Analyzing Life Expectancy by Race and Education

Sonam T Gurung

Part 2: Analyzing Life Expectancy by Race and Education

```
library(tidyverse)
library(knitr) # Integration of latex and R
library(kableExtra) # For a good-looking table
```

```
aggregated_data <- read_csv("data-task-year-race-education-collapse.csv")

# Check for missing values in the aggregated data
sum(is.na(aggregated_data))
```

```
## [1] 0
```

```
# No missing values in the aggregated dataset

# Count the number of White and Black individuals
aggregated_data %>%
  group_by(race) %>%
  summarize(count = n())
```

```
## # A tibble: 2 x 2
##   race  count
##   <chr> <int>
## 1 Black  4620
## 2 White  4620
```

```
# Equal - 4620 each
```

Function to find the life expectanct at age 30

```
# The function requires the dataset and the mortality_rate as input parameters, which can
# be either smoothed or raw

get_life_exp_at_30 <- function(data, mort){
  data %>%
    mutate(survival_rate = 1 - mort,
           cumulative_survival = cumprod(survival_rate)) %>%
    summarize(life_exp_at_30 = round(sum(cumulative_survival) + 30, 2))

}
```

Using the provided aggregated data, calculate life expectancy at age 30 for each combination of race and education level for the year 2003. I will be using mortality rates directly without smoothing

```
# Apply the life expectancy at 30 function to the aggregated data for each
# combination of race and education level for the year 2003
# The pick(everything()) step selects all columns from the dataset
# as it has been filtered and grouped so far. This ensures that
# get_life_exp_at_30() uses the manipulated dataset up to this point

life_exp_race_edu_2003 <- aggregated_data %>%
  filter(year == 2003) %>%
  group_by(race, education) %>%
  summarize(life_exp_at_30 = get_life_exp_at_30(pick(everything()), mortality)$life_exp_at_30) %>%
  ungroup()

# Change column names to be a presentable name for the table
colnames(life_exp_race_edu_2003) <- c("Race", "Education Level", "Life Expectancy at Age 30")

# Create a polished table to display the life expectancy for the knitted pdf
kable(life_exp_race_edu_2003, format = "latex", booktabs = TRUE,
      caption = "Life Expectancy at Age 30 by Race and Education Level (2003)") %>%
  kable_styling(latex_options = c("hold_position", "striped"),
                font_size = 10)
```

Table 1: Life Expectancy at Age 30 by Race and Education Level (2003)

Race	Education Level	Life Expectancy at Age 30
Black	B.A.+	76.28
Black	H.S.	69.67
Black	Less than H.S.	68.15
Black	Some college	76.04
White	B.A.+	78.82
White	H.S.	74.18
White	Less than H.S.	70.49
White	Some college	78.52

Calculate and plot life expectancy by race over time from 2003 to 2019, allowing the education distribution to vary between racial groups:

For each race, calculate the distribution of education levels within that race.

```
# Calculate the distribution of education levels within each race
# Steps:
# Group by race and education
# Count the number of people for each race-education combination
# Ungroup to reset the group
# Regroup by race to calculate proportions, but only by race,
# to calculate proportions within each race
# Compute the proportion of each education level within the race
# Remove the count column
edu_distrib_by_race <- aggregated_data %>%
  group_by(race, education) %>%
```

```
  summarize(count = sum(population)) %>%
  ungroup() %>%
  group_by(race) %>%
  mutate(proportion = count / sum(count)) %>%
  select(-count)
```

