

Stat Project

Ema, Eileen, Elsie

Initial setup:

Introduction

1. General Life expectancy is primarily regarded as the main indicator for views on the general health of certain populations. A major part of evaluating life expectancy is knowing which other variables could have an impact on the life expectancy, and how changes in one variable affects life expectancy. The primary goal of this project is to evaluate which demographic / social health-related factors are associated with variation in life expectancy across different countries and years. Specifically, there are many individual factors such as years in schooling, percentage of measles, and economy status. Previous research, including a study published in 2018 by Daniel et. al, have emphasized the role addressing social determinants can improve patient care. Daniel et. al claimed that there is a “15-year difference in life expectancy between the most advantaged and disadvantaged citizens”. Based on this information and the dataset we chose, we wanted to explore along a similar line if there are specific factors that could affect life expectancy more than other factors.
2. Dataset The dataset used in this project came originally from Kaggle, which was sourced from the WHO (World Health Organization). The actual data was compiled from across three sources: WHO, World Bank data, and Our World in Data - a University of Oxford project. Each row represents a country and all the data collected from one year in that country across all the other variables. Even though the dataset comes from multiple sources, they are all adjusted and standardized. The observational unit was a country-year. All the dataset variables are listed below:

Dataset variables:

Country: A list of 179 distinct countries included in the dataset. Region: Categorizes the 179 countries into 9 geographical regions, such as Africa, Asia, Oceania, and the European Union. Year: The observed year, ranging from 2000 to 2015. Infant_deaths: Represents the number of infant deaths per 1,000 population. Under_five_deaths: Represents the number of deaths of

children under five years old per 1,000 population. Adult_mortality: Represents the number of deaths of adults per 1,000 population. Alcohol_consumption: Records alcohol consumption in litres of pure alcohol per capita for individuals aged 15 years and over. Hepatitis_B: Represents the percentage of coverage for Hepatitis B (HepB3) immunisation among 1-year-olds. Measles: Represents the percentage of coverage for Measles containing vaccine first dose (MCV1) immunisation among 1-year-olds. BMI: Body Mass Index, a measure of nutritional status in adults (defined as a person's weight in kilograms divided by the square of that person's height in meters). Polio: Represents the percentage of coverage for Polio (Pol3) immunisation among 1-year-olds. Diphtheria: Represents the percentage of coverage for Diphtheria tetanus toxoid and pertussis (DTP3) immunisation among 1-year-olds. Incidents_HIV: Represents the incidents of HIV per 1,000 population aged 15-49.

Response Variable: Life_expectancy: The average life expectancy for both genders across different years, from 2000 to 2015.

Predictor Variables: Economic and Demographic –

GDP_per_capita: Gross Domestic Product per capita in current US Dollars. Population_mln: Total population of a country in millions. Schooling: Average years individuals aged 25 and over have spent in formal education. Economy_status_Developed: A binary indicator (0 or 1) denoting whether a country is classified as 'Developed'. Economy_status_Developing: A binary indicator (0 or 1) denoting whether a country is classified as 'Developing'.

Lifestyle – Alcohol_consumption: Records alcohol consumption in litres of pure alcohol per capita for individuals aged 15 years and over. BMI: Body Mass Index, a measure of nutritional status in adults (defined as a person's weight in kilograms divided by the square of that person's height in meters). Thinness_ten_nineteen_years: Prevalence of thinness among adolescents aged 10-19 years (specifically, BMI < -2 standard deviations below the median). Thinness_five_nine_years: Prevalence of thinness among children aged 5-9 years (specifically, BMI < -2 standard deviations below the median).

