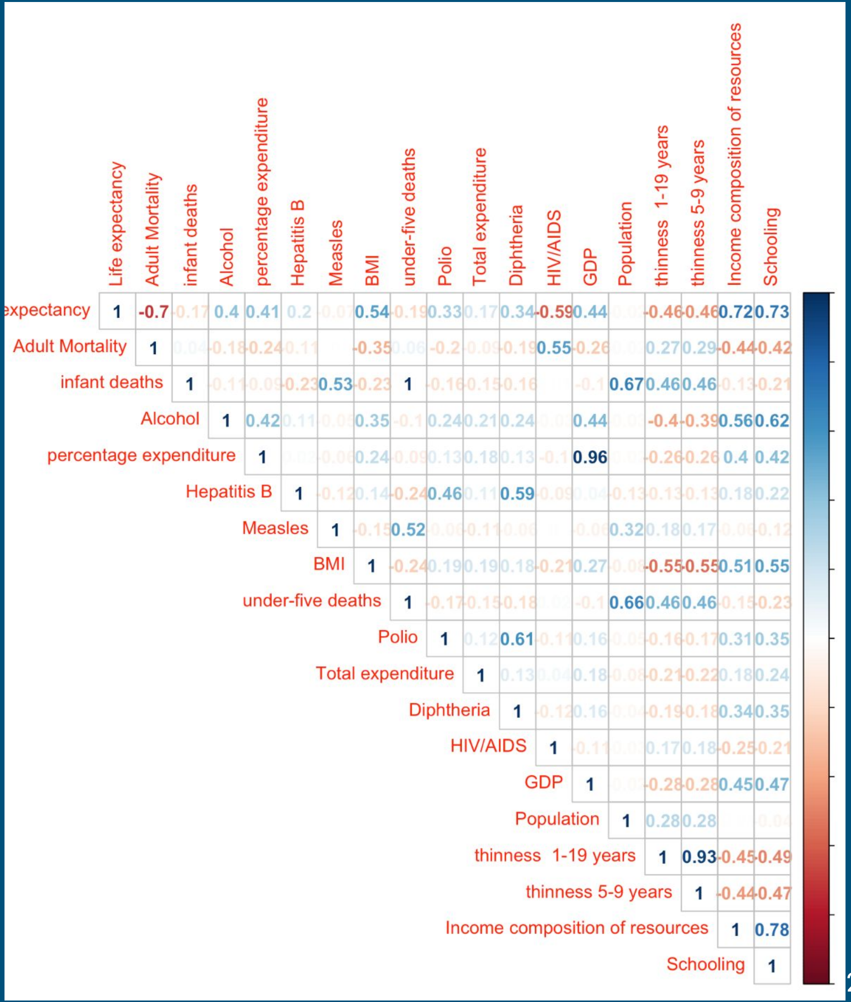
How Can Countries Increase Their Life Expectancy

