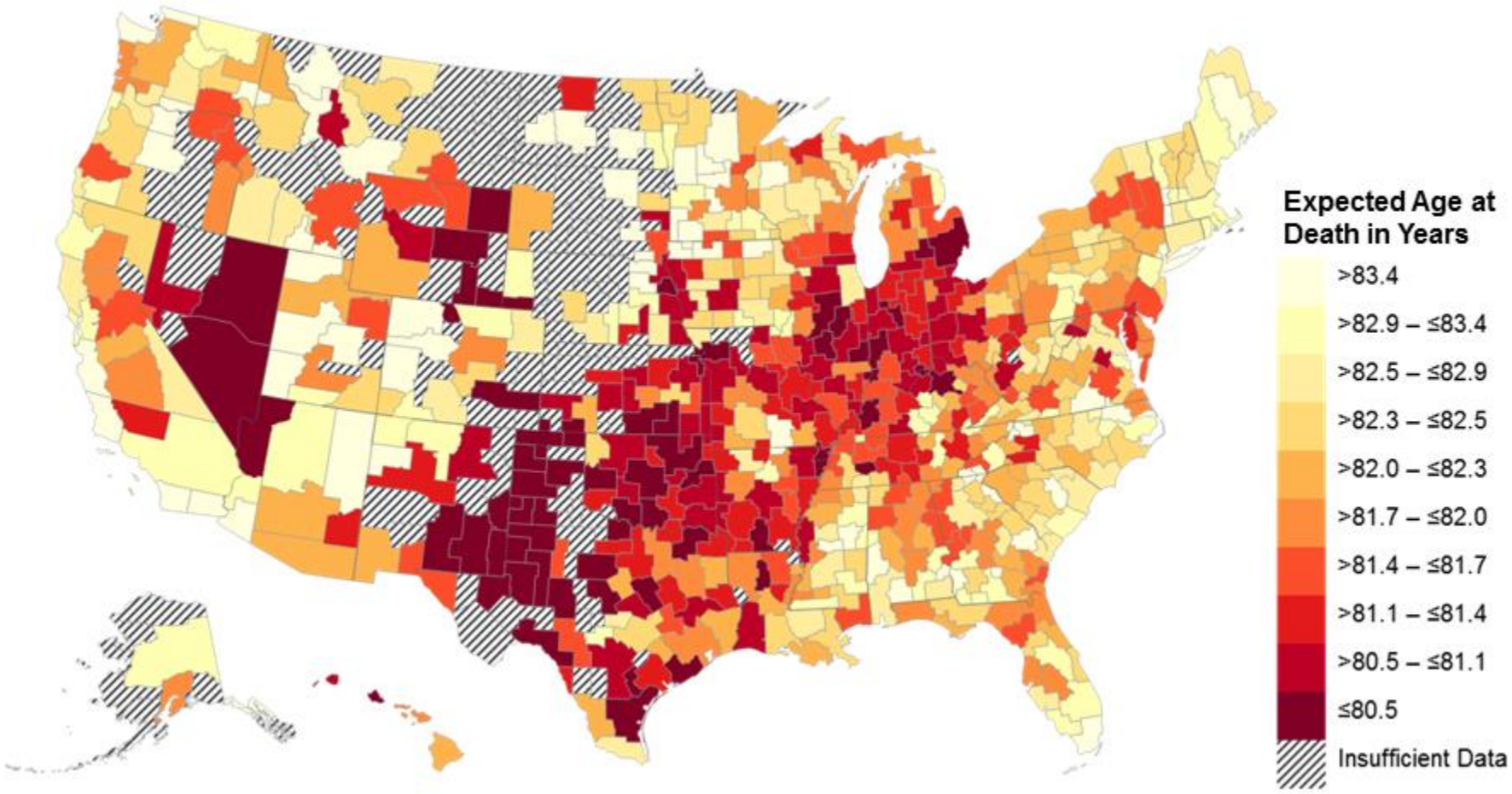
Note: Lighter Colors Represent Areas with Higher Life Expectancy

Expected Age at Death for 40 Year Old Men
Pooling All Income Groups



Note: Lighter Colors Represent Areas with Higher Life Expectancy

Expected Age at Death for 40 Year Old Women
Bottom Quartile of U.S. Income Distribution



