Regression Analysis for Life Expectancy Dataset from WHO

Kyr Nastahunin, Boshika Tara, Liang Wu

2022-08-01

Introduction

The goal of this analysis is to understand Life Expectancy in different developed and developing countries and how social, economic, and environmental factors in these countries affect life expectancy.

Some of the questions, we have explored in this study are:

1. Major predictors of life expectancy
2. Life expectancy in developed countries vs developing countries
3. Effects of health and social factors like immunization, or schooling on LE

Goal is to build a linear regression model that will predict the life expectancy at birth in different countries based on various economic and cultural factors within that country. We want to build a lean model, where we can narrow down to the most salient factors, that are most important in predicting Life Expectancy

We will use a dataset found on Kaggle https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who, created by WHO (World Health Organization).

Methods & Results

Data Cleaning

Let's load the dataset and inspect the first 10 entries.

```
data = read.csv("Life Expectancy Data.csv")
head(data, 5)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
## 2 Afghanistan 2014 Developing            59.9             271            64
## 3 Afghanistan 2013 Developing            59.9             268            66
## 4 Afghanistan 2012 Developing            59.5             272            69
## 5 Afghanistan 2011 Developing            59.2             275            71
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01                71.280          65    1154 19.1                83
## 2    0.01                73.524          62     492 18.6                86
## 3    0.01                73.219          64     430 18.1                89
## 4    0.01                78.184          67    2787 17.6                93
## 5    0.01                 7.097          68    3013 17.2                97
```

```
##   Polio Total.expenditure Diphtheria HIV.AIDS    GDP Population
## 1     6              8.16         65      0.1 584.26   33736494
## 2    58              8.18         62      0.1 612.70     327582
## 3    62              8.13         64      0.1 631.74   31731688
## 4    67              8.52         67      0.1 669.96    3696958
## 5    68              7.87         68      0.1  63.54    2978599
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                 17.2               17.3                           0.479
## 2                 17.5               17.5                           0.476
## 3                 17.7               17.7                           0.470
## 4                 17.9               18.0                           0.463
## 5                 18.2               18.2                           0.454
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
```

We see that the dataset includes data over several years, which will get in the way when we're building our model. We will keep the latest year which in this case is 2015. Additionally each observation contains the name of the country. We will not need the country names for the purposes of this project, so we will drop this column.

```
data = data[data$Year == 2015, ]
data = subset(data, select = -c(Country, Year))
```

```
head(data, 5)
```

```
##        Status Life.expectancy Adult.Mortality infant.deaths Alcohol
## 1  Developing            65.0             263            62    0.01
## 17 Developing            77.8              74             0    4.60
## 33 Developing            75.6              19            21      NA
## 49 Developing            52.4             335            66      NA
## 65 Developing            76.4              13             0      NA
##    percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths Polio
## 1                   71.28         65    1154 19.1                83     6
## 17                 364.98         99       0 58.0                 0    99
## 33                   0.00         95      63 59.5                24    95
## 49                   0.00         64     118 23.3                98     7
## 65                   0.00         99       0 47.7                 0    86
##    Total.expenditure Diphtheria HIV.AIDS     GDP Population
## 1               8.16         65      0.1   584.3   33736494
## 17              6.00         99      0.1  3954.2      28873
## 33                NA         95      0.1  4132.8   39871528
## 49                NA         64      1.9  3695.8    2785935
## 65                NA         99      0.2 13567.0         NA
##    thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                  17.2               17.3                           0.479
## 17                  1.2                1.3                           0.762
## 33                  6.0                5.8                           0.743
## 49                  8.3                8.2                           0.531
## 65                  3.3                3.3                           0.784
```

```
##    Schooling
## 1       10.1
## 17      14.2
## 33      14.4
## 49      11.4
## 65      13.9
```

Looking at the data again we notice that the `Alcohol` and `Total.expenditure` columns contain too many `NA` values, so it will be better to drop them. We can also notice that the `percentage.expenditure` has too many 0 values.

```
data = subset(data, select = -c(Alcohol, Total.expenditure, percentage.expenditure))
head(data, 5)
```

```
##        Status Life.expectancy Adult.Mortality infant.deaths Hepatitis.B Measles
## 1  Developing            65.0             263            62          65    1154
## 17 Developing            77.8              74             0          99       0
## 33 Developing            75.6              19            21          95      63
## 49 Developing            52.4             335            66          64     118
## 65 Developing            76.4              13             0          99       0
##     BMI under.five.deaths Polio Diphtheria HIV.AIDS     GDP Population
## 1  19.1                83     6         65      0.1   584.3   33736494
## 17 58.0                 0    99         99      0.1  3954.2      28873
## 33 59.5                24    95         95      0.1  4132.8   39871528
## 49 23.3                98     7         64      1.9  3695.8    2785935
## 65 47.7                 0    86         99      0.2 13567.0         NA
##    thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                  17.2               17.3                           0.479
## 17                  1.2                1.3                           0.762
## 33                  6.0                5.8                           0.743
## 49                  8.3                8.2                           0.531
## 65                  3.3                3.3                           0.784
##    Schooling
## 1       10.1
## 17      14.2
## 33      14.4
## 49      11.4
## 65      13.9
```

Now let's drop all the observations that have `NA` left in them to avoid any errors when building models.

```
data = na.omit(data)
nrow(data)
```

```
## [1] 130
```

Let's check if our categorical variables are factors.

```
is.factor(data$Status)
```

```
## [1] FALSE
```

```
data$Status = as.factor(data$Status)
```