Presented by : Abraca-Data
Prachi Patel, Evan Meade, Alejandro De La Cruz

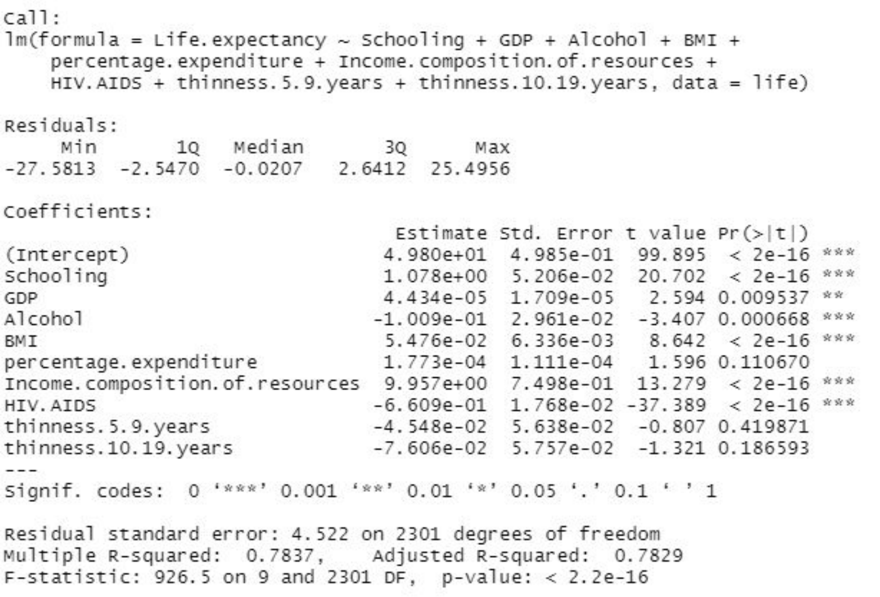
The Data

- Sourced from the **World Health Organization** (WHO)
- Contains 2,938 observations of 22 variables
  - 193 countries
  - 2000 - 2015
- Selected variables
  - **ID (3):** Year, Country, Development Status
  - **Response (1):** Life Expectancy
  - **Predictors (10):** Schooling, GDP, Alcohol, BMI, Percentage Expenditure, Income composition of resources, HIV/AIDS, Thinness 10-19 years (and 5-9 years), adult mortality
- Correlation plot (right)



	Life expectancy	Adult Mortality	infant deaths	Alcohol	percentage expenditure	Hepatitis B	Measles	BMI	under-five deaths	Polio	Total expenditure	Diphtheria	HIV/AIDS	GDP	Population	thinness 1-19 years	thinness 5-9 years	Income composition of resources	Schooling
expectancy	1	-0.7	-0.17	0.4	0.41	0.2	-0.07	0.54	-0.19	0.33	0.17	0.34	-0.59	0.44	[illegible]	-0.46	-0.46	0.72	0.73
Adult Mortality		1	0.04	-0.18	-0.24	-0.11		-0.35	0.06	-0.2	-0.09	-0.19	0.55	-0.26	[illegible]	0.27	0.29	-0.44	-0.42
infant deaths			1	-0.11	-0.09	-0.23	0.53	-0.23	1	-0.16	-0.15	-0.16		-0.1	0.67	0.46	0.46	-0.13	-0.21
Alcohol				1	0.42	0.11	-0.05	0.35	-0.1	0.24	0.21	0.24	[illegible]	0.44	[illegible]	-0.4	-0.39	0.56	0.62
percentage expenditure					1	[illegible]	-0.06	0.24	-0.09	0.13	0.18	0.13	-0.1	0.96	[illegible]	-0.26	-0.26	0.4	0.42
Hepatitis B						1	-0.12	0.14	-0.24	0.46	0.11	0.59	-0.09	0.04	-0.13	-0.13	-0.13	0.18	0.22
Measles							1	-0.15	0.52	-0.06	-0.11	-0.06		-0.06	0.32	0.18	0.17	-0.06	-0.12
BMI								1	-0.24	0.19	0.19	0.18	-0.21	0.27	-0.08	-0.55	-0.55	0.51	0.55
under-five deaths									1	-0.17	-0.15	-0.18	[illegible]	-0.1	0.66	0.46	0.46	-0.15	-0.23
Polio										1	0.12	0.61	-0.11	0.16	-0.05	-0.16	-0.17	0.31	0.35
Total expenditure											1	0.13	0.04	0.18	-0.08	-0.21	-0.22	0.18	0.24
Diphtheria												1	-0.12	0.16	0.04	-0.19	-0.18	0.34	0.35
HIV/AIDS													1	-0.11	[illegible]	0.17	0.18	-0.25	-0.21
GDP														1	[illegible]	-0.28	-0.28	0.45	0.47
Population															1	0.28	0.28		-0.04
thinness 1-19 years																1	0.93	-0.45	-0.49
thinness 5-9 years																	1	-0.44	-0.47
Income composition of resources																		1	0.78
Schooling																			1

