MADA Course Project

Risk Factors Affecting Life Expectancy

Annabella Hines

5/5/23

# 1. Abstract

Life expectancy is an important general metric in comparing countries’ health of their populations on the same scale, but the factors that affect it are numerous and nuanced.The goal is to create a model to better understand which risks are the most influential on life expectancy by combining two data sets containing information on life expectancy, demographics, and prevalent risk factors. Through analysis of data collected by the World Health Organization on the average life expectancy and most prevalent risk factors for death for each country over the years, I aim to draw initial associations that may be areas for further study or targets of public health measures. Through a surface level exploration of the data, the trends in life expectancy compared to other factors can be observed and then further examined with simple statistical analyses. After that point, machine learning models are employed to more accurately predict life expectancy from provided variables.

# 2. Introduction

## 2.1 General Background Information

With new unique risks arising to human health, along with older ones becoming less prevalent, the factors that tie in to life expectancy are changing over time. Broad vaccine coverage has virtually eliminated the presence of diseases like polio and measles in most developed countries and populations, while air pollution and poor nutrition/weight issues have become an increasing problem for many people(Goodson et al. 2017) (Yin et al. 2020). The risk exposures that have increased the most globally from 2010 to 2019 are ambient particulate matter pollution, drug use, high fasting plasma glucose, and high body-mass index while the ones that have decreased are household air pollution, unsafe water/sanitation/handwashing, child growth failure, tobacco smoking, and lead exposure (Murray et al. 2020). Although the leading risk factors resulting in death vary by age and location, the overall highest were high systolic blood pressure and and tobacco use(Murray et al. 2020).

Many factors can play a part in life expectancy in different countries from availability of medical care and vaccines to mental health and happiness. Factors may differ in the weight of their effect depending on developmental status of the country as well, and some previous research has shown the more obvious variables like GDP or health expenditure may not always be as influential as one may assume in these cases (Kabir 2008). There have been attempts to create formulas to mathematically determine a population’s life expectancy based on risk factors within that population but each are so specific that its difficult to create an overall generalizable model.

## 2.2 Description of data and data source

I decided to analyze two data sets, one is a life expectancy data set collected by the World Health organization which I obtained at this link https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who from Kaggle. This data specifically looks at immunization rates and the human development index for each country along with their overall life expectancy from. The other data set is Worldwide Deaths by country and risk factor which was also downloaded from Kaggle at https://www.kaggle.com/datasets/varpit94/worldwide-deaths-by-risk-factors. This data is as the title states, with the number of deaths for different risk factors in countries by year. Some examples include unsafe water sources and hygiene, child wasting and thinness, lifestyle and behavioral factors, and pollution. The summary table of the risk factor data set (**table-1?**) and life expectancy data set (**table-2?**) can be seen in the exploratory and descriptive analysis section of the manuscript.

## 2.3 Questions/Hypotheses to be addressed

Which variables are the most highly correlated to life expectancy? What risk factors or demographics are the best predictors of life expectancy? Are there differences in the most associated factors between the years 2000 and 2015?

# 3. Methods

The data for each set was separated by country and year, with the metrics represented in either percentages or number per 100,000 people. The full explanation of each variable can be found in the supplementary material. Cleaning the data involved matching the years of both data sets and combining them, along with renaming certain variables for more clarification. The analysis starts with some simple one variable linear models focused on the association to the variable of interest, life expectancy. Then a machine learning approaches was used, specifically the decision tree model.

## 3.1 Data import and cleaning

The cleaning of the raw data can be found in datacleaning.qmd file in the code folder and processing\_code subfolder. I filtered down the risk factor data set to only include the years from 2000 to 2015 in order to match the life expectancy data. I also renamed variables in each set to make the information clearer and more ubiquitous, such as ‘Entity’ to ‘Country’ in the risk factor data.

## 3.2 Statistical analysis

Initially some generalized linear models were ran with specific variables that would seem the relevant logically. After finding those results, a machine learning method, a decision tree, was used for more specialized conclusions and predictive modeling. The decision tree [Figure 4](#fig-result5) was developed initially that illustrated simple predictive variables leading to the strata of life expectancy. A multivariate linear regression was ran with all of the variables to determine which were the most significant, than that subset of variables was run again on their own.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

First, here are summary tables for each data set to see an overview of the data.