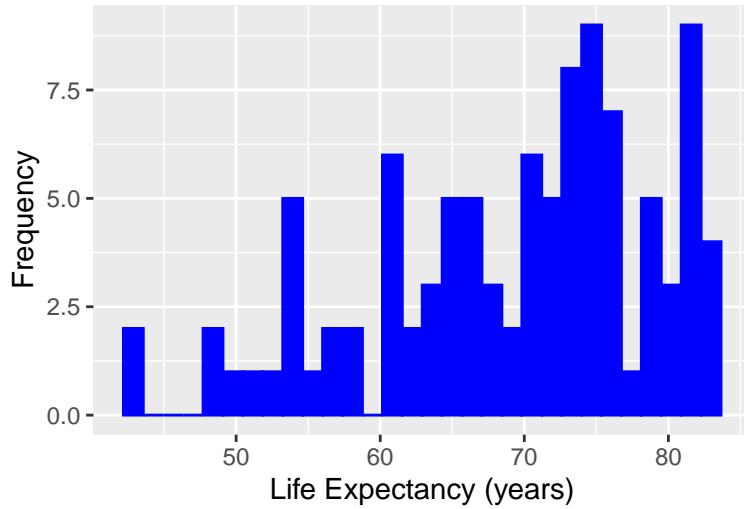
Mortality and Disease – Infant_deaths: Represents the number of infant deaths per 1,000 population. Under_five_deaths: Represents the number of deaths of children under five years old per 1,000 population. Adult_mortality: Represents the number of deaths of adults per 1,000 population. Hepatitis_B: Represents the percentage of coverage for Hepatitis B (HepB3) immunisation among 1-year-olds. Measles: Represents the percentage of coverage for Measles containing vaccine first dose (MCV1) immunisation among 1-year-olds. Polio: Represents the percentage of coverage for Polio (Pol3) immunisation among 1-year-olds. Diphtheria: Represents the percentage of coverage for Diphtheria tetanus toxoid and pertussis (DTP3) immunisation among 1-year-olds. Incidents_HIV: Represents the incidents of HIV per 1,000 population aged 15-49.

Observational Units: Country: A list of 179 distinct countries included in the dataset. Region: Categorises the 179 countries into 9 geographical regions, such as Africa, Asia, Oceania, and the European Union. Year: The observed year, ranging from 2000 to 2015.

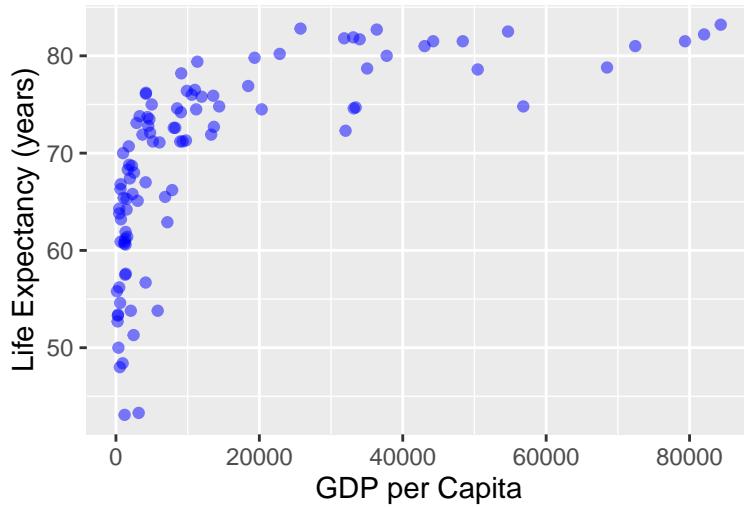
3. Rationale The reason for this data analysis is that it helps identify which health and economic variables most strongly explain differences in life expectancy. It also helps shows how different diseases / vaccinations and child and adult mortality can contribute to the overall health of a nation. This data analysis can also help us make conclusions on what types of health policy can be recommended or suggested in order to target preventable health risks in certain countries.