Variable Selection



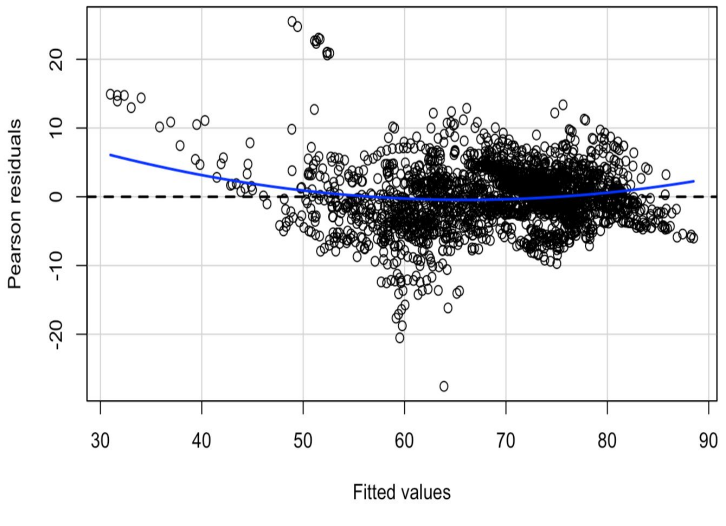
```
Call:
lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI + 
    percentage.expenditure + Income.composition.of.resources + 
    HIV.AIDS + thinness.5.9.years + thinness.10.19.years, data = life)

Residuals:
     Min       1Q   Median       3Q      Max 
-27.5813  -2.5470  -0.0207   2.6412  25.4956 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      4.980e+01  4.985e-01  99.895  < 2e-16 ***
Schooling                        1.078e+00  5.206e-02  20.702  < 2e-16 ***
GDP                              4.434e-05  1.709e-05   2.594 0.009537 ** 
Alcohol                         -1.009e-01  2.961e-02  -3.407 0.000668 ***
BMI                              5.476e-02  6.336e-03   8.642  < 2e-16 ***
percentage.expenditure           1.773e-04  1.111e-04   1.596 0.110670    
Income.composition.of.resources  9.957e+00  7.498e-01  13.279  < 2e-16 ***
HIV.AIDS                        -6.609e-01  1.768e-02 -37.389  < 2e-16 ***
thinness.5.9.years              -4.548e-02  5.638e-02  -0.807 0.419871    
thinness.10.19.years            -7.606e-02  5.757e-02  -1.321 0.186593    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.522 on 2301 degrees of freedom
Multiple R-squared:  0.7837,	Adjusted R-squared:  0.7829 
F-statistic: 926.5 on 9 and 2301 DF,  p-value: < 2.2e-16
```

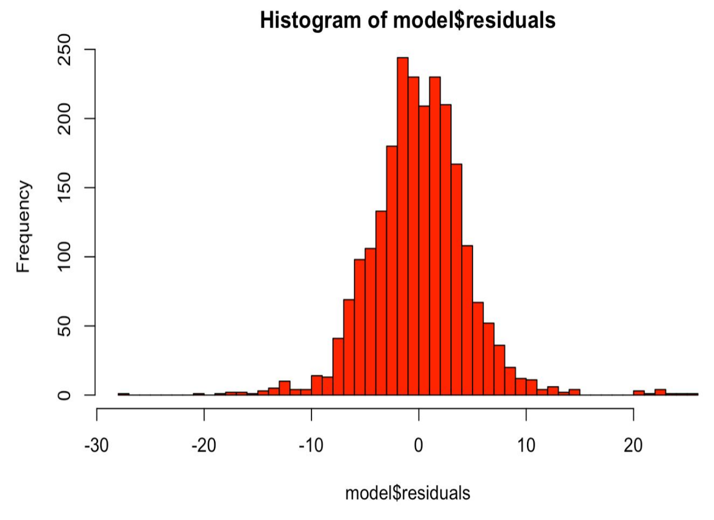
Initial model with all 9 predictors (2 later trimmed)