# A tibble: 31 × 17  
 skim\_type skim\_variable n\_missing complete\_rate character.min character.max  
 \* <chr> <chr> <int> <dbl> <int> <int>  
 1 character Country 0 1 4 48  
 2 numeric Year 0 1 NA NA  
 3 numeric Unsafe.water.s… 0 1 NA NA  
 4 numeric Unsafe.sanitat… 0 1 NA NA  
 5 numeric No.access.to.h… 0 1 NA NA  
 6 numeric Household.air.… 0 1 NA NA  
 7 numeric Non.exclusive.… 0 1 NA NA  
 8 numeric Discontinued.b… 0 1 NA NA  
 9 numeric Child.wasting 0 1 NA NA  
10 numeric Child.stunting 0 1 NA NA  
# ℹ 21 more rows  
# ℹ 11 more variables: character.empty <int>, character.n\_unique <int>,  
# character.whitespace <int>, numeric.mean <dbl>, numeric.sd <dbl>,  
# numeric.p0 <dbl>, numeric.p25 <dbl>, numeric.p50 <dbl>, numeric.p75 <dbl>,  
# numeric.p100 <dbl>, numeric.hist <chr>

# A tibble: 22 × 20  
 skim\_type skim\_variable n\_missing complete\_rate character.min character.max  
 \* <chr> <chr> <int> <dbl> <int> <int>  
 1 character Country 0 1 4 52  
 2 factor Status 0 1 NA NA  
 3 numeric Year 0 1 NA NA  
 4 numeric Life.expectancy 10 0.997 NA NA  
 5 numeric Adult.Mortality 10 0.997 NA NA  
 6 numeric Infant.deaths 0 1 NA NA  
 7 numeric Alcohol 194 0.934 NA NA  
 8 numeric Percentage.exp… 0 1 NA NA  
 9 numeric Hepatitis.B 553 0.812 NA NA  
10 numeric Measles 0 1 NA NA  
# ℹ 12 more rows  
# ℹ 14 more variables: character.empty <int>, character.n\_unique <int>,  
# character.whitespace <int>, factor.ordered <lgl>, factor.n\_unique <int>,  
# factor.top\_counts <chr>, numeric.mean <dbl>, numeric.sd <dbl>,  
# numeric.p0 <dbl>, numeric.p25 <dbl>, numeric.p50 <dbl>, numeric.p75 <dbl>,  
# numeric.p100 <dbl>, numeric.hist <chr>

|  |
| --- |
| Figure 1: Life Expectancy and Adult Mortality 2000 |

|  |
| --- |
| Figure 2: Life Expectancy and Adult Mortality 2015 |

The trend of life expectancy decreasing as adult mortality increases is similar in 2000 and 2015, with developed countries shown at the end with high life expectancy and low adult mortality.

|  |
| --- |
| Figure 3: Life Expectancy Status Boxplot in 2000 and 2015. |

The life expectancy distribution for developed and developing countries were similar in 2000 and 2015, with developed countries having an overall higher distribution.

## 4.2 Basic statistical analysis

Here I explored some factors that would seem to be reasonably significant with simple single predictor models to life expectancy, the outcome of interest.

First, a logistic regression with country status and life expectancy.

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 30.2 1.39 21.7 6.03e-104  
2 Life.expectancy -0.383 0.0181 -21.1 9.39e- 99

Life expectancy appears to be negatively correlated with the country status of developing.

Then a linear regression comparing year to life expectancy.

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) -636. 75.5 -8.42 5.96e-17  
2 Year 0.351 0.0376 9.33 1.96e-20

Year is positively correlated with life expectancy.

Another linear regression with BMI.

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 59.0 0.314 188. 0   
2 BMI 0.270 0.00727 37.1 8.92e-247

In this instance, BMI is positively correlated with life expectancy.

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 53.7 0.571 94.0 0   
2 Polio 0.189 0.00666 28.4 1.96e-156

Polio doesn’t seem significantly correlated.

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 71.0 0.155 459. 0   
2 HIV.AIDS -1.04 0.0288 -36.2 7.67e-238

Deaths from HIV/AIDS is negatively correlated with life expectancy.

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 69.4 0.944 73.5 1.75e-91  
2 Air.pollution -0.00000317 0.00000599 -0.530 5.97e- 1

Air pollution is negatively correlated with life expectancy.