Note: Lighter Colors Represent Areas with Higher Life Expectancy

Expected Age at Death for 40 Year Olds in Bottom Quartile
Top 10 and Bottom 10 CZs Among 100 Largest CZs

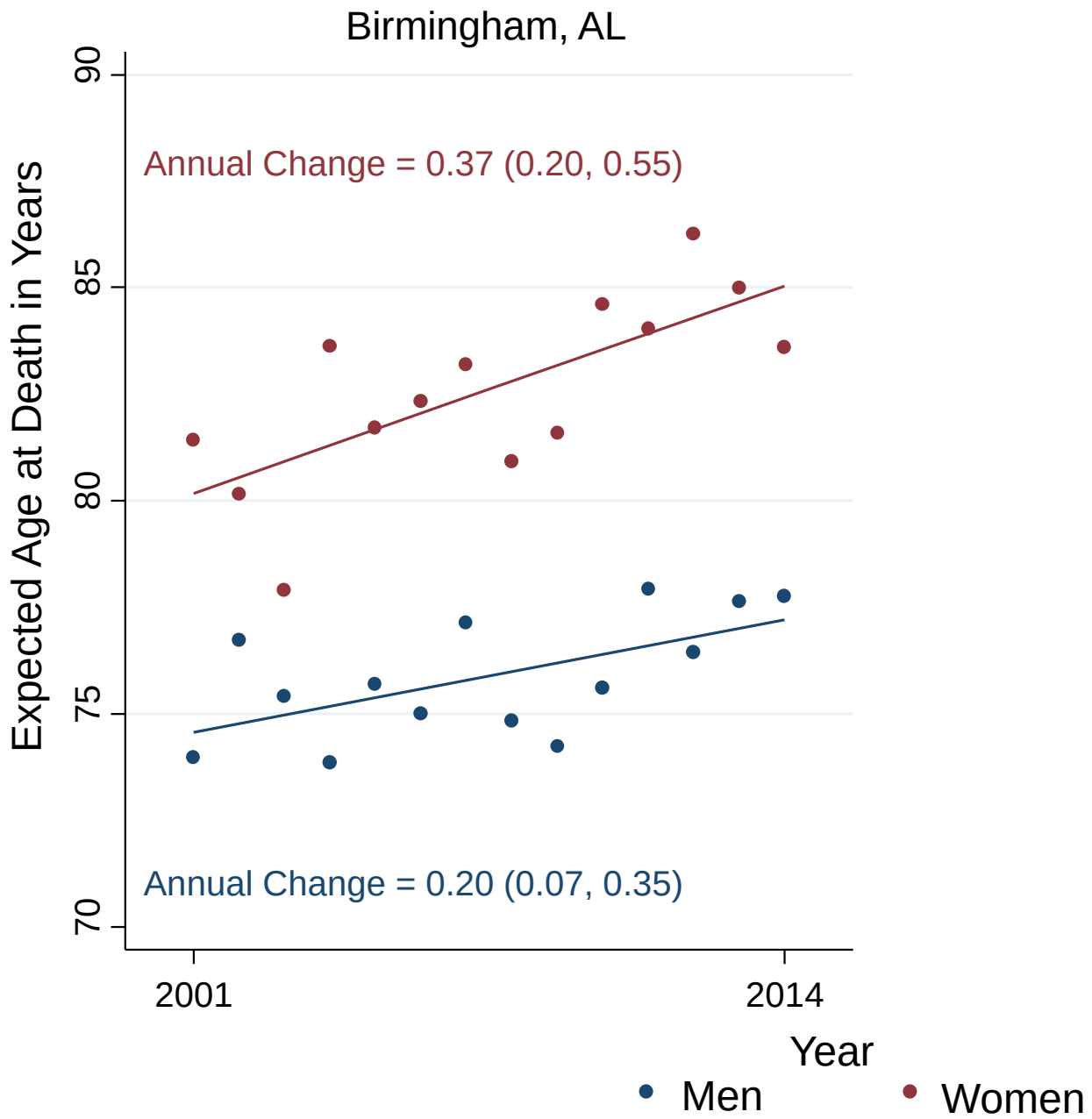
Top 10 CZs			Bottom 10 CZs		
Rank	CZ	Expected Age at Death	Rank	CZ	Expected Age at Death
1	New York, NY	81.8 (81.6, 82.0)	91	San Antonio, TX	78.0 (77.6, 78.4)
2	Santa Barbara, CA	81.7 (81.3, 82.1)	92	Louisville, KY	77.9 (77.7, 78.2)
3	San Jose, CA	81.6 (81.2, 82.0)	93	Toledo, OH	77.9 (77.6, 78.2)
4	Miami, FL	81.2 (80.9, 81.6)	94	Cincinnati, OH	77.9 (77.7, 78.1)
5	Los Angeles, CA	81.1 (80.9, 81.4)	95	Detroit, MI	77.7 (77.5, 77.8)
6	San Diego, CA	81.1 (80.8, 81.4)	96	Tulsa, OK	77.6 (77.4, 77.9)
7	San Francisco, CA	80.9 (80.6, 81.3)	97	Indianapolis, IN	77.6 (77.4, 77.8)
8	Santa Rosa, CA	80.8 (80.5, 81.2)	98	Oklahoma City, OK	77.6 (77.3, 77.8)
9	Newark, NJ	80.7 (80.5, 80.9)	99	Las Vegas, NV	77.6 (77.4, 77.8)
10	Port St. Lucie, FL	80.7 (80.5, 80.9)	100	Gary, IN	77.4 (77.1, 77.8)

