DATA 602 Project - Due on October 20, 2021, 11:59pm

By: Sallene Wong (10122532), Jordan Swanson (10005366), Alberto Ávila García (30153763)

2021-10-20

-

Contents

Motivation

Now more than ever, health data is a major topic in society. Since the beginning of the COVID-19 pandemic, we've been inundated with statistics on infection rates, death rates, R_0 values, and many, many more. It has turned many people (for the better or for the worse) into armchair statisticians and data experts. Part of the increase is the wide availability of published data from various forms of governments, as well as the increased availability of tools and the resources to learn how to use them (e.g. R). The effect of these changes required only a catalyst, which COVID has served as, to allow people of all stripes to try their hand at more casual health data investigation.

Though we're investigating our chosen dataset with a less casual interest, the data we're looking at is widely available and has been for decades. Both governments and non-governmental organizations appreciate the value of population data, and there are no shortage of reliable repositories from which we can access the data. Given that we have the tools, the knowledge, and an ongoing grim reminder of the importance of public health, we've chosen exactly that as our data of interest.

Pushing COVID-19 from our minds for a minute, our focus is on public health in non-pandemic years. One of the primary measures when looking at population-level health data is life expectancy, both for its availability and ease of measurement. Though alternatives have been proposed (Stiefel, 2010), they don't have the widespread adoption, nor the clout. In addition to life expectancy, we'd like to investigate several other measurements that we could reasonably expect to be associated with increased longevity. Some of these factors are direct measures of public health, others are influenced heavily by government policy, and more yet are included to allow us to compare countries across the world. Thankfully this data has been collected for decades by the United Nations (UN) and it's related organizations.

When talking about countries at the world level, it would be difficult to compare them without mentioning basic metrics, such as population and gross domestic product (GDP). We hypothesize that GDP, expressed as per capita, will have a significant influence on many of the other public health measures. Availability of education is another metric that we're quite interested in, suspecting that it may be related to a country's GDP, though the direction of the relationship we don't want to speculate on just yet. Aside from GDP, the UN also provides an indicator known as Human Development Index (HDI) that brings together many of these variables. We won't be utilizing HDI, other than its role in classifying "developed" versus "developing" nations.

All of the data we're looking at is sourced from different UN agencies: the World Health Organization, the Global Health Observatory, and the World Bank. These agencies provide high-quality, easily accessible data from the UN's member states. Using this data, we want to analyze the factors that make a country desirable in each health data category, and see if we can identify any correlations. Once we've found variables of interest, we want to see if we can utilize linear regression to predict any missing data values we have, with a focus on life expectancy. The specific measurements we'll be looking at are:

```
##  [1] "Country"
##  [2] "Year"
##  [3] "BMI"
##  [4] "Adult mortality"
##  [5] "Status"
##  [6] "Health expenditure as a percentage of GDP"
##  [7] "Life Expectancy"
##  [8] "Population"
##  [9] "Mean years of schooling"
## [10] "GDP per capita (USD)"
```

Data Description

Some definitions for the data measurements we'll be investigating:

Country - Common name of country of interest

Year - Calendar year of data collection

Status - UN classification of a "developed" vs "developing" country - Human Development Index score of > 0.800 = "Developed"; < 0.800 = "Developing"

Adult Mortality - Rate of both sexes' probability of dying between 15 and 60 years of age (per 1000 population)

Health Expenditure - Expenditure on health as a percentage of GDP (per capita)

Body Mass Index (BMI) - Average body mass index of the population (both sexes) in kg/mˆ2

Gross Domestic Product (GDP) - Gross domestic product of a country (in 2019 USD)

Life Expectancy - Life expectancy at birth, total years (both sexes)

Population - Total population of a country

Schooling - Average number of years of education received by people ages 25 and older, converted from education attainment levels using official durations of each level

Correlation Plots

First off, we'll construct a correlation plot to identify correlations between our variables of interest. This will allow us to find meaningful relationships and narrow our focus on the measurements that we should investigate

TABLE 1 - correlation plot between all listed variables

```r
# Filter numerics, print example data
num_data <- master_data[, c(3,4,6,7,8,9,10)]
head(num_data, 6)
```

```
##    BMI Adult mortality Health expenditure as a percentage of GDP
## 1 21.5        316.0496                                        NA
## 2 21.6        307.2416                                        NA
## 3 21.7        292.3430                                   8.98703
## 4 21.8        286.4569                                   5.57518
## 5 21.9        281.8943                                   6.98491
## 6 22.0        277.1813                                   5.49062
##   Life Expectancy Population Mean years of schooling GDP per capita (USD)
## 1          55.841   20779957                     2.2                   NA
## 2          56.308   21606992                     2.2                   NA
## 3          56.784   22600774                     2.3             179.4266
## 4          57.271   23680871                     2.4             190.6838
## 5          57.772   24726689                     2.5             211.3821
## 6          58.290   25654274                     2.6             242.0313
```

```r
# Construct correlation matrices, drop NAs

cat("This table chows the numeric correlations between the variables:")
```

```
## This table chows the numeric correlations between the variables:
```

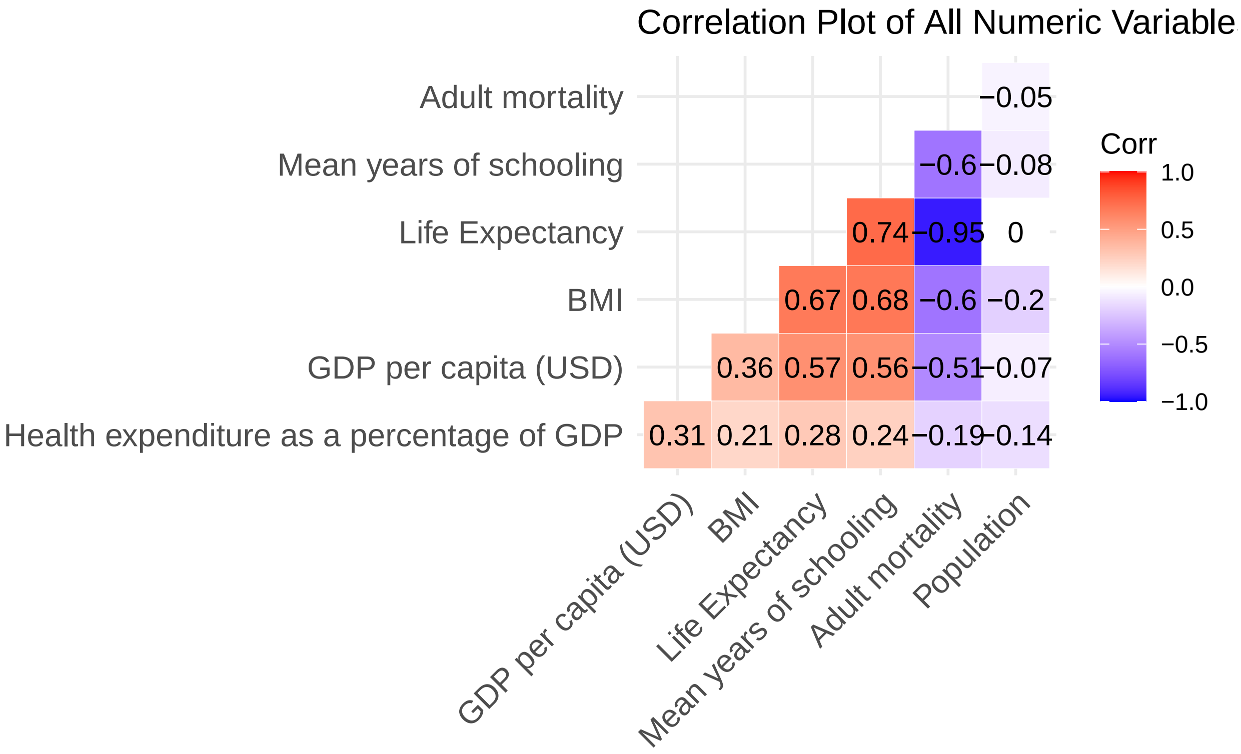
```r
cor_chart <- cor(num_data, method = "pearson", use = "complete.obs")
round(cor_chart, 2)
```

```
##                                             BMI Adult mortality
## BMI                                        1.00           -0.60
## Adult mortality                           -0.60            1.00
## Health expenditure as a percentage of GDP  0.21           -0.19
## Life Expectancy                            0.67           -0.95
## Population                                -0.20           -0.05
## Mean years of schooling                    0.68           -0.60
## GDP per capita (USD)                       0.36           -0.51
##                                           Health expenditure as a percentage of GDP
## BMI                                                                            0.21
## Adult mortality                                                               -0.19
## Health expenditure as a percentage of GDP                                      1.00
## Life Expectancy                                                                0.28
## Population                                                                    -0.14
## Mean years of schooling                                                        0.24
## GDP per capita (USD)                                                           0.31
##                                           Life Expectancy Population
## BMI                                                  0.67      -0.20
## Adult mortality                                     -0.95      -0.05
## Health expenditure as a percentage of GDP            0.28      -0.14
## Life Expectancy                                      1.00       0.00
## Population                                           0.00       1.00
## Mean years of schooling                              0.74      -0.08
## GDP per capita (USD)                                 0.57      -0.07
##                                           Mean years of schooling
## BMI                                                          0.68
## Adult mortality                                             -0.60
## Health expenditure as a percentage of GDP                    0.24
## Life Expectancy                                              0.74
## Population                                                  -0.08
## Mean years of schooling                                      1.00
## GDP per capita (USD)                                         0.56
##                                           GDP per capita (USD)
## BMI                                                       0.36
## Adult mortality                                          -0.51
## Health expenditure as a percentage of GDP                 0.31
## Life Expectancy                                           0.57
## Population                                               -0.07
## Mean years of schooling                                   0.56
## GDP per capita (USD)                                      1.00
```