Calculate life expectancy for each combination of race and education level for each year from 2003 to 2019.

```
# Create an empty data frame to store life expectancy for each
# combination of race and education level
# for each year from 2003 to 2019.

# For each year from 2003 to 2019:
# 1. Group the data by race and education level.
# 2. Calculate the life expectancy at age 30 using the get_life_exp_at_30 function
# 3. Bind the rows for that specific year to the main data frame

life_expectancy_2003_2019 <- data.frame()

# Loop through each year from 2003 to 2019
for (current_year in 2003:2019) {
  # Calculate life expectancy at age 30 for each group
  life_expectancy_year <- aggregated_data %>%
    filter(year == current_year) %>%
    group_by(race, education) %>%
    arrange(race, education, age) %>%  # Sort by age within each group
    summarize(life_exp_at_30 = get_life_exp_at_30(pick(everything()), mortality)$life_exp_at_30) %>%
    ungroup() %>%
    mutate(year = current_year)  # Add a column for the current year

  # Append the results to the final data frame
  life_expectancy_2003_2019 <- bind_rows(life_expectancy_2003_2019, life_expectancy_year)
}
```

Use these race-specific education distributions to compute weighted averages of the life expectancies across education levels for each race.

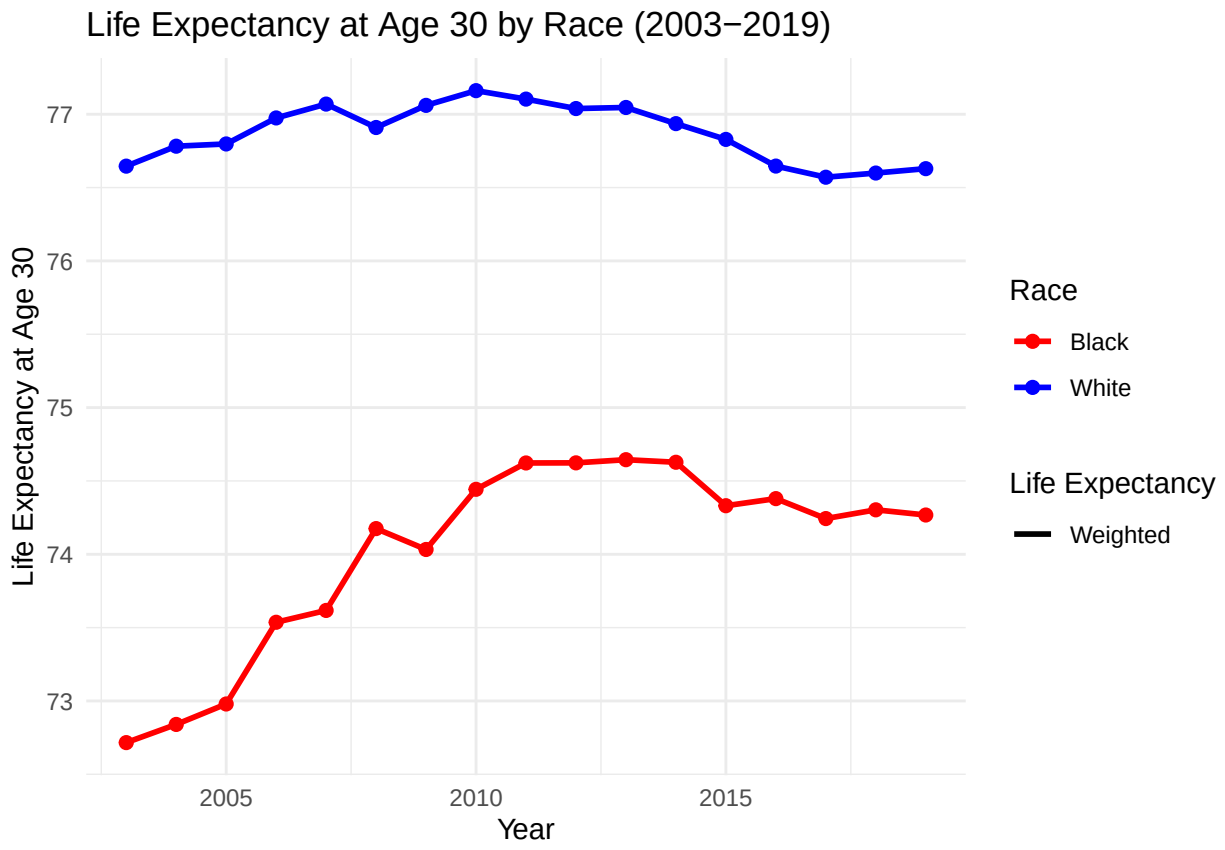
```
# Merge education proportions by race into life_expectancy_2003_2019,
# linking by race and education to include life expectancy, race,
# education, and the proportion of each education level within that race
life_expectancy_2003_2019 <- left_join(life_expectancy_2003_2019, edu_distrib_by_race,
                                       by = c("race", "education"))

# Calculate weighted averages of life expectancy for each race and year
# By grouping race and year and summarizing the weighted averages of the life expectancy across
# education levels for each race.
weighted_avg_life_exp <- life_expectancy_2003_2019 %>%
  group_by(race, year) %>%
  summarize(weighted_life_expectancy = sum(life_exp_at_30 * proportion)) %>%
  ungroup()
```

