Analysing Life Expectancy With Linear Regression

Rinku Bajaj (s3672522), Arion Barzoucas-Evans (s3650046), Deepika Joshi (s3672595)

05/06/2019

Contents

Introduction **3**

Data Pre-Processing and Data Exploration **4**

Methodology **11**
- Linear Regression for Developed Countries . . . 12
  - Baseline Model . . . 12
  - Model Building Comparison using AIC Values . . . 16
  - Model Building using R Squared Values . . . 17
- Linear Regression for Developing Countries - Baseline Model . . . 19
  - Baseline Model . . . 19
  - Model Building Comparison . . . 23
  - Model Building using R Squared Values . . . 24

Results **26**
- For developed nations . . . 26
- For developing nations . . . 27
- Factors affecting Life Expectancy . . . 29
- Impact of Healthcare Expenditure on Life Expectancy . . . 29
- Impact of Alcohol Consumption on Life Expectancy . . . 29
- Impact of Immunization Coverage on Life Expectancy . . . 29

Conclusion **30**

Introduction

The main aim of this project is to identify the most influential factors affecting life expectancy in developed and developing countries from 2000-2015. The project considers factors traditionally studied for life expectancy predictions such as demographic variables, mortality rate and income composition as well as the effects of new factors such as immunization and human development index. The data is sourced from Kaggle Repository and is a combination of life-expectancy and health data from Global Health Observatory (GHO) published by World Health Organisation (WHO) and economic data published by United Nations. It consists of data spanning over 15 years from 2000-2015 for 193 countries and can be broadly divided into Economic Factors, Social Factors, Immunization Factors and Mortality Factors. The dataset has 2938 observations for 20 predictor variables describing the life expectancy of each nation for that year.

Linear regression models will be used to predict the life expectancy based on the variables present in the data. This analysis will help in determining the influential factors different than those traditionally used by implementing different methods involved in a regression analysis.

Data Pre-Processing and Data Exploration

The first step after getting the data was data cleaning. Some basic cleaning steps were necessary like removing the NAs present in the data.

```
# load libraries
library(ggplot2)
library(data.table)
library(knitr)
library(corrplot)
library(car)
library(TSA)
library(leaps)
```

```
# read data
data = setDT(read.csv("Life Expectancy Data.csv"))

# replace/remove NAs
data = data[!is.na(Life.expectancy)]
cols = names(data)[!grepl("Country|Status", names(data))]
data[, `:=`((cols), lapply(.SD, function(x) {
    ifelse(is.na(x), -1, x)
})), .SDcols = cols]
```

Data exploration was implemented by means of various visualisations which gave insights into the data and made next steps clearer.

Figure 1 shows the correlation between the numeric variables in the dataset. According to this, life expectancy has a high negative correlation with adult mortality, HIV, and thinness. Conversely, there is a high positive correlation between life expectancy and BMI, schooling, polio, Diphtheria, and GDP. Additionally, many of the independent variables are correlated to each other and may lead to multi-collinearity issues.

Figure 2 displays the distribution of the life expectancy variable in the dataset. This shows that life expectancy ranges from around 40 to 90 years with 50% of the values being over 72 years. This indicates that the variable is slightly skewed to the left. The bar chart in Figure 3 shows that the majority of countries in the dataset are developing countries with only around 30 countries being developed.

```
# correlation plot
corr = cor(data[, c(2, 4:22)], use = "complete.obs")
corrplot(corr, type = "lower")
```

```
# life expectancy
ggplot(data, aes(x = Life.expectancy)) + geom_histogram(fill = "aquamarine4", colour = "white") +
    geom_vline(xintercept = quantile(data$Life.expectancy, 0.5), linetype = "dashed",
        colour = "red", size = 1) + annotate("text", x = quantile(data$Life.expectancy,
    0.5) - 3, y = 400, label = "Median", colour = "red") + theme_minimal() + labs(x = "Life Expectancy"
    y = "Count", title = "Life expectancy distribution")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

There will be differences in results based on the fact that whether the country is a developed or a developing one. Therefore, different statistics were observed across status of the countries.

Life expectancy is strongly correlated to a country's status (Fig 4). It can be seen that developed countries generally have a life expectancy around 80 with little variation. On the other hand, developing countries have an average life expectancy of just under 70 with much greater variance and several outliers with life expectancy of around 40. Life expectancy for both developed and developing countries has an upward trend having a 5 year improvement from 2000 to 2015 (Fig 5). Figure 6 shows a strong negative correlation between

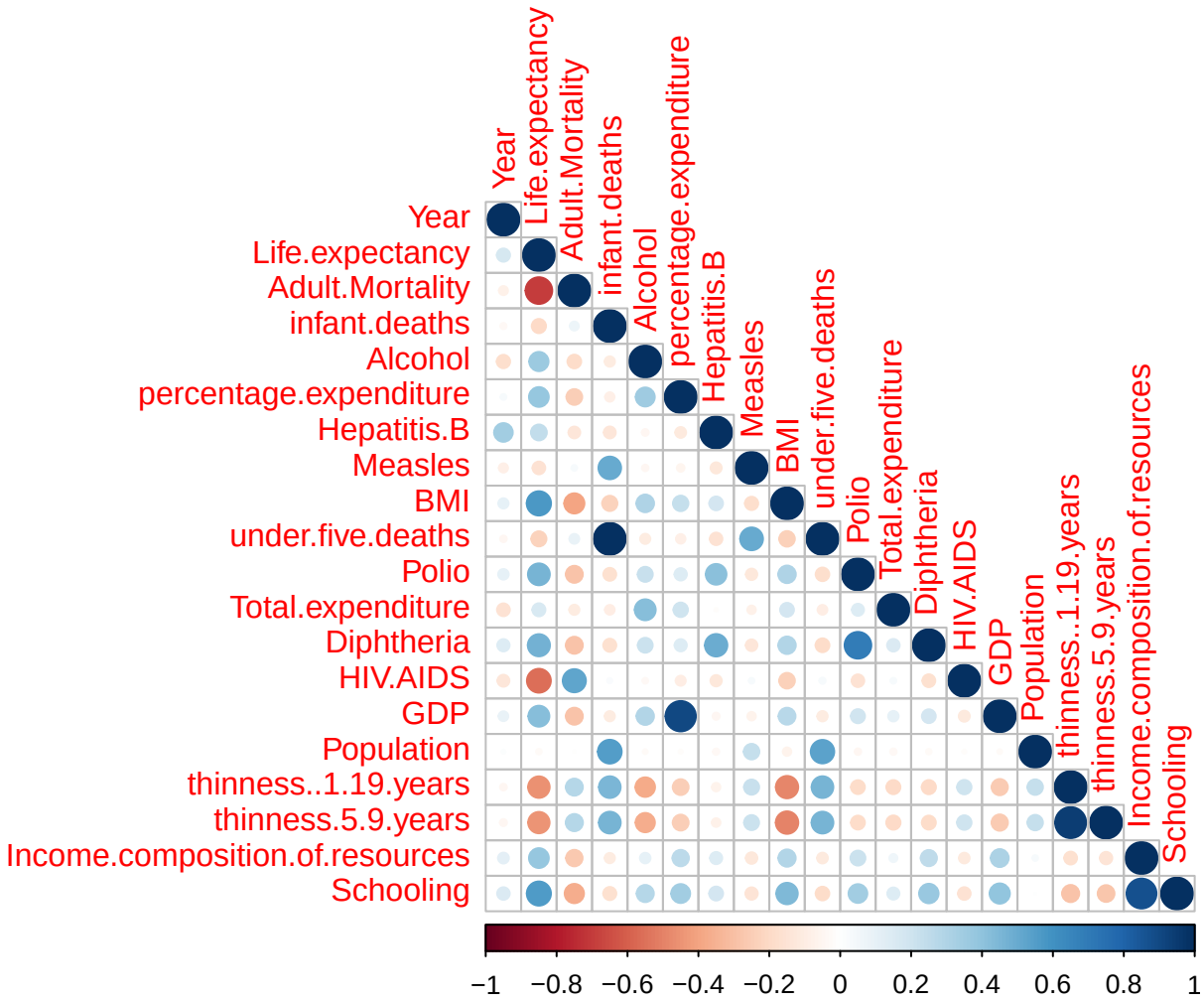


Figure 1: Correlation plot for all numeric variables in the dataset.

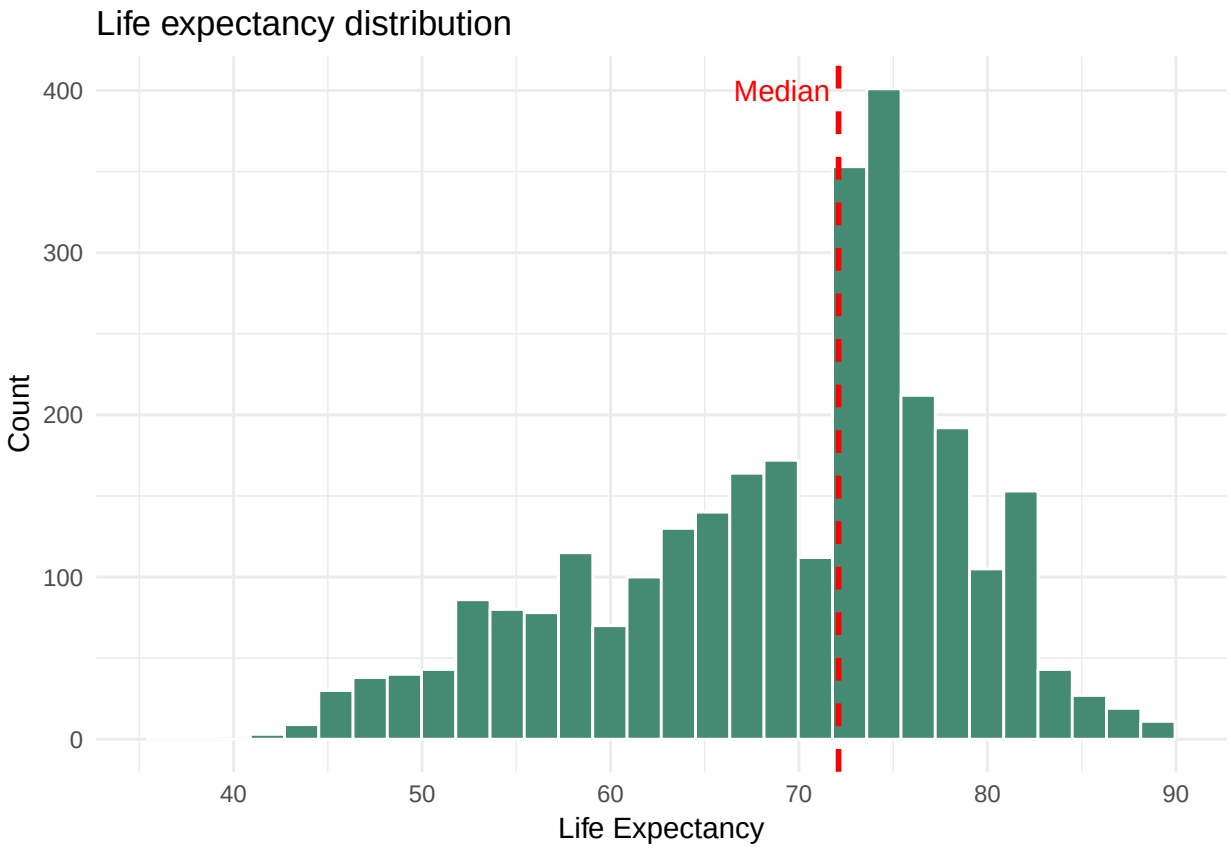


Figure 2: Status distribution in the dataset.

life expectancy and adult mortality for both developed and developing countries. However, there are many cases of countries with a low adult mortality rate that still have low life expectancy. This may indicate that adult mortality alone is not a sufficient indicator of life expectancy.

