Khang Luu
STATS 101A, Winter 2024
16 March 2024

Life Lines: Predicting Life Expectancy From Demographic, Economic, and Health Factors

INTRODUCTION

The author of this paper, upon discovering the stark contrast in life expectancy ranking between the United States (47th for 79.74 years), Canada (19th for 83.02 years), and Mexico (92nd for 75.04 years) (Worldometer), who are all in geographical proximity and share borders, was motivated to predict life expectancy for all countries. Unlike prior studies, which were limited in time span and number of countries, this research uses data from 193 countries over 15 years (2000 to 2015) for greater generalizability.

RESEARCH QUESTION

To what extent can demographic indicators, economic performance, and health factors predict a nation's average life expectancy?

DATASET BACKGROUND

This research uses the Life Expectancy (WHO) dataset from Kaggle. This dataset pooled health data from the Global Health Observatory's data repository and economic data from the United Nation. The dataset includes 22 variables and 2,938 observations.

DESCRIPTIVE ANALYSIS

Though the dataset initially includes 21 predictors, the author runs pairs() function to produce a correlation scatter plot to identify and eliminate variables with little correlation with life expectancy. The chosen predictors and the response variable (life expectancy) are as follows:

- ***Life Expectancy***: The average age to which a person is expected to live.
- ***Adult Mortality***: The rate of death among adults aged 15 to 60 per 1,000 population.
- ***Alcohol***: The per capita consumption of alcohol, measured in liters of pure alcohol.
- ***BMI***: The average Body Mass Index of the population.
- ***GDP***: Gross Domestic Product per capita measured in USD.
- ***Population***: The total population.
- ***Schooling***: The average number of years of schooling completed by the population.

```
life_expectancy adult_mortality    alcohol           BMI
Min.   :44.0    Min.   :  1.0   Min.   : 0.010   Min.   : 2.00
1st Qu.:64.4    1st Qu.: 77.0   1st Qu.: 0.810   1st Qu.:19.50
Median :71.7    Median :148.0   Median : 3.790   Median :43.70
Mean   :69.3    Mean   :168.2   Mean   : 4.533   Mean   :38.13
3rd Qu.:75.0    3rd Qu.:227.0   3rd Qu.: 7.340   3rd Qu.:55.80
Max.   :89.0    Max.   :723.0   Max.   :17.870   Max.   :77.10

      GDP              population          schooling
Min.   :     1.68   Min.   :3.400e+01   Min.   : 4.20
1st Qu.:   462.15   1st Qu.:1.919e+05   1st Qu.:10.30
Median :  1592.57   Median :1.420e+06   Median :12.30
Mean   :  5566.03   Mean   :1.465e+07   Mean   :12.12
3rd Qu.:  4718.51   3rd Qu.:7.659e+06   3rd Qu.:14.00
Max.   :119172.74   Max.   :1.294e+09   Max.   :20.70
```

Figure 1: Summary statistics, variances, and standard deviations for all variables.

MODEL FITTING

```
lm(formula = life_expectancy ~ adult_mortality + alcohol + BMI +
    GDP + population + schooling, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-24.7011  -2.2545   0.4575   2.8819  13.6082

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.551e+01  6.937e-01  80.028  < 2e-16 ***
adult_mortality -3.177e-02  1.004e-03 -31.640  < 2e-16 ***
alcohol         -8.914e-02  3.633e-02  -2.454   0.0142 *
BMI              5.453e-02  6.873e-03   7.934 3.88e-15 ***
GDP              7.864e-05  1.130e-05   6.959 4.93e-12 ***
population      -2.838e-11  1.585e-09  -0.018   0.9857
schooling        1.404e+00  6.103e-02  23.012  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.512 on 1642 degrees of freedom
Multiple R-squared:  0.7378, Adjusted R-squared:  0.7369
F-statistic: 770.2 on 6 and 1642 DF,  p-value: < 2.2e-16
```

Figure 2: A Linear Regression model is initially fitted to the 6 predictors.