- We began with the 10 predictors most correlated with life expectancy (simple $R^2 > 0.4$)
    - Later excluded adult mortality, as it is part of the life expectancy calculation
- Then we employed **backward selection** to trim down the model until all predictors had a significance < 0.05
    - This was in line with removing the most mutually correlated variables (highest VIFs ~ 7)
- The best overall model had **7 predictors**, though the other 2 predictors were still useful for other parts of the analysis
- Observations containing missing values were excluded for simplicity and consistency
    - Our subset therefore had 2,311 observations of 13 variables

Question 3

Given these predictors, how accurately can one predict life expectancy with a linear model?



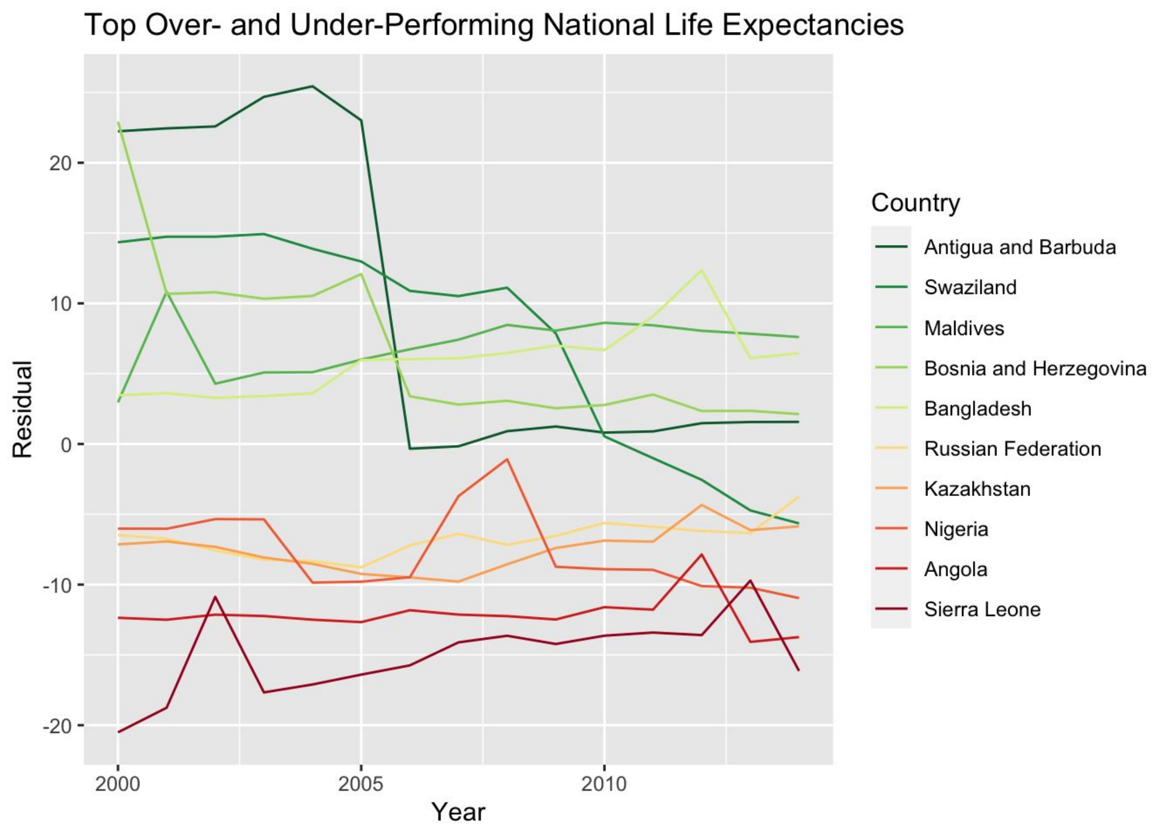


- We have obtained an accuracy 78.28% from the trimmed model.
- We can note that that the points (left) are “equally” spread across the x-axis indicating that our model does not have any non-linear relationships.
- Constructed a histogram (right), from the linear model that the residuals are normally distributed.