```
# status
ggplot(data[, length(unique(Country)), Status], aes(x = Status, y = V1)) + geom_bar(stat = "identity",
    fill = "aquamarine4") + theme_minimal() + labs(y = "Countries", title = "Distribution of countries b
```

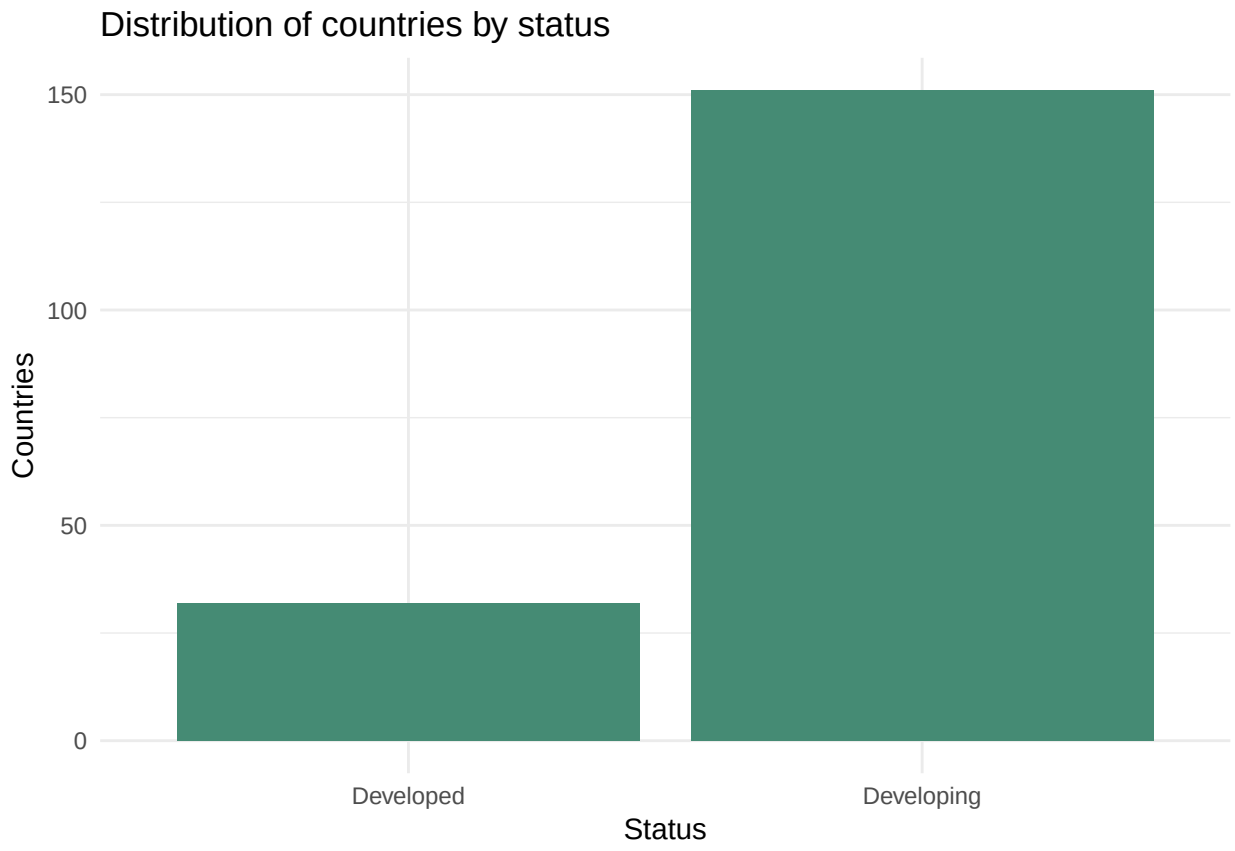


Figure 3: Life expectancy distribution in the dataset.

```
# status vs expectancy
ggplot(data, aes(x = Status, y = Life.expectancy)) + geom_boxplot(fill = "aquamarine4") +
    theme_minimal() + labs(y = "Life Expectancy", title = "Life expectancy by country status")
```

```
# timeseries by status
ggplot(data[, mean(Life.expectancy, na.rm = T), .(Year, Status)], aes(x = Year, y = V1,
    colour = Status, group = Status)) + geom_point() + geom_line() + theme_minimal() +
    labs(y = "Average Life Expectancy", title = "Life expectancy timeseries by country status",
        x = "Year")
```

