Data Analysis of Life Expectancy in 2007

Yechen Cao, Luning Ding, Xianya Fu, Yuer Tang

1. Introduction

As we are experiencing a better lifestyle now, we do care more about how well we can maintain that quality of life and how long we can live to enjoy our lives. For this report, we want to study how the status of developed/developing countries, BMI, alcohol consumption, general government expenditure on health, thinness among children and adolescents for Age 10 to 19, and Hepatitis B coverage affected the life expectancy in 2007 to understand different factors contributing to health problems globally.

This dataset originates from the Global Health Observatory (GHO) data repository, under the auspices of the World Health Organization (WHO), a research project conducted from 2000 to 2015 across 193 countries. The choice to focus on data from the year 2007 was strategic, as it contains the fewest missing values (NAs) and may offer a more comprehensive and reliable basis for analysis within the specified timeframe.

	Life.expectancy <dbl>	Status <int>	Alcohol <dbl>	Hepatitis.B <int>	BMI <dbl>	Total.expenditure <dbl>	Thinness <dbl>
1	58.1	1	0.03	64	15.7	8.33	18.8
2	75.3	1	5.61	99	52.6	5.87	1.6
3	74.1	1	0.46	91	51.8	4.20	6.0
4	48.7	1	7.07	69	19.3	3.84	9.5
5	75.2	1	8.27	98	43.2	4.69	3.4
6	75.4	1	8.41	9	58.6	6.66	1.0

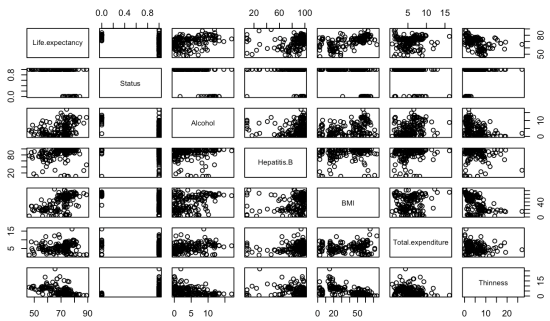
For this dataset, we chose Life.expectancy, which is the life expectancy in age, as our Y value. And we want to model a multiple linear regression: $Y = B0 + B1x1 + B2x2 + B3x3 + B4x4 + B5x5 + B6x6 + e$. X1 is the status, which is the developed or developing status of a country; x2 is Alcohol, which is recorded per capita (15+) consumption (in liters of pure alcohol); x3 is Hepatitis.B, which is the Hepatitis B immunization coverage among 1-year-olds; x4 is BMI, which is the average Body Mass Index of entire population; x5 is Total.expenditure, which is the general government expenditure on health as a percentage of total government expenditure (%); x6 is Thinness, which is the prevalence of thinness among children and adolescents for Age 10 to 19 (%).

This report looks closely at the basic setup of our raw data, providing a detailed analysis of both full models and transformed models. It evaluates the goodness of fit across various variance selection methods to identify the most suitable final model to explain our dataset. Finally, we'll discuss how we can use our findings in real life situations and their limitations.

2. Data Description

Variable correlation:

	Life.expectancy	Status	Alcohol	Hepatitis.B	BMI	Total.expenditure	Thinness
Life.expectancy	1.0000	-0.3885	0.3656	0.2439	0.5555	0.0909	-0.4846
Status	-0.3885	1.0000	-0.5579	-0.1255	-0.3907	-0.2237	0.3221
Alcohol	0.3656	-0.5579	1.0000	0.0352	0.3709	0.2590	-0.4274
Hepatitis.B	0.2439	-0.1255	0.0352	1.0000	0.1886	0.1632	-0.1349
BMI	0.5555	-0.3907	0.3709	0.1886	1.0000	0.1532	-0.5678
Total.expenditure	0.0909	-0.2237	0.2590	0.1632	0.1532	1.0000	-0.2172
Thinness	-0.4846	0.3221	-0.4274	-0.1349	-0.5678	-0.2172	1.0000