We can also visually demonstrate the correlations in a correlation plot, by combining data for all countries and all years:

FIGURE 1 - Correlation Plot of All Numeric Variables

```
ggcorrplot(cor_chart, hc.order = TRUE, type = "lower",
           outline.col = "white",
           lab = TRUE,
           title = "Correlation Plot of All Numeric Variables")
```



From the above plot, it's clear that the most positively correlated measurement with life expectancy is mean years of schooling, with a correlation coefficient of **0.74** (The highest absolute correlation with life expectancy, adult mortality, is expected, as the two measurements are slightly different ways of measuring the same outcome).

Within our data, the countries have been stratified by their deveopment status (i.e. developing versus developed). There are many factors that go in to how these classifications are made (United Nations, 2020), but we're interested in seeing the relationship between development status and life expectancy:

Life Expectancy and Development Status

TABLE 2 - Development status

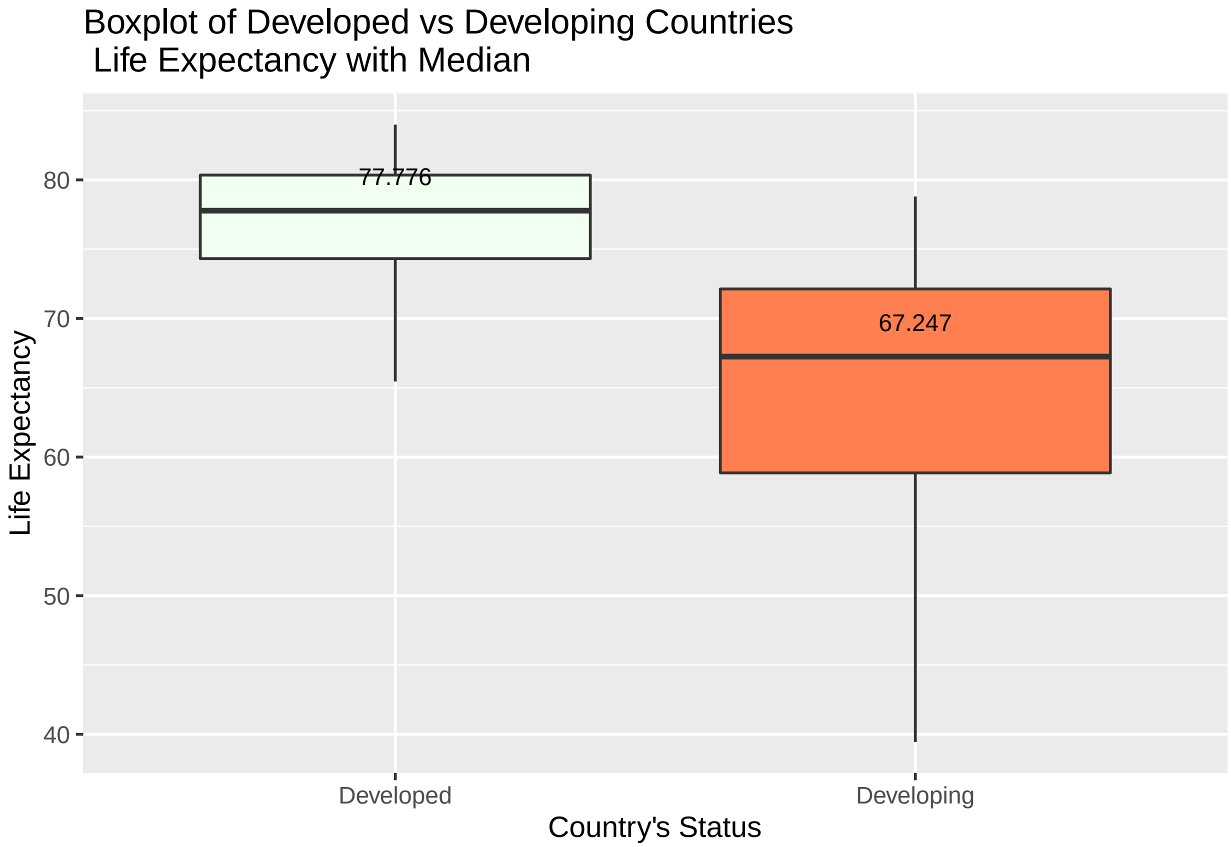
```
# Split data by development status
splitted <- master_data %>%
  select(`Life Expectancy`, Status,`Mean years of schooling`)

splitted = na.omit(splitted)

detach("package:plyr", unload = TRUE) ##have to unload or summarize won't work

# Show the percentage of data contribution from Developing vs Developed
splitted  %>%
  group_by(Status) %>%
  summarize(count = n()) %>%
  mutate(percentage = paste0(round(count/sum(count)*100, 2), "%"))
```

```
## # A tibble: 2 x 3
##   Status     count percentage
##   <chr>      <int> <chr>
## 1 Developed    951 35.64%
## 2 Developing  1717 64.36%
```

To focus on the comparison of developed vs developing countries summary, we have chosen to omit the NA values in the “splitted”" dataframe. This dataframe groups the countries by development staus and includes the two highly correlated variables life expectancy and mean years of schooling. Above is the count of listed developing countries versus developed countries, where the developing count is significantly higher than developed countries even after dropping NA values

FIGURE 2 - Boxplot displaying Median of Developing vs Developed Country Status and Life Expectancy

```
library(plyr)
# Plot box plots of the life expectancy data for developed/developing countries
p_meds <- ddply(splitted, .(Status), summarise, med = median(`Life Expectancy`))


plot1 <-  ggplot(splitted, aes(x=Status, y =`Life Expectancy` , fill = Status)) +
                geom_boxplot() +
                scale_fill_manual(values=c("honeydew", "coral")) +
                labs(x = "Country's Status", y = "Life Expectancy") +
                theme(legend.position = "none") +
    geom_text (data = p_meds, aes(x = Status, y = med, label = med), size = 3, vjust = -1.5)  + ggtitle


# show the plots
plot1
```



Based off the Median line on developed countries in the light green boxplot at 77.8, it is clear that developed countries have a much higher mean, median, and inter-qaurtile range for life expectancy.