Note: 95% confidence intervals shown in parentheses

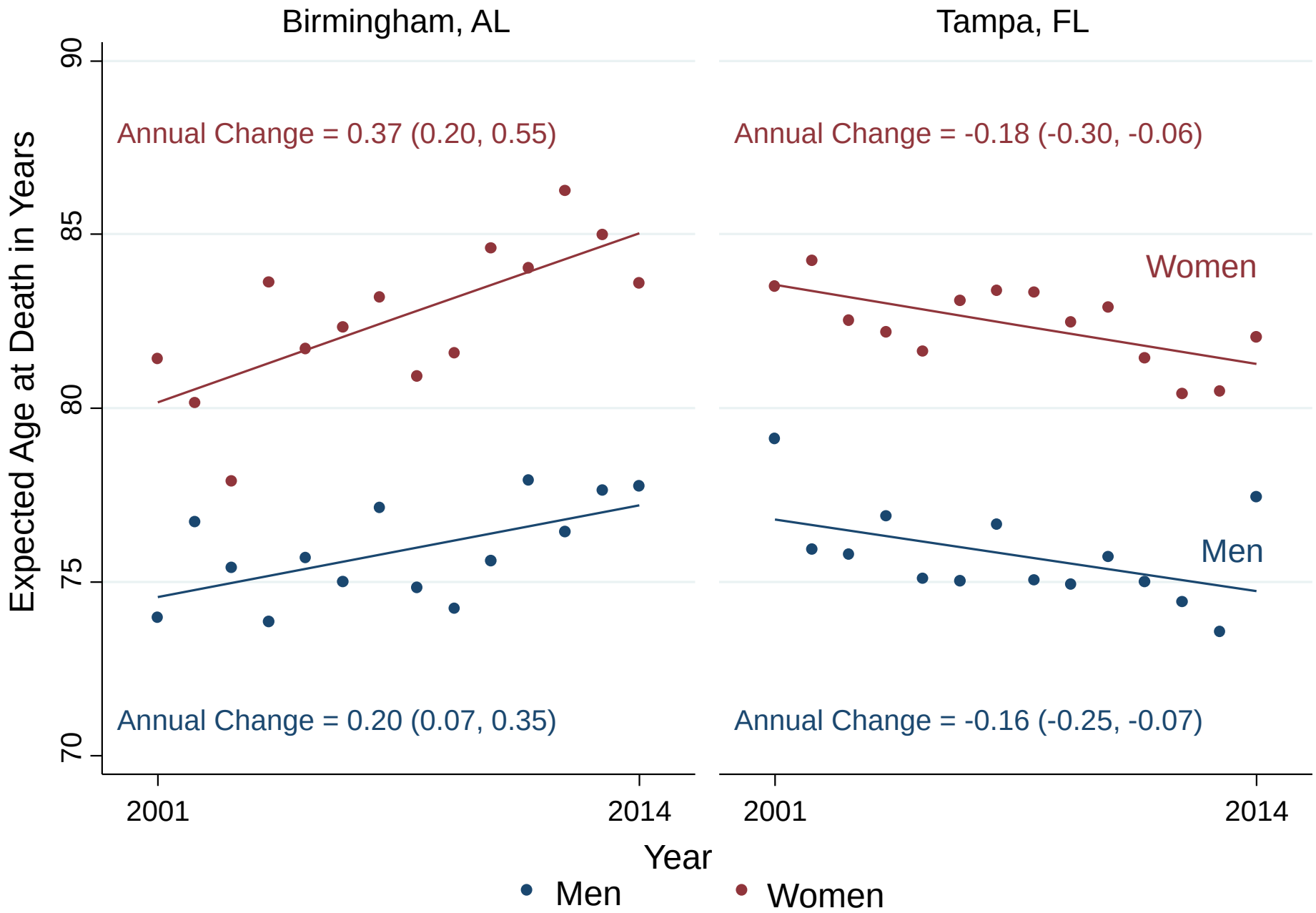
Local Area Variation in Trends

- Next, analyze how *trends* in life expectancy vary across areas

Change in Expected Age at Death in Bottom Quartile



Change in Expected Age at Death in Bottom Quartile



Change in Expected Age at Death in Bottom Quartile
Top 10 and Bottom 10 CZs Among 100 Largest CZs

Top 10 CZs			Bottom 10 CZs		
Rank	CZ	Change over Decade	Rank	CZ	Change over Decade
1	Toms River, NJ	3.8 (2.4, 5.2)	91	Cape Coral, FL	-0.7 (-2.1, 0.6)
2	Birmingham, AL	2.9 (1.8, 4.1)	92	Miami, FL	-0.7 (-1.4, -0.1)
3	Richmond, VA	2.6 (1.3, 3.9)	93	Tucson, AZ	-0.7 (-2.0, 0.5)
4	Syracuse, NY	2.5 (1.1, 4.0)	94	Albuquerque, NM	-0.8 (-2.2, 0.6)
5	Cincinnati, OH	2.4 (1.5, 3.4)	95	Sarasota, FL	-0.8 (-2.0, 0.3)
6	Fayetteville, NC	2.4 (1.0, 3.8)	96	Des Moines, IA	-1.0 (-3.0, 0.8)
7	Springfield, MA	2.3 (0.6, 4.1)	97	Bakersfield, CA	-1.2 (-2.8, 0.3)
8	Gary, IN	2.2 (0.8, 3.8)	98	Knoxville, TN	-1.2 (-2.6, 0.1)
9	Scranton, PA	2.1 (0.8, 3.4)	99	Pensacola, FL	-1.5 (-3.0, -0.2)
10	Honolulu, HI	2.1 (0.5, 3.8)	100	Tampa, FL	-1.7 (-2.5, -0.9)