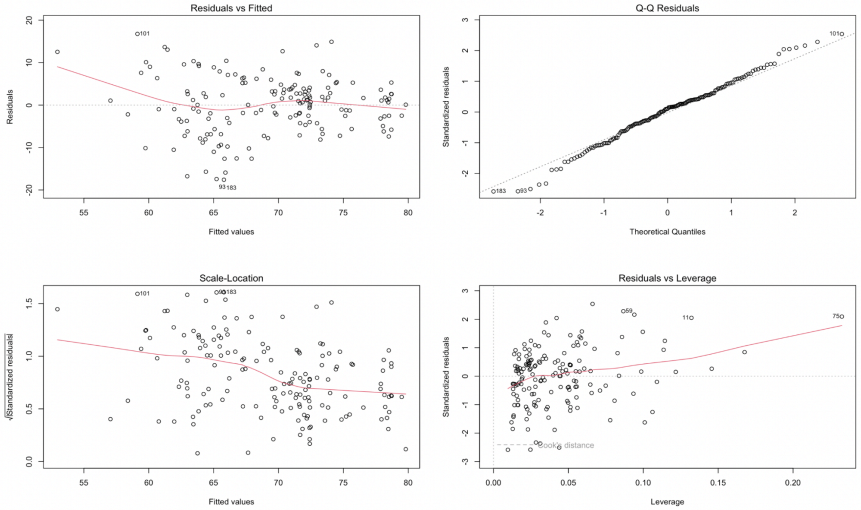
From the correlation table and graph, we can see that the majority of the predictor variables (X variables) are relatively independent from one another, with the exception of Thinness and BMI, which show a correlation of about 0.57. This level of correlation is notable but not exceptionally high.

Despite the general independence among the X variables, there is a potential problem with the low correlation between the dependent variable (Y) and the predictors. Only the correlation between life expectancy (Y) and BMI is higher than 0.5, suggesting that life expectancy may not have a strong linear relationship with most of the variables in their current form within the raw data. This observation shows that we might potentially need some transformation of the full model we intend to use to better explain the relationships within our dataset.

Full Model summary:

```
Call:
lm(formula = Life.expectancy ~ Status + Alcohol + Hepatitis.B +
    BMI + Total.expenditure + Thinness, data = life_new)

Residuals:
     Min       1Q   Median       3Q      Max
-17.5927  -4.0000   0.7234   3.8625  16.7722

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       64.84518    4.00510  16.191  < 2e-16 ***
Status            -3.61542    1.99869  -1.809  0.07242 .
Alcohol            0.19390    0.17192   1.128  0.26112
Hepatitis.B        0.05962    0.02637   2.261  0.02518 *
BMI                0.14302    0.03403   4.203 4.45e-05 ***
Total.expenditure -0.29049    0.22664  -1.282  0.20186
Thinness          -0.41640    0.15729  -2.647  0.00895 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We used t-tests and diagnostic plots to check the validity of the full model. The summary table shows that only three predictors, Hepatitis B, BMI, and Thinness, are statistically significant, with their p-values less than 0.05. This indicates these variables have a meaningful impact on Y, whereas other predictors fail to reject the null hypothesis that their coefficients are zeros. The model's R-squared value is 0.4025, suggesting that only about 40% of the variance in the Y variable is explained by the model. While the R-squared value is not high, the F-statistic suggests that this model still provides a better fit than a simplistic model with only the intercept.