# A tibble: 2 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 69.1 0.967 71.4 3.26e-90  
2 Low.physical.activity 0.0000368 0.0000482 0.763 4.47e- 1

## 4.3 Full analysis

Here is a decision tree model with the predictions for different stratum of life expectancy.

|  |
| --- |
| Figure 4: Decision Tree |

The initial split was on high numbers of HIV and AIDS,then polio into the lower life expectancy strata. If there are low numbers of HIV and AIDS, it further is split into income composition of resources and then adult mortality and child stunting. This a simpler and less nuanced model but has an overall look at the predictive value of the data.

Out of curiosity, here is also a decision tree for only the data from the year 2015, which had a different configuration with the addition of polio and child stunting but the overall most influential variables were HIV/AIDS, income composition of resources, and adult mortality.

|  |
| --- |
| Figure 5: Decision tree 2015 |

Next, a multivariate regression analysis was performed with all of the variables to determine which were the most significant in predicting life expectancy.

Call:  
lm(formula = Life.expectancy ~ ., data = combo\_dropped)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-9.3416 -1.9204 -0.1148 1.7419 11.6604   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)  
(Intercept) 2.921e+02 1.099e+02 2.658 0.008230  
Year -1.176e-01 5.491e-02 -2.142 0.032879  
StatusDeveloping -1.144e+00 6.997e-01 -1.635 0.102850  
Adult.Mortality -1.046e-02 1.888e-03 -5.540 5.97e-08  
Infant.deaths 1.306e-01 1.110e-01 1.177 0.239923  
Alcohol -1.417e-01 7.799e-02 -1.817 0.070029  
Percentage.expenditure -1.931e-04 4.773e-04 -0.405 0.685975  
Hepatitis.B -6.643e-03 8.232e-03 -0.807 0.420202  
Measles -2.083e-05 2.208e-05 -0.943 0.346099  
BMI 5.502e-03 1.137e-02 0.484 0.628682  
Under.five.deaths -1.807e-01 6.977e-02 -2.591 0.009985  
Polio 7.412e-03 1.011e-02 0.733 0.463753  
Total.expenditure 2.947e-01 8.513e-02 3.462 0.000602  
Diphtheria 1.198e-02 1.053e-02 1.138 0.255914  
HIV.AIDS -5.164e-01 4.669e-02 -11.059 < 2e-16  
GDP 9.581e-05 7.814e-05 1.226 0.220961  
Population 1.213e-08 8.534e-09 1.421 0.156266  
Thinness.Age1to19 -1.282e-01 1.519e-01 -0.844 0.399347  
Thinness.Age5to9 -1.069e-02 1.463e-01 -0.073 0.941795  
Income.composition.of.resources 6.231e+00 1.311e+00 4.753 2.94e-06  
Years.of.education 9.626e-01 1.206e-01 7.984 2.08e-14  
Unsafe.water.source -9.502e-05 2.679e-04 -0.355 0.723047  
Unsafe.sanitation 8.295e-05 4.062e-04 0.204 0.838303  
No.access.to.handwashing.facility 7.360e-05 4.312e-04 0.171 0.864566  
Household.air.pollution.from.solid.fuels -4.402e-04 8.436e-04 -0.522 0.602151  
Non.exclusive.breastfeeding 9.524e-05 8.324e-04 0.114 0.908974  
Discontinued.breastfeeding 6.244e-04 5.269e-03 0.118 0.905752  
Child.wasting -1.136e-04 1.619e-04 -0.702 0.483351  
Child.stunting 3.139e-04 2.186e-04 1.436 0.151846  
Low.birth.weight.for.gestation 4.626e-04 1.504e-04 3.076 0.002264  
Secondhand.smoke 8.801e-05 3.314e-04 0.266 0.790754  
Alcohol.use 7.694e-06 3.992e-05 0.193 0.847265  
Drug.use 6.020e-04 2.074e-04 2.902 0.003941  
Diet.low.in.fruits -4.081e-05 9.835e-05 -0.415 0.678431  
Diet.low.in.vegetables -8.800e-06 1.104e-04 -0.080 0.936512  
Unsafe.sex -6.985e-06 1.404e-05 -0.498 0.619063  
Low.physical.activity 2.071e-04 3.697e-04 0.560 0.575818  
High.fasting.plasma.glucose 4.519e-05 2.371e-05 1.906 0.057480  
High.total.cholesterol -4.852e-05 7.839e-05 -0.619 0.536321  
High.body.mass.index -1.594e-04 6.079e-05 -2.622 0.009116  
High.systolic.blood.pressure 8.617e-05 6.058e-05 1.422 0.155830  
Smoking -1.281e-05 3.765e-05 -0.340 0.733881  
Iron.deficiency -1.212e-03 1.198e-03 -1.012 0.312253  
Vitamin.A.deficiency 3.921e-05 2.613e-04 0.150 0.880801  
Low.bone.mineral.density 6.273e-05 5.179e-04 0.121 0.903651  
Air.pollution 3.655e-04 8.917e-04 0.410 0.682154  
Outdoor.air.pollution -2.633e-04 8.083e-04 -0.326 0.744765  
Diet.high.in.sodium -7.293e-05 7.071e-05 -1.032 0.303009  
Diet.low.in.whole.grains -1.125e-04 1.462e-04 -0.769 0.442118  
Diet.low.in.nuts.and.seeds -1.723e-04 2.328e-04 -0.740 0.459692  
   