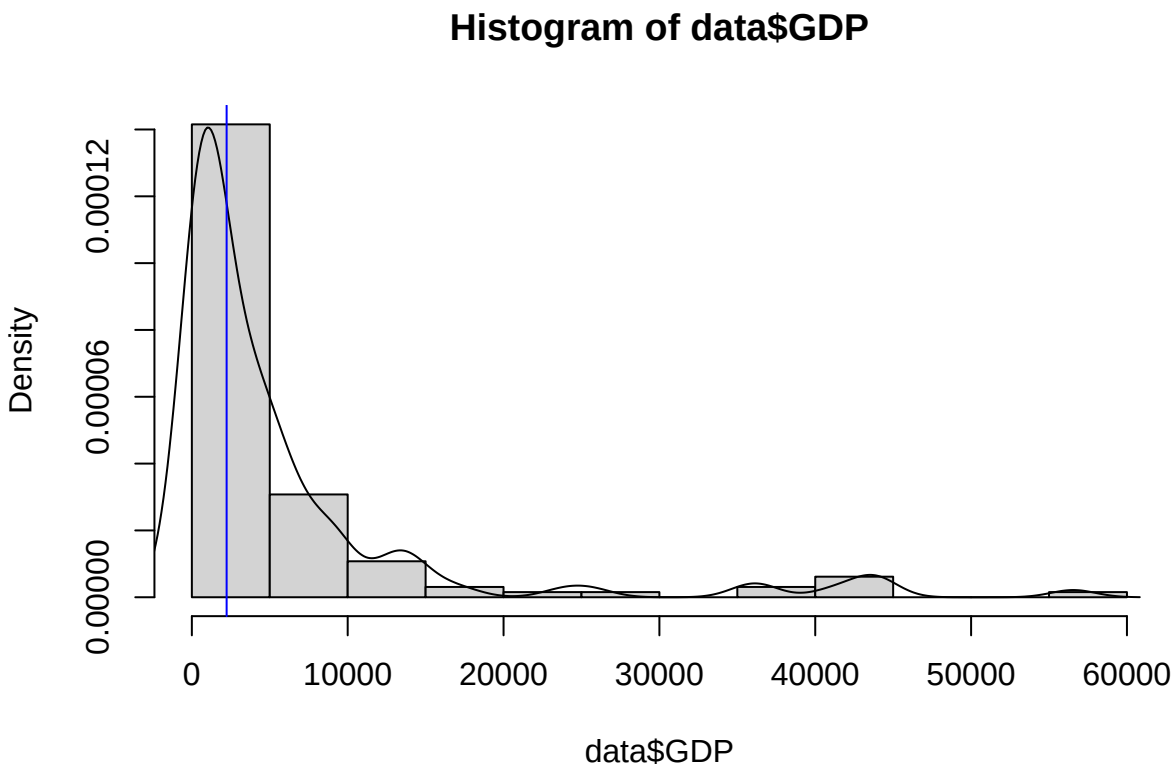
Exploratory Data Analysis

Now that we have our dataset cleaned, and prepare let's perform some basic data analysis. The first thing I will look at is how the features are distributed. To do this we would use density plots. Specifically, we would like to look at the life expectancy, GDP, BMI, and schooling. The World Health Organization describes these as the most important factors in determining health of an individual.

```
hist(data$Life.expectancy, prob = TRUE)
lines(density(data$Life.expectancy))
```

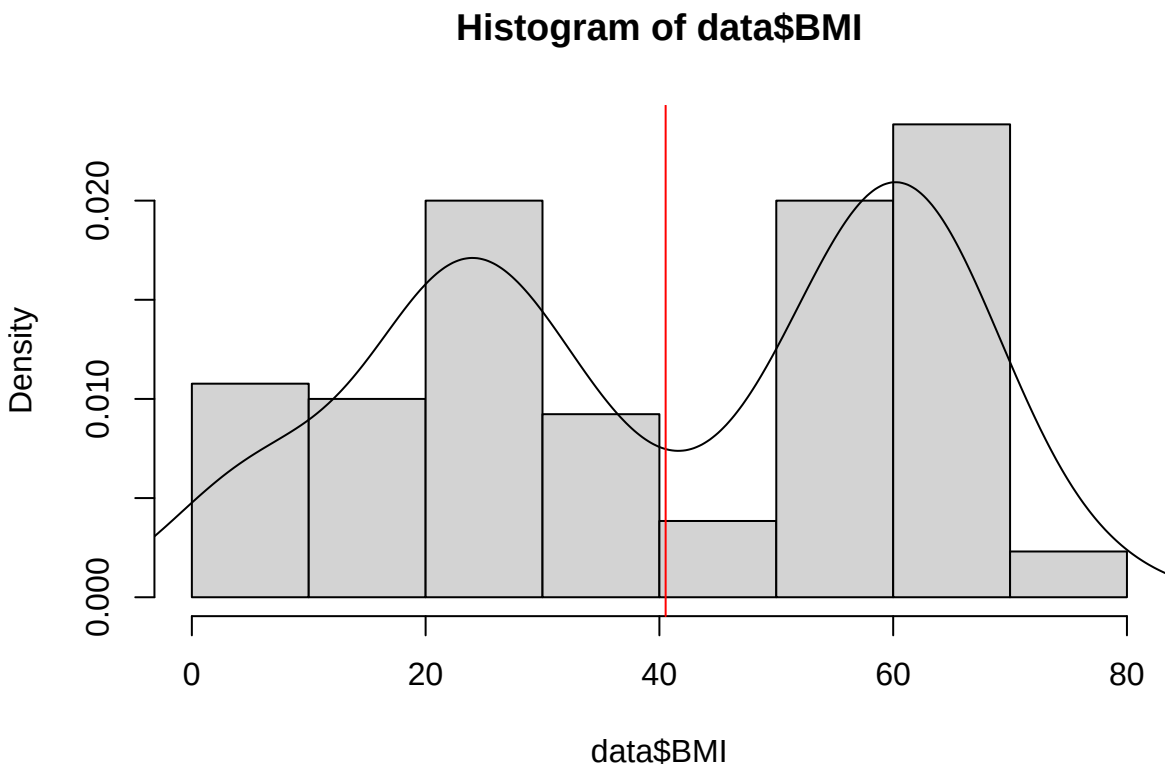
We can see that most countries have a life expectancy of greater than 65 yrs old. Next, we will look at GDP. Since incomes are often skewed. We will use the median as our measure of spread (noted by the line).

```
med = median(data$GDP)
hist(data$GDP, prob = TRUE)
lines(density(data$GDP))
abline(v = med, col="blue")
```