```
outlier <- which(abs(standres) >2)
outlier
```

```
##   4  11  32  54  66  82  88  90 129 158 161
##   4  11  32  54  66  82  88  90 129 158 161
```

```
leverage <- which(lev > 2*7/n)
leverage
```

```
##   1   6  11  32  51  66  76  96 103 123 130 137 145 153
##   1   6  11  32  51  66  76  96 103 123 130 137 145 153
```

```
influen <- which(cook > 4/(n-2))
influen
```

```
##  11  13  32  33  54  66  76  88  90 129 130 137 145 158
##  11  13  32  33  54  66  76  88  90 129 130 137 145 158
```

From the residual plots, we can see that Y and Xs are not very linearly related, and the variance is not perfectly constant. From the Q-Q plot, we know that the error is mostly normal except for both tails. And there are some outliers, leverage points, and influential points in the dataset. There are even some bad leverage points—both leverage and outlier points—that are worth looking into. However, since this data is from 2007 and covers all different countries with potentially other factors involved, we would not consider removing any outliers to avoid making an unreal analysis. We would consider transformation to make our regression fit better.

3. Result and Interpretation:

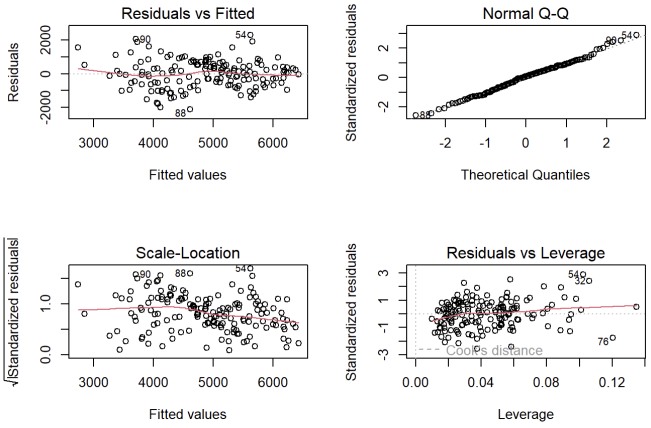
A. Transformation

After checking the diagnostic diagram, we found that the model still needs some improvement. Therefore, we will first transform the predictors and response variables of this model. We choose to change the multiple linear regression model using the Box-Cox Method, which will transform Y, X1,... Xp simultaneously to have joint normality.

```
## bcPower Transformations to Multinormality
##                    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Life.expectancy       2.8336          2.00        1.8425        3.8246
## Alcohol               0.4036          0.50        0.2971        0.5101
## Hepatitis.B           3.4860          3.49        2.8652        4.1067
## BMI                   1.1131          1.00        0.8830        1.3432
## Total.expenditure     0.5685          0.50        0.3247        0.8122
## Thinness              0.2052          0.21        0.0924        0.3180
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                        LRT df       pval
## LR test, lambda = (0 0 0 0 0 0) 576.8202  6 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                      LRT df       pval
## LR test, lambda = (1 1 1 1 1 1) 363.741  6 < 2.22e-16
```

According to the output, we find that the p-value for lambda = (0 0 0 0 0 0 0) is smaller than 0.05, which rejects the null hypothesis that transforming all of the variables with log is appropriate. The p-value for lambda = (1 1 1 1 1 1) is also smaller than 0.05, p-value for lambda = (1 1 1 1 1 1 1) is smaller than 0.05, which rejects the null hypothesis that no transformation is needed for this data. We choose rounded lambdas to transform each variable. Our final transformed model is: Life Expectancy^2 = β_0 + β_1 * Status + β_2 * Hepatitis.B$^{3.49}$ + β_3 * BMI + β_4 * Total Expenditure$^{0.5}$ + β_5 * Thinness$^{0.21}$ + e.

