The World Health Organization: Factors that Influence Life Expectancy

Lily Geiser, Hannah Ashburn, Angel Baeza

2023-12-16

I. Introduction

The World Health Organization (WHO) is a specialized agency under the United Nations, its parent organization. Established on April 7, 1948, the WHO merged the assets, personnel, and duties of the League of Nations Health Organization with Paris's Office International d'Hygiène Publique, which included the International Classification of Diseases (ICD).

While their primary objective is to help ensure the global population attains the highest level of health possible, the WHO has also been actively involved in gathering a wide range of data from member countries. This data, most of which is publicaly available, has been used to conduct a variety of research projects in hopes of solving the world's most pressing global health challenges. The metrics collected by the WHO's Global Health Observatory (GHO) include a broad range of health indicators such as mortality, disease prevalence, health system performance, and other social determinants of health. They also collect more qualitative data through the World Health Survey, which collects voluntarily submitted insights into people's perceived access to healthcare, health-related behaviors, and other personal information that would not be available otherwise.

By analyzing this data, not only can we add to a growing body of research, we can contribute our findings towards creating a better understanding of health trends and disparities across the globe. Furthermore, any remarkable trends identified in this data could have significant implications on policy development, resource allocation, and disease prevention and control– all of which are vastly important areas. Such findings could bolster and facilitate efforts to increase the global population's quality of life and overall life expectancy, generating a significant impact on society as we know it.

II. Model Creation and Analysis: Full Model

The first model we created was a multiple regression model using data from only the year 2015. Our goal in creating this model is to see how life expectancy is correlated with the other variables in our dataset. We used data from only 2015 so that the different life expectancies reported varied by country alone, not by country and. year. Further, this model could then be applied to other years to verify that its accuracy; if this model is not accurate from year to year, this could imply to us that there are other confounding variables not included in the data, or that the explanatory factors could have varying impacts by country.