Plot these race-specific life expectancies from 2003 to 2019.

```r
# Base plot with race-specific life expectancy lines
original_plot <- ggplot(weighted_avg_life_exp, aes(x = year)) +
  geom_line(aes(y = weighted_life_expectancy, color = race, linetype = "Weighted"), linewidth = 1) +
  geom_point(aes(y = weighted_life_expectancy, color = race), size = 2) +
  labs(x = "Year",
       y = "Life Expectancy at Age 30",
       title = "Life Expectancy at Age 30 by Race (2003-2019)",
       color = "Race",
       linetype = "Life Expectancy") +
  scale_linetype_manual(values = c("Weighted" = "solid", "Reweighted" = "dashed")) +
  scale_color_manual(values = c("White" = "blue",
                                "Black" = "red")) +
  theme_minimal()

original_plot
```



Implement a reweighting method to control for differences in educational attainment between racial groups:

For each year, calculate the overall distribution of education levels in the entire population, regardless of race.

```r
# Calculate the distribution of education levels in the entire population
# Steps:
# Group by year and education
# Count the number of people for each race-year combination
# Ungroup to reset the group
```

```r
# Regroup by year to calculate education proportions within each year
# Compute the proportion of each education level for each year
# Remove the count column
edu_distrib_by_yr <- aggregated_data %>%
  group_by(year, education) %>%
  summarize(count = sum(population)) %>%
  ungroup() %>%
  group_by(year) %>%
  mutate(proportion_reweighted = count / sum(count)) %>%
  select(-count)
```

For each race and year, reweight the life expectancies of different education levels using this overall education distribution.