Analysis of Variance

Test for estimating life expectancy variable changes according a country's development status. This will test whether there is a difference in means of the groups at each level of the independent variable. Test at signficance level of $\alpha = 0.05$.

H_0: There is no difference in a country's development status and its corresponding mean life expectancy
H_a: There is a difference in a country's development status and its corresponding mean life expectancy

TABLE 3 - ANOVA of Life Expectancy and Development Status

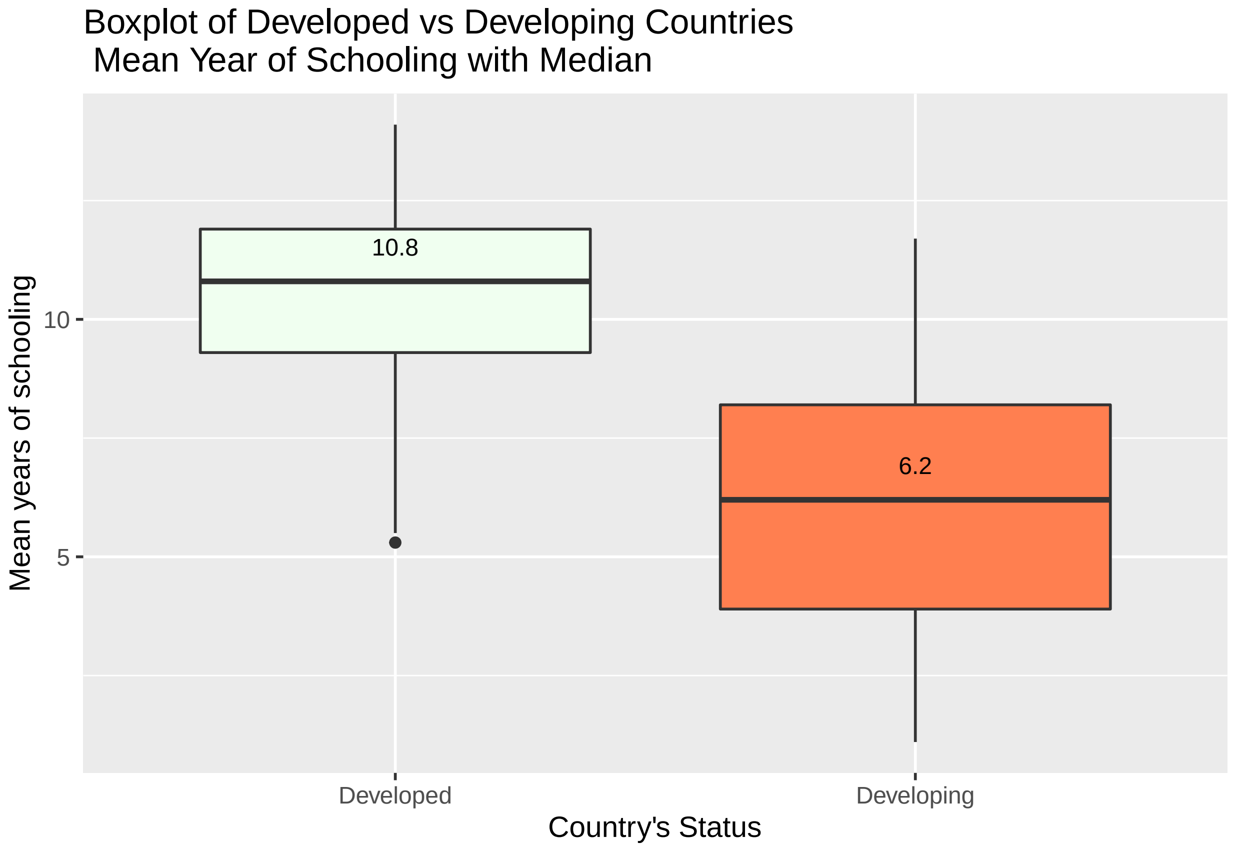
```
# average varaince summary
summary(aov(`Life Expectancy` ~ Status, data = splitted ))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## Status          1  89637   89637    1712 <2e-16 ***
## Residuals    2666 139614      52
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The resulting p-value is at $p < 0.001$ from the ANOVA test, which is vanishingly small and far less than the significance level of $\alpha = 0.05$. Therefore, we can reject null hypothesis and conclude that there is a significant difference in the life expectancy of developed versus developing countries.

FIGURE 3 - Box Plots of Mean Years of Schooling and Development Status

```
q_meds <- ddply(splitted, .(Status), summarise, med = median(`Mean years of schooling`))


plot2 <-  ggplot(splitted, aes(x=Status, y =`Mean years of schooling`  ,  fill = Status)) +
                geom_boxplot() +
                scale_fill_manual(values=c("honey dew", "coral")) +
                labs(x = "Country's Status", y = "Mean years of schooling") +
                theme(legend.position = "none") +
    geom_text (data = q_meds, aes(x = Status, y = med, label = med), size = 3, vjust = -1.5) + ggtitle(


# show the plots
plot2
```



The above box plots show a stark difference in mean years of schooling between developed countries and developing countries, with the median years of schooling being over 4.5 years higher in developed countries (6.2 years versus 10.8 years). Similar to the life expectancy box plots, the mean schooling data shows that even developed countries in the 25th percentile receive more education (on average) than developing countries in the 75th percentile.

TABLE 4 - ANOVA of Mean Years of Schooling and Development Status

H_0: There is no difference in a country's development status and its mean mean years of schhooling
H_a: There is a difference in a country's development status and its mean years of schooling

```
summary(aov(`Mean years of schooling` ~ Status, data = splitted ))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)    
## Status          1  11733   11733    2043 <2e-16 ***
## Residuals    2666  15312       6                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As seen above, the ANOVA p-value of $p < 0.001$ is far smaller than the significance level of $\alpha = 0.05$, and we can reject null hypothesis and conclude that there is a significant difference in mean years of schooling in developed versus developing countries.

What is the significance of these tests?

From these simple tests, we want to prove that:
- 1)From the corrrealation plot, the amount of Schooling shows a positive correaltion to Life Expectancy, meaning that a higher education suggests a higher mortality.
- 2) We now want to see the difference of Life Expectancy between a Developing Country and a Developed County, in other words, we identified that generally speaking Developed country have a longer lifespan.
- 3) Then, if point two is true, then the average mean years of schooling must also be higher for Developed countries, and with no surprises, we proved that Developed countries recieves a higher education as well.

Linear Regression - Life Expectancy in Developed vs Developing Countries

In order to truly drive home the relationship between life expectancy and schooling from our dataset, we'll build a linear regression model:

$$LifeExpectancy_{World,i} = \alpha + \beta * Schooling_{World,i} + e_i$$

```
# Prepare data

splitted2 <- master_data %>%
  select( Country, Year, `Life Expectancy`, Status,`Mean years of schooling`)

splitted2 = na.omit(splitted2)

head(splitted2, 10)
```