```
# status & adult mortality
ggplot(data, aes(x = Adult.Mortality, y = Life.expectancy, colour = Status)) + geom_point() +
    theme_minimal() + labs(y = "Life Expectancy", title = "Life Expectancy by Adult Mortality & Country
    x = "Adult Mortality")
```

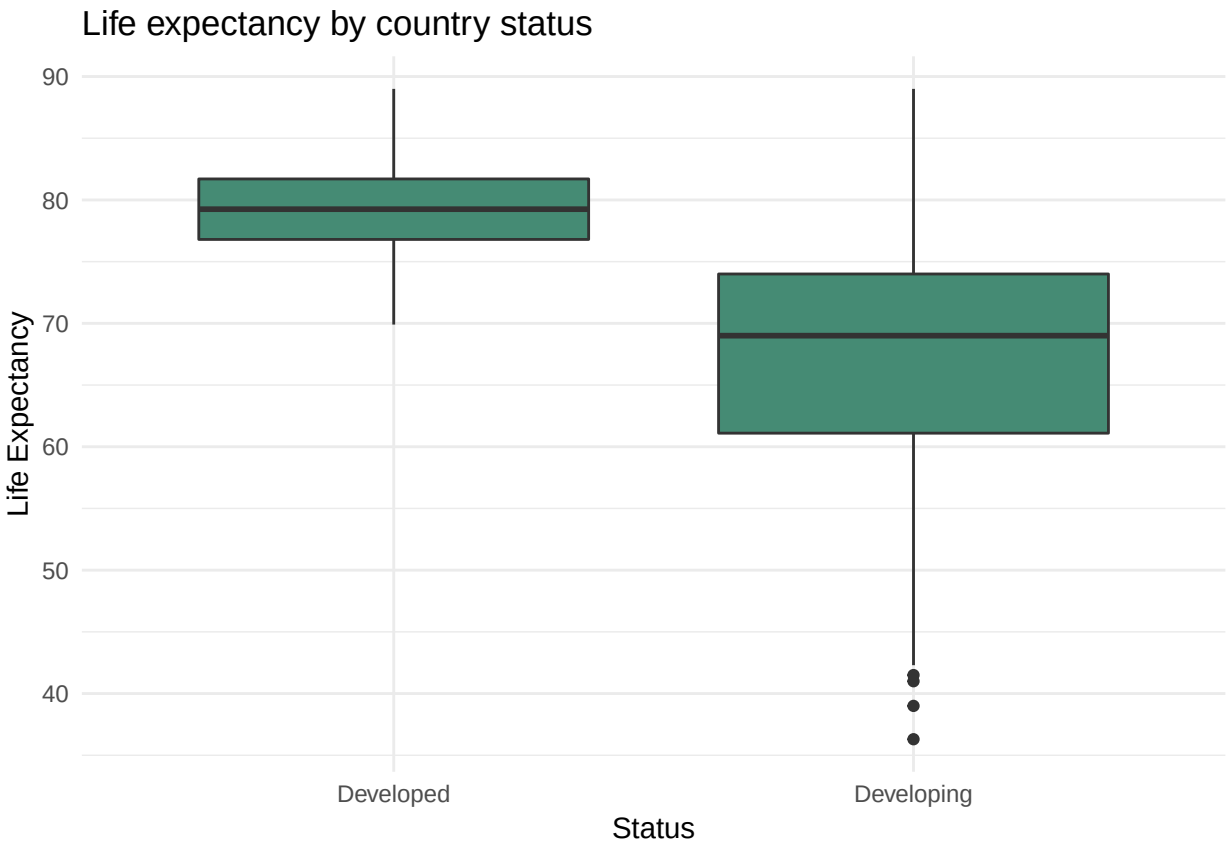


Figure 4: Life expectancy by status.

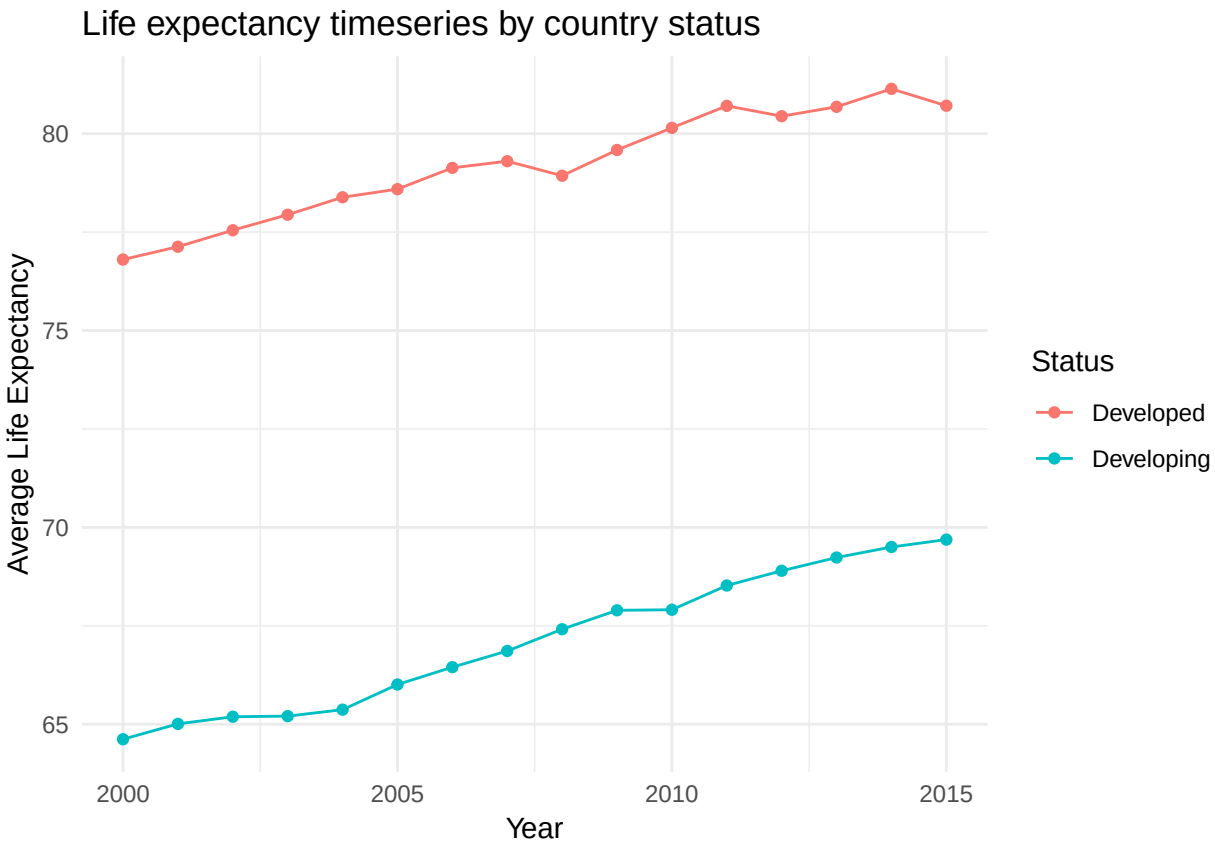


Figure 5: Historical life expectancy by status.

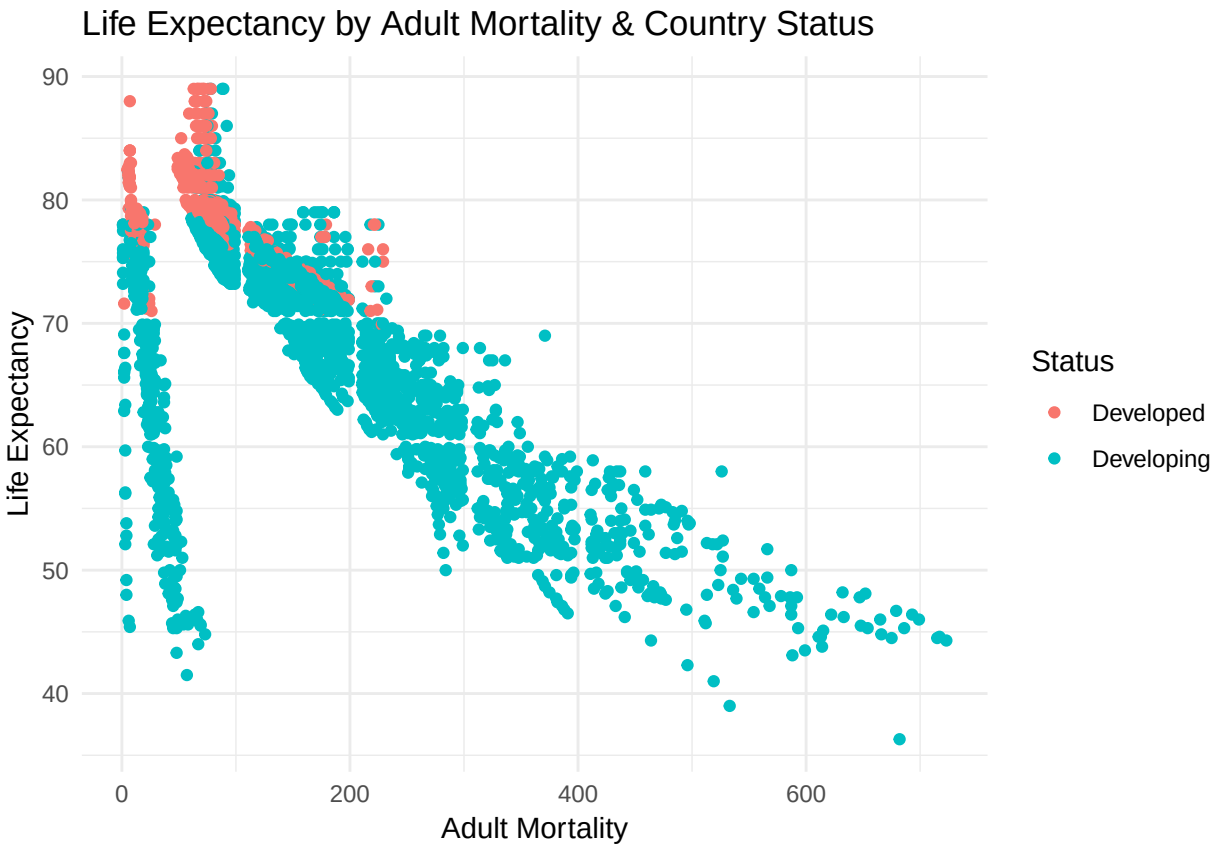


Figure 6: Life expectancy by adult mortality and status.

Methodology

The countries present in the dataset will be categorised into developed and developing. Different linear regression models will be fitted to each group using the traditionally considered factors, the extended factors, and using stepwise regression. The adequacy of the fitted models will be assessed with a combination of residual diagnostics and statistical tests. Using the models with the best fit to the data, the most influential factors will be identified for both developed and developing countries.

Splitting dataset into developed and developing nations