```
##
## Call:
## lm(formula = Trans_Life.expectancy ~ Status + Trans_Alcohol +
##     Trans_Hepatitis.B + BMI + Trans_Total.expenditure + Trans_Thinness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2116.4  -555.5    60.2   533.7  2298.7
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 6.999e+03  7.503e+02   9.328  < 2e-16 ***
## Status                     -4.069e+02  2.371e+02  -1.716   0.0881 .
## Trans_Alcohol               7.884e+01  7.909e+01   0.997   0.3204
## Trans_Hepatitis.B           1.134e-04  2.547e-05   4.453 1.62e-05 ***
## BMI                         1.093e+01  4.470e+00   2.445   0.0156 *
## Trans_Total.expenditure    -3.315e+02  1.306e+02  -2.538   0.0121 *
## Trans_Thinness             -1.725e+03  3.533e+02  -4.884 2.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 840.8 on 154 degrees of freedom
## Multiple R-squared:  0.4951, Adjusted R-squared:  0.4754
## F-statistic: 25.17 on 6 and 154 DF,  p-value: < 2.2e-16
```



Next, we use t-tests and diagnosis plots to analyze the validity of the model. There are only two predictor variables that aren't significant. At the same time, R^2 = 0.4951 has also been improved significantly. For the overall f-test, p-value < 0.05, which means that our transformed model is significant. Then look at four diagnosis plots. The residual plot shows that the transformed model has less heteroscedasticity and a better linear relationship. Normal Q-Q also shows better normality. At the same time, this model has fewer leverage points and influential points.

All in all, the model after transformation has been significantly improved compared with the original one. But because there are still significant variables, the model can still be optimized. Our next step is to check out multicollinearity and decide whether to remove any variables.

B. Subset Variable Selection

```
Subset selection object
6 Variables  (and intercept)
                        Forced in Forced out
Status                      FALSE      FALSE
trans_Hepatitis.B           FALSE      FALSE
BMI                         FALSE      FALSE
trans_Total.expenditure     FALSE      FALSE
trans_Thinness              FALSE      FALSE
trans_Alcohol               FALSE      FALSE
1 subsets of each size up to 6
Selection Algorithm: exhaustive
         Status trans_Hepatitis.B BMI trans_Total.expenditure trans_Thinness trans_Alcohol
1  ( 1 ) " "    " "               " " " "                    "*"             " "
2  ( 1 ) " "    "*"               " " " "                    "*"             " "
3  ( 1 ) " "    "*"               "*" " "                    "*"             " "
4  ( 1 ) " "    "*"               "*" "*"                    "*"             " "
5  ( 1 ) "*"    "*"               "*" "*"                    "*"             " "
6  ( 1 ) "*"    "*"               "*" "*"                    "*"             "*"
```

Size	Rad	AIC	AICc	BIC
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.3290142	2210.053	2210.204	2216.216
2	0.4209419	2187.315	2187.568	2196.559
3	0.4493289	2180.200	2180.582	2192.526
4	0.4623314	2177.324	2177.863	2192.731
5	0.4754146	2174.323	2175.045	2192.811
6	0.4753927	2175.287	2176.222	2196.857

First, we consider using the subset selection method to select variables. According to the R output, there's a table showing the combinations of variables needed for each size of the model. We then used adjusted R squared, AIC, AICc, and BIC to evaluate which combination of variables yields the best model fit based on these criteria. Higher adjusted R squared and lower AIC, AICc, and BIC indicate a better fit. Adjusted R squared, AIC, and AICc suggest that the subset with the size of 5 is the best, while BIC suggests that the subset with the size of 4 is the best. Thus, we should choose the model with a size of 5, since three out of four metrics suggest it.

C. Stepwise Selection

To gain further insights, we explored stepwise selection methods, starting with backward selection.

Backward Selection:

```
Start:  AIC=2175.29
trans_Life.expectancy ~ Status + trans_Alcohol + trans_Hepatitis.B +
    BMI + trans_Total.expenditure + trans_Thinness

                       Df Sum of Sq       RSS    AIC
- trans_Alcohol         1    702284 109559734 2174.3
<none>                              108857450 2175.3
- Status                1   2082169 110939619 2176.3
- BMI                   1   4226992 113084442 2179.4
- trans_Total.expenditure  1   4554852 113412302 2179.9
- trans_Hepatitis.B     1  14019349 122876798 2192.8
- trans_Thinness        1  16860547 125717997 2196.5