The full model we created it shown below. As discussed, the independent variable is life expectancy; all other variables in the dataset are used, excluding country (as each life expectancy comes from a different country, therefore this variable could predict the life expectancy with 100% accuracy), year (as all data is from 2015), and Economy_status_Developing (as this is simply the reverse of Economy_status_Developed, which is included). Our null hypothesis associated with this model is that none of the explanatory variables have any correlation with life expectancy (i.e. that their coefficients equal 0). Our alternative hypothesis is that at least one of the explanatory variables has a coefficient that does not equal zero. For this model and all models used throughout this report, we will be using a significance level of 0.05.

```r
#importing data and creating 2015 dataset
data <- read.csv("cleandata.csv")
data.2015 <- data[data$Year == 2015, ]

#creating regression model and outputting summary data
reg_2015 <- lm(Life_expectancy~as.factor(Region)+Infant_deaths+Under_five_deaths+
                 Adult_mortality+Alcohol_consumption+Hepatitis_B+Measles+BMI+Polio+
                 Diphtheria+Incidents_HIV+GDP_per_capita+Population_mln+
                 Thinness_ten_nineteen_years+Thinness_five_nine_years+Schooling+
                 as.factor(Economy_status_Developed), data=data.2015)
print(summary(reg_2015))
```

```
##
## Call:
## lm(formula = Life_expectancy ~ as.factor(Region) + Infant_deaths +
##     Under_five_deaths + Adult_mortality + Alcohol_consumption +
##     Hepatitis_B + Measles + BMI + Polio + Diphtheria + Incidents_HIV +
##     GDP_per_capita + Population_mln + Thinness_ten_nineteen_years +
##     Thinness_five_nine_years + Schooling + as.factor(Economy_status_Developed),
##     data = data.2015)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9233 -0.8085  0.0111  0.6475  3.6337
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                 8.428e+01  2.365e+00  35.632
## as.factor(Region)Asia                       2.464e-01  4.255e-01   0.579
## as.factor(Region)Central America and Caribbean  1.799e+00  4.834e-01   3.722
## as.factor(Region)European Union             -7.480e-01  6.876e-01  -1.088
## as.factor(Region)Middle East                1.204e-01  5.340e-01   0.226
## as.factor(Region)North America              -1.769e-01  9.619e-01  -0.184
## as.factor(Region)Oceania                    -1.078e+00  5.891e-01  -1.831
## as.factor(Region)Rest of Europe             1.407e-01  5.500e-01   0.256
## as.factor(Region)South America              1.849e+00  5.292e-01   3.494
## Infant_deaths                               -3.149e-02  3.852e-02  -0.817
## Under_five_deaths                           -6.452e-02  2.631e-02  -2.452
## Adult_mortality                             -4.996e-02  2.992e-03 -16.698
## Alcohol_consumption                         -1.137e-02  4.914e-02  -0.231
## Hepatitis_B                                 -2.112e-02  2.444e-02  -0.864
## Measles                                     1.168e-02  8.017e-03   1.456
## BMI                                         -1.414e-01  8.248e-02  -1.715
## Polio                                       -4.576e-03  2.242e-02  -0.204
## Diphtheria                                  1.564e-02  2.792e-02   0.560
## Incidents_HIV                               2.071e-01  9.027e-02   2.294
## GDP_per_capita                              2.476e-05  8.902e-06   2.782
## Population_mln                              -1.112e-04  7.092e-04  -0.157
## Thinness_ten_nineteen_years                 -1.251e-01  1.127e-01  -1.109
## Thinness_five_nine_years                    1.078e-01  1.124e-01   0.959
## Schooling                                   7.727e-02  7.408e-02   1.043
## as.factor(Economy_status_Developed)1        2.679e+00  6.394e-01   4.189
##                                             Pr(>|t|)
```

```
## (Intercept)                                      < 2e-16 ***
## as.factor(Region)Asia                           0.563373
## as.factor(Region)Central America and Caribbean 0.000277 ***
## as.factor(Region)European Union                 0.278369
## as.factor(Region)Middle East                    0.821848
## as.factor(Region)North America                  0.854297
## as.factor(Region)Oceania                        0.069058 .
## as.factor(Region)Rest of Europe                 0.798480
## as.factor(Region)South America                  0.000621 ***
## Infant_deaths                                   0.414921
## Under_five_deaths                               0.015324 *
## Adult_mortality                                  < 2e-16 ***
## Alcohol_consumption                             0.817268
## Hepatitis_B                                     0.388847
## Measles                                         0.147345
## BMI                                             0.088415 .
## Polio                                           0.838578
## Diphtheria                                      0.576235
## Incidents_HIV                                   0.023123 *
## GDP_per_capita                                  0.006083 **
## Population_mln                                  0.875556
## Thinness_ten_nineteen_years                     0.268963
## Thinness_five_nine_years                        0.338990
## Schooling                                       0.298511
## as.factor(Economy_status_Developed)1            4.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.247 on 154 degrees of freedom
## Multiple R-squared:  0.9781, Adjusted R-squared:  0.9746
## F-statistic: 286.1 on 24 and 154 DF,  p-value: < 2.2e-16
```

There are several insights we can obtain from the above model. Firstly, we can say that this model appears to be a good fit and does an excellent job of predicting life expectancy. Firstly, we can observe that the p-value associated with the overall model is extremely close to 0, at less than 2.2e-16. This tells us that we can certainly reject our null hypothesis and accept the alternative hypothesis. Further, we can see that the multiple R-squared value is 0.9781; this is a value incredibly close to 1, which tells us that nearly all of the variation in life expectancy can be explained by the variation in the explanatory variables.

If we look at the coefficients associated with the variables, we will observe that multiple of them are statistically significant; specifically, the regions Central America and the Caribbean, and South America; the under five mortality rate; adult mortality; HIV incidence; GDP per capita; and whether the country is classified as Developed or Developing all have statistically significant coefficients. The coefficient associated with the intercept is also significant. This tells us there is evidence that these variables specifically have nonzero coefficents; there is not sufficient evidence to come to this conclusion for the other variables, as they all have associated p-values of greater than 0.05.

III. Explanatory Variable Correlation and a Reduced Model

While our model above appears to be an incredibly good fit, we wanted to see if this model could be improved upon with a reduced model. Firstly, we wanted to see whether any of the variables above were highly correlated. To accomplish this, we first obtained the associated VIF values:

```r
library(car)
```

```
## Loading required package: carData
```

```r
print(vif(reg_2015))
```

```
##                                       GVIF Df GVIF^(1/(2*Df))
## as.factor(Region)                78.924227  8        1.313940
## Infant_deaths                    78.408752  1        8.854872
## Under_five_deaths                82.240022  1        9.068628
## Adult_mortality                   8.291875  1        2.879562
## Alcohol_consumption               3.872895  1        1.967967
## Hepatitis_B                      13.718816  1        3.703892
## Measles                           1.927141  1        1.388215
## BMI                               3.739444  1        1.933764
## Polio                             9.762729  1        3.124537
## Diphtheria                       19.264854  1        4.389175
## Incidents_HIV                     2.451655  1        1.565776
## GDP_per_capita                    2.847603  1        1.687484
## Population_mln                    1.235429  1        1.111499
## Thinness_ten_nineteen_years      24.635592  1        4.963425
## Thinness_five_nine_years         25.470867  1        5.046867
## Schooling                         6.219399  1        2.493872
## as.factor(Economy_status_Developed)  7.715681  1        2.777712
```