```
# Create Developing and Developed Nations Datasets
data_developing = subset(data, Status == "Developing")
data_developed = subset(data, Status == "Developed")

# Refactor the levels in the Dataset
data_developing$Status <- NULL
data_developed$Status <- NULL
# Discussion
```

Creating a function for Regression Model Analysis

```
regression.analysis <- function(regression_model) {
    # regression_model = lm.fit_1

    # Residual check.options
    par(mfrow = c(2, 2))
    plot(regression_model)

    # Evaluate homoscedasticity non-constant error variance test
    cat("\nHomoskedascity Check:\n")
    print(ncvTest(regression_model))

    # Test for Autocorrelated Errors
    cat("\nAutoCorrelation Check:\n")
    print(durbinWatsonTest(regression_model))

    # Test for Normally Distributed Errors
    cat("\nNormality Check:\n")
    print(shapiro.test(regression_model$residuals))

    # Check multicollinearity
    cat("\nMulticollinearity Check:\n")
    kable(vif(lm.fit_1), col.names = "VIF")
}
```

Linear Regression for Developed Countries

Baseline Model

First a baseline model is created for developed countries. This baseline model includes all features from the data. But on further processing it was observed that the feature for HIV AIDs when included in the model, produces NAs. Therefore, this feature was removed from the model. The variables Country and Year were excluded as well.

From the summary of the model it can be seen that, 51.59% of the variation in life expectancy is explained by the features used in the model. Also, the p-value of the regression is very low and hence it can be concluded that the regression is significant.

```
# Linear Regression Model
lm.fit_1 <- lm(Life.expectancy ~ . - (Country + Year + HIV.AIDS), data_developed)
summary(lm.fit_1)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ . - (Country + Year + HIV.AIDS),
##     data = data_developed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9487 -1.6294 -0.4584  1.0129  9.7749
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     8.295e+01  1.684e+00  49.252  < 2e-16 ***
## Adult.Mortality                -1.646e-02  2.955e-03  -5.568 4.23e-08 ***
## infant.deaths                  -1.424e+00  4.218e-01  -3.376 0.000794 ***
## Alcohol                        -2.602e-01  3.995e-02  -6.512 1.83e-10 ***
## percentage.expenditure          1.457e-04  8.850e-05   1.646 0.100374
## Hepatitis.B                    -2.927e-03  3.064e-03  -0.955 0.340026
## Measles                         1.702e-05  5.053e-05   0.337 0.736324
## BMI                            -2.010e-02  7.478e-03  -2.688 0.007422 **
## under.five.deaths               1.141e+00  3.560e-01   3.206 0.001431 **
## Polio                           2.417e-03  1.449e-02   0.167 0.867590
## Total.expenditure              -2.554e-02  4.392e-02  -0.582 0.561149
## Diphtheria                      2.424e-02  1.246e-02   1.945 0.052303 .
## GDP                             1.010e-05  1.486e-05   0.680 0.496915
## Population                      1.145e-08  9.569e-09   1.197 0.231996
## thinness..1.19.years           -2.007e+00  1.349e+00  -1.488 0.137402
## thinness.5.9.years             -2.762e-01  1.240e+00  -0.223 0.823806
## Income.composition.of.resources -1.930e+00  9.308e-01  -2.074 0.038618 *
## Schooling                       1.811e-01  9.542e-02   1.898 0.058282 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.735 on 494 degrees of freedom
## Multiple R-squared:  0.5321, Adjusted R-squared:  0.5159
## F-statistic: 33.04 on 17 and 494 DF,  p-value: < 2.2e-16
```

Checking significance of regression co-efficients

ANOVA test was performed to get significance of the model parameters. The hypotheses for this test are:

$H_0 : \beta_i = 0$ $H_a : \beta_i \neq 0$ where i runs from 1:17 and i represents each predictor.

For this model it is observed that, at 95% significance level, estimates for adult mortality, infant deaths, alcohol, percentage expenditure, Hepatitis B, under five deaths, population and thinness 1-19 years are statistically significant, when used with all other predictors in the model.

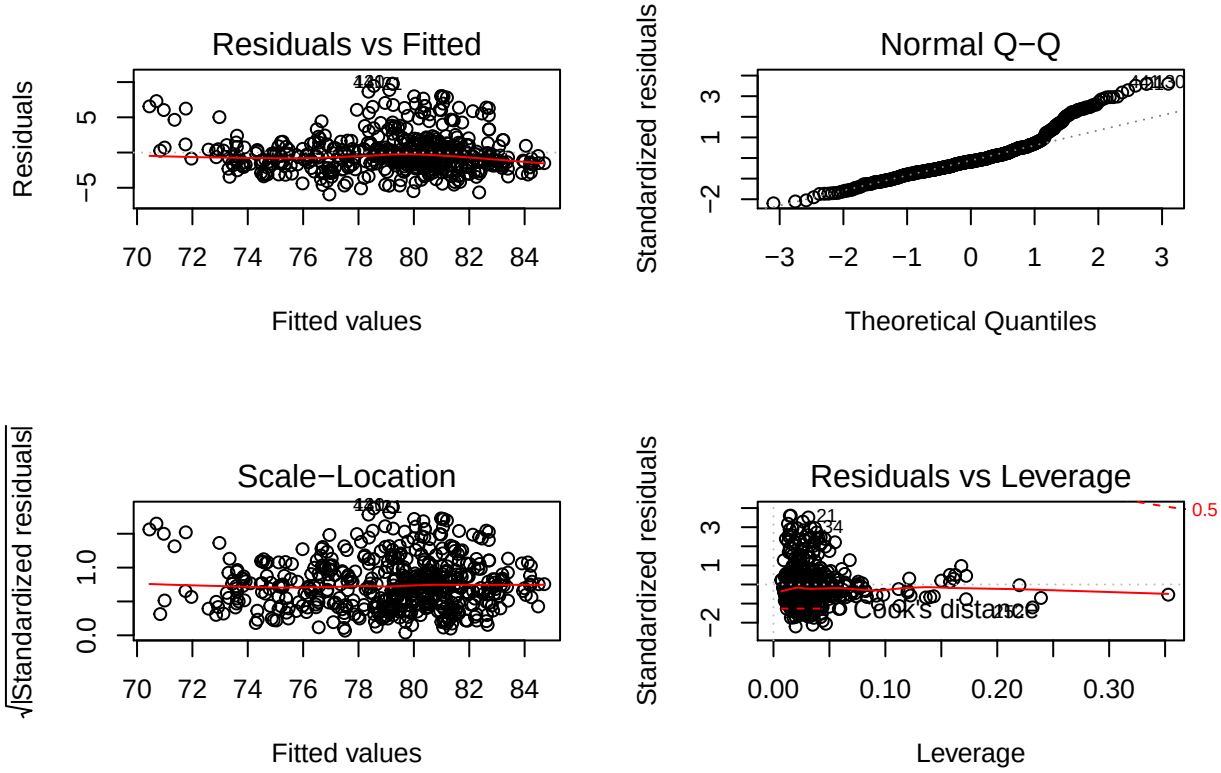
```
# Using ANOVA to check significance of variables
anova(lm.fit_1)
```

```
## Analysis of Variance Table
##
## Response: Life.expectancy
##                                  Df Sum Sq Mean Sq  F value    Pr(>F)
## Adult.Mortality                   1 1861.1 1861.11 248.8219 < 2.2e-16 ***
## infant.deaths                     1   48.3   48.30   6.4574 0.0113536 *
## Alcohol                           1  360.4  360.44  48.1895 1.222e-11 ***
## percentage.expenditure            1  603.6  603.62  80.7011 < 2.2e-16 ***
## Hepatitis.B                       1   46.5   46.52   6.2193 0.0129626 *
## Measles                           1    0.0    0.03   0.0039 0.9502457
## BMI                               1    3.8    3.77   0.5034 0.4783421
## under.five.deaths                 1  101.9  101.88  13.6211 0.0002485 ***
## Polio                             1    3.3    3.34   0.4464 0.5043390
## Total.expenditure                 1    5.6    5.58   0.7455 0.3883235
## Diphtheria                        1    1.0    0.97   0.1300 0.7185952
## GDP                               1    3.0    2.99   0.4000 0.5273981
## Population                        1   44.6   44.57   5.9582 0.0149988 *
## thinness..1.19.years              1 1084.2 1084.20 144.9526 < 2.2e-16 ***
## thinness.5.9.years                1    1.5    1.45   0.1939 0.6598936
## Income.composition.of.resources   1    5.4    5.45   0.7283 0.3938423
## Schooling                         1   26.9   26.94   3.6023 0.0582820 .
## Residuals                       494 3695.0    7.48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual analysis for validating Model