4. Exploratory Data Analysis

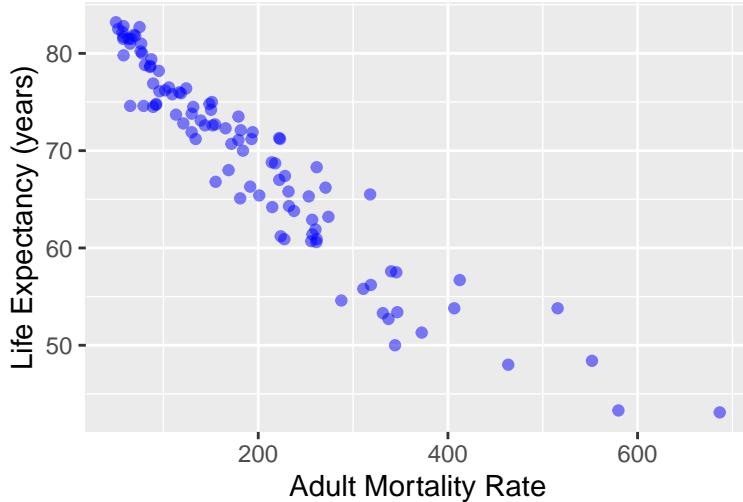
Distribution of Life Expectancy



GDP per Capita vs. Life Expectancy



Adult Mortality Rate vs. Life Expectancy



The EDA helps us visualize some of the aspects of this dataset. It helps look at life expectancy as a whole across all the countries, and to help see if it is normally distributed. Additionally, the two scatter plots visually confirm whether economic indicators (GDP per capita) and health indicators (adult mortality) are strong predictors of life expectancy.

Methodology

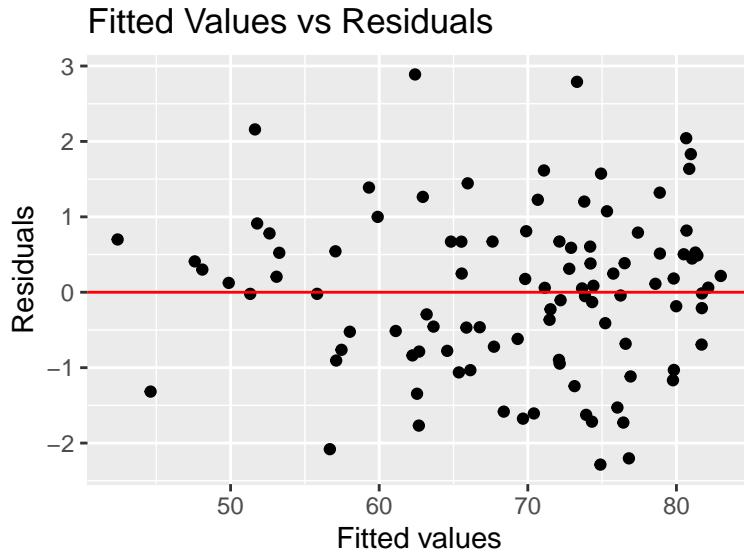
The goal of this analysis was to identify how the different variables in the dataset impacted life expectancy across different countries and years. We selected a linear regression model over other models because the outcome variable (life expectancy) is continuous and the research question focuses on estimating relationships between predictors and mean life expectancy, while holding other variables constant. A multiple linear regression is appropriate for this dataset because the outcome is continuous, the predictors are a mix of numerical and categorical variables, and the goal is to estimate how each variable relates to life expectancy. Other methods such as logistic regression, ANOVA, and t-tests, were not appropriate here because they are designed for categorical outcomes or for comparing a small number of group means. Our dataset includes a continuous response and many quantitative predictors, so linear regression is the best and most efficient method that fits our research question.