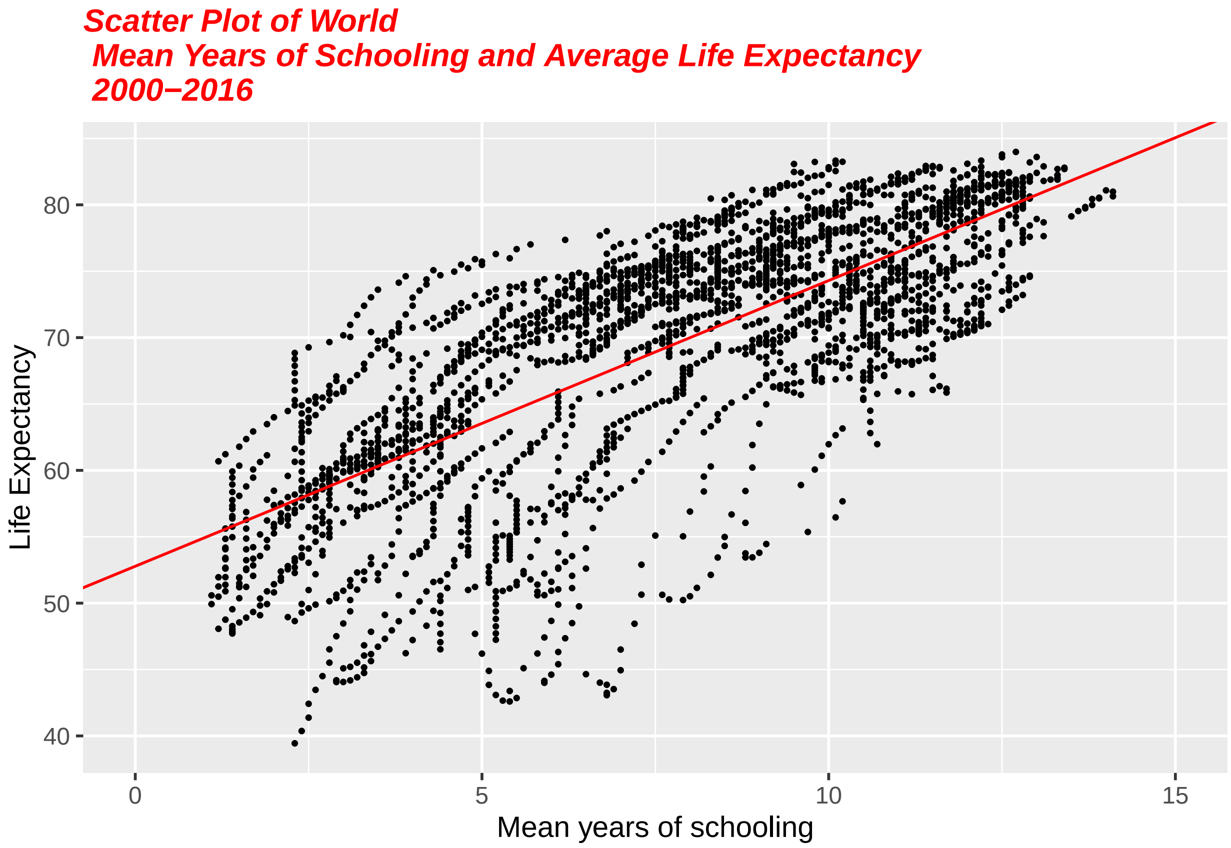
```
##        Country Year Life Expectancy     Status Mean years of schooling
## 1  Afghanistan 2000          55.841 Developing                     2.2
## 2  Afghanistan 2001          56.308 Developing                     2.2
## 3  Afghanistan 2002          56.784 Developing                     2.3
## 4  Afghanistan 2003          57.271 Developing                     2.4
## 5  Afghanistan 2004          57.772 Developing                     2.5
## 6  Afghanistan 2005          58.290 Developing                     2.6
## 7  Afghanistan 2006          58.826 Developing                     2.7
## 8  Afghanistan 2007          59.375 Developing                     2.9
## 9  Afghanistan 2008          59.930 Developing                     3.0
## 10 Afghanistan 2009          60.484 Developing                     3.1
```

World Linear regression model predicting Life Expectancy(y) by Mean years of Schooling(x)

```
# Make plots and the regression line

world_fit = lm(`Life Expectancy`~ `Mean years of schooling`, data = splitted2)
print(world_fit)
```

```
##
## Call:
## lm(formula = ‘Life Expectancy‘ ~ ‘Mean years of schooling‘, data = splitted2)
##
## Coefficients:
##                (Intercept)  ‘Mean years of schooling‘
##                     52.778                      2.152
```

```
ggplot(splitted2, aes(`Mean years of schooling`, `Life Expectancy`))  +
  geom_point(size = 0.5) +
  xlim(0, 15) +
  ggtitle("Scatter Plot of World \n Mean Years of Schooling and Average Life Expectancy \n 2000-2016")
  theme (
plot.title = element_text(color="red",
                          size=12,
                          face="bold.italic")) + geom_abline(slope = coef(world_fit)[[2]], intercept =
```



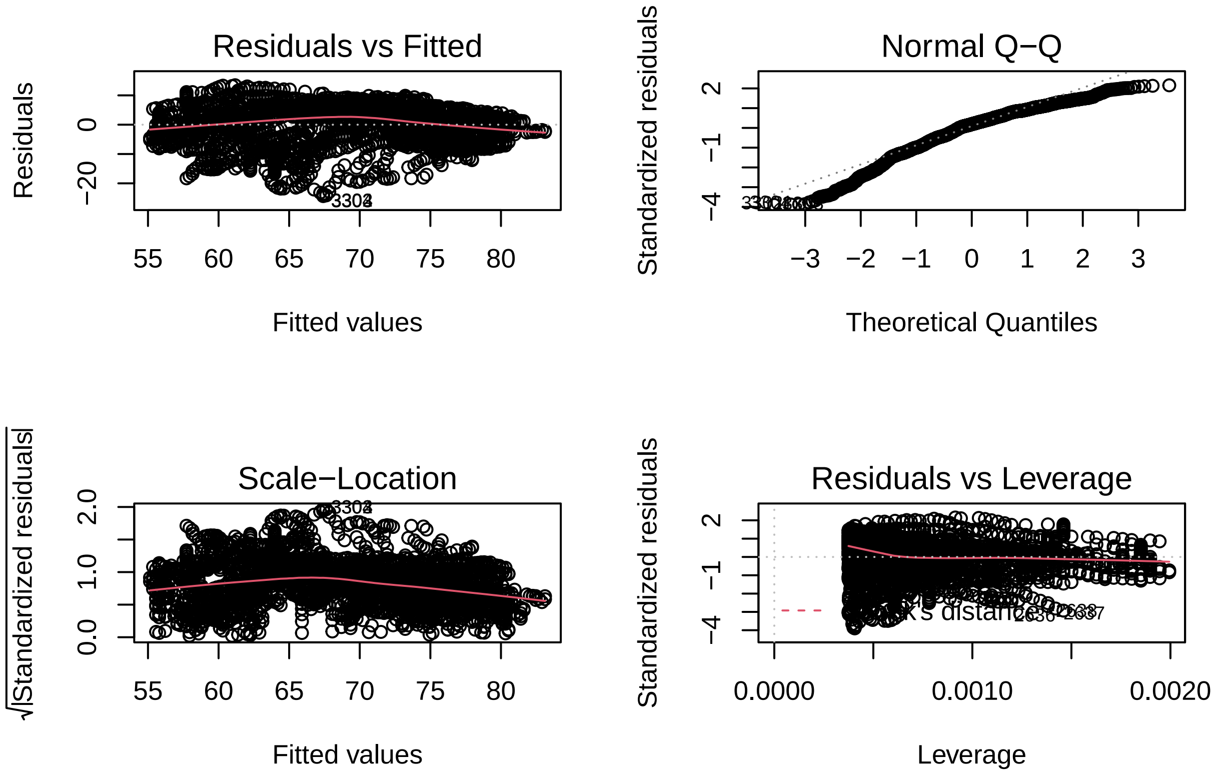
From the scatter plot and linear regression line we see a increase in life expectancy across the world as the level of education (as mean years of schooling) increases. Although there are few outliers, none seem to be high leverage, so this seems to be an appropriate fit for this model.

All the same, we'll run model diagnostics to verify it:

```
summary(world_fit)
```

```
##
## Call:
## lm(formula = ‘Life Expectancy‘ ~ ‘Mean years of schooling‘, data = splitted2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.346  -3.527   1.196   4.650  13.451
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               52.77798    0.31709  166.44   <2e-16 ***
## ‘Mean years of schooling‘  2.15188    0.03798   56.66   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.246 on 2666 degrees of freedom
## Multiple R-squared:  0.5463, Adjusted R-squared:  0.5461
## F-statistic:  3210 on 1 and 2666 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))  # set up a 2 rows and 2 columns plotting environment

plot(world_fit)
```



The top left, and bottom left plots check the linearity assumption and the equal variance assumption. When these assumptions are met, we expect the plotted points evenly distributed in a rectangular region. We see a very small deviation from horizontal, but nothing of major concern.