Here, the function created for residual analysis is used.

```
regression.analysis(lm.fit_1)
```




```
##
## Homoskedascity Check:
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.007319894    Df = 1     p = 0.9318191
##
## AutoCorrelation Check:
##  lag Autocorrelation D-W Statistic p-value
##    1       0.5877129     0.8232728       0
##  Alternative hypothesis: rho != 0
##
## Normality Check:
##
##  Shapiro-Wilk normality test
##
## data:  regression_model$residuals
## W = 0.91705, p-value = 3.76e-16
##
##
## Multicollinearity Check:
```

	VIF
Adult.Mortality	1.367702
infant.deaths	255.594031
Alcohol	1.450732
percentage.expenditure	7.825414

	VIF
Hepatitis.B	1.318336
Measles	1.115706
BMI	1.129846
under.five.deaths	250.918986
Polio	1.667478
Total.expenditure	1.666328
Diphtheria	1.665702
GDP	7.702266
Population	1.156524
thinness..1.19.years	71.143549
thinness.5.9.years	72.205097
Income.composition.of.resources	17.439966
Schooling	16.784452

1) residuals v/s fitted plot

This plot is used for testing the linear approximation assumption. In this case, the residuals seem to be randomly distributed.

2) QQ plot

This plot is used to test the normality assumption. There seems to be deviation of the upper tail from the straight line.

3) scale location plot

This plot is used to check homoskedasticity. There is a horizontal line with randomly spaced points.

4) residual v/s leverage plot

This plot is used to find influential cases. Here, all the points are within cook's distance.

Homoskedasticity test:

H_0: Errors have constant variance H_a: Errors have non-constant variance

The nCV test has pvalue > 0.05. Therefore, we fail to reject the null hypothesis. So the constant error variance assumption is not violated.

Test for autocorrelated errors:

H_0: Errors are uncorrelated H_1: Errors are correlated

p-value is > 0.05. So, we don't have enough evidence to reject H_0. Therefore, uncorrelated error assumption is not violated.

Test for normality of residuals:

H_0: errors are normally distributed H_a: errors aren't normally distributed

p-value < 0.05. Therefore, we reject the null hypothesis. This indicates that the normality assumption is violated.

Test for multicollinearity:

Very high values of VIF indicate strong correlation. Hence, it can be said that infant death and under five deaths are highly correlated. Also, 'thinness 1-19 yrs' and 'thinness 5-9 yrs' are correlated.

Some assumptions are violated. Some sort of non-linear transformations can be applied on the data to fix this.

Model Building Comparison using AIC Values

Model building comparison for backward elimination, forward selection and stepwise regression using AIC values

In forward selection, initially there are no variables and then variables are added one at a time. Based on AIC value, the best model is selected. In case of backward elimination, a full model (model with all features included) is created and then variables are examined and removed sequentially. Stepwise is a combination of both forward selection and backward elimination. Here contribution of each feature is checked using partial F test. Akaike Information Criteria (AIC) penalises based on the number of parameters in the model. Smaller value of AIC is desirable.

The coefficients of the parameters for the best model selected using backward elimination are:

```
# Backward Elimination
b1 = step(lm.fit_1, data = data_developed, direction = "backward", trace = FALSE)

kable(b1$coefficients, col.names = "Backward Elimination Parameter Estimates")
```

	Backward Elimination Parameter Estimates
(Intercept)	83.0325163
Adult.Mortality	-0.0166258
infant.deaths	-1.4159206
Alcohol	-0.2793219
percentage.expenditure	0.0002058
BMI	-0.0206657
under.five.deaths	1.1266582
Diphtheria	0.0258218
thinness..1.19.years	-2.3160206
Income.composition.of.resources	-1.8975011
Schooling	0.1817349

The coefficients of the parameters for the best model selected using forward selection are:

```
# null model contains no variable
null_data1 = lm(Life.expectancy ~ 1, data = data_developed)

# forward selection using AIC values
f1 = step(null_data1, scope = list(lower = null_data1, upper = lm.fit_1), direction = "forward",
    trace = FALSE)
kable(f1$coefficients, col.names = "Forward Selection Parameter Estimates")
```

	Forward Selection Parameter Estimates
(Intercept)	84.1376432
thinness.5.9.years	-2.2027031
Alcohol	-0.2564736
GDP	0.0000089
Adult.Mortality	-0.0164504
infant.deaths	-1.3484573
under.five.deaths	1.0778448
BMI	-0.0162380
Diphtheria	0.0199727
percentage.expenditure	0.0001444
Population	0.0000000

The coefficients of the parameters for the best model selected using stepwise regression are:

```
# stepwise regression using AIC values
s1 = step(null_data1, scope = list(upper = lm.fit_1), data = data_developed, direction = "both",
    trace = FALSE)
kable(s1$coefficients, col.names = "Stepwise Regression Parameter Estimates")
```

	Stepwise Regression Parameter Estimates
(Intercept)	84.2388162
thinness.5.9.years	-2.1987648
Alcohol	-0.2630321
Adult.Mortality	-0.0165789
infant.deaths	-1.3362721
under.five.deaths	1.0656474
BMI	-0.0166584
Diphtheria	0.0202817
percentage.expenditure	0.0001927
Population	0.0000000

Model parameters differ a bit based on the chosen method. Features like adult mortality, alcohol, diphtheria, infant deaths, percentage expenditure, BMI, under 5 deaths are significant across all the 3 methods.

Model Building using R Squared Values

Here, best model is determined by using adjusted r2 as the selection criteria. Model with the highest adjusted R2 value is chosen. Also, adjusted R2 is always a better metric to compare models instead of R2.

Using 'regsubsets' function to determine the best subset model.

```
new_model_developed <- regsubsets(Life.expectancy ~ . - (Country + Year + HIV.AIDS),
    data = data_developed)
nm1 = summary(new_model_developed)
nm1
```

```
## Subset selection object
## Call: regsubsets.formula(Life.expectancy ~ . - (Country + Year + HIV.AIDS),
##     data = data_developed)
## 17 Variables  (and intercept)
##                                Forced in Forced out
## Adult.Mortality                    FALSE      FALSE
## infant.deaths                      FALSE      FALSE
## Alcohol                            FALSE      FALSE
## percentage.expenditure             FALSE      FALSE
## Hepatitis.B                        FALSE      FALSE
## Measles                            FALSE      FALSE
## BMI                                FALSE      FALSE
## under.five.deaths                  FALSE      FALSE
## Polio                              FALSE      FALSE
## Total.expenditure                  FALSE      FALSE
## Diphtheria                         FALSE      FALSE
## GDP                                FALSE      FALSE
## Population                         FALSE      FALSE
## thinness..1.19.years               FALSE      FALSE
## thinness.5.9.years                 FALSE      FALSE
```

```
## Income.composition.of.resources     FALSE      FALSE
## Schooling                           FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Adult.Mortality infant.deaths Alcohol percentage.expenditure
## 1  ( 1 ) " "             " "           " "     " "
## 2  ( 1 ) " "             " "           "*"     " "
## 3  ( 1 ) " "             " "           "*"     " "
## 4  ( 1 ) "*"             " "           "*"     "*"
## 5  ( 1 ) "*"             "*"           "*"     "*"
## 6  ( 1 ) "*"             "*"           "*"     "*"
## 7  ( 1 ) "*"             "*"           "*"     "*"
## 8  ( 1 ) "*"             "*"           "*"     "*"
##          Hepatitis.B Measles BMI under.five.deaths Polio Total.expenditure
## 1  ( 1 ) " "         " "     " " " "               " "   " "
## 2  ( 1 ) " "         " "     " " " "               " "   " "
## 3  ( 1 ) " "         " "     " " " "               " "   " "
## 4  ( 1 ) " "         " "     " " " "               " "   " "
## 5  ( 1 ) " "         " "     " " " "               " "   " "
## 6  ( 1 ) " "         " "     " " "*"               " "   " "
## 7  ( 1 ) " "         " "     "*" "*"               " "   " "
## 8  ( 1 ) " "         " "     "*" "*"               " "   " "
##          Diphtheria GDP Population thinness..1.19.years thinness.5.9.years
## 1  ( 1 ) " "        " " " "        " "                  "*"
## 2  ( 1 ) " "        " " " "        " "                  "*"
## 3  ( 1 ) " "        "*" " "        " "                  "*"
## 4  ( 1 ) " "        " " " "        " "                  "*"
## 5  ( 1 ) " "        " " " "        " "                  "*"
## 6  ( 1 ) " "        " " " "        " "                  "*"
## 7  ( 1 ) " "        " " " "        "*"                  " "
## 8  ( 1 ) "*"        " " " "        "*"                  " "
##          Income.composition.of.resources Schooling
## 1  ( 1 ) " "                             " "
## 2  ( 1 ) " "                             " "
## 3  ( 1 ) " "                             " "
## 4  ( 1 ) " "                             " "
## 5  ( 1 ) " "                             " "
## 6  ( 1 ) " "                             " "
## 7  ( 1 ) " "                             " "
## 8  ( 1 ) " "                             " "
```

Selecting the best subset model based on highest Adjusted R^2 value

```
adjr2_developed = which.max(nm1$adjr2)
adjr2_developed
```

```
## [1] 8
```

The model with the highest adjusted R2 value has adult mortality, infant deaths, alcohol, percentage expenditure, BMI, under five deaths, diphtheria, thinness 1-19 in it.