(Intercept) \*\*   
Year \*   
StatusDeveloping   
Adult.Mortality \*\*\*  
Infant.deaths   
Alcohol .   
Percentage.expenditure   
Hepatitis.B   
Measles   
BMI   
Under.five.deaths \*\*   
Polio   
Total.expenditure \*\*\*  
Diphtheria   
HIV.AIDS \*\*\*  
GDP   
Population   
Thinness.Age1to19   
Thinness.Age5to9   
Income.composition.of.resources \*\*\*  
Years.of.education \*\*\*  
Unsafe.water.source   
Unsafe.sanitation   
No.access.to.handwashing.facility   
Household.air.pollution.from.solid.fuels   
Non.exclusive.breastfeeding   
Discontinued.breastfeeding   
Child.wasting   
Child.stunting   
Low.birth.weight.for.gestation \*\*   
Secondhand.smoke   
Alcohol.use   
Drug.use \*\*   
Diet.low.in.fruits   
Diet.low.in.vegetables   
Unsafe.sex   
Low.physical.activity   
High.fasting.plasma.glucose .   
High.total.cholesterol   
High.body.mass.index \*\*   
High.systolic.blood.pressure   
Smoking   
Iron.deficiency   
Vitamin.A.deficiency   
Low.bone.mineral.density   
Air.pollution   
Outdoor.air.pollution   
Diet.high.in.sodium   
Diet.low.in.whole.grains   
Diet.low.in.nuts.and.seeds   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 3.174 on 349 degrees of freedom  
 (3297 observations deleted due to missingness)  
Multiple R-squared: 0.881, Adjusted R-squared: 0.8643   
F-statistic: 52.73 on 49 and 349 DF, p-value: < 2.2e-16

The variables marked as significant when all of them are included in predicting life expectancy are year, adult mortality, under five deaths, total expenditure, HIV/AIDS, income composition of resources, years of education, low birth weight for gestation, drug use, and high body mass index. The ones that were the most significant of those are adult mortality, total expenditure, HIV/AIDS, income composition of resources, and years of education. Here is a regression run with those variables deemed significant by the previous model.

Call:  
lm(formula = Life.expectancy ~ Year + Adult.Mortality + Under.five.deaths +   
 Total.expenditure + HIV.AIDS + Income.composition.of.resources +   
 Years.of.education + Low.birth.weight.for.gestation + Drug.use +   
 High.body.mass.index, data = combo\_dropped)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-21.8357 -2.1326 -0.1695 2.1933 23.3532   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 1.984e+02 4.029e+01 4.924 9.08e-07 \*\*\*  
Year -7.250e-02 2.012e-02 -3.604 0.000320 \*\*\*  
Adult.Mortality -1.625e-02 9.187e-04 -17.685 < 2e-16 \*\*\*  
Under.five.deaths -4.147e-02 3.421e-03 -12.123 < 2e-16 \*\*\*  
Total.expenditure 1.577e-01 3.756e-02 4.198 2.79e-05 \*\*\*  
HIV.AIDS -5.877e-01 2.245e-02 -26.183 < 2e-16 \*\*\*  
Income.composition.of.resources 9.410e+00 6.673e-01 14.102 < 2e-16 \*\*\*  
Years.of.education 1.127e+00 4.316e-02 26.111 < 2e-16 \*\*\*  
Low.birth.weight.for.gestation 1.949e-04 1.777e-05 10.967 < 2e-16 \*\*\*  
Drug.use 9.612e-05 2.755e-05 3.489 0.000493 \*\*\*  
High.body.mass.index -1.483e-05 5.107e-06 -2.904 0.003717 \*\*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 4.095 on 2376 degrees of freedom  
 (1309 observations deleted due to missingness)  
Multiple R-squared: 0.8169, Adjusted R-squared: 0.8162   
F-statistic: 1060 on 10 and 2376 DF, p-value: < 2.2e-16