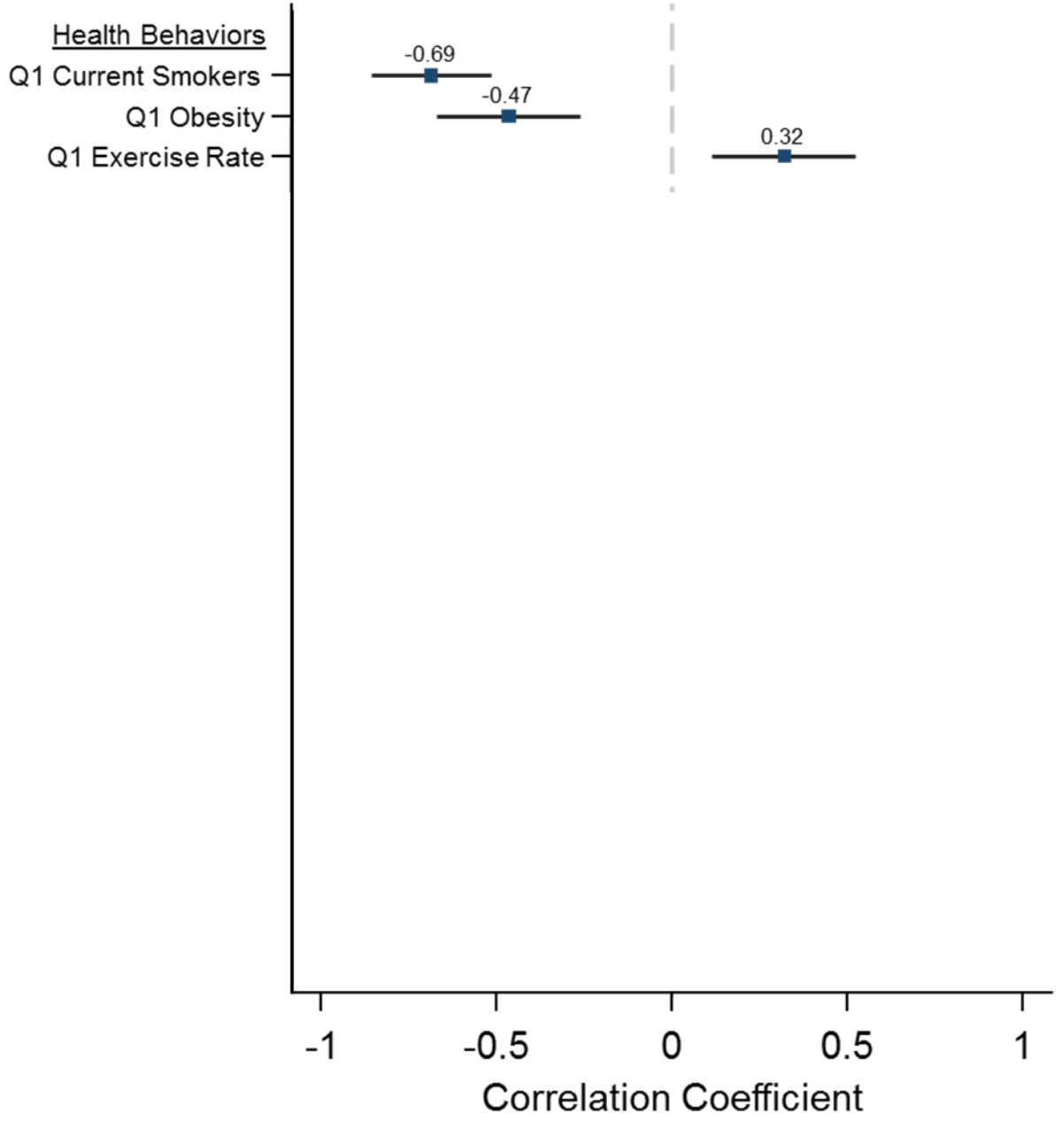
Note: 95% confidence intervals shown in parentheses

