Data Analysis Report for the Life Expectancy Study
BIOST 512 Winter 2024 (Last Modified: 7th March 2024)
Authors: Prity Lata Chakraborty, Gaetan Khim, Sophie Whikehart

Background

Life expectancy is a key metric for assessing population health and has consistently increased in the last 100 years for most countries. Life expectancy has however risen unequally in different areas of the world and may be affected by metrics such as a country's income and a country's expenditure on health as a percentage of its GDP. Our study aims to analyze these associations to clarify further the roles that income, health expenditure, and 2001 life expectancy play in determining life expectancy in 2019.

Methods

For our descriptive analysis, we developed a table measuring the proportion of occurrence of each factor variable for country income classification per region, which included data missingness, (Table 1) to give us a general idea of trends based on health expenditure, the percentage of income category per region, and identifying potential limitations in the data. A box plot showing the life expectancy distribution in 2001 and 2019 by income group (Figure 1) provides a visual representation of the data for ease of interpretation. To test for association between countries' income group and 2019 life expectancy, we used one-way ANOVA, which we expressed as: $\mathbb{E}$**[life expectancy 2019|countries' income groups]** $= \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2$, with the hypothesis: **H0:** $\beta_1 = \beta_2 =$ **0, H1:** $\beta_i \neq$ **0 for any i = 1,2.** We used ANCOVA to model categorical and continuous predictors, which we expressed as: $\mathbb{E}$**[life expectancy 2019|2019 health expenditures, countries' income groups]** $= \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{Z}_2 + \beta_3 \mathbf{Z}_3 + \beta_4 \mathbf{X}\mathbf{Z}_2 + \beta_5 \mathbf{X}\mathbf{Z}_3$. We tested 2019 health expenditure for interaction prior to adjustment with the hypothesis: **H0:** $\beta_4 = \beta_5 =$ **0, H1:** $\beta_i \neq$ **0 for any i = 4,5.** If we cannot reject H0 for interaction, we would test for the adjusted model with the following: $\mathbb{E}$**[life expectancy 2019|2019 health expenditures, countries' income groups]** $= \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{Z}_2 + \beta_3 \mathbf{Z}_3$. To test for interaction between 2001 life expectancy and countries' income groups adjusted for 2019 health expenditures we used ANCOVA, which we expressed as: $\mathbb{E}$**[life expectancy 2019|2001 health expenditures, countries' income groups, 2019 health expenditures]** $= \beta_0 + \beta_1 \mathbf{X} + \beta_2 \mathbf{Z}_2 + \beta_3 \mathbf{Z}_3 + \beta_4 \mathbf{W} + \beta_5 \mathbf{X}\mathbf{Z}_2 + \beta_6 \mathbf{X}\mathbf{Z}_3$ with the following hypothesis: **H0:** $\beta_5 = \beta_6 =$ **0, H1:** $\beta_i \neq$ **0 for any i = 5,6**.

Results

Our descriptive analysis estimates that the median 2019 life expectancy in high, middle, and low-income countries is 82, 63, and 73 years old respectively. It is important to note that missing data on 2019 life expectancy (n=10) may be a potential source of bias leading to incorrect estimations of life expectancy by region. Based on the multiple F-test, with a p-value of less than 0.001, we reject the null hypothesis of no association between income group and life expectancy in 2019. At the 5% significance level, we fail to reject the null hypothesis of no interaction between health expenditure in 2019 and income group, meaning that there is no statistically significant evidence that health expenditure is an effect modifier (multiple F-statistic = 2.37, p-value = 0.096). In the unadjusted model, the coefficient estimates differ from those in the adjusted model by about 7% (Middle income: from -18.35 to -17.17) and 14% (Low income: from -8.99 to -7.88), so we are not worried about confounding (Table 2). At the 5% significance level, we reject the null hypothesis that adjusting for 2019 health expenditure, there is no interaction between life expectancy in 2001 and income group, this means that there is statistically significant evidence that life expectancy in 2001 is an effect modifier (multiple F-statistic = 95.93, p-value < 0.001).