# 5. Discussion

## 5.1 Summary and Interpretation

After cleaning the data and combining the sets, the variables were explored to get an estimate of possible associations to look into further. Status of developing or developed appeared highly correlated with life expectancy, along with infant deaths and year. After getting an overall idea of what the data looked like and loose associations, statistical tests were ran to get a more concrete idea of the actual association. These tests and models provided output on what variables had the closest relevance and association to the provided life expectancy.They didn’t contradict the initial visual associations but revealed more significant correlations such as income composition of resources and adult mortality. It makes logical sense that adult mortality would be tightly associated with life expectancy as it is the rate of adult deaths. Income composition is an interesting variable, as it is described as a version of human development index, which is a kind of quantification of development status. This makes sense with how the status variable itself was correlated highly as well. The other factor deemed significant by both the multivariate regression and the decision tree was the number of deaths per 1,000 live births due to HIV/AIDS for children 0-4 years of age. This is reasonable as it indicates the level of HIV or AIDS in the total population as the most common mode that children contract HIV is through the mother. It illustrates deaths in one of the most vulnerable parts of the population and indicates that it may be in a community without resources for infant prophylaxis, possibly a lower income or developing country (Levy 1993).

## 5.2 Strengths and Limitations

Some strengths include how large and expansive the data set is, as it has many variables for nearly every country and with a large range of years. It also comes from a very reliable source, the World Health Organization. A weakness that goes along with that is that it was difficult to work with such a large data set and pick out what the most important things to explore were. Also life expectancy is a nuanced estimate that has to do with a combination of multiple factors and also things that are not included in the available data, so a truly accurate analysis may be out of my reach. There was also a fair amount of missing data for some variables, also making it hard to have a fully formed analysis.

## 5.3 Conclusions

Both full statistical analyses were in general agreement of adult mortality, income composition of resources, and HIV/AIDS being the most predictive of life expectancy. These could be factors to explore more in the future for more concrete correlations and associations and possibly eventually the targets for improving life expectancy in various parts of the world. The purpose of these results is to be purely exploratory and it has answered the original questions.

# 6. References

Goodson, James L., James P. Alexander, Robert W. Linkins, and Walter A. Orenstein. 2017. “Measles and Rubella Elimination: Learning from Polio Eradication and Moving Forward with a Diagonal Approach.” *Expert Review of Vaccines* 16 (12): 1203–16. <https://doi.org/10.1080/14760584.2017.1393337>.

Kabir, Mahfuz. 2008. “Determinants of Life Expectancy in Developing Countries.” *The Journal of Developing Areas* 41 (2): 185–204. <https://doi.org/10.1353/jda.2008.0013>.

Levy, J A. 1993. “Pathogenesis of Human Immunodeficiency Virus Infection.” *Microbiological Reviews* 57 (1): 183–289. <https://doi.org/10.1128/mr.57.1.183-289.1993>.

Murray, Christopher J L, Aleksandr Y Aravkin, Peng Zheng, Cristiana Abbafati, Kaja M Abbas, Mohsen Abbasi-Kangevari, Foad Abd-Allah, et al. 2020. “Global Burden of 87 Risk Factors in 204 Countries and Territories, 19902019: A Systematic Analysis for the Global Burden of Disease Study 2019.” *The Lancet* 396 (10258): 1223–49. <https://doi.org/10.1016/s0140-6736(20)30752-2>.

Yin, Peng, Michael Brauer, Aaron J Cohen, Haidong Wang, Jie Li, Richard T Burnett, Jeffrey D Stanaway, et al. 2020. “The Effect of Air Pollution on Deaths, Disease Burden, and Life Expectancy Across China and Its Provinces, 19902017: An Analysis for the Global Burden of Disease Study 2017.” *The Lancet Planetary Health* 4 (9): e386–98. <https://doi.org/10.1016/s2542-5196(20)30161-3>.