Correlates of Spatial Variation in Life Expectancy

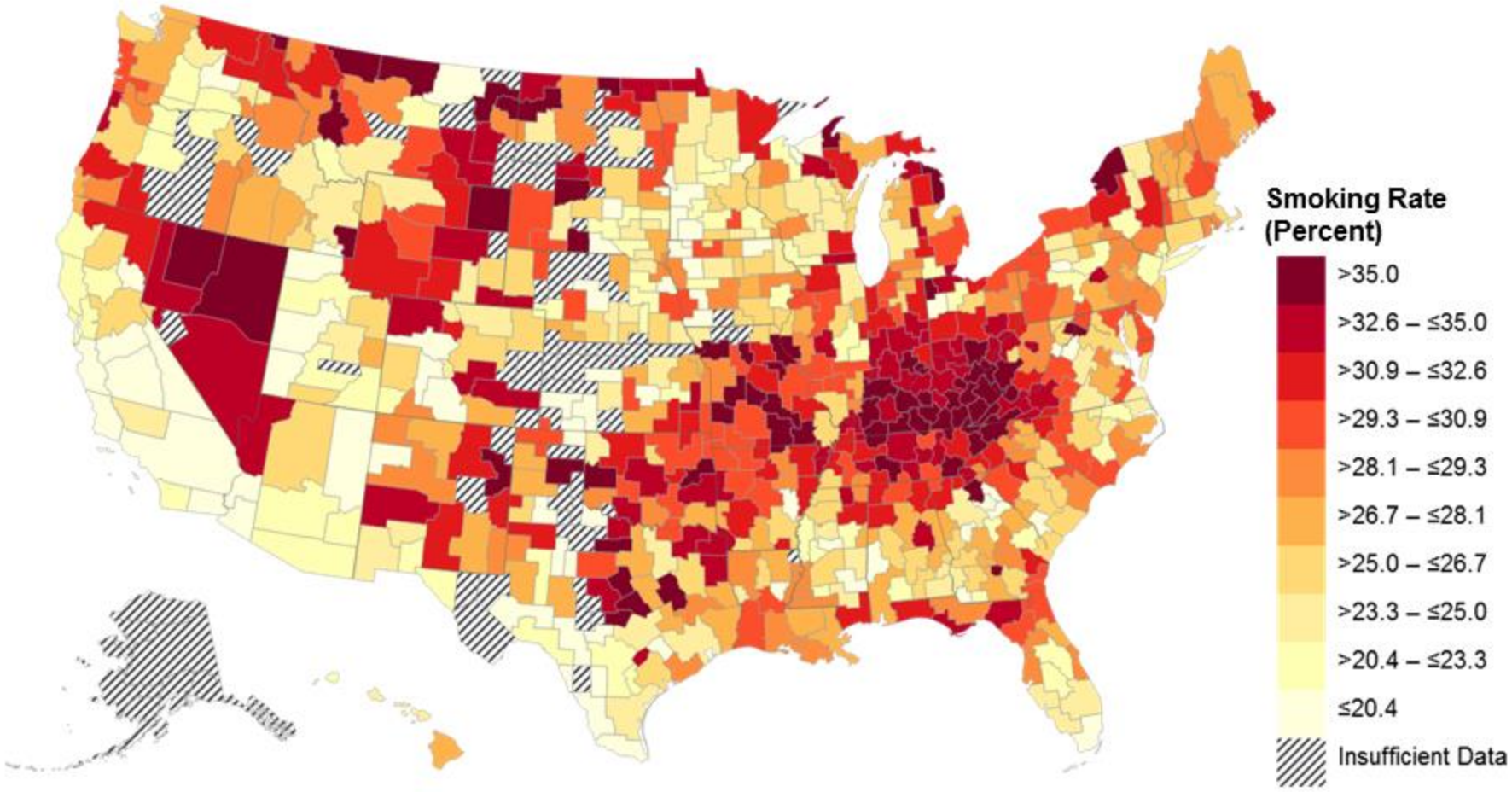
Why Does Life Expectancy Vary Across Areas?

- Finally, we characterize the features of areas with high vs. low life expectancy conditional on income

Correlations of Expected Age at Death with Health and Social Factors For Individuals in Bottom Quartile of Income Distribution