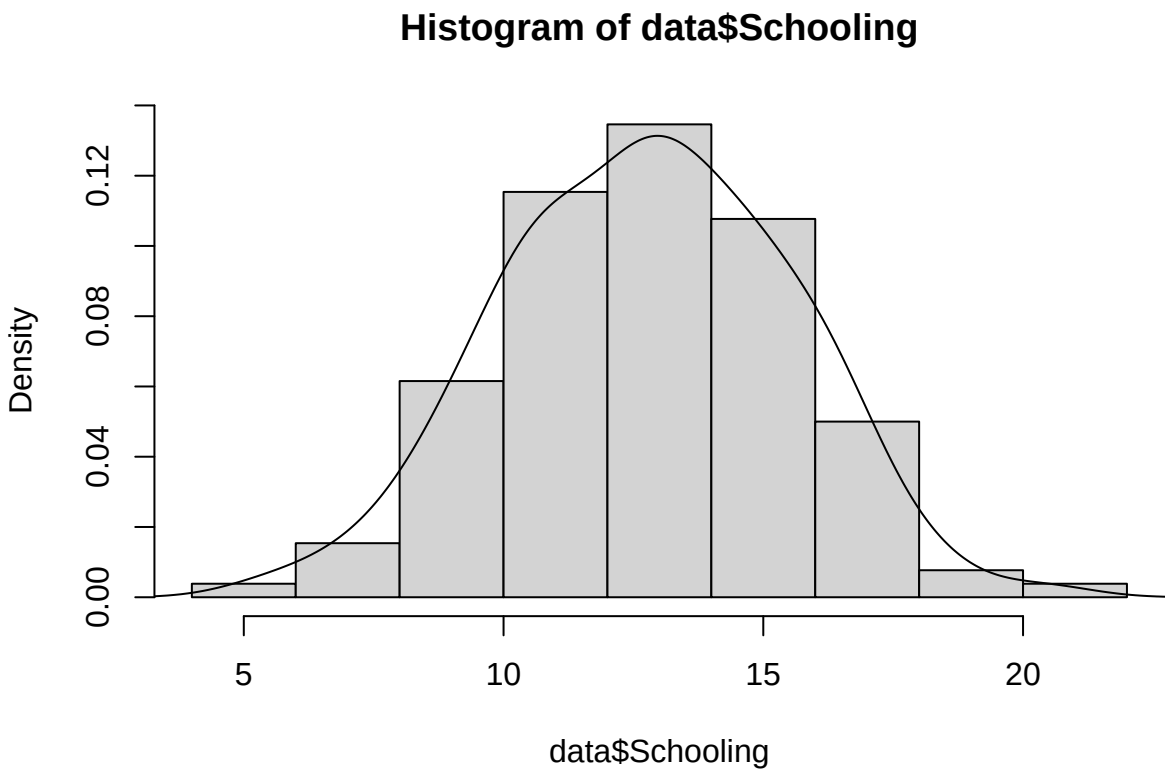
Looks like the population of most countries make below $5,000 a year on average. Next we will take a look at BMI.

```
md = mean(data$BMI)
hist(data$BMI, prob = TRUE)
lines(density(data$BMI))
abline(v = md, col="red")
```



Most of the data lie between 20-30 BMI and then 50-70 BMI. The mean is at ~40 denoted by the red line. Next we will take a look at the education.

```
hist(data$Schooling, prob = TRUE)
lines(density(data$Schooling))
```

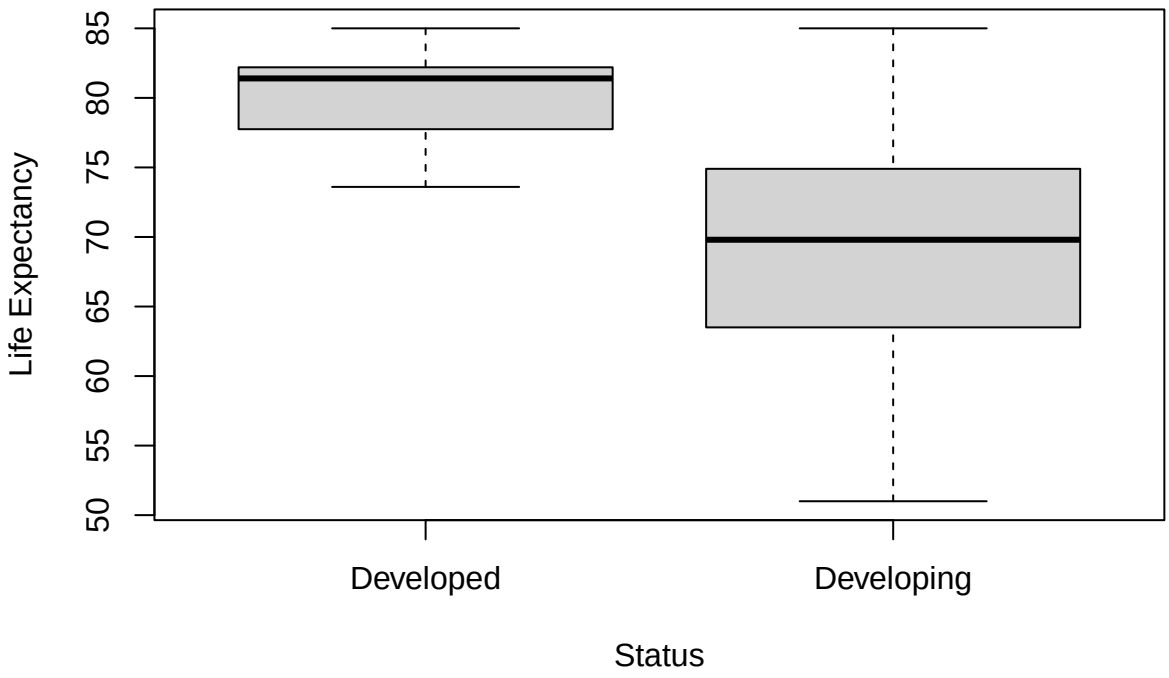


Most individuals have a decent number of years in education. Most data lie between 10-15 years of schooling.