The top right plot checks the normality assumption. If the assumption is met, we expect to see a straight line pattern. We see very minimal departures at the lower and upper tails, but overall a relatively straight line.

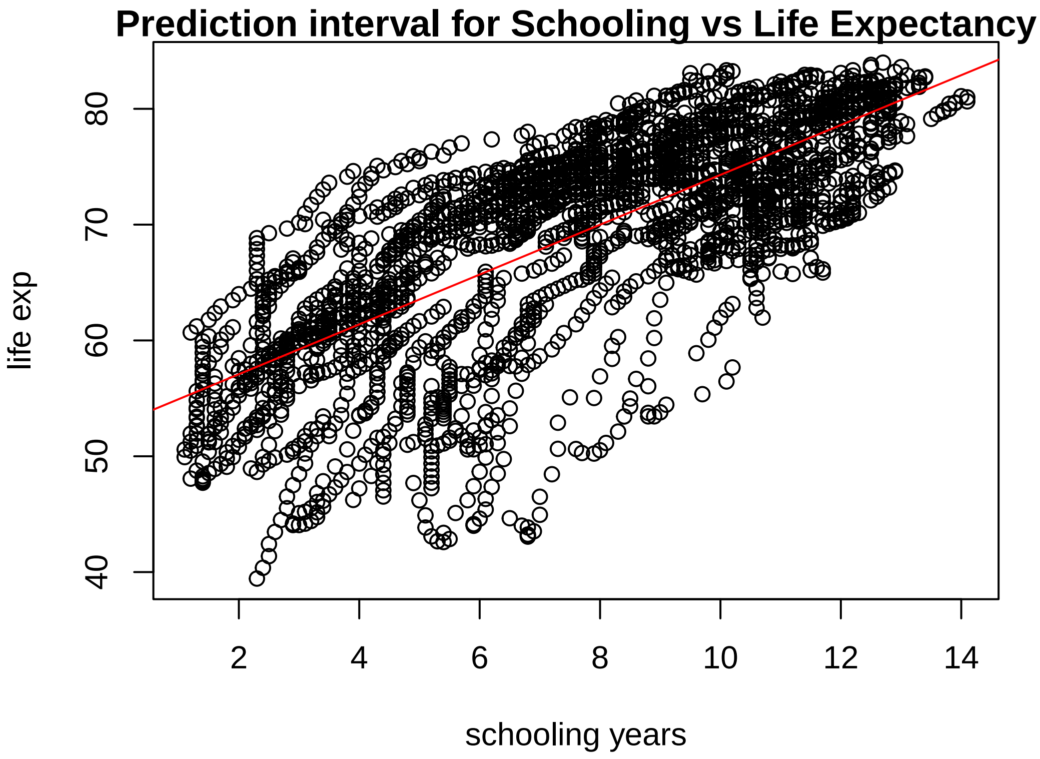
The bottom right plot checks for influential data points, of which none seem to be too worrisome.

Regarding the test statistics, the overall F-test (F-statistic $=$ **3210** on 1 and 2666 DoF, $p < 0.0001$ and the individual t tests (p-value for intercept is $p < 0.0001$, p-value for slope is $p < 0.0001$) all suggest that the fitted model is highly statistically significant. The coefficient of determination is $R^2 = 0.5463$.

Prediction intervals

Given a random country with 12 mean years of schooling, can we predict their life expectancy?

```
# vars definition
df = splitted2
n = nrow(splitted2)

S_xx = sum((df$`Mean years of schooling`-mean(df$`Mean years of schooling`))^2)
S_xy = sum((df$`Mean years of schooling`-mean(df$`Mean years of schooling`))*(df$`Life Expectancy`-mean
S_yy = sum((df$`Life Expectancy`-mean(df$`Life Expectancy`))^2)

b1 = S_xy / S_xx
b0 = mean(df$`Life Expectancy`)-mean(df$`Mean years of schooling`)*b1
sse = S_yy - S_xy^2/S_xx
s = sqrt(sse/(n-2))
r = S_xy/sqrt(S_xx*S_yy)

# plotting the data
par(oma=c(1.5,1.5,1.5,1.5)); par(mar=c(4,5.5,1,2))
plot(df$`Mean years of schooling`, df$`Life Expectancy`,
               ylab = "life exp",
               xlab = "schooling years",
               main="Prediction interval for Schooling vs Life Expectancy")
text(25, 40, col="red", "y = b0 + b1 x")
lines(c(0, 250), c(b0,b0+b1*250), col="red")
```



```
cat("b0 =",round(b0,4),"\tb1 =",round(b1,4))
```

```
b0 = 52.778     b1 = 2.1519
```

```
# prediction interval
xp = 12
x = round(mean(df$`Mean years of schooling`),4)
a = 0.05
t = qt(1-a/2, n-2)
est = b0 + b1*xp
c = t*s*sqrt(1 + (1/n) + (xp-mean(df$`Mean years of schooling`))^2/S_xx)
lo = est - c
up = est + c

cat(" Resulting interval is (", round(lo,4), ",", round(up,4), "),
    this is the 95% prediction interval \n at 12 years of schooling.")
```

```
 Resulting interval is ( 66.3462 , 90.855 ),
    this is the 95% prediction interval
 at 12 years of schooling.
```

Multilinear Regression: Developed vs Developing

Since we can clearly identify a strong linear relationship between mean years of schooling across all countries using the linear model world_fit, let's dive into adding GDP per capita(USD) to predict the life expectancy of a developing versus developed country along with the mean years of schooling.

```
# Prep the data for Developed country

status_df <- master_data %>%
  select(`Life Expectancy`, Status,`Mean years of schooling`, `GDP per capita (USD)`)

status_df = na.omit(status_df)

# Split the date in two groups

status_df = split(status_df, status_df$Status)

# Prepare variables by development status

developed_life = status_df$Developed$`Life Expectancy`
developed_school = status_df$Developed$`Mean years of schooling`
developed_gdp = status_df$Developed$`GDP per capita (USD)`

# Make the multiple regression model

multi_model <- lm(developed_life ~ developed_school + developed_gdp , data = status_df$Developed)

summary(multi_model)
```

```
##
## Call:
## lm(formula = developed_life ~ developed_school + developed_gdp,
##     data = status_df$Developed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0239 -1.8808  0.2356  2.0406  6.0925
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.289e+01  5.446e-01 133.854   <2e-16 ***
## developed_school 1.337e-01  5.430e-02   2.462    0.014 *
## developed_gdp    1.129e-04  4.421e-06  25.540   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.687 on 931 degrees of freedom
## Multiple R-squared:  0.4715, Adjusted R-squared:  0.4703
## F-statistic: 415.2 on 2 and 931 DF,  p-value: < 2.2e-16
```

```
summary(multi_model)$coefficient
```

```
##                      Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)      7.289383e+01 5.445762e-01 133.854228  0.000000e+00
## developed_school 1.336983e-01 5.429738e-02   2.462333  1.398382e-02
## developed_gdp    1.129044e-04 4.420626e-06  25.540373 1.808352e-109
```

The overall F-test (F-statistic: 415.2 on 2 and 931 DF, $p < 0.0001$) and the individual t tests (p-value for intercept is $p < 0.0001$, p-value for slope is $p < 0.0001$) all suggest that the fitted model is highly statistically significant. The coefficient of determination is $R^2 = 0.4715$ using both mean year of schooling and GDP per capita to estimate life expectancy for developed countries. This means 47.15% of the variance in the measure of developed country’s life expectancy can be predicted by developed country’s mean years of schooling and GDP per capita, which is quite high.

Using the data as an example: the estimate GDP per capita is 1.129e-04, meaning an increase in average GDP of $20,000 adds 2.26 years to life expectancy.