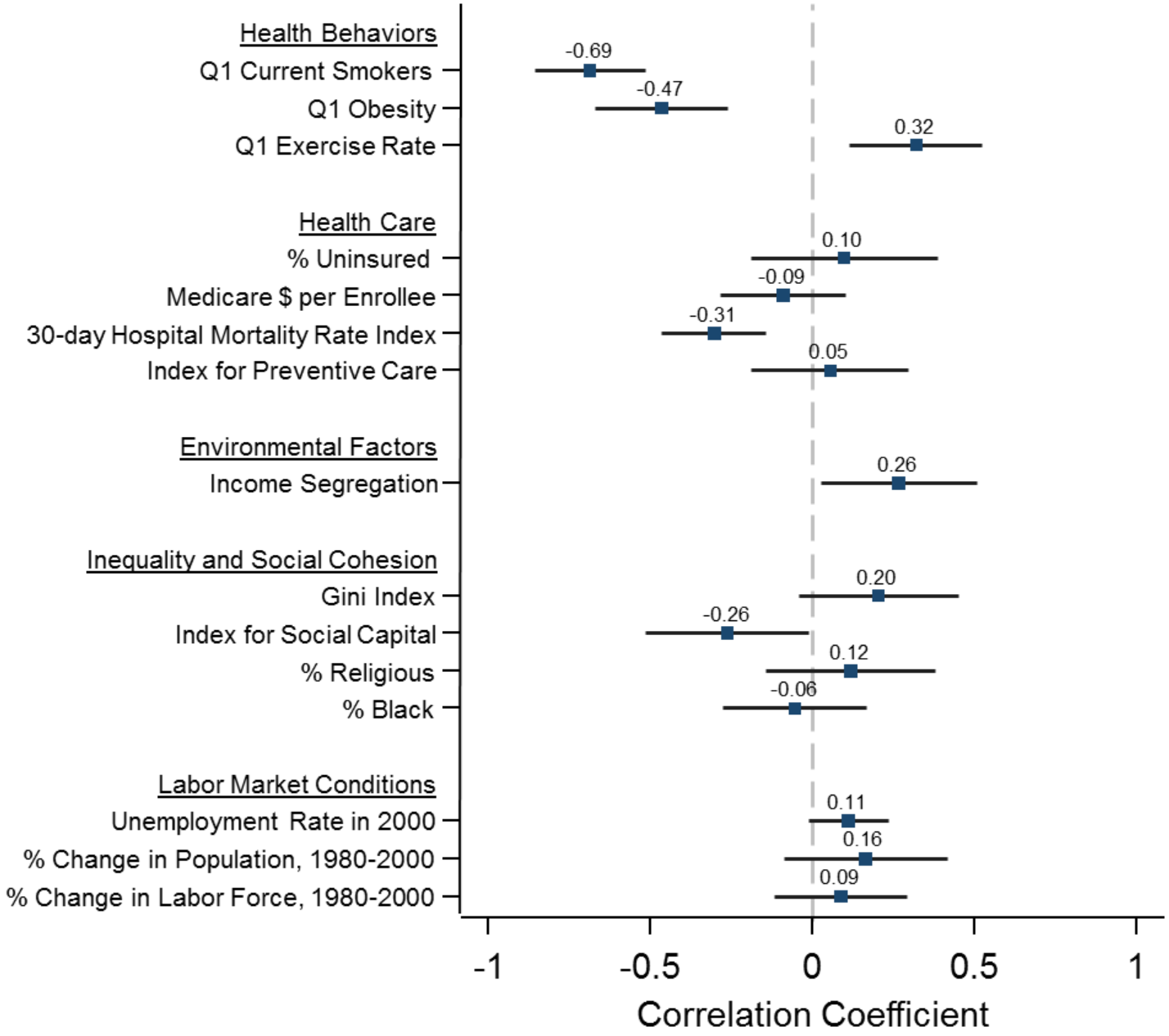


Smoking Rates for Individuals in Bottom Income Quartile

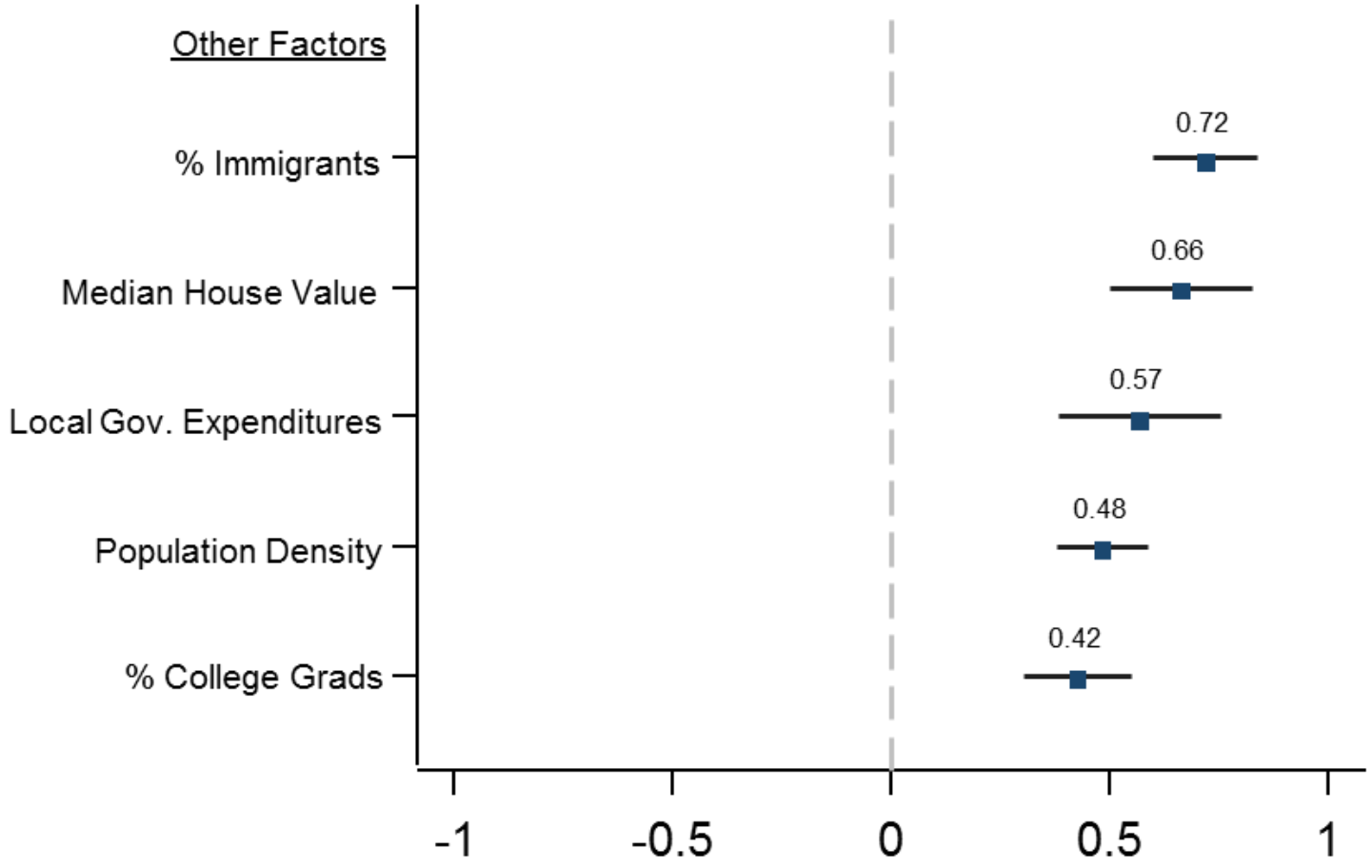


Note: Lighter Colors Represent Areas Lower Smoking Rates

Correlations of Expected Age at Death with Health and Social Factors For Individuals in Bottom Quartile of Income Distribution



Correlations of Expected Age at Death with Other Factors For Individuals in Bottom Quartile of Income Distribution



Correlations: Summary

- General pattern: Low-income people in affluent, educated cities live longer (and have healthier behaviors)

- Why is this the case?

  - Spillovers from rich to poor: regulation, public revenues/transfers

  - Exposure to people with healthier behaviors

  - Sorting: low-income people who live in expensive cities are a selected group with different characteristics

Income and Life Expectancy: Lessons

1. Disparities in life expectancy are large and growing, but not immutable:
   - Some areas in the U.S. have relatively small and shrinking gaps
2. Reducing health disparities likely to require local interventions
   - Ex: targeted efforts to improve health among low-income individuals in specific cities such as Las Vegas and Detroit