Linear Regression for Developing Countries - Baseline Model

Baseline Model

LM Model and Summary

The exact same procedure is followed for developing countries as well. First, baseline model with all the features included (except country and year) is built. Summary of the model shows that regression is significant as p-value is very small. Adjusted R2 shows that this baseline model is able to explain 76.8% variability in life expectancy using the predictors in the model.

```
# Linear Regression Model
lm.fit_2 <- lm(Life.expectancy ~ . - (Country + Year), data_developing)
summary(lm.fit_2)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ . - (Country + Year), data = data_developing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.826  -2.354   0.086   2.497  19.025
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.872e+01  5.199e-01 112.953  < 2e-16 ***
## Adult.Mortality                -1.908e-02  8.750e-04 -21.807  < 2e-16 ***
## infant.deaths                   1.235e-01  8.941e-03  13.811  < 2e-16 ***
## Alcohol                         1.864e-01  3.046e-02   6.120 1.09e-09 ***
## percentage.expenditure          1.135e-03  2.057e-04   5.519 3.78e-08 ***
## Hepatitis.B                     1.792e-02  3.087e-03   5.805 7.30e-09 ***
## Measles                        -1.373e-05  8.249e-06  -1.665  0.09613 .
## BMI                             6.614e-02  5.992e-03  11.038  < 2e-16 ***
## under.five.deaths              -9.137e-02  6.559e-03 -13.930  < 2e-16 ***
## Polio                           2.455e-02  4.890e-03   5.020 5.54e-07 ***
## Total.expenditure              -1.123e-01  3.553e-02  -3.162  0.00159 **
## Diphtheria                      2.157e-02  5.280e-03   4.085 4.55e-05 ***
## HIV.AIDS                       -4.929e-01  1.899e-02 -25.952  < 2e-16 ***
## GDP                             3.950e-06  2.129e-05   0.186  0.85281
## Population                     -1.046e-09  1.818e-09  -0.575  0.56518
## thinness..1.19.years           -1.200e-01  5.369e-02  -2.235  0.02552 *
## thinness.5.9.years              2.486e-02  5.314e-02   0.468  0.63998
## Income.composition.of.resources -2.541e+00  4.554e-01  -5.579 2.69e-08 ***
## Schooling                       6.873e-01  5.102e-02  13.471  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.329 on 2397 degrees of freedom
## Multiple R-squared:  0.7706, Adjusted R-squared:  0.7689
## F-statistic: 447.4 on 18 and 2397 DF,  p-value: < 2.2e-16
```

Checking significance of regression co-efficients

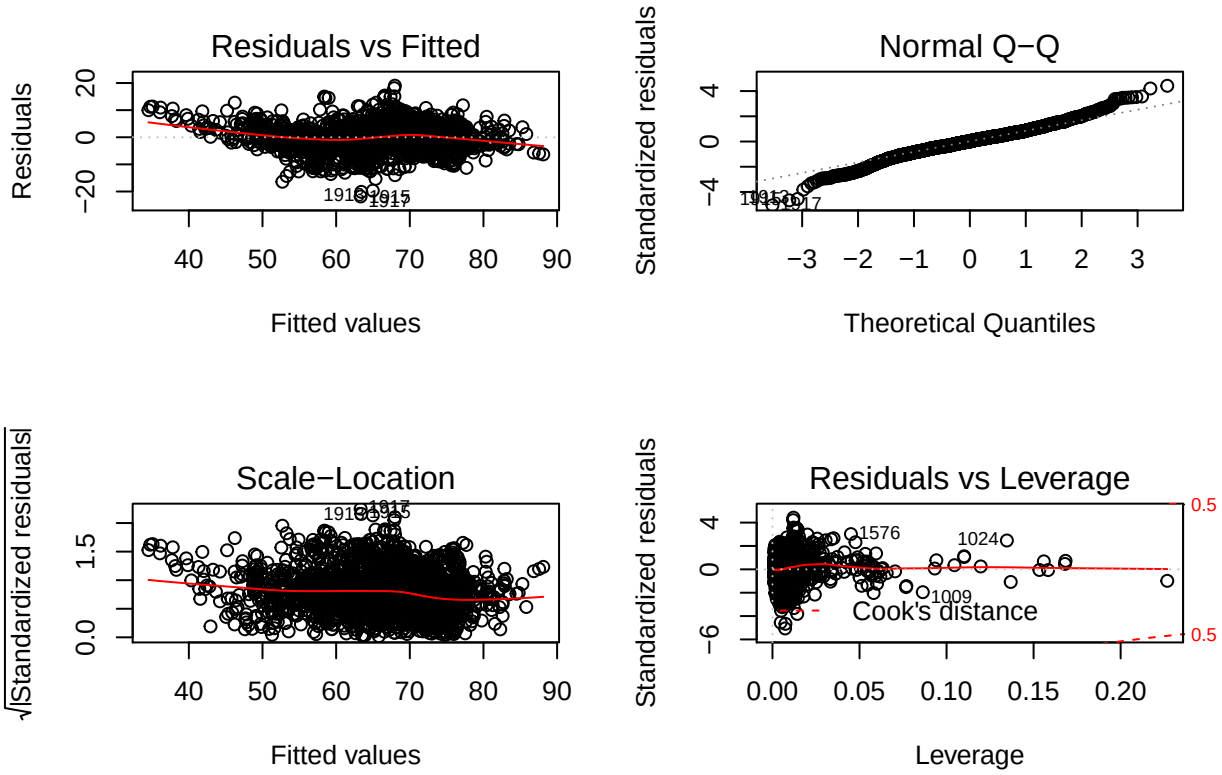
ANOVA test is used to check significance of model parameters. It is observed that at 5% significance level, except features like GDP, Population and thinness 5-9 years, rest all the feature parameters are significant when considered with rest of the features.

```r
# Using ANOVA to check significance of variables
anova(lm.fit_2)
```

```
## Analysis of Variance Table
##
## Response: Life.expectancy
##                                  Df Sum Sq Mean Sq   F value    Pr(>F)
## Adult.Mortality                   1  85542   85542 4563.9288 < 2.2e-16
## infant.deaths                     1   3612    3612  192.7107 < 2.2e-16
## Alcohol                           1   4593    4593  245.0676 < 2.2e-16
## percentage.expenditure            1   5877    5877  313.5666 < 2.2e-16
## Hepatitis.B                       1  13459   13459  718.0636 < 2.2e-16
## Measles                           1    668     668   35.6362 2.731e-09
## BMI                               1  10119   10119  539.8884 < 2.2e-16
## under.five.deaths                 1   6602    6602  352.2269 < 2.2e-16
## Polio                             1   1799    1799   95.9591 < 2.2e-16
## Total.expenditure                 1    453     453   24.1912 9.314e-07
## Diphtheria                        1    710     710   37.8690 8.841e-10
## HIV.AIDS                          1  12122   12122  646.7694 < 2.2e-16
## GDP                               1     32      32    1.6825  0.194716
## Population                        1      2       2    0.0860  0.769304
## thinness..1.19.years              1    190     190   10.1485  0.001463
## thinness.5.9.years                1     49      49    2.6272  0.105176
## Income.composition.of.resources   1   1724    1724   91.9636 < 2.2e-16
## Schooling                         1   3401    3401  181.4672 < 2.2e-16
## Residuals                      2397  44927      19
##
## 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                        ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual analysis for validating Model