Correlations

Now that we have some basic understanding of our data. Let's take a look at the correlations between the features. There are several ways to do this, but we prefer to use the correlation matrix. Since `Status` is a categorical field we will visualize it separately.

```
plot(data$Status,data$Life.expectancy, xlab ="Status", ylab="Life Expectancy")
```



Seems like developed countries has a higher life expectancy and a smaller spread compared to developing countries. We do see a strong correlation here. Developed countries tend to have a higher life expectancy. Now we will create a correlation matrix for the reminder features.

```
dataCorr = subset(data, select = -c(Status))
round(cor(dataCorr),2)
```

```
##                                 Life.expectancy Adult.Mortality infant.deaths
## Life.expectancy                            1.00           -0.73         -0.21
## Adult.Mortality                           -0.73            1.00          0.15
## infant.deaths                             -0.21            0.15          1.00
## Hepatitis.B                                0.37           -0.13         -0.08
## Measles                                   -0.05            0.03          0.82
## BMI                                        0.54           -0.35         -0.21
## under.five.deaths                         -0.24            0.18          0.99
## Polio                                      0.49           -0.30         -0.12
## Diphtheria                                 0.47           -0.23         -0.11
## HIV.AIDS                                  -0.62            0.63          0.07
## GDP                                        0.49           -0.31         -0.12
## Population                                -0.03            0.03          0.27
## thinness..1.19.years                      -0.46            0.25          0.56
## thinness.5.9.years                        -0.45            0.26          0.56
## Income.composition.of.resources            0.90           -0.59         -0.20
## Schooling                                  0.81           -0.47         -0.22
##                                 Hepatitis.B Measles   BMI under.five.deaths
## Life.expectancy                        0.37   -0.05  0.54             -0.24
```

```
## Adult.Mortality                       -0.13    0.03 -0.35             0.18
## infant.deaths                         -0.08    0.82 -0.21             0.99
## Hepatitis.B                            1.00    0.03  0.15            -0.09
## Measles                                0.03    1.00 -0.13             0.79
## BMI                                    0.15   -0.13  1.00            -0.22
## under.five.deaths                     -0.09    0.79 -0.22             1.00
## Polio                                  0.50   -0.01  0.20            -0.14
## Diphtheria                             0.90    0.02  0.17            -0.13
## HIV.AIDS                              -0.34   -0.04 -0.27             0.10
## GDP                                    0.09   -0.07  0.39            -0.12
## Population                            -0.05    0.13  0.01             0.31
## thinness..1.19.years                  -0.04    0.38 -0.49             0.55
## thinness.5.9.years                    -0.09    0.37 -0.51             0.54
## Income.composition.of.resources        0.28   -0.06  0.62            -0.22
## Schooling                              0.30   -0.06  0.61            -0.24
##                                  Polio Diphtheria HIV.AIDS   GDP Population
## Life.expectancy                   0.49      0.47    -0.62  0.49      -0.03
## Adult.Mortality                  -0.30     -0.23     0.63 -0.31       0.03
## infant.deaths                    -0.12     -0.11     0.07 -0.12       0.27
## Hepatitis.B                       0.50      0.90    -0.34  0.09      -0.05
## Measles                          -0.01      0.02    -0.04 -0.07       0.13
## BMI                               0.20      0.17    -0.27  0.39       0.01
## under.five.deaths                -0.14     -0.13     0.10 -0.12       0.31
## Polio                             1.00      0.58    -0.38  0.22      -0.23
## Diphtheria                        0.58      1.00    -0.41  0.20      -0.05
## HIV.AIDS                         -0.38     -0.41     1.00 -0.19       0.02
## GDP                               0.22      0.20    -0.19  1.00       0.07
## Population                       -0.23     -0.05     0.02  0.07       1.00
## thinness..1.19.years             -0.18     -0.08     0.17 -0.29      -0.01
## thinness.5.9.years               -0.18     -0.13     0.15 -0.29      -0.02
## Income.composition.of.resources   0.44      0.40    -0.48  0.57       0.03
## Schooling                         0.39      0.39    -0.39  0.57       0.05
##                                  thinness..1.19.years thinness.5.9.years
## Life.expectancy                                 -0.46              -0.45
## Adult.Mortality                                  0.25               0.26
## infant.deaths                                    0.56               0.56
## Hepatitis.B                                     -0.04              -0.09
## Measles                                          0.38               0.37
## BMI                                             -0.49              -0.51
## under.five.deaths                                0.55               0.54
## Polio                                           -0.18              -0.18
## Diphtheria                                      -0.08              -0.13
## HIV.AIDS                                         0.17               0.15
## GDP                                             -0.29              -0.29
## Population                                      -0.01              -0.02
## thinness..1.19.years                             1.00               0.97
## thinness.5.9.years                               0.97               1.00
## Income.composition.of.resources                 -0.51              -0.50
## Schooling                                       -0.50              -0.49
##                                  Income.composition.of.resources Schooling
## Life.expectancy                                             0.90      0.81
## Adult.Mortality                                            -0.59     -0.47
## infant.deaths                                              -0.20     -0.22
## Hepatitis.B                                                 0.28      0.30
```

```
## Measles                                -0.06     -0.06
## BMI                                     0.62      0.61
## under.five.deaths                      -0.22     -0.24
## Polio                                   0.44      0.39
## Diphtheria                              0.40      0.39
## HIV.AIDS                               -0.48     -0.39
## GDP                                     0.57      0.57
## Population                              0.03      0.05
## thinness..1.19.years                   -0.51     -0.50
## thinness.5.9.years                     -0.50     -0.49
## Income.composition.of.resources         1.00      0.92
## Schooling                               0.92      1.00
```

As you can see, Income composition of resources, GDP and education are positively correlated to Life expectancy, while many of the health features such as (BMI, thinness, and diseases) have a negative correlation to life expectancy. Surprisingly, population has little to do with life expectancy.