As we can see, some of these variables have incredibly high VIF values; specifically, Region, Infant_deaths, Under_five_deaths, Hepatitis_B, Diphtheria, Thinness_ten_nineteen_years, and Thinness_five_nine_years all have GVIF values above 10. It is important to note that the adjusted GVIF values found in the righthand column are signficantly lower (for example, due to the multiple degrees of freedom in the Region variable, the adjusted GVIF is substantially lower). However, our goal in creating the reduced model was to reduce multicollinearity as much as possible.

Our next step was to determine which specific variables are closely related to each other. The code to accomplish this is below; the output of this code is which cells of the covariance matrix have values greater than 0.8, and the associated variables. Categorical variables were excluded from the generated covariance matrix.

```r
#create covariance matrix
matrix <- cor(data.2015[, sapply(data.2015, is.numeric)])
```

```
## Warning in cor(data.2015[, sapply(data.2015, is.numeric)]): the standard
## deviation is zero
```

```r
#find high correlation values and their corresponding row and column names
high_correlation_indices <- which(matrix > 0.8 & matrix < 1, arr.ind = TRUE)
high_correlation_values <- matrix[high_correlation_indices]
row_names <- rownames(matrix)[high_correlation_indices[, 1]]
col_names <- colnames(matrix)[high_correlation_indices[, 2]]

#export information to a datafram and print in order of correlation
result_df <- data.frame(
  Row = row_names,
```

```
  Column = col_names,
  Correlation = high_correlation_values
)
result_df <- result_df[order(-result_df$Correlation), ]
print(result_df)
```

```
##                            Row                    Column Correlation
## 1            Under_five_deaths             Infant_deaths   0.9899859
## 3                Infant_deaths         Under_five_deaths   0.9899859
## 13    Thinness_five_nine_years Thinness_ten_nineteen_years   0.9731355
## 14 Thinness_ten_nineteen_years    Thinness_five_nine_years   0.9731355
## 8                   Diphtheria                 Hepatitis_B   0.9438458
## 11                Hepatitis_B                  Diphtheria   0.9438458
## 10                  Diphtheria                       Polio   0.9296290
## 12                       Polio                  Diphtheria   0.9296290
## 7                        Polio                Hepatitis_B   0.8886645
## 9                 Hepatitis_B                       Polio   0.8886645
## 2              Adult_mortality             Infant_deaths   0.8537298
## 5                Infant_deaths            Adult_mortality   0.8537298
## 4              Adult_mortality         Under_five_deaths   0.8422934
## 6            Under_five_deaths            Adult_mortality   0.8422934
```

As we can see above, there are multiple variables which are very strongly correlated. For example, Under_five_deaths and Infant_deaths have a near perfect correlation, with a covariance value of nearly one.

Based on the above information, we created the reduced model coded below:

```
reg_2015_red <- lm(Life_expectancy~Adult_mortality
                   +Alcohol_consumption+Measles+BMI+
                   Diphtheria+Incidents_HIV+GDP_per_capita+Population_mln+
                   Thinness_ten_nineteen_years+Schooling+
                   as.factor(Economy_status_Developed), data=data.2015)
print(summary(reg_2015_red))
```

```
##
## Call:
## lm(formula = Life_expectancy ~ Adult_mortality + Alcohol_consumption +
##     Measles + BMI + Diphtheria + Incidents_HIV + GDP_per_capita +
##     Population_mln + Thinness_ten_nineteen_years + Schooling +
##     as.factor(Economy_status_Developed), data = data.2015)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.646 -1.098 -0.022  1.058  5.239
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        7.816e+01  2.691e+00  29.044  < 2e-16 ***
## Adult_mortality                   -7.208e-02  2.898e-03 -24.868  < 2e-16 ***
## Alcohol_consumption                1.450e-01  5.478e-02   2.647 0.008904 **
## Measles                            7.219e-03  1.064e-02   0.678 0.498413
## BMI                               -6.087e-02  8.871e-02  -0.686 0.493592
## Diphtheria                         2.746e-02  1.105e-02   2.485 0.013938 *
```

```
## Incidents_HIV                           3.777e-01  1.090e-01   3.465 0.000673 ***
## GDP_per_capita                          1.505e-05  1.103e-05   1.365 0.174239
## Population_mln                          1.448e-04  9.423e-04   0.154 0.878047
## Thinness_ten_nineteen_years            -7.239e-02  4.377e-02  -1.654 0.100011
## Schooling                               3.443e-01  8.269e-02   4.164 4.99e-05 ***
## as.factor(Economy_status_Developed)1    3.622e-02  5.487e-01   0.066 0.947452
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.712 on 167 degrees of freedom
## Multiple R-squared:  0.9552, Adjusted R-squared:  0.9522
## F-statistic: 323.4 on 11 and 167 DF,  p-value: < 2.2e-16
```