From the statistical findings above we now build our model equation as follows:

$$LifeExpectancy_{Developed,i} = 72.8938 + 0.1337 * Schooling_{Developed,i} + 0.0001 * GDPperCapDeveloped, i$$

```
# The confidence interval of the model coefficient

confint(multi_model)
```

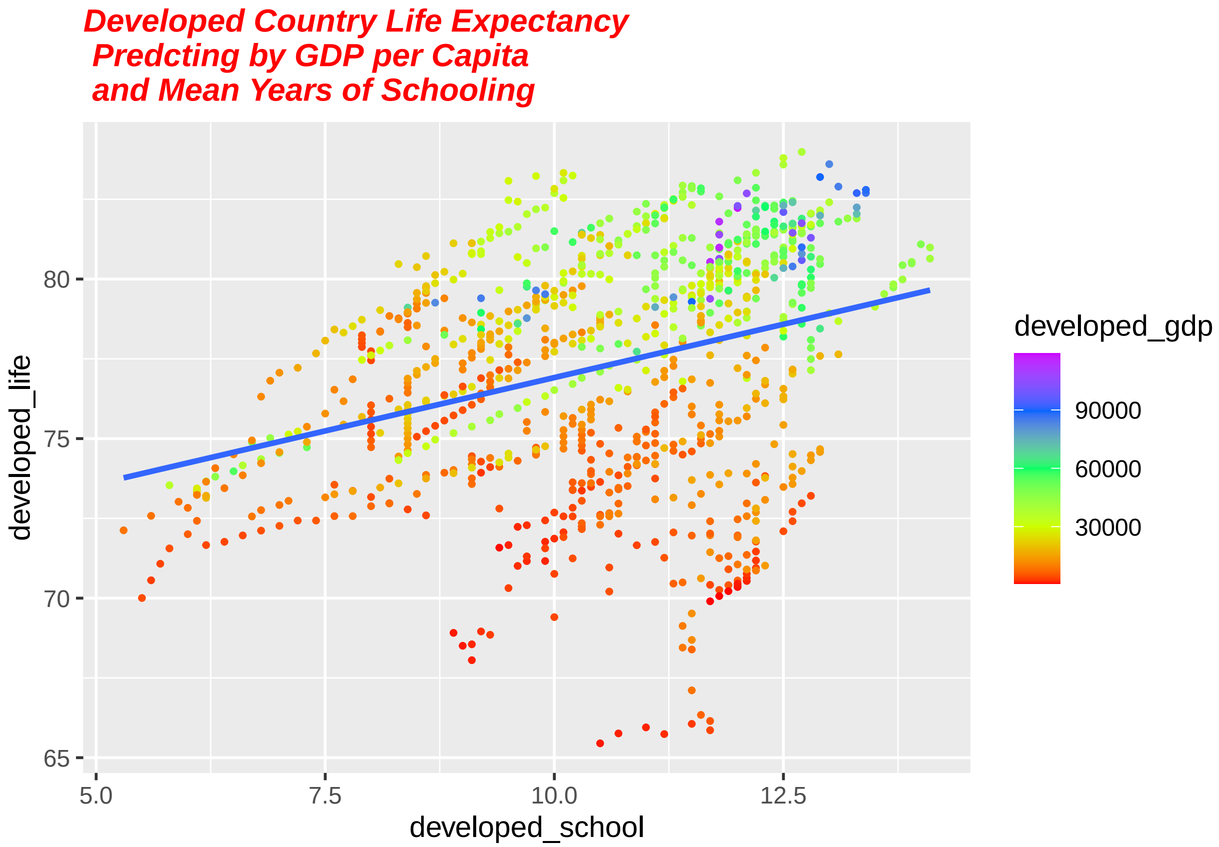
```
##                         2.5 %      97.5 %
## (Intercept)      7.182509e+01 73.96256650
## developed_school 2.713881e-02  0.24025770
## developed_gdp    1.042289e-04  0.00012158
```

```
#Residual Standard Error (RSE)

cat("The standard error is:",
    round(sigma(multi_model)/mean(status_df$Developed$`Life Expectancy`), 4),
    "which proves this model accuracy is reliable." )
```

```
## The standard error is: 0.0348 which proves this model accuracy is reliable.
```

```
# plot the model

ggplot(status_df$Developed,aes(y=developed_life,x=developed_school,color=developed_gdp))+geom_point(size
plot.title = element_text(color="red", size=12, face="bold.italic")) + scale_color_gradientn(colours = r
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Above is a representation of the multiple regression model, using colour as a scale to show GDP per capita.

```
# prep the data for Developing country

developing_life = status_df$Developing$`Life Expectancy`
developing_school = status_df$Developing$`Mean years of schooling`
developing_gdp = status_df$Developing$`GDP per capita (USD)`


#make the multilinear regression model

multi_model2 <- lm(developing_life ~ developing_school + developing_gdp, data = status_df$Developing)

summary(multi_model2)
```

```
##
## Call:
## lm(formula = developing_life ~ developing_school + developing_gdp,
##     data = status_df$Developing)
##
## Residuals:
##     Min       1Q  Median      3Q     Max
## -21.670  -3.474   1.018   5.018  12.555
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.358e+01  4.136e-01 129.533   <2e-16 ***
## developing_school 1.602e+00  6.951e-02  23.040   <2e-16 ***
## developing_gdp    5.511e-04  5.683e-05   9.697   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.717 on 1688 degrees of freedom
## Multiple R-squared:  0.3931, Adjusted R-squared:  0.3924
## F-statistic: 546.7 on 2 and 1688 DF,  p-value: < 2.2e-16
```

```
summary(multi_model2)$coefficient
```

```
##                        Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)       5.357838e+01 4.136270e-01 129.533070  0.000000e+00
## developing_school 1.601586e+00 6.951333e-02  23.039991 2.342996e-102
## developing_gdp    5.510592e-04 5.682789e-05   9.696984  1.127923e-21
```

```
cat("The overall F-test ( F-statistic: 546.7 on 2 and 1688 DF,  p-value: < 2.2e-16) and the individual *
```

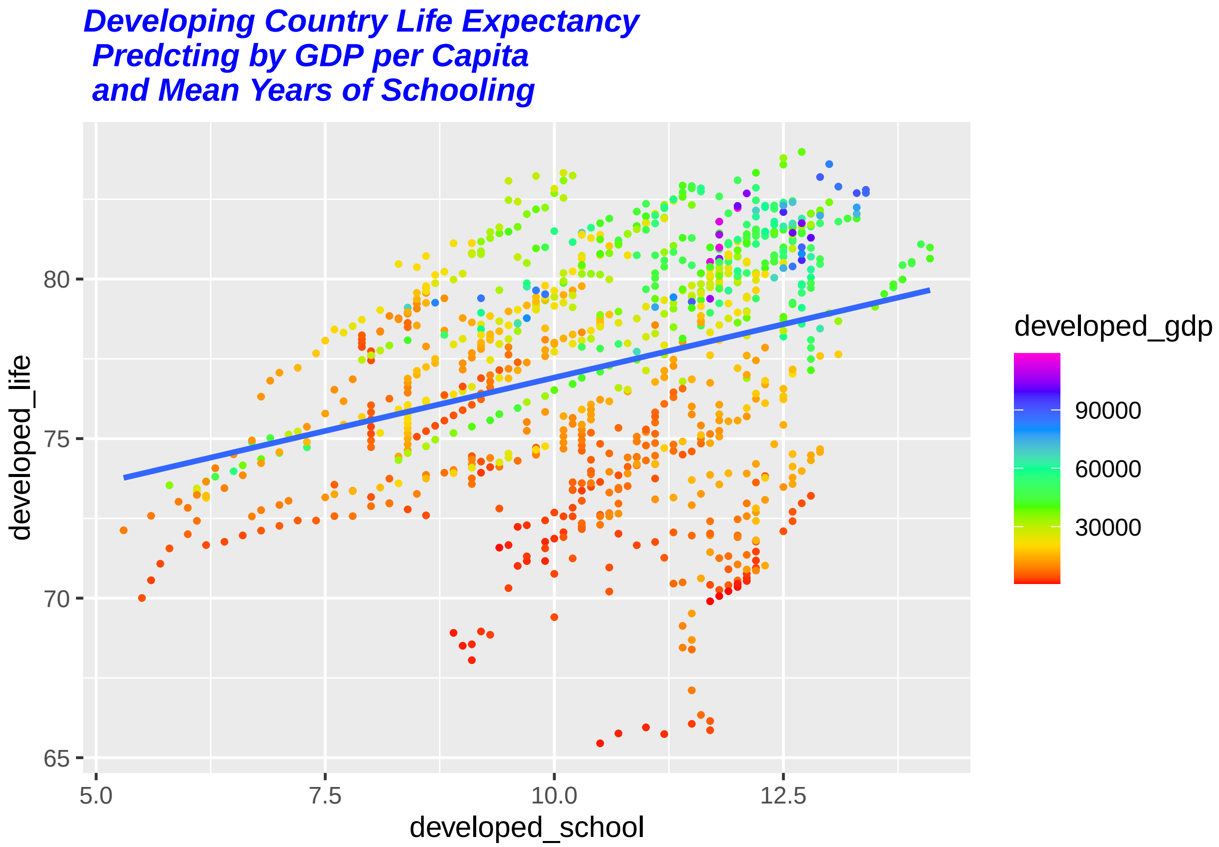
```
## The overall F-test ( F-statistic: 546.7 on 2 and 1688 DF,  p-value: < 2.2e-16) and the individual *t*
```

From the statistical findings above we now build our model equation as follows:

$$LifeExpectancy_{Developing,i} = 53.5783 + 1.601 * Schooling_{Developing,i} + 0.0006 * GDPperCapDeveloping, i$$