Prepping data for modeling

We will split the data into train and test data in order to verify our results.

```
set.seed(42)

train_size = 100
trn_idx = sample(nrow(data), train_size)
train = data[trn_idx, ]
test = data[-trn_idx, ]
test = test[-c(19), ] # remove a data point that is causing additional trouble
```

Regression Modeling

First, let's define the measure we will use to evaluate the success of our models on the test data. We will evaluate the models using the percent error function defined below. The smaller the error, the better the model performs.

$$\frac{1}{n}\sum_{i}\frac{|\text{predicted}_i - \text{actual}_i|}{\text{predicted}_i} \times 100$$

```
percent_error = function(model, test, predictor){
  prediction = predict(model, newdata = test)
  (1 / nrow(test)) * sum((abs(prediction - predictor) / prediction) * 100)
}
```

Let's fit a simple additive model and see how it performs.

```
additive = lm(Life.expectancy ~ ., data = train)
summary(additive)$coef
```

```
##                                Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)                   5.347e+01  3.047e+00 17.54471 1.165e-29
```

```
## StatusDeveloping                -2.508e-02  9.177e-01 -0.02733 9.783e-01
## Adult.Mortality                 -2.716e-02  4.667e-03 -5.81964 1.073e-07
## infant.deaths                    6.143e-02  3.954e-02  1.55358 1.241e-01
## Hepatitis.B                      1.600e-02  3.175e-02  0.50380 6.157e-01
## Measles                         -2.056e-05  6.103e-05 -0.33685 7.371e-01
## BMI                             -1.551e-02  1.803e-02 -0.86000 3.923e-01
## under.five.deaths               -4.735e-02  2.931e-02 -1.61516 1.101e-01
## Polio                            7.658e-03  1.414e-02  0.54157 5.896e-01
## Diphtheria                       1.965e-02  3.651e-02  0.53827 5.918e-01
## HIV.AIDS                        -3.682e-01  2.580e-01 -1.42693 1.574e-01
## GDP                             -1.533e-05  3.613e-05 -0.42435 6.724e-01
## Population                       8.971e-09  1.787e-08  0.50188 6.171e-01
## thinness..1.19.years            -2.754e+00  1.680e+00 -1.63933 1.049e-01
## thinness.5.9.years               2.564e+00  1.655e+00  1.54917 1.251e-01
## Income.composition.of.resources  3.105e+01  5.919e+00  5.24632 1.167e-06
## Schooling                       -5.194e-02  2.840e-01 -0.18288 8.553e-01
```

We can see right away the some of the predictors fail the $H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$ test. We will use the techniques we learned in lectures to find which predictors we should keep.

Let's see how this basic model performs on our test data.

```
percent_error(additive, test, test$Life.expectancy)
```

```
## [1] 4.229
```

The observed error is already small enough to call our model successfull, however we know that it can be even better.

Let's eliminate unnecessary predictors from our model using `backward AIC` built into `R`. We will compare this model to the full additive model that we created in the previous step.

```
additive_aic = step(additive, direction = "backward", trace = 0)
summary(additive_aic)$coef
```

```
##                                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                     53.64303   2.538491  21.132 7.289e-37
## Adult.Mortality                 -0.02753   0.004273  -6.444 5.442e-09
## infant.deaths                    0.04282   0.024414   1.754 8.283e-02
## Hepatitis.B                      0.03446   0.012771   2.698 8.308e-03
## under.five.deaths               -0.03382   0.018726  -1.806 7.417e-02
## HIV.AIDS                        -0.39031   0.244900  -1.594 1.145e-01
## thinness..1.19.years            -2.80748   1.552769  -1.808 7.390e-02
## thinness.5.9.years               2.65085   1.523150   1.740 8.517e-02
## Income.composition.of.resources 29.81660   2.738025  10.890 3.594e-18
```

The step function selected the most relative predictors for us. We can see that all of the `p` values are lower than 0.1, which is what we want.

Now compare this model to the full additive model that we created in the previous step.

```
anova(additive_aic, additive)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ Adult.Mortality + infant.deaths + Hepatitis.B +
##     under.five.deaths + HIV.AIDS + thinness..1.19.years + thinness.5.9.years +
##     Income.composition.of.resources
## Model 2: Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
##     Hepatitis.B + Measles + BMI + under.five.deaths + Polio +
##     Diphtheria + HIV.AIDS + GDP + Population + thinness..1.19.years +
##     thinness.5.9.years + Income.composition.of.resources + Schooling
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1     91 627
## 2     83 609  8      17.9 0.31   0.96
```

The model picked by the `step` function is much better than the initial full additive model that we have created. Let's see how it performs on the test data.

```
percent_error(additive_aic, test, test$Life.expectancy)
```

```
## [1] 4.105
```

This is a slight improvement compared to the initial model we made.

We want to try creating an interaction model from the predictors that the `step` function picked for us in the previous step. This model will include all possible 2-way interactions between the predictors.

```
interaction = lm(Life.expectancy ~ (Adult.Mortality + infant.deaths +
                  Hepatitis.B + under.five.deaths + HIV.AIDS + thinness..1.19.years + thinness.5.9.years
                  Income.composition.of.resources) ^ 2, data = train)
```

Once again we have a lot of predictors that become unnecessary now that we added interactions, we will take care of them soon, but first let's evaluate the performance of this model on the test data.

```
percent_error(interaction, test, test$Life.expectancy)
```

```
## [1] 3.547
```

This model's performance on the test data is better than the performance of the additive model we picked before. Let's pick the best predictors from this model and compare it to the `additive_aic` model.

```
interaction_aic = step(interaction, direction = "backward", trace = 0)
anova(additive_aic, interaction_aic)
```

```
## Analysis of Variance Table
##
## Model 1: Life.expectancy ~ Adult.Mortality + infant.deaths + Hepatitis.B +
##     under.five.deaths + HIV.AIDS + thinness..1.19.years + thinness.5.9.years +
##     Income.composition.of.resources
## Model 2: Life.expectancy ~ Adult.Mortality + infant.deaths + Hepatitis.B +
##     under.five.deaths + HIV.AIDS + thinness..1.19.years + thinness.5.9.years +
##     Income.composition.of.resources + Adult.Mortality:Hepatitis.B +
##     Adult.Mortality:HIV.AIDS + Adult.Mortality:thinness..1.19.years +
```

```
##     Adult.Mortality:thinness.5.9.years + infant.deaths:Hepatitis.B +
##     infant.deaths:under.five.deaths + infant.deaths:thinness.5.9.years +
##     Hepatitis.B:HIV.AIDS + Hepatitis.B:thinness..1.19.years +
##     Hepatitis.B:thinness.5.9.years + Hepatitis.B:Income.composition.of.resources +
##     under.five.deaths:thinness.5.9.years + under.five.deaths:Income.composition.of.resources +
##     HIV.AIDS:thinness..1.19.years + HIV.AIDS:thinness.5.9.years +
##     HIV.AIDS:Income.composition.of.resources + thinness..1.19.years:thinness.5.9.years +
##     thinness..1.19.years:Income.composition.of.resources + thinness.5.9.years:Income.composition.of.
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     91 627
## 2     72 221 19       406 6.95 5.8e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `p-value` of the ANOVA test indicates that the `interaction` model performs much better on the training data than the `additive` model. Let's check the performance on the test data.