   - Changing health behaviors at local level likely to be important
3. Trends imply that indexing eligibility for Social Security and Medicare to *average* life expectancy will amplify inequality

How Can We Improve Population Health?

- Two types of approaches, with different questions and methods:

  1. Public health interventions to change behaviors such as smoking, exercise, water quality, and spread of diseases

  2. Provision of health care and health insurance

- Illustrate frontier of research using big data in each of these areas using recent examples

Forecasting Pandemics

- Classic problem in population health: forecasting and preventing health pandemics
- Contagious diseases like flu spread exponentially → large returns to taking action quickly when disease emerges
- Most common method to monitor contagious diseases in developed countries: aggregated data from local clinics
- Problem: slow reporting and small samples → data not very fine-grained

Forecasting Pandemics: Google Flu Trends

- Ginsberg et al. (2009) propose a new data source to monitor spread of the flu: Google search data

- Idea: people often search for terms like “antibiotics” or “how to treat cough” when getting sick

- Use aggregated search data to get predictions of spread of flu that are (a) more timely and (b) available at fine geographies

Forecasting Pandemics: Google Flu Trends

- Method: predictive modeling
  - Get historical data on truth from CDC and estimate a statistical model using Google search data to predict that data
  - Then evaluate the model using future data that was not used for estimation to evaluate model's predictive accuracy

Google Flu Trends: Methodology

- Data to be predicted: 1,152 observations from CDC on flu incidence
  - Weekly data from 9 regions of the U.S. from 2004-2007
- Data used for prediction: counts of Google search data
  - Weekly data on Google search counts for 50 million terms by state from 2004-2007

Google Flu Trends: Overfitting Problem

- This is an example of “wide data”
  - Many more variables than number of observations
  - Overfitting problem: can fit the data perfectly using 1,152 explanatory variables → cannot use traditional statistical methods like regression
- Solve this problem using *out-of-sample validation*
  - Idea: use separate samples to estimate the model and evaluate its predictive accuracy