Transforming to Stay Alive

Katelyn Patricio

2025-07-01

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
GDP <- read.csv('World Development Indicators GDP Per Capita.csv')
life_exp <- read.csv('LifeExpectancy-1.csv')
```

```
head(GDP)
```

```
##                X    X1990    X2000    X2010    X2020 X.1
## 1    Afghanistan       ..       ..    554.6    516.9  NA
## 2        Albania    617.2  1,126.7  4,094.3  5,343.0  NA
## 3        Algeria  2,431.6  1,780.4  4,495.9  3,354.2  NA
## 4 American Samoa       ..       .. 10,446.9 15,501.5  NA
## 5        Andorra 19,208.7 21,620.5 48,237.9 37,207.2  NA
## 6         Angola    949.3    556.9  3,496.8  1,503.0  NA
```

```r
head(life_exp)
```

```
##           Country.Name Country.Code                    Region
## 1          Afghanistan          AFG                South Asia
## 2               Angola          AGO        Sub-Saharan Africa
## 3              Albania          ALB     Europe & Central Asia
## 4              Andorra          AND     Europe & Central Asia
## 5 United Arab Emirates          ARE Middle East & North Africa
## 6            Argentina          ARG  Latin America & Caribbean
##           IncomeGroup Year Life.Expectancy.World.Bank
## 1          Low income 2001                     56.308
## 2 Lower middle income 2001                     47.059
## 3 Upper middle income 2001                     74.288
## 4         High income 2001                         NA
## 5         High income 2001                     74.544
## 6 Upper middle income 2001                     73.755
##   Prevelance.of.Undernourishment    CO2 Health.Expenditure..
## 1                           47.8    730                   NA
## 2                           67.5  15960              4.483516
## 3                            4.9   3230              7.139524
## 4                             NA    520              5.865939
## 5                            2.8  97200              2.484370
## 6                            3.0 125260              8.371798
##   Education.Expenditure.. Unemployment Corruption Sanitation   Injuries
## 1                      NA       10.809         NA         NA 2179727.10
## 2                      NA        4.004         NA         NA 1392080.71
## 3                 3.45870       18.575         NA   40.52090  117081.67
## 4                      NA           NA         NA   21.78866    1697.99
## 5                      NA        2.493         NA         NA  144678.14
## 6                 4.83374       17.320         NA   48.05400 1397676.07
##   Communicable NonCommunicable
## 1   9689193.70      5795426.38
## 2  11190210.53      2663516.34
## 3    140894.78       532324.75
## 4       695.56        13636.64
## 5     65271.91       481740.70
## 6   1507068.98      8070909.52
```

```r
colnames(GDP)
```

```
## [1] "X"     "X1990" "X2000" "X2010" "X2020" "X.1"
```

```r
colnames(life_exp)
```

```
##  [1] "Country.Name"                   "Country.Code"
##  [3] "Region"                         "IncomeGroup"
##  [5] "Year"                           "Life.Expectancy.World.Bank"
##  [7] "Prevelance.of.Undernourishment" "CO2"
##  [9] "Health.Expenditure.."           "Education.Expenditure.."
## [11] "Unemployment"                   "Corruption"
## [13] "Sanitation"                     "Injuries"
## [15] "Communicable"                   "NonCommunicable"
```

```r
cleaned <- GDP %>%
  rename(Country.Name = X)

reshaped <- cleaned %>%
  pivot_longer(cols = starts_with("X"),
               names_to = "Year",
               values_to = "GDP") %>%
  mutate(Year = as.numeric(gsub("X", "", Year)),
         GDP = gsub(",", "", GDP),
         GDP = ifelse(GDP %in% c("", "NA", "N/A", "..", "--"), NA, GDP),
         GDP = as.numeric(GDP))

merged <- left_join(life_exp, reshaped, by = c("Country.Name", "Year"))
```

In order to find the relationship between gross domestic product (GDP) and life expectancy, we need to merge the data. Before that is done, the data needs to be cleaned and transformed. Here I am changing the GDP dataframe to match the life expetancy with its country name column as well as transform it from a wide format to a long format. I then merge the data by country name and year.

```r
model <- lm(Life.Expectancy.World.Bank ~ GDP, data = merged)
summary(model)
```

Running the model, we see a strong relationship between GDP and life expectancy. For every one-unit increase of GDP, life expectancy increases by about 0.0002677. Although a small unit change, the p-value indicates that this is not by chance.

```
##
## Call:
## lm(formula = Life.Expectancy.World.Bank ~ GDP, data = merged)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.385  -4.689   2.335   5.494  10.736
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.619e+01  7.062e-01  93.732   <2e-16 ***
## GDP         2.677e-04  2.884e-05   9.284   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.401 on 160 degrees of freedom
##   (3144 observations deleted due to missingness)
## Multiple R-squared:  0.3501, Adjusted R-squared:  0.346
## F-statistic: 86.19 on 1 and 160 DF,  p-value: < 2.2e-16
```

```r
life_exp_grouped <- life_exp %>%
  mutate(YearGroup = case_when(
    Year >= 2000 & Year <= 2009 ~ 2000,
    Year >= 2010 & Year <= 2014 ~ 2010,
    Year >= 2015 & Year <= 2020 ~ 2020
  )) %>%
  filter(!is.na(YearGroup)) %>%  # Drop any NA (doesn't fit into category)
  group_by(Country.Name, YearGroup) %>%
  summarise(
    Life.Expectancy.World.Bank = mean(Life.Expectancy.World.Bank, na.rm = TRUE),
    .groups = "drop"
  )
```

```r
merged_grouped <- left_join(life_exp_grouped, reshaped, by = c("Country.Name" = "Country.Name", "YearGr
```

```r
model2 <- lm(Life.Expectancy.World.Bank ~ GDP, data = merged_grouped)
summary(model2)
```

I wanted to further my analysis as the data may have some limitation due to the differences in the data itself. GDP only corresponds to four years (1990, 2000, 2010, and 2020) while life expectancy captures all dates in between. In order to address this I decided to group the data and then merge. Ignoring 1990 as life expectancy starts at 2001, I group the life expectancy data to match those within GDP. Since we are finding life expectancy for multiple years, we find the mean.

```
##
## Call:
## lm(formula = Life.Expectancy.World.Bank ~ GDP, data = merged_grouped)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.608  -4.079   2.328   5.444  10.301
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.694e+01  3.999e-01  167.37   <2e-16 ***
## GDP         2.860e-04  1.819e-05   15.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.362 on 481 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.3395, Adjusted R-squared:  0.3382
## F-statistic: 247.3 on 1 and 481 DF,  p-value: < 2.2e-16
```

Conclusion: Both models suggest a statistically positive relationship between GDP and life expectancy. Pulling from both models, for every one unit increase in GDP, life expectancy

increases about 0.00027. Both have very small p-values (< 0.001) suggesting this relationship is statistically significant. In conclusion, both models propose that countries who are wealthier tend to have higher life expectancy. Despite this relationship, there are many other factors that could be contributing to life expectancy such as health, education, and unemployment. However, I believe GDP is the most important as it could impact these factors as well.