```
percent_error(interaction_aic, test, test$Life.expectancy)
```

```
## [1] 3.811
```

After using backward AIC on the interaction model the performance on the test data noticeably decreased. Let's try backward BIC instead and see if it performs the elimination better.

```
interaction_bic = step(interaction, direction = "backward", k = log(nrow(train)), trace = 0)
percent_error(interaction_bic, test, test$Life.expectancy)
```

```
## [1] 4.453
```

Looks like backward elimination with BIC made the interaction model worse than the initial additive model we have had, so we will disregard the BIC model.

Using Social and Health Determinants of Life Expectancy to build the model

We know from studies conducted, that health and social factors play an important role in determining Life Expectancy. Here we aim to fit multiple additive models to see how they affect life expectancy predictions, we fit models with different health and immunization indicating parameters.

```
select_factors <- lm(Life.expectancy ~ Adult.Mortality +  HIV.AIDS + Income.composition.of.resources, da
percent_error(select_factors, test, test$Life.expectancy)
```

```
## [1] 3.239
```

```
select_factors2 <- lm(Life.expectancy ~ Adult.Mortality +  Income.composition.of.resources, data = trai
percent_error(select_factors2, test, test$Life.expectancy)
```

```
## [1] 3.369
```

```
select_factors3 <- lm(Life.expectancy ~ Adult.Mortality + Income.composition.of.resources + HIV.AIDS + H
percent_error(select_factors3, test, test$Life.expectancy)
```

```
## [1] 3.064
```

```
select_factors4 <- lm(Life.expectancy ~ Adult.Mortality + Income.composition.of.resources + HIV.AIDS + H
percent_error(select_factors4, test, test$Life.expectancy)
```

```
## [1] 3.144
```

```
select_factors5 <- lm(Life.expectancy ~ Adult.Mortality + Income.composition.of.resources + HIV.AIDS + H
percent_error(select_factors5, test, test$Life.expectancy)
```

```
## [1] 3.141
```

Looks like one of our models performed noticeably better on the test data than all of the other models we have tried so far. Let's explore that model closer.

```
additive__factors3_aic = step(select_factors3, direction = "backward", trace = 0)
summary(additive__factors3_aic)$coef
```

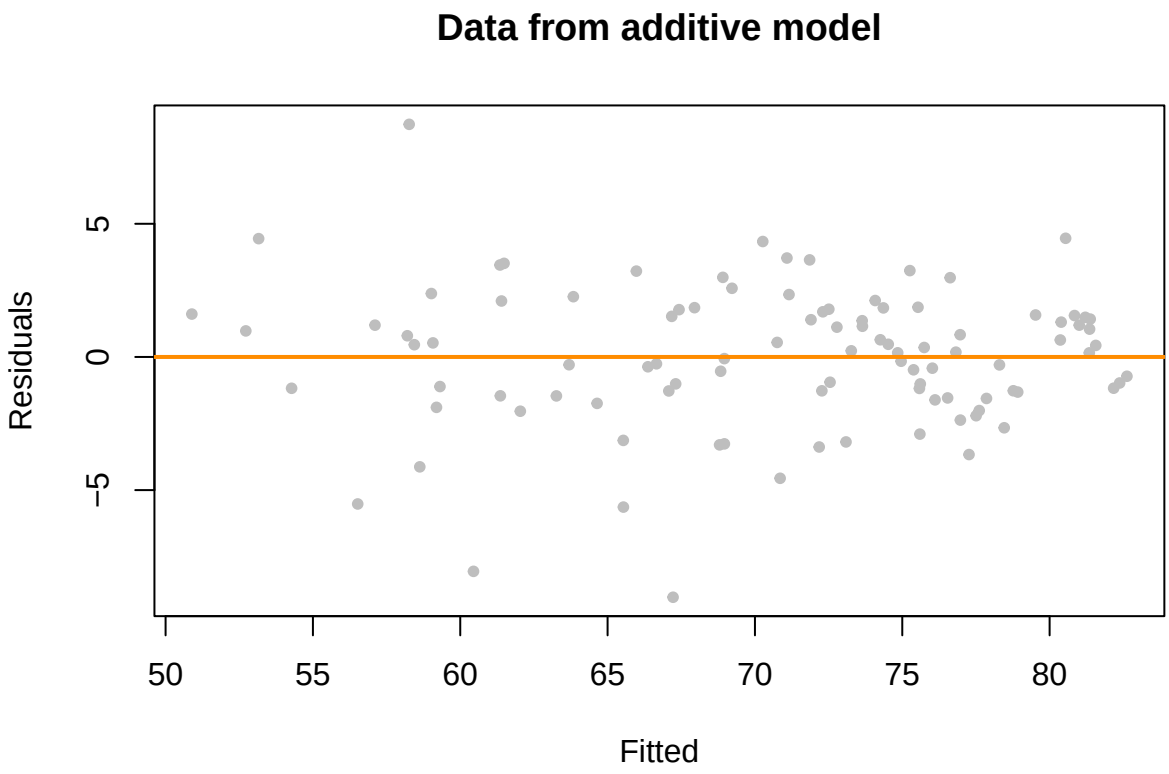
```
##                                 Estimate Std. Error t value  Pr(>|t|)
## (Intercept)                     50.10480   2.157067  23.228 5.904e-41
## Adult.Mortality                 -0.02576   0.004245  -6.069 2.619e-08
## Income.composition.of.resources 33.47151   2.328814  14.373 1.454e-25
## HIV.AIDS                        -0.49373   0.246445  -2.003 4.798e-02
## Hepatitis.B                      0.03344   0.012802   2.613 1.045e-02
```

This model has only 4 predictors out of the 18 available, and yet it performs best.

Model Diagnostics

Now that we settled with a preferred model for predicting the `Life Expectancy` we need to make sure that our model does not violate any assumption of the regression model. Let's begin by plotting fitted vs residuals of the `interactive` model.