Conclusions

Our results show an association between income group and 2019 life expectancy; when adjusted for 2019 health expenditure we do not observe large differences between the unadjusted and adjusted model. Finally, this adjusted association differs between income group strata when we include 2001 life expectancy in the model, which acts as an effect modifier on countries' income groups. Limitations to this report include missing data and unexpected disparities. Future research could include longitudinal analysis, qualitative inquiry, cross-country comparisons, and policy implications.

Authors: Prity Lata Chakraborty, Gaetan Khim, Sophie Whikehart

Table and Figures:

Table 1:

Table 1. Proportion of Occurence by Region on Income Group, Health Expenditure2019 and Life Expectancy 2019

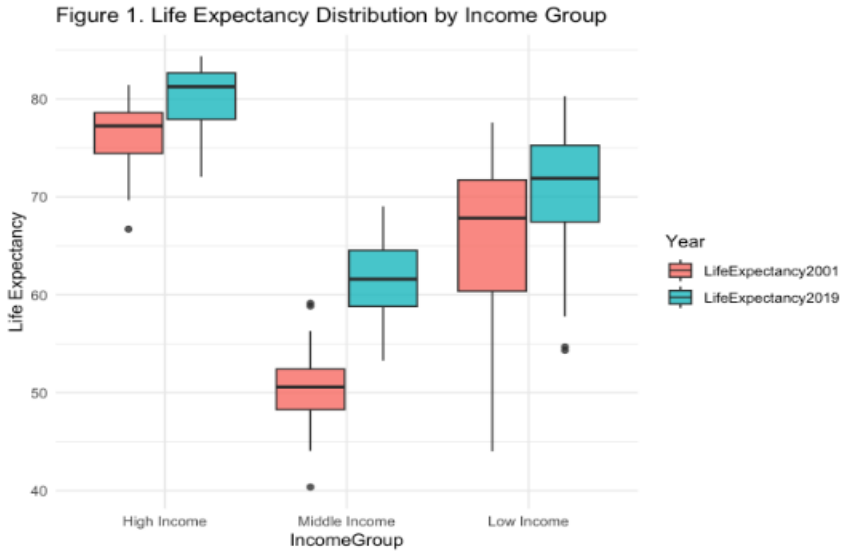
Region	Low Income "%"	Middle Income "%"	High Income "%"	Median HealthExpenditure2019	Median LifeExpectancy2019	Missing Health Expenditure 2019 N/A Count	Missing Life Expectancy 2019 N/A Count
East Asia & Pacific	74.07%	0%	25.93%	5.12	73.00	3	6
Europe & Central Asia	34.04%	0%	65.96%	7.13	78.50	2	3
Latin America & Caribbean	75.86%	0%	24.14%	7.09	75.16	1	1
Middle East & North Africa	50%	0%	50%	5.50	76.79	1	0
North America	0%	0%	100%	13.81	81.87	1	0
South Asia	87.5%	12.5%	0%	3.84	71.28	0	0
Sub-Saharan Africa	47.73%	50%	2.27%	5.19	62.89	1	0

Table 2:

Table 2. Unadjusted VS. Adjusted Table

X	Unadjusted.model	Adjusted.model.for.2019.health.expenditure
Intercept	80.04 (95% CI: 79.2 to 80.89)	75.31 (95% CI: 72.09 to 78.53)
Middle income group 2019 life expectancy	-18.35 (95% CI: -16.43 to -20.28)	-17.17 (95% CI: -14.95 to -19.38)
Low income group 2019 life expectancy	-8.99 (95% CI: -7.51 to -10.48)	-7.88 (95% CI: -6.15-9.60)

Figure 1:



Code:

```
knitr::opts_chunk$set(echo = TRUE,
                      include = TRUE,
                      warning = FALSE,
                      comment = "")

# load libraries
library(psych
library(rigr)
library(lattice)
library(dplyr)
library(tidyr)
library(kableExtra)
library(ggplot2)
library(ggthemes)

# read in the dataset
dat <- read.csv("BIOST512-subset-LifeExpectancy.csv")

##Creation of summary table

# create factors [categorical] variables
dat$CountryName <- as.factor(dat$CountryName)
dat$Region <- as.factor(dat$Region)
dat$IncomeGroup <- as.factor(dat$IncomeGroup)
levels(dat$IncomeGroup) <- c("High Income", "Middle Income", "Low Income")
# group data by region and income classification and then calculate counts
region_income_counts <- dat %>%
  group_by(Region, IncomeGroup) %>%
  summarise(count = n())

# spread the data to wide format to get counts for each income category
region_income_counts_wide <- spread(region_income_counts, key = IncomeGroup,
value = count, fill = 0)

# group data by region and income classification, then calculate counts
region_income_counts <- region_income_counts_wide %>%
  group_by(Region) %>%
  summarise(`Low Income` = sum(`Low Income`),
            `Middle Income` = sum(`Middle Income`),
            `High Income` = sum(`High Income`))

# caclulate total counts for each region
region_income_counts$Total <- rowSums(region_income_counts[, -1])

# calculate percentages
region_income_percentages <- region_income_counts %>%
  mutate(across(-c(Region, Total), ~paste0(round(./Total * 100, 2), "%")))
```

```
# merge the datasets on region
merged_data <- merge(region_income_percentages, dat, by = 'Region')


# group by region and present mean values to summarise
df_grouped <- merged_data %>%
  group_by(Region) %>%
  summarise(
      `Low Income "%"` = toString(unique(`Low Income`), na.rm = TRUE),
      `Middle Income "%"` = toString(unique(`Middle Income`), na.rm = TRUE),
      `High Income "%"` = toString(unique(`High Income`), na.rm = TRUE),
      `Median HealthExpenditure2019` = round(median(`HealthExpenditure2019`,
na.rm = TRUE), 2),
      `Median LifeExpectancy2019` = round(median(`LifeExpectancy2019`, na.rm
= TRUE), 2)
  )

# calculate missing value counts for income group and health expenditure by
region
missing_value_counts <- merged_data %>%
  group_by(Region) %>%
  summarise(
      `Missing Health Expenditure 2019 N/A Count` =
sum(is.na(HealthExpenditure2019)),
      `Missing Life Expectancy 2019 N/A Count` =
sum(is.na(LifeExpectancy2019)))

# Merge missing value counts with the grouped data
df_grouped_with_missing <- merge(df_grouped, missing_value_counts)

# print as pretty table
kable(df_grouped_with_missing, align = "c", caption = "**Table 1.**
Proportion of Occurence by Region on Income Group, Health Expenditure2019 and
Life Expectancy 2019" ) %>%
  kable_classic_2()

# remove rows with missing values
df <- na.omit(dat[, c('LifeExpectancy2001', 'LifeExpectancy2019',
'IncomeGroup')])

# reshape data to long format
df_long <- pivot_longer(df, cols = c(LifeExpectancy2001, LifeExpectancy2019),
names_to = "Year", values_to = "LifeExpectancy")

##Creation of box plot

# Create combined plot
ggplot(df_long, aes(x = IncomeGroup, y = LifeExpectancy, fill = Year)) +
  geom_boxplot(position = position_dodge(width = 0.8), alpha = 0.8) +
  labs(title = "Figure 1. Life Expectancy Distribution by Income Group",
```

```
      y = "Life Expectancy",
      fill = "Year") +
  theme_minimal()



#scientific question #1

## ANOVA

fit1 <- regress("mean", LifeExpectancy2019 ~ IncomeGroup, data=dat)

Fit1

#scientific question #2

## ANCOVA

fit2 <- regress("mean", LifeExpectancy2019 ~ IncomeGroup +
HealthExpenditure2019 + IncomeGroup:HealthExpenditure2019, data=dat)
fit2

fit3 <- regress("mean", LifeExpectancy2019 ~ IncomeGroup +
HealthExpenditure2019, data=dat)
fit3

##scientific question #3

## 2001 life expectancy is an effect modifier of income group association
with life expectancy in 2019 adjusted for health expenditure in 2019. For
each year of life expectancy in 2001, there is a positive association between
income and life expectancy in 2019. The level of association within each
subgroup is statistically different.

# ANCOVA

 fit4 <- regress("mean", LifeExpectancy2019 ~ IncomeGroup +
HealthExpenditure2019 + IncomeGroup:LifeExpectancy2001, data=dat)
Fit4

Table <- read.csv(file = "Unadjusted_adjusted table - Sheet1.csv")

kable(table)

kable(table, align = "c", caption = "**Table 2. Unadjusted VS. Adjusted
Table")

%>%

kable_classic_2()
```