```r
regression.analysis(lm.fit_2)
```




```
##
## Homoskedascity Check:
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 98.17636    Df = 1     p = 3.827325e-23
##
## AutoCorrelation Check:
##  lag Autocorrelation D-W Statistic p-value
##    1        0.669526     0.6577663       0
##  Alternative hypothesis: rho != 0
##
## Normality Check:
##
##  Shapiro-Wilk normality test
##
## data:  regression_model$residuals
## W = 0.98232, p-value < 2.2e-16
##
##
## Multicollinearity Check:
```

	VIF
Adult.Mortality	1.367702
infant.deaths	255.594031
Alcohol	1.450732
percentage.expenditure	7.825414

	VIF
Hepatitis.B	1.318336
Measles	1.115706
BMI	1.129846
under.five.deaths	250.918986
Polio	1.667478
Total.expenditure	1.666328
Diphtheria	1.665702
GDP	7.702266
Population	1.156524
thinness..1.19.years	71.143549
thinness.5.9.years	72.205097
Income.composition.of.resources	17.439966
Schooling	16.784452

1) residuals v/s fitted plot

The residuals seem to be randomly distributed.

2) QQ plot

There seems to be slight deviation from the straight line.

3) scale location plot

This plot is used to check homoskedasticity. There is a horizontal line with randomly spaced points.

4) residual v/s leverage plot

All the points are within cook's distance.

Homoskedasticity test:

H_0: Errors have constant variance H_a: Errors have non-constant variance

The nCV test has pvalue < 0.05. Therefore, reject the null hypothesis. So the constant error variance assumption is violated.

Test for autocorrelated errors:

H_0: Errors are uncorrelated H_1: Errors are correlated

p-value is > 0.05. So, we don't have enough evidence to reject H_0. Therefore, uncorrelated error assumption is not violated.

Test for normality of residuals:

H_0: errors are normally distributed H_a: errors aren't normally distributed

p-value < 0.05. Therefore, we reject the null hypothesis. This indicates that the normality assumption is violated.

Test for multicollinearity:

Very high values of VIF indicate strong correlation. Hence, it can be said that infant death and under five deaths are highly correlated. Also, 'income composition of resources' and 'schooling' are correlated.

If needed, the models can be improved further using more complex methods to get fix the problem of assumptions being violated.

Model Building Comparison

Best models based on AIC values from backward elimination, forward selection and stepwise regression are compared.

```
# Backward Elimination
b2 = step(lm.fit_2, data = data_developing, direction = "backward", trace = FALSE)
kable(b2$coefficients, col.names = "Backward Elimination Parameter Estimates")
```

	Backward Elimination Parameter Estimates
(Intercept)	58.7468874
Adult.Mortality	-0.0190778
infant.deaths	0.1227101
Alcohol	0.1853281
percentage.expenditure	0.0011651
Hepatitis.B	0.0179230
Measles	-0.0000137
BMI	0.0658982
under.five.deaths	-0.0909685
Polio	0.0246080
Total.expenditure	-0.1143888
Diphtheria	0.0215884
HIV.AIDS	-0.4925487
thinness..1.19.years	-0.0976929
Income.composition.of.resources	-2.5442213
Schooling	0.6879433

```
# null model contains no variable
null_data2 = lm(Life.expectancy ~ 1, data = data_developing)

# forward selection using AIC values
f2 = step(null_data2, scope = list(lower = null_data2, upper = lm.fit_2), direction = "forward",
    trace = FALSE)
kable(f2$coefficients, col.names = "Forward Selection Parameter Estimates")
```

	Forward Selection Parameter Estimates
(Intercept)	57.5586999
Adult.Mortality	-0.0198014
Schooling	0.7231942
HIV.AIDS	-0.5070142
BMI	0.0682506
Diphtheria	0.0290133
percentage.expenditure	0.0011390
Income.composition.of.resources	-2.5681629
Hepatitis.B	0.0184523
Polio	0.0281819
thinness..1.19.years	-0.0905599
Alcohol	0.1374685
Total.expenditure	-0.1120023
Measles	-0.0000227

```r
# stepwise regression using AIC values
s2 = step(null_data2, scope = list(upper = lm.fit_2), data = data_developed, direction = "both",
    trace = FALSE)
kable(s2$coefficients, col.names = "Stepwise Regression Parameter Estimates")
```

	Stepwise Regression Parameter Estimates
(Intercept)	57.5586999
Adult.Mortality	-0.0198014
Schooling	0.7231942
HIV.AIDS	-0.5070142
BMI	0.0682506
Diphtheria	0.0290133
percentage.expenditure	0.0011390
Income.composition.of.resources	-2.5681629
Hepatitis.B	0.0184523
Polio	0.0281819
thinness..1.19.years	-0.0905599
Alcohol	0.1374685
Total.expenditure	-0.1120023
Measles	-0.0000227

Across all 3 methods, features like adult mortality, schooling, HIV AIDS, BMI, Diphtheria, percentage expenditure, income composition of resources, hepatitis B‘, polio, thinness 1-19 years and total expenditure seem to be influential.

Model Building using R Squared Values

Using ‘regsubsets’ function to determine the best subset model

```r
new_model_developing <- regsubsets(Life.expectancy ~ . - (Country + Year), data = data_developing)
nm2 = summary(new_model_developing)
nm2
```

```
## Subset selection object
## Call: regsubsets.formula(Life.expectancy ~ . - (Country + Year), data = data_developing)
## 18 Variables  (and intercept)
##                                 Forced in Forced out
## Adult.Mortality                     FALSE      FALSE
## infant.deaths                       FALSE      FALSE
## Alcohol                             FALSE      FALSE
## percentage.expenditure              FALSE      FALSE
## Hepatitis.B                         FALSE      FALSE
## Measles                             FALSE      FALSE
## BMI                                 FALSE      FALSE
## under.five.deaths                   FALSE      FALSE
## Polio                               FALSE      FALSE
## Total.expenditure                   FALSE      FALSE
## Diphtheria                          FALSE      FALSE
## HIV.AIDS                            FALSE      FALSE
## GDP                                 FALSE      FALSE
## Population                          FALSE      FALSE
## thinness..1.19.years                FALSE      FALSE
```

```
## thinness.5.9.years                  FALSE      FALSE
## Income.composition.of.resources     FALSE      FALSE
## Schooling                           FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Adult.Mortality infant.deaths Alcohol percentage.expenditure
## 1  ( 1 ) "*"             " "           " "     " "
## 2  ( 1 ) " "             " "           " "     " "
## 3  ( 1 ) "*"             " "           " "     " "
## 4  ( 1 ) "*"             " "           " "     " "
## 5  ( 1 ) "*"             " "           " "     " "
## 6  ( 1 ) "*"             " "           " "     "*"
## 7  ( 1 ) "*"             "*"           " "     " "
## 8  ( 1 ) "*"             "*"           " "     "*"
##          Hepatitis.B Measles BMI under.five.deaths Polio Total.expenditure
## 1  ( 1 ) " "         " "     " " " "               " "   " "
## 2  ( 1 ) " "         " "     " " " "               " "   " "
## 3  ( 1 ) " "         " "     " " " "               " "   " "
## 4  ( 1 ) " "         " "     "*" " "               " "   " "
## 5  ( 1 ) " "         " "     "*" " "               " "   " "
## 6  ( 1 ) " "         " "     "*" " "               " "   " "
## 7  ( 1 ) " "         " "     "*" "*"               " "   " "
## 8  ( 1 ) " "         " "     "*" "*"               " "   " "
##          Diphtheria HIV.AIDS GDP Population thinness..1.19.years
## 1  ( 1 ) " "        " "      " " " "        " "
## 2  ( 1 ) " "        "*"      " " " "        " "
## 3  ( 1 ) " "        "*"      " " " "        " "
## 4  ( 1 ) " "        "*"      " " " "        " "
## 5  ( 1 ) "*"        "*"      " " " "        " "
## 6  ( 1 ) "*"        "*"      " " " "        " "
## 7  ( 1 ) "*"        "*"      " " " "        " "
## 8  ( 1 ) "*"        "*"      " " " "        " "
##          thinness.5.9.years Income.composition.of.resources Schooling
## 1  ( 1 ) " "                " "                             " "
## 2  ( 1 ) " "                " "                             "*"
## 3  ( 1 ) " "                " "                             "*"
## 4  ( 1 ) " "                " "                             "*"
## 5  ( 1 ) " "                " "                             "*"
## 6  ( 1 ) " "                " "                             "*"
## 7  ( 1 ) " "                " "                             "*"
## 8  ( 1 ) " "                " "                             "*"
```

Selecting the best subset model based on highest Adjusted R^2 value

The model with highest R2 has parameters like schooling, HIV AIDS, Diphtheria, under five deaths, BMI, percentage expenditure, infant deaths, adult mortality.