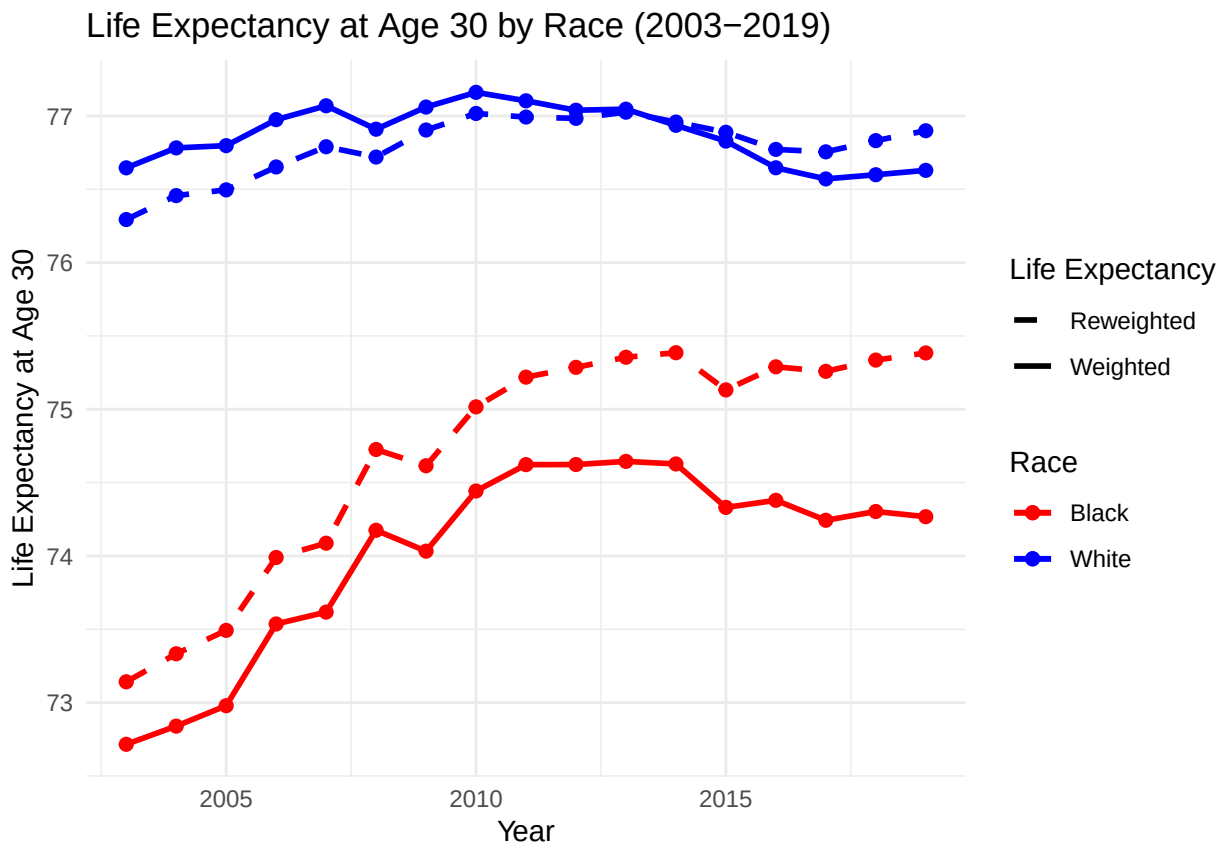
```r
# Join the rewighted education proportion to the life expectancy dataset
# We use "year" and "education" as keys to merge the reweighted proportion of each education level
# within each year into the life expectancy dataset.
life_expectancy_2003_2019  <- left_join(life_expectancy_2003_2019, edu_distrib_by_yr,
                                       by = c("year", "education"))

# Calculate reweighted life expectancy for each race and year
# By grouping race and year and summarizing the reweighted averages of the life expectancy across
# education levels for each race.
reweighted_life_expectancy <- life_expectancy_2003_2019 %>%
  group_by(race, year) %>%
  summarize(reweighted_life_exp = sum(life_exp_at_30 * proportion_reweighted)) %>%
  ungroup()
```

Create a plot showing both the original and reweighted life expectancy by race over time.

```r
# Add reweighted lines to the original weighted life expectancy plot
final_plot <- original_plot +
  geom_line(data = reweighted_life_expectancy,
            aes(y = reweighted_life_exp, color = race,  linetype = "Reweighted"),
            linewidth = 1) +
  geom_point(data = reweighted_life_expectancy, aes(y = reweighted_life_exp, color =  race),
             size = 2)

final_plot
```



White individuals have a higher original life expectancy compared to Black individuals, and this trend persists even after reweighting for educational differences. However, the gap between races narrows in the reweighted data. This reduction in the gap highlights the association between education and life expectancy, as White individuals have a higher proportion of B.A.+ and Some college education compared to Black individuals. Educational differences play a key role in driving racial disparities in life expectancy. By controlling for education through reweighting, we observe that a portion of the racial gap in life expectancy is attributable to differences in education.