Question 5

Which countries over- or under-perform in life expectancy relative to what our linear model would predict? What might account for this difference?

- Overall residual distribution is **approximately normal**
- The distribution of average national errors suggests outliers*
- We calculated p-values for each nation's average error (equivalent to sampling a mean under H_0)
- There were **43 significant outliers** for average error
- The top over- and under-performers are shown here
- Only **external conditions** can account for these consistent differences from the model



$$H_0 : \epsilon \sim N(0, \sigma^2) \text{ for all countries}$$

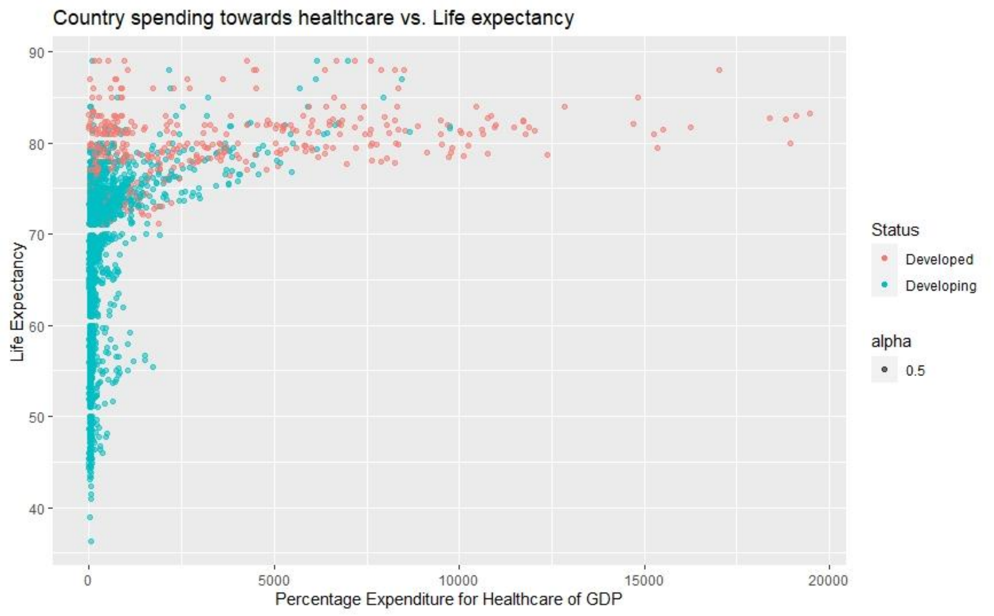
$$H_1 : \epsilon \nsim N(0, \sigma^2) \text{ for at least one country}$$

$$z_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \epsilon = \frac{1}{n_j} \sum_{i=1}^{n_j} N(0, \sigma^2) \sim N(0, \sigma^2 / n_j) \text{ for each country } j$$

* residual variance did not significantly decrease when averaged by nation, in disagreement with the CLT

Question 7

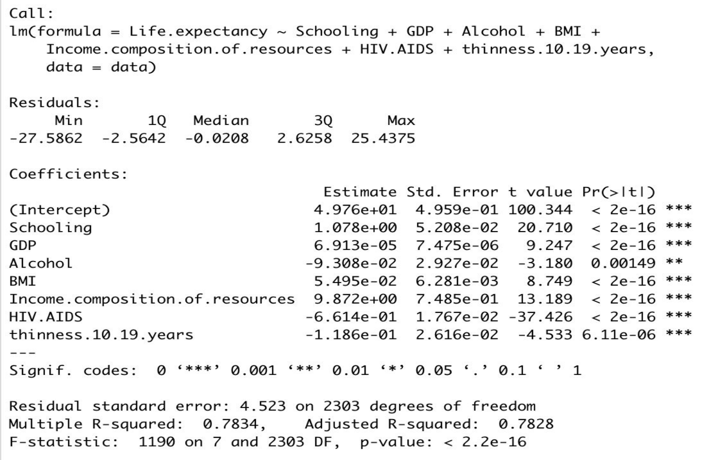
To improve the lifespan of a country with low life expectancy(<65), should they improve on their healthcare expenditure? Does it differ between developing and developed countries?