```
ggplot(status_df$Developed,aes(y=developed_life,x=developed_school,color=developed_gdp))+geom_point(size
plot.title = element_text(color="blue", size=12, face="bold.italic")) + scale_color_gradientn(colours =
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Conclusion

It would appear that there indeed a very strong link between GDP, schooling, and life expectancy. The interaction seems to be quite intricate, as all three variables seem to increase (or decrease) hand-in-hand. As these are some of the main conditions that make a country developed versus developing, it should come as no surprise that there’s a stark difference between the two statuses of countries.

References

Bhaskaran, K., dos-Santos-Silva, I., Leon, D. A., Douglas, I. J., & Smeeth ,L. (2018). Association of BMI with overall and cause specific mortality: a population-based cohort study of 3 · 6 million adults in the UK. *The Lancet Diabetes & Endocrinology, (6)* 12, 944-953, https://doi.org/10.1016/S2213-8587(18)30288-2

Global Health Observatory Data Repository (2018), *Global Health Observatory indicator views*, https://apps.who.int/gho/data/node.imr. Retrieved September 21st, 2021

Rajarshi, K. (2018, February). Life Expectancy (WHO): Statistical Analysis on factors influencing Life Expectancy, Version 1. Retrieved September 21, 2021 from https://www.kaggle.com/kumarajarshi/life-expectancy-who/version/1.

Stiefel, M. C., Perla, R. J., & Zell, B. L. (2010). A healthy bottom line: healthy life expectancy as an outcome measure for health improvement efforts. *The Milbank quarterly, 88* (1), 30–53. https://doi.org/10.1111/j.1468-0009.2010.00588.x

United Nations (2020), *Human Development Report Office*, http://hdr.undp.org/en/data. Retrieved October 6th, 2021

World Bank (2021, September), *World Development Indicators*, https://datacatalog.worldbank.org/search/dataset/0037712. Retrieved October 6th, 2021