We can see based on the above that this model is still an incredibly good fit; the multiple R-squared value is still extremely close to 1 at 0.9552, and the p-value associated with the overall model is still extremely small. This model also has a significantly reduced amount of multicollinearity, as shown below:

```
library(car)
print(vif(reg_2015_red))
```

```
##                       Adult_mortality                   Alcohol_consumption
##                              4.127313                              2.552720
##                               Measles                                   BMI
##                              1.800868                              2.295263
##                             Diphtheria                         Incids_HIV
##                              1.600795                              1.896535
##                          GDP_per_capita                       Population_mln
##                              2.320639                              1.157087
##          Thinness_ten_nineteen_years                          Schooling
##                              1.970545                              4.111272
## as.factor(Economy_status_Developed)
##                              3.014722
```

All of the VIF values above are quite low, demonstrating that we have managed to reduce collinearity.

This reduced model also provides some different insights into the data. For example, alcohol consumption, Diphtheria, and Schooling now all have significant p-values associated with their coefficients. This could either be because they truly do have a significant relationship with life expectancy, which was masked in the full model, or because we removed too many variables and they are confounding the information we have above.

To determine whether we should use the full model or reduced model, we will perform an ANOVA test:

```
anova(reg_2015_red, reg_2015, test = "F")
```

```
## Analysis of Variance Table
##
## Model 1: Life_expectancy ~ Adult_mortality + Alcohol_consumption + Measles +
##     BMI + Diphtheria + Incids_HIV + GDP_per_capita + Population_mln +
##     Thinness_ten_nineteen_years + Schooling + as.factor(Economy_status_Developed)
## Model 2: Life_expectancy ~ as.factor(Region) + Infant_deaths + Under_five_deaths +
##     Adult_mortality + Alcohol_consumption + Hepatitis_B + Measles +
##     BMI + Polio + Diphtheria + Incids_HIV + GDP_per_capita +
```

```
##     Population_mln + Thinness_ten_nineteen_years + Thinness_five_nine_years +
##     Schooling + as.factor(Economy_status_Developed)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    167 489.59
## 2    154 239.52 13    250.07 12.368 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output above tells us that, although we had initially seen that the reduced model is still a pretty good fit, our full model is in fact substantially better than the reduced model. The p-value associated with this ANOVA test is < 2.2e-16, which tells us there is a statistically significant difference in the performance of the two models. Further, we can see that the RSS associated with the reduced model is over twice the value of the full model (489.59 v. 239.52). Despite the good fit of the reduced model, we can therefore conclude based on this that the full model should give us more accurate results.

IV. Testing our Model on 2014 Data

To further confirm that our full model is more accurate than our reduced model, we are going to use these two models to predict the life expectancy for the different countries in 2014, and see which yields more accurate results. The code for this trial can be found below:

```
#creating 2014 data
data.2014 <- data[data$Year == 2014, ]

#predicting 2014 life expectancy with different models
predict_full <- predict(reg_2015, data.2014)
predict_red <- predict(reg_2015_red, data.2014)

#find MSE for both models
mse_full <- mean((predict_full - data.2014$Life_expectancy)^2)
mse_red <- mean((predict_red - data.2014$Life_expectancy)^2)
print(mse_full)
```

```
## [1] 1.354277
```

```
print(mse_red)
```

```
## [1] 2.834837
```

We can see based on the above that our test MSE for the full model is approximately 1.35, while the test MSE for the reduced model is about 2.83. As the full model produces a lower MSE, we can once again conclude that the full model produces a more accurate prediction of life expectancy than the reduced model.