```
plot(fitted(select_factors3), resid(select_factors3), col = "grey", pch = 20,
     xlab = "Fitted", ylab = "Residuals", main = "Data from additive model")
abline(h = 0, col = "darkorange", lwd = 2)
```



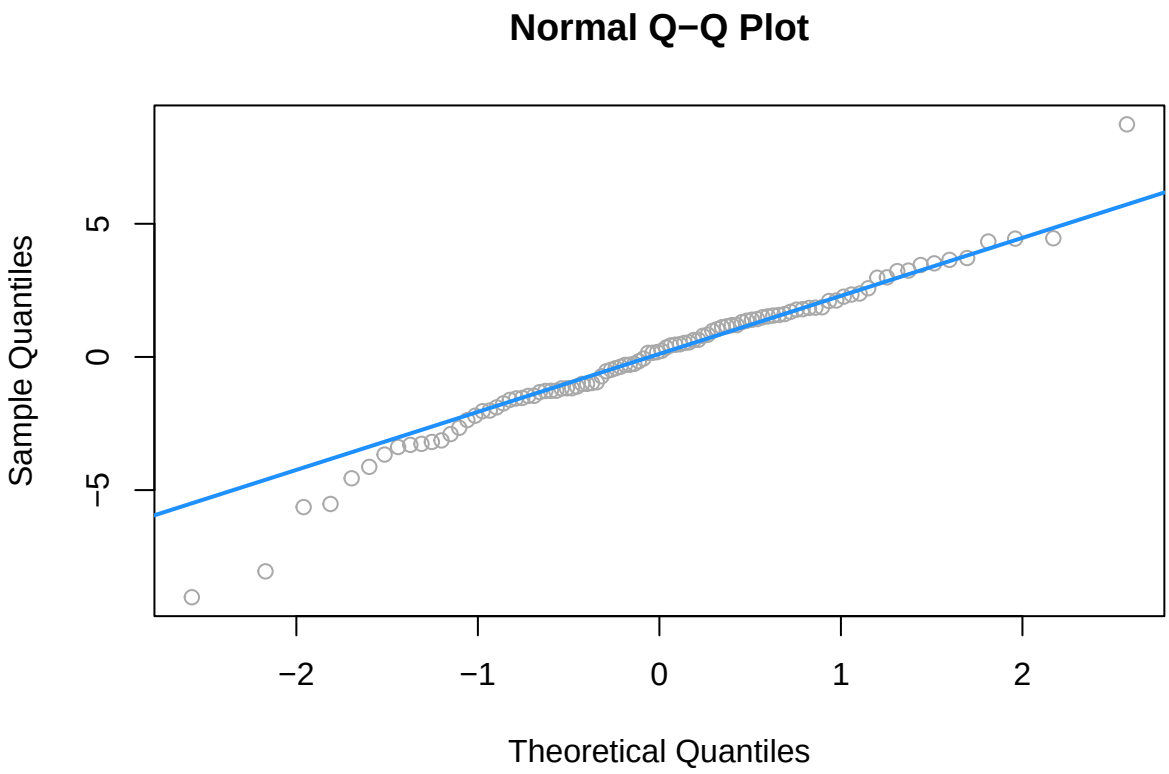
The plot does not suggest there are any violations within the model. However, let's run studentized Breusch-Pagan test on the model to verify the constant variance assumption.

```
library(lmtest)
bptest(select_factors3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  select_factors3
## BP = 18, df = 4, p-value = 0.001
```

We see a small p-value, so we can reject the null of homoscedasticity. The constant variance assumption might be violated. However, we suspect that there are few influential points that make the p-value so small, and we should continue using this model. Let's create a Q-Q Plot to test other normality assumptions.

```
qqnorm(resid(select_factors3), main = "Normal Q-Q Plot", col = "darkgrey")
qqline(resid(select_factors3), col = "dodgerblue", lwd = 2)
```



The Normal Q-Q plot looks reasonable and does not indicate any violations within our model. In order to confirm that we will run the `Shapiro-Wilk` test on the residuals of this model

```
shapiro.test(resid(select_factors3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(select_factors3)
## W = 0.97, p-value = 0.02
```

The null hypothesis of the `Shapiro-Wilk` assumes the data were sampled from a normal distribution, thus a high p-value indicates we believe there is a good probability the data could have been sampled from a normal distribution.

Let's check for multicollinearity of our model. To do this we will use the condition number and eigen value. A condition number between 10 to 30 indicates there is a presence of multicollinearity. Values greater than 30 are problematic according to our source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6900425/#:~:text=Condition%20Number%20and%20Condition%20Index&text=The%20largest%20condition%20index%20is,multicollinearity%20is%20regarded%20as%20strong

```
library(olsrr)
round(ols_eigen_cindex(select_factors3),5)
```

```
##   Eigenvalue Condition Index intercept Adult.Mortality
```

```
## 1    3.97225           1.000   0.00095         0.00573
## 2    0.84718           2.165   0.00066         0.00924
## 3    0.12645           5.605   0.00021         0.55821
## 4    0.04368           9.536   0.02483         0.01420
## 5    0.01043          19.511   0.97335         0.41261
##   Income.composition.of.resources HIV.AIDS Hepatitis.B
## 1                         0.00157  0.00968     0.00347
## 2                         0.00362  0.32939     0.00594
## 3                         0.02562  0.64139     0.00714
## 4                         0.16443  0.01949     0.92910
## 5                         0.80476  0.00005     0.05436
```

Since our condition number is between 10-30 and we have small portions of variance ($< .5$), we can conclude that multicollinearity exist within our model but is not problematic.

Unusual Observations

Let's see how many observations in our train dataset have high leverage, as defined in the lectures during the class.