Step:  AIC=2174.32
trans_Life.expectancy ~ Status + trans_Hepatitis.B + BMI + trans_Total.expenditure +
    trans_Thinness

                       Df Sum of Sq       RSS    AIC
<none>                              109559734 2174.3
- Status                1   3456904 113016638 2177.3
- trans_Total.expenditure  1   4103285 113663019 2178.2
- BMI                   1   4303121 113862855 2178.5
- trans_Hepatitis.B     1  14042512 123602246 2191.7
- trans_Thinness        1  19339550 128899284 2198.5
```

Forward Selection:

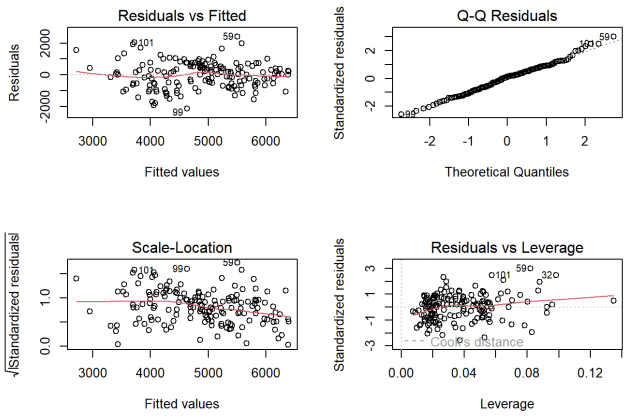
```
Start:  AIC=2273.3
trans_Life.expectancy ~ 1

                          Df Sum of Sq       RSS    AIC
+ trans_Thinness           1  71835369 143751890 2210.1
+ BMI                      1  67550091 148037168 2214.8
+ Status                   1  35734151 179853108 2246.1
+ trans_Hepatitis.B        1  28742713 186844546 2252.3
+ trans_Alcohol            1  27471825 188115434 2253.4
<none>                                215587259 2273.3
+ trans_Total.expenditure  1    799023 214788236 2274.7

Step:  AIC=2210.05
trans_Life.expectancy ~ trans_Thinness

                          Df Sum of Sq       RSS    AIC
+ trans_Hepatitis.B        1  20474817 123277074 2187.3
+ BMI                      1  13807143 129944747 2195.8
+ Status                   1   7703795 136048095 2203.2
+ trans_Alcohol            1   2709218 141042672 2209.0
<none>                                143751890 2210.1
+ trans_Total.expenditure  1   1728213 142023677 2210.1
```

```
Step:  AIC=2187.32
trans_Life.expectancy ~ trans_Thinness + trans_Hepatitis.B

                          Df Sum of Sq       RSS    AIC
+ BMI                      1   6785366 116491708 2180.2
+ trans_Total.expenditure  1   4452683 118824391 2183.4
+ Status                   1   4261548 119015526 2183.7
<none>                                123277074 2187.3
+ trans_Alcohol            1   1513955 121763119 2187.3

Step:  AIC=2180.2
trans_Life.expectancy ~ trans_Thinness + trans_Hepatitis.B +
    BMI

                          Df Sum of Sq       RSS    AIC
+ trans_Total.expenditure  1   3475070 113016638 2177.3
+ Status                   1   2828689 113663019 2178.2
<none>                                116491708 2180.2
+ trans_Alcohol            1   1159285 115332423 2180.6

Step:  AIC=2177.32
trans_Life.expectancy ~ trans_Thinness + trans_Hepatitis.B +
    BMI + trans_Total.expenditure

                 Df Sum of Sq       RSS    AIC
+ Status          1   3456904 109559734 2174.3
+ trans_Alcohol   1   2077020 110939619 2176.3
<none>                        113016638 2177.3

Step:  AIC=2174.32
trans_Life.expectancy ~ trans_Thinness + trans_Hepatitis.B +
    BMI + trans_Total.expenditure + Status