- Choice is seemingly obvious before analysis
- After analysis, results are mixed
- After certain point in spending, country guaranteed high life expectancy
- Pattern differs between developing and developed countries
- Already high life expectancy for developed countries

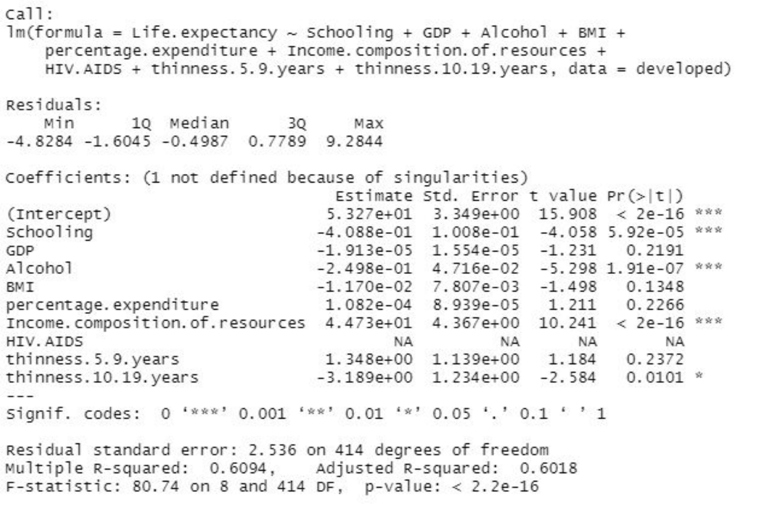
Conclusion

Overall Question this analysis is solving: *What are the key factors that improve life expectancy?*



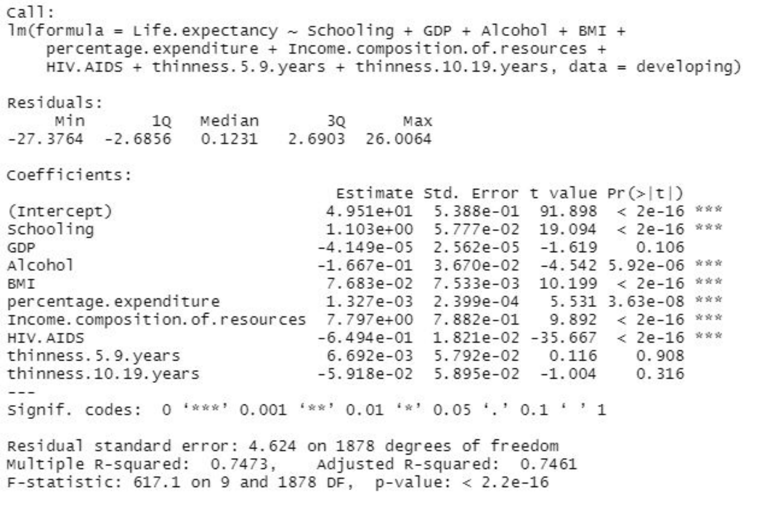
```
Call:
lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI + 
    Income.composition.of.resources + HIV.AIDS + thinness.10.19.years, 
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-27.5862  -2.5642  -0.0208   2.6258  25.4375 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      4.976e+01  4.959e-01 100.344  < 2e-16 ***
Schooling                        1.078e+00  5.208e-02  20.710  < 2e-16 ***
GDP                              6.913e-05  7.475e-06   9.247  < 2e-16 ***
Alcohol                         -9.308e-02  2.927e-02  -3.180  0.00149 ** 
BMI                              5.495e-02  6.281e-03   8.749  < 2e-16 ***
Income.composition.of.resources  9.872e+00  7.485e-01  13.189  < 2e-16 ***
HIV.AIDS                        -6.614e-01  1.767e-02 -37.426  < 2e-16 ***
thinness.10.19.years            -1.186e-01  2.616e-02  -4.533 6.11e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.523 on 2303 degrees of freedom
Multiple R-squared:  0.7834,	Adjusted R-squared:  0.7828 
F-statistic:  1190 on 7 and 2303 DF,  p-value: < 2.2e-16
```