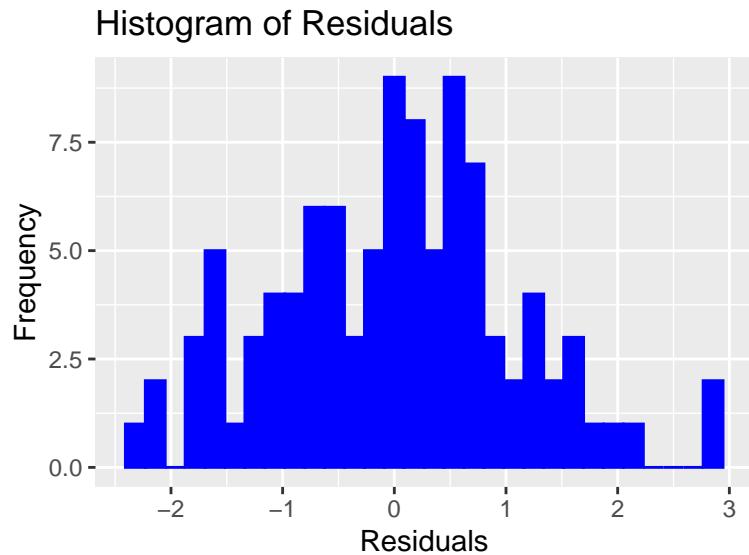
For replication, the linear regression uses the equation shown below as a guide, with life expectancy as the output, and the other variables as its predictors. When replicated, it should be ensured the data is loaded into R, and that when inputting this model into R with the lm() function, that life expectancy is the output, and that all other variables are quantitative except Economy Status Developed and Region, which are factors and therefore dummy variables in this linear regression. Results are shown using the summary() function. P-values and slopes were assessed for the goals of this project.

$$Y_i = 0 + 1 (\text{Year}_i) + 2 (\text{InfantDeath}_i) + 3 (\text{Under5Death}_i) + 4 (\text{AdultMortality}_i) + 5 (\text{AlcoholConsumption}_i) + 6 (\text{HepatitisB}_i) + 7 (\text{Measles}_i) + 8 (\text{BMI}_i) + 9 (\text{Polio}_i) + 10 (\text{Diphtheria}_i) + 11 (\text{HIVIncidence}_i) + 12 (\text{GDPperCapita}_i) + 13 (\text{PopulationMillions}_i) + 14 (\text{Thinness10to19}_i) + 15 (\text{Thinness5to9}_i) + 16 (\text{Schooling}_i) + 17 (\text{EconomyStatusDeveloped}_i) + r = 1 \ K - 1 \ r (\text{Region}_i) + i$$

The residual plot is symmetrically distributed around the horizontal axis, showing that our data satisfied the assumption of linearity. The residuals appear to be roughly vertically evenly spaced along the y axis, satisfying the assumption of constant variance of the errors. The histogram of the residuals also appears to be roughly symmetrical, satisfying the assumption of a normal distribution of errors. Finally, for independence in this dataset, each row represents a specific country in a specific year. Since each country operates independently with its own health system, economy, and demographic conditions, it makes sense to assume that the error for one observation doesn't directly affect the error for another. Because the same country can appear multiple times (for different years), it's possible that a country's values might be somewhat related over time. However, the model treats each country-year as its own observation, and we also included variables like Region and Economy Status, which already help capture similarities between countries that might otherwise create dependence. Because of this, even though small within-country patterns over time could exist, assuming that the errors are independent overall is still a reasonable and practical assumption for this analysis.

Assumptions:





Results

```
# A tibble: 25 x 3
  Variable      Estimate   P_value
  <chr>        <dbl>     <dbl>
1 (Intercept) -162.       0.032
2 Year          0.126     0.001
3 Infant_deaths -0.127    0.03
4 Under_five_deaths -0.014  0.702
5 Adult_mortality -0.044   0
6 Alcohol_consumption -0.041  0.588
7 Hepatitis_B    -0.022    0.162
8 Measles         0.01      0.315
9 BMI            -0.3       0.035
10 Polio          0.056     0.407
# i 15 more rows
```

Discussion

Citations

<https://www.acpjournals.org/doi/10.7326/M17-2441> https://www.opendatabay.com/data/ai-ml/0bdbea7e-f40b-4c41-b010-37537d03a723?utm_source