```
adjr2_developing = which.max(nm2$adjr2)
adjr2_developing
```

```
## [1] 8
```

Results

For developed nations

Traditionally, features like demographic values like population, mortality rates and income compositions are the only factors considered influential for life expectancy. This analysis shows that along with these 3 factors, there are other factors that are influential as well in determining life expectancy of a country. This result for observed for both developed as well as developing nations.

Model is built with only those factors considered influential traditionally. The summary of the model shows that even though the regression is significant, it is able to explain just 25% variability in life expectancy.

Model with only traditional variables (Developed Country):

```
traditional_developed <- lm(Life.expectancy ~ Adult.Mortality + Population + Income.composition.of.resou
    data = data_developed)
summary(traditional_developed)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Population +
##     Income.composition.of.resources, data = data_developed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7580  -1.9649  -0.0767   1.5871   9.6595
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     8.167e+01  3.441e-01 237.304  < 2e-16 ***
## Adult.Mortality                -4.023e-02  3.143e-03 -12.798  < 2e-16 ***
## Population                      1.417e-08  1.116e-08   1.270 0.204723
## Income.composition.of.resources  9.692e-01  2.797e-01   3.465 0.000575 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.397 on 508 degrees of freedom
## Multiple R-squared:  0.2578, Adjusted R-squared:  0.2534
## F-statistic: 58.81 on 3 and 508 DF,  p-value: < 2.2e-16
```

Then, using features found to be significant using all the above methods, model is built and from its summary it can be clearly seen that including variables other than the traditionally chosen ones, the results are far better. The new model is able to explain 51% of the variation in life expectancy. This value is exactly double of the traditional method model. Including features like infant deaths, alcohol, percentage expenditure, BMI, under five deaths, Diphtheria, thinness has improved the model significantly.

Model with highest R^2 analysis:

```
final_developed <- lm(Life.expectancy ~ Adult.Mortality + infant.deaths + Alcohol +
    percentage.expenditure + BMI + under.five.deaths + Diphtheria + thinness..1.19.years,
    data = data_developed)
summary(final_developed)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
##     Alcohol + percentage.expenditure + BMI + under.five.deaths +
```

```
##     Diphtheria + thinness..1.19.years, data = data_developed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0485 -1.6860 -0.3929  1.0440  9.4681
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              8.463e+01  1.031e+00  82.075  < 2e-16 ***
## Adult.Mortality         -1.676e-02  2.910e-03  -5.761 1.46e-08 ***
## infant.deaths           -1.364e+00  4.000e-01  -3.411 0.000699 ***
## Alcohol                 -2.731e-01  3.375e-02  -8.090 4.49e-15 ***
## percentage.expenditure   1.967e-04  3.411e-05   5.766 1.42e-08 ***
## BMI                     -1.872e-02  7.184e-03  -2.606 0.009443 **
## under.five.deaths        1.094e+00  3.408e-01   3.210 0.001410 **
## Diphtheria               2.267e-02  9.885e-03   2.293 0.022260 *
## thinness..1.19.years    -2.423e+00  1.938e-01 -12.503  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.733 on 503 degrees of freedom
## Multiple R-squared:  0.5242, Adjusted R-squared:  0.5166
## F-statistic: 69.27 on 8 and 503 DF,  p-value: < 2.2e-16
```

For developing nations

The model built with just the features from the traditional approach like mortality, population and income composition is able to explain just 50% of the variability in life expectancy.

Model with only traditional variables (Developing Country):

```
traditional_developing <- lm(Life.expectancy ~ Adult.Mortality + Population + Income.composition.of.res
    data = data_developing)
summary(traditional_developing)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + Population +
##     Income.composition.of.resources, data = data_developing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.763  -2.441   1.156   3.760  24.203
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     7.176e+01  3.264e-01 219.871   <2e-16 ***
## Adult.Mortality                -4.150e-02  1.061e-03 -39.118   <2e-16 ***
## Population                     -5.261e-09  2.198e-09  -2.394   0.0167 *
## Income.composition.of.resources  5.891e+00  3.513e-01  16.769   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.397 on 2412 degrees of freedom
```

```
## Multiple R-squared:  0.4962, Adjusted R-squared:  0.4955
## F-statistic: 791.8 on 3 and 2412 DF,  p-value: < 2.2e-16
```

When additional predictors like infant.deaths, percentage.expenditure, Schooling, BMI, under.five.deaths, Diphtheria and HIV.AIDS were included in the model, the results improve significantly. This new model is able to explain 75% of the variability. This is huge improvement over the traditional approach. Therefore, this is an indication of a good model.

Model with highest R^2 analysis (Developing Country):

```
final_developing <- lm(Life.expectancy ~ Adult.Mortality + infant.deaths + percentage.expenditure +
    Schooling + BMI + under.five.deaths + Diphtheria + HIV.AIDS, data = data_developing)
summary(final_developing)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + infant.deaths +
##     percentage.expenditure + Schooling + BMI + under.five.deaths +
##     Diphtheria + HIV.AIDS, data = data_developing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.9740  -2.5569   0.0239   2.5986  22.3040
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            58.9523142  0.4504796  130.87   <2e-16 ***
## Adult.Mortality        -0.0190058  0.0009022  -21.07   <2e-16 ***
## infant.deaths           0.1168309  0.0089885   13.00   <2e-16 ***
## percentage.expenditure  0.0012202  0.0001151   10.60   <2e-16 ***
## Schooling               0.4894724  0.0295521   16.56   <2e-16 ***
## BMI                     0.0857047  0.0056815   15.09   <2e-16 ***
## under.five.deaths      -0.0884421  0.0066374  -13.32   <2e-16 ***
## Diphtheria              0.0543058  0.0040040   13.56   <2e-16 ***
## HIV.AIDS               -0.5019348  0.0194197  -25.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.493 on 2407 degrees of freedom
## Multiple R-squared:  0.752,  Adjusted R-squared:  0.7512
## F-statistic: 912.3 on 8 and 2407 DF,  p-value: < 2.2e-16
```

Factors affecting Life Expectancy

The influential factors in determining the life expectancy in developed nations are:

1. Adult Mortality
2. Infant Deaths
3. Alcohol
4. Percentage expenditure - percentage expenditure on health as percent of GDP
5. BMI
6. Under five deaths
7. Diphtheria - Percentage of Diphtheria immunization coverage among 1 year-olds
8. thinness 1-19 years

The linear regression model used to predict life expectancy for developed nations is:

Life expectancy = 8.463 - 0.0167 Adult.mortality - 1.364 infant.deaths - 0.273 Alcohol + 0.001967 percentage.expenditure - 0.01872 BMI - 1.094 under.five.deaths + 0.02267 Diphtheria - 2.423 thinness.1.19.years

Similarly, influential factors in case of developing nations are:

1. Adult Mortality
2. Infant Deaths
3. Percentage expenditure - percentage expenditure on health as percent of GDP
4. Schooling
5. BMI
6. Under five deaths
7. Diphtheria - Percentage of Diphtheria immunization coverage among 1 year olds
8. thinness 1-19 years

The linear regression model used to predict life expectancy for developing nations is:

Life expectancy = 58.95 - 0.019 Adult.mortality - 0.1168 infant.deaths + 0.0012 percentage.expenditure - 0.01872 BMI + 0.4894 Schooling + 0.85 BMI - 0.008 under.five.deaths+ 0.0543 Diphtheria - 0.0501 thinness.1.19.years

Therefore, new predictors that improved model efficiency common across both types of countries are infant deaths, percentage expenditure, BMI, under five deaths, Diphtheria and thinness 1-19 years.

Impact of Healthcare Expenditure on Life Expectancy

Healthcare expenditure has an impact on life expectancy. The positive coefficient implies healthcare expenditure has a positive relationship with life expectancy. Therefore, as expenditure on healthcare increases, life expectancy of that country increases. This is true for both developed and developing nations.

Impact of Alcohol Consumption on Life Expectancy

An interesting observation was made when it comes to alcohol consumption. This is a predictor only for developed nations and not for developing nations. It has a negative relation with life expectancy.

Impact of Immunization Coverage on Life Expectancy

Positive regression coefficient for Diphtheria immunization coverage implies positive relation between immunization coverage and life expectancy. So as immunization increases life expectancy increases as well.

Conclusion

Life expectancy of a country can be determined using various features other than those used traditionally. It was observed that model built using just the traditional predictors was resulting in a weak model which was able to account for just 25% (for developed nations) or 50% (for developing nations) of the variability in life expectancy. But the model improved significantly when other features like healthcare expenditure, alcohol consumption, immunization coverage where included as predictors.

Both healthcare spend and immunization of children have positive relation with life expectancy. Logically, this relationship makes perfect sense as well. If a country has better vaccination rates and higher spend on healthcare, people will be immune against variety of diseases and they also will be privy to good health care facilities thus resulting in increased life expectancy.

In developed nations, alcohol consumption has negative impact on life expectancy. Logically speaking, heavy alcohol consumption leads to various ailments some of which result in pre-mature deaths, can also result in accidents and so on. Therefore, results from the model make sense.

Multiple linear regression model was satisfactorily able to explain influential factors for life expectancy.