- ➢ Key factors for all countries are schooling, GDP, Alcohol, BMI, income composition of resources, and HIV/AIDS
- ➢ Removed percentage.expenditure and the thinness 5-9 categories, as they are not significant to the model.
- ➢ Very minute changes to R-squared value



```
Call:
lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
    percentage.expenditure + Income.composition.of.resources +
    HIV.AIDS + thinness.5.9.years + thinness.10.19.years, data = developed)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8284 -1.6045 -0.4987  0.7789  9.2844

Coefficients: (1 not defined because of singularities)
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.327e+01  3.349e+00  15.908  < 2e-16 ***
Schooling                      -4.088e-01  1.008e-01  -4.058 5.92e-05 ***
GDP                            -1.913e-05  1.554e-05  -1.231   0.2191
Alcohol                        -2.498e-01  4.716e-02  -5.298 1.91e-07 ***
BMI                            -1.170e-02  7.807e-03  -1.498   0.1348
percentage.expenditure          1.082e-04  8.939e-05   1.211   0.2266
Income.composition.of.resources 4.473e+01  4.367e+00  10.241  < 2e-16 ***
HIV.AIDS                               NA         NA      NA       NA
thinness.5.9.years              1.348e+00  1.139e+00   1.184   0.2372
thinness.10.19.years           -3.189e+00  1.234e+00  -2.584   0.0101 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.536 on 414 degrees of freedom
Multiple R-squared:  0.6094,    Adjusted R-squared:  0.6018
F-statistic: 80.74 on 8 and 414 DF,  p-value: < 2.2e-16
```

Developed



```
Call:
lm(formula = Life.expectancy ~ Schooling + GDP + Alcohol + BMI +
    percentage.expenditure + Income.composition.of.resources +
    HIV.AIDS + thinness.5.9.years + thinness.10.19.years, data = developing)

Residuals:
     Min       1Q   Median       3Q      Max
-27.3764  -2.6856   0.1231   2.6903  26.0064

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.951e+01  5.388e-01  91.898  < 2e-16 ***
Schooling                       1.103e+00  5.777e-02  19.094  < 2e-16 ***
GDP                            -4.149e-05  2.562e-05  -1.619    0.106
Alcohol                        -1.667e-01  3.670e-02  -4.542 5.92e-06 ***
BMI                             7.683e-02  7.533e-03  10.199  < 2e-16 ***
percentage.expenditure          1.327e-03  2.399e-04   5.531 3.63e-08 ***
Income.composition.of.resources 7.797e+00  7.882e-01   9.892  < 2e-16 ***
HIV.AIDS                       -6.494e-01  1.821e-02 -35.667  < 2e-16 ***
thinness.5.9.years              6.692e-03  5.792e-02   0.116    0.908
thinness.10.19.years           -5.918e-02  5.895e-02  -1.004    0.316
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.624 on 1878 degrees of freedom
Multiple R-squared:  0.7473,    Adjusted R-squared:  0.7461
F-statistic: 617.1 on 9 and 1878 DF,  p-value: < 2.2e-16
```

Developing

- Developed countries have fewer key factors to improve life expectancy
- Schooling, Alcohol, and income composition of resources
- average BMI, percentage expenditure on healthcare, income composition of resources, and HIV/AIDS cases

THANK YOU