Appendix A - Data Wrangling

```r
# Read in the adult mortality data, filter data only for both sexes, remove columns
adult_mortality = read.csv("Adult Mortality.csv")
adult_mortality = adult_mortality[adult_mortality$Sex == "Both sexes",]
adult_mortality = adult_mortality[c("Year", "Country", "Numeric")]
colnames(adult_mortality)[3] = "Adult mortality"
```

```r
# Read in BMI data, filter data only for both sexes, remove unneeded years and columns
bmi = read.csv("BMI.csv")
bmi = bmi[bmi$Sex == "Both sexes",]
bmi = bmi[bmi$Year >= 2000, ]
bmi = bmi[c("Year", "Country", "Numeric")]
colnames(bmi)[3] = "BMI"
```

```r
# Read in HDI data. Determine developed vs developing based on HDI. Rename columns
country_status = read.csv("Country Status.csv")
country_status = transform(country_status, Status=ifelse(hdi2019 >= 0.800, "Developed", "Developing"))
country_status = country_status[c("ï..country", "Status")]
colnames(country_status)[1] = "Country"
# Change country names to allow dataframe merging later
country_status$Country =  gsub("Bolivia", "Bolivia (Plurinational State of)", country_status$Country)
country_status$Country = gsub("Czech Republic", "Czechia", country_status$Country)
country_status$Country = gsub("Egypt, Arab Rep.", "Egypt", country_status$Country)
country_status$Country = gsub("Iran, Islamic Rep.", "Iran (Islamic Republic of)", country_status$Countr
country_status$Country = gsub("South Korea", "Republic of Korea", country_status$Country)
country_status$Country = gsub("Kyrgyz Republic", "Kyrgyzstan", country_status$Country)
country_status$Country = gsub("Lao PDR", "Lao People's Democratic Republic", country_status$Country)
country_status$Country = gsub("Micronesia, Fed. Sts.", "Micronesia (Federated States of)", country_statu
country_status$Country = gsub("Moldova", "Republic of Moldova", country_status$Country)
```

```r
country_status$Country = gsub("St. Kitts and Nevis", "Saint Kitts and Nevis", country_status$Country)
country_status$Country = gsub("St. Lucia", "Saint Lucia", country_status$Country)
country_status$Country = gsub("St. Vincent and the Grenadines", "Saint Vincent and the Grenadines", coun
country_status$Country = gsub("Slovak Republic", "Slovakia", country_status$Country)
country_status$Country = gsub("Tanzania (United Republic of)", "United Republic of Tanzania", country_st
country_status$Country = gsub("United Kingdom", "United Kingdom of Great Britain and Northern Ireland",
country_status$Country = gsub("United States", "United States of America", country_status$Country)
country_status$Country = gsub("Venezuela, RB", "Venezuela (Bolivarian Republic of)", country_status$Cou
country_status$Country = gsub("Vietnam", "Viet Nam", country_status$Country)
country_status$Country = gsub("Yemen, Rep.", "Yemen", country_status$Country)
```

```r
# Read in health expenditure data, drop and rename columns
health_exp = read.csv("Expenditure on health as a percentage of Gross Domestic Product per capita(%).cs
health_exp = health_exp[c("Year", "Country", "Numeric")]
colnames(health_exp)[3] = "Health expenditure as a percentage of GDP"
```

```r
# Read in GDP data. Remove empty columns, rename remaining columns
gdp = read.csv("GDP.csv", header=FALSE, skip=4)
gdp = gdp[c(1, 45:65)]
colnames(gdp) = gdp[1, ]
colnames(gdp)[1] = "Country"
gdp = pivot_longer(gdp, !"Country", names_to="Year", values_to="GDP per capita (USD)")
gdp = gdp[gdp$"Country" != "Country Name",]
# Change country names to allow for dataframe merging later
gdp$Country =  gsub("Bolivia", "Bolivia (Plurinational State of)", gdp$Country)
gdp$Country = gsub("Czech Republic", "Czechia", gdp$Country)
gdp$Country = gsub("Egypt, Arab Rep.", "Egypt", gdp$Country)
gdp$Country = gsub("Iran, Islamic Rep.", "Iran (Islamic Republic of)", gdp$Country)
gdp$Country = gsub("Korea, Rep.", "Republic of Korea", gdp$Country)
gdp$Country = gsub("Kyrgyz Republic", "Kyrgyzstan", gdp$Country)
gdp$Country = gsub("Lao PDR", "Lao People's Democratic Republic", gdp$Country)
gdp$Country = gsub("Micronesia, Fed. Sts.", "Micronesia (Federated States of)", gdp$Country)
gdp$Country = gsub("Moldova", "Republic of Moldova", gdp$Country)
gdp$Country = gsub("St. Kitts and Nevis", "Saint Kitts and Nevis", gdp$Country)
gdp$Country = gsub("St. Lucia", "Saint Lucia", gdp$Country)
gdp$Country = gsub("St. Vincent and the Grenadines", "Saint Vincent and the Grenadines", gdp$Country)
gdp$Country = gsub("Slovak Republic", "Slovakia", gdp$Country)
gdp$Country = gsub("Tanzania (United Republic of)", "United Republic of Tanzania", gdp$Country)
gdp$Country = gsub("United Kingdom", "United Kingdom of Great Britain and Northern Ireland", gdp$Countr
gdp$Country = gsub("United States", "United States of America", gdp$Country)
gdp$Country = gsub("Venezuela, RB", "Venezuela (Bolivarian Republic of)", gdp$Country)
gdp$Country = gsub("Vietnam", "Viet Nam", gdp$Country)
gdp$Country = gsub("Yemen, Rep.", "Yemen", gdp$Country)
```

```r
# Read in life expectancy data, rename columns
life_expectancy = read.csv("life-expectancy-at-birth-total-years.csv")
colnames(life_expectancy) = c("Country", "Code", "Year", "Life Expectancy")
life_expectancy = life_expectancy[life_expectancy$Year >= 2000, ]
life_expectancy = life_expectancy[c("Country", "Year", "Life Expectancy")]
# Change country names to allow for dataframe merging later
life_expectancy$Country = gsub("Micronesia (country), Rep.", "Micronesia (Federated States of)", life_e
# Read in schooling data, drop unneeded columns. Trim whitespace from data
schooling = read.csv("Mean years of schooling (years).csv",header=FALSE, skip=5, nrows=191)
schooling = schooling[c(2, 23, 25, 27, 29, 31, 33, 35, 37, 39, 41, 43, 45, 47, 49, 51, 53, 55, 57, 59, 
colnames(schooling) = c("Country", 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 20
schooling = schooling[-1, ]
schooling = pivot_longer(schooling, !"Country", names_to="Year", values_to="Mean years of schooling")
schooling$Country = trimws(schooling$Country, which="both")
schooling$Year = trimws(schooling$Year, which="both")
schooling$`Mean years of schooling` = trimws(schooling$`Mean years of schooling`, which="both")
# Change country names to allow for dataframe merging later
schooling$Country = gsub("Egypt, Arab Rep.", "Egypt", schooling$Country)
schooling$Country = gsub("Iran, Islamic Rep.", "Iran (Islamic Republic of)", schooling$Country)
schooling$Country = gsub("Korea (Republic of)", "Republic of Korea", schooling$Country)
schooling$Country = gsub("Kyrgyz Republic", "Kyrgyzstan", schooling$Country)
schooling$Country = gsub("Lao PDR", "Lao People's Democratic Republic", schooling$Country)
schooling$Country = gsub("Micronesia, Fed. Sts.", "Micronesia (Federated States of)", schooling$Country
schooling$Country = gsub("Moldova", "Republic of Moldova", schooling$Country)
schooling$Country = gsub("St. Kitts and Nevis", "Saint Kitts and Nevis", schooling$Country)
schooling$Country = gsub("St. Lucia", "Saint Lucia", schooling$Country)
schooling$Country = gsub("St. Vincent and the Grenadines", "Saint Vincent and the Grenadines", schoolin
schooling$Country = gsub("Slovak Republic", "Slovakia", schooling$Country)
schooling$Country = gsub("Tanzania (United Republic of)", "United Republic of Tanzania", schooling$Coun
schooling$Country = gsub("United Kingdom", "United Kingdom of Great Britain and Northern Ireland", scho
schooling$Country = gsub("United States", "United States of America", schooling$Country)
schooling$Country = gsub("Venezuela, RB", "Venezuela (Bolivarian Republic of)", schooling$Country)
schooling$Country = gsub("Vietnam", "Viet Nam", schooling$Country)
schooling$Country = gsub("Yemen, Rep.", "Yemen", schooling$Country)
```

```r
# Read in population data, drop unneeded columns, rename remaining columns
population = read.csv("Population.csv", header=FALSE, skip=4)
population = population[c(1, 45:65)]
colnames(population) = population[1, ]
colnames(population)[1] = "Country"
population = pivot_longer(population, !"Country", names_to="Year", values_to="Population")
population = population[population$"Country" != "Country Name",]
# Change country names to allow for dataframe merging later
population$Country = gsub("Bolivia", "Bolivia (Plurinational State of)", population$Country)
population$Country = gsub("Cote d'Ivoire", "CÃ´te d'Ivoire", population$Country)
population$Country = gsub("Czech Republic", "Czechia", population$Country)
population$Country = gsub("Egypt, Arab Rep.", "Egypt", population$Country)
population$Country = gsub("Iran, Islamic Rep.", "Iran (Islamic Republic of)", population$Country)
population$Country = gsub("Korea, Rep.", "Republic of Korea", population$Country)
population$Country = gsub("Kyrgyz Republic", "Kyrgyzstan", population$Country)
population$Country = gsub("Lao PDR", "Lao People's Democratic Republic", population$Country)
population$Country = gsub("Micronesia, Fed. Sts.", "Micronesia (Federated States of)", population$Count
population$Country = gsub("Moldova", "Republic of Moldova", population$Country)
population$Country = gsub("St. Kitts and Nevis", "Saint Kitts and Nevis", population$Country)
population$Country = gsub("St. Lucia", "Saint Lucia", population$Country)
population$Country = gsub("St. Vincent and the Grenadines", "Saint Vincent and the Grenadines", populat
population$Country = gsub("Slovak Republic", "Slovakia", population$Country)
population$Country = gsub("Tanzania", "United Republic of Tanzania", population$Country)
population$Country = gsub("United Kingdom", "United Kingdom of Great Britain and Northern Ireland", popu
population$Country = gsub("United States", "United States of America", population$Country)
```

```
population$Country = gsub("Venezuela, RB", "Venezuela (Bolivarian Republic of)", population$Country)
population$Country = gsub("Vietnam", "Viet Nam", population$Country)
population$Country = gsub("Yemen, Rep.", "Yemen", population$Country)
```

```
# Create master dataframe from individual datasets through left joins
master_data = merge(bmi,adult_mortality, by=c("Country", "Year"), all.x=TRUE)
master_data = merge(master_data,country_status, by="Country", all.x=TRUE)
master_data = merge(master_data,health_exp, by=c("Country", "Year"), all.x=TRUE)
master_data = merge(master_data,life_expectancy, by=c("Country", "Year"), all.x=TRUE)
master_data = merge(master_data,population, by=c("Country", "Year"), all.x=TRUE)
master_data = merge(master_data, schooling, by=c("Country", "Year"), all.x=TRUE)
master_data = merge(master_data,gdp, by=c("Country", "Year"), all.x=TRUE)
# Correct to numeric datatypes
master_data$`Mean years of schooling` <- as.numeric(master_data$`Mean years of schooling`)
```

```
## Warning: NAs introduced by coercion
```

```
master_data$`Year` <- as.character.Date(master_data$`Year`)

#sapply(master_data, class)

# head(master_data, 4)

colnames(master_data) #Colnames will return column names present in the dataset,df=DataFrame name
```

```
##  [1] "Country"
##  [2] "Year"
##  [3] "BMI"
##  [4] "Adult mortality"
##  [5] "Status"
##  [6] "Health expenditure as a percentage of GDP"
##  [7] "Life Expectancy"
##  [8] "Population"
##  [9] "Mean years of schooling"
## [10] "GDP per capita (USD)"
```