                 Df Sum of Sq       RSS    AIC
<none>                        109559734 2174.3
+ trans_Alcohol   1    702284 108857450 2175.3
```

Subset selection, forward selection, and backward selection all suggest the model with 5 predictors to be the best. Thus, we chose the model incorporating transformed life expectancy, hepatitis B, BMI, total expenditure, thinness, and status.

Final Model:
Consequently, the preferred final model is as follows:
$Life.expectancy^2 = 7273 + 1.135 * 10^{-4} * Hepatitis.B^{3.49} + 11.03 * BMI - 310.5 * Total.expenditure^{0.5} - 109* Thinness^{0.21} + 490.5* Status$

```
##
## Call:
## lm(formula = trans_Life.expectancy ~ trans_Hepatitis.B + BMI +
##     trans_Total.expenditure + trans_Thinness + Status, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2130.76  -549.32    94.87   539.67  2418.11
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.273e+03  6.983e+02  10.415  < 2e-16 ***
## trans_Hepatitis.B        1.135e-04  2.547e-05   4.457 1.58e-05 ***
## BMI                      1.103e+01  4.469e+00   2.467   0.0147 *
## trans_Total.expenditure -3.105e+02  1.289e+02  -2.409   0.0172 *
## trans_Thinness          -1.803e+03  3.447e+02  -5.231 5.40e-07 ***
## Status                  -4.905e+02  2.218e+02  -2.211   0.0285 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 840.7 on 155 degrees of freedom
##   (22 observations deleted due to missingness)
## Multiple R-squared:  0.4918, Adjusted R-squared:  0.4754
## F-statistic:    30 on 5 and 155 DF,  p-value: < 2.2e-16
```



In the final model, every predictor proved to be statistically significant, as indicated by p-values for F-statistics below 0.05, affirming the overall model's significance. The summary for the final model looks better than the original model, with an improved adjusted R-squared of 0.4754. From the diagnosis plot, we see some improvement in the linearity of Y and X's, the normality of error terms, and the normal variances. However, the improvement is not significant, and there is still room for improvement. That being said, we would still consider the function to be a valid model.

4. Discussion:

We created a final model based on the dataset's characteristics, tests, and plots, and we can use some of these predictors to explain the predictor variable. For example: Each unit increase in Body Mass Index (BMI) increases the square of life expectancy by 11.03 units; when a country changes from developed to developing status, the square of life expectancy decreases by 490.5 units.

Unfortunately, this final model is still not applicable to real life, as total expenditure should not hurt life expectancy in real life. We have to admit that we struggle with choosing between the original model and the backward selection model. We still choose the final model over the original model because 3 out of 6 predictor variables in the original model are statistically insignificant. Even though the original model is easier to interpret, the large proportion of insignificant variables suggests messy correlations among them. This evolution underscores the trade-off between model simplicity and the statistical significance of its predictors. However, even when choosing a model where every predictor variable is significant, the occurrence of incorrect signs still suggests underlying issues.

We suspect the wrong sign of the coefficients of our predictions is because of the high degree of collinearity among our variables. One interpretation might be that life expectancy is highly correlated with each factor, making it hard to separate them. For example, the status of a country, whether developed

or developing, will affect general thinness, expenditure, alcohol consumption, and a lot more. That might be why, even with variable selection, there might still be a high level of collinearity.

Besides the wrong sign, there are other drawbacks and limitations to our final model. On the dataset side, the data we selected is outdated, from 2007, which may not be representative of or have some deviations from the current situation. Furthermore, many country data are deleted because they may contain NA values in one predictor variable. On the model side, our model is hard to interpret and apply to real life since it involves power transformation and variable selection. If we want to improve these issues, we may need to seek out more specialized determinants to replace some of our current predictor variables or even use another regression method. Future research should consider these methods to address the challenges of collinearity and enhance the model's applicability to real-world scenarios.