```
sum(hatvalues(select_factors3) > 2 * mean(hatvalues(select_factors3)))
```

```
## [1] 12
```

More than 10% of the observations are considered high leverage, which is undesirable, let's see how this affects influence.

```
cd_select_factors3 = cooks.distance(select_factors3)
sum(cd_select_factors3 > 4 / length(cd_select_factors3))
```

```
## [1] 7
```

There are 7 observations with high influence in our training dataset, let's inspect what effects it may have on the model.

As a reminder the accuracy of this model with the influential points is

```
percent_error(select_factors3, test, test$Life.expectancy)
```

```
## [1] 3.064
```

We will refit this model without the influential points and try again

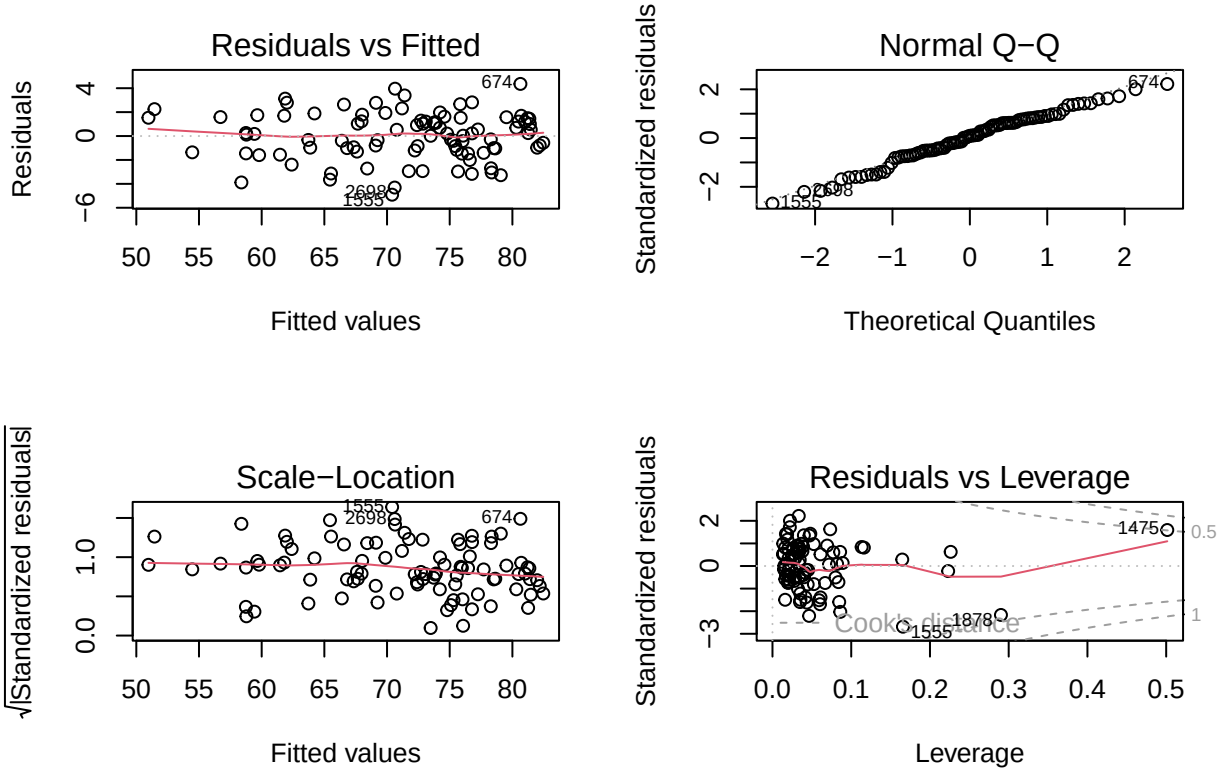
```
interaction_no_inf = lm(Life.expectancy ~ Adult.Mortality + Income.composition.of.resources + HIV.AIDS +
    subset = cd_select_factors3 <= 4 / length(cd_select_factors3))
percent_error(interaction_no_inf, test, test$Life.expectancy)
```

```
## [1] 3.433
```

Surprisingly, by getting rid of the infulential points the performance of the model on the test data actually decreased. This can be explained by the size of our model, which is very small, considering how complex the question of life expectancy is. However, getting rid of the influential points in other models did yield better results, they still weren't as good.

Let's plot the final diagnostics of the model without the influential points.

```
par(mfrow = c(2, 2))
plot(interaction_no_inf)
```



Those plots look a little better compared to what we have seen before, however there aren't any practical gains from that.

Outliers diagnostics

The STAT 420 book mentions that the standardized residuals greater than 2 in magnitude should only happen approximately 5 percent of the time. We will check the percentage of outliers in our model.

```
length(rstandard(select_factors3)[abs(rstandard(select_factors3)) > 2]) / length(rstandard(select_factor
```

```
## [1] 0.05
```

We observe a percentage similar to what we expect to see according to the book. Therefore we assume that we don't need to investigate the outliers any further.

Discussion

In this study, we have successfully explored the major predictors of life expectancy, and what effects health and social factors like immunization or schooling have on life expectancy in developed or developing countries. We have used the techniques learned in STAT 420 to pick the models, perform model diagnosis and evaluate the success of our models.

We can conclude that major predictors of life expectancy are Adult.Mortality, Income.composition.of.resources, HIV.AIDS, and Hepatitis.B. And each one of these feature have a direct correlation to our response variable (life expectancy). Interestingly, the status of the country (Developing vs Developed) was not picked as one of the predictors for the any of our successful models, probably due to collegiality with other predictors.

In our exploratory data analysis we have looked at the predictor and response variables and explored their values. Most importantly we have looked at how the Status of the country relates to it's life expectancy.

Regarding our final model selection, we used backwards AIC and BIC to select both our additive and interaction models, however our best performing model was picked by hand, looking at the collinearity matrix. This is an unexpected result, since we know that AIC and BIC use statistical methods to pick the best performing models.