With an R^2 value of 0.7369, the first model seems promising, suggesting that 73.69% of the variation in life expectancy could be explained by the predictors. However, the predictor population is not statistically significant. Furthermore, model assumptions, like linearity, normality, and constant variance need to be verified.

MODEL ASSUMPTIONS VERIFICATION

From the diagnostic tools in Figure 3, we observe a violation of linearity because the red line in the Residuals vs. Fitted plot is not linear. From the Q-Q Residuals plot, we observe a violation of normality of the error term because the points at the left-tail do not align to the straight line. From the Standardized residuals vs. Leverages plot, we observe many points that could be outliers. Fortunately, constant variance is observed and multicollinearity is not present because VIF values for predictors are all less than 5. Nevertheless, power transformation is still needed.

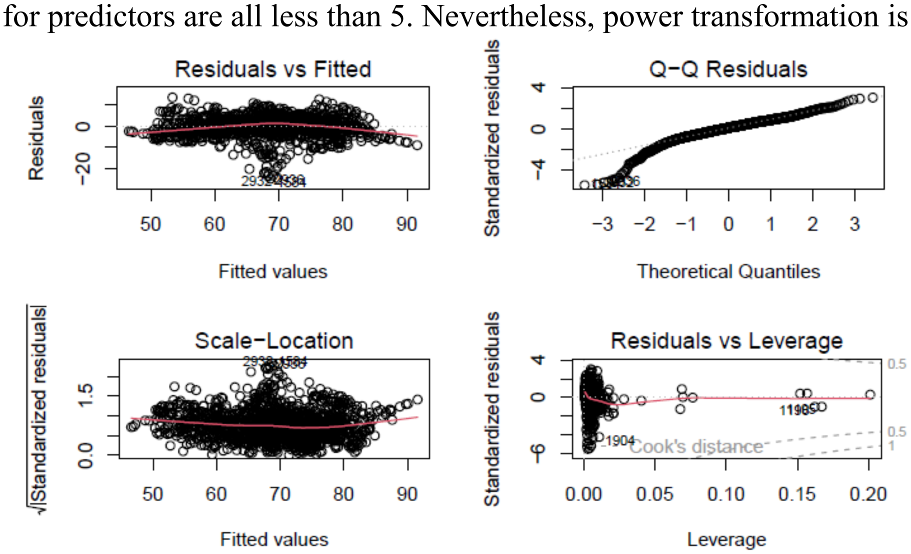

Figure 3: Diagnostic plots for the first model.

MODEL TRANSFORMATION

Box-Cox Transformation is applied to all variables based upon the power recommended by powerTransformed function for each variable, following the formula:

$$y(\lambda) \;=\; \frac{y^{\lambda}-1}{\lambda} \; if\; \lambda \neq 0,\; else\; log(y).$$

In the transformed model, predictors GDP and population undergo log transformation. Other variables like life expectancy, mortality, alcohol, BMI, and schooling undergo power transformation with values 2.87, 0.64, 0.44, 1.10, and 1.33 respectively.

```
Residuals:
   Min     1Q Median     3Q    Max 
-49515  -7335    789   7920  46602 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 35654.200   2425.730  14.698  < 2e-16 ***
t_mortality  -418.891     17.450 -24.005  < 2e-16 ***
t_alcohol      -7.724    183.523  -0.042    0.966    
t_BMI          95.674     13.492   7.091 1.97e-12 ***
log_GDP      1699.201    222.917   7.623 4.18e-14 ***
log_pop      -151.305    112.561  -1.344    0.179    
t_schooling  1703.072     76.516  22.258  < 2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1

Residual standard error: 12450 on 1642 degrees of freedom
Multiple R-squared:  0.7116,	Adjusted R-squared:  0.7105 
F-statistic: 675.1 on 6 and 1642 DF,  p-value: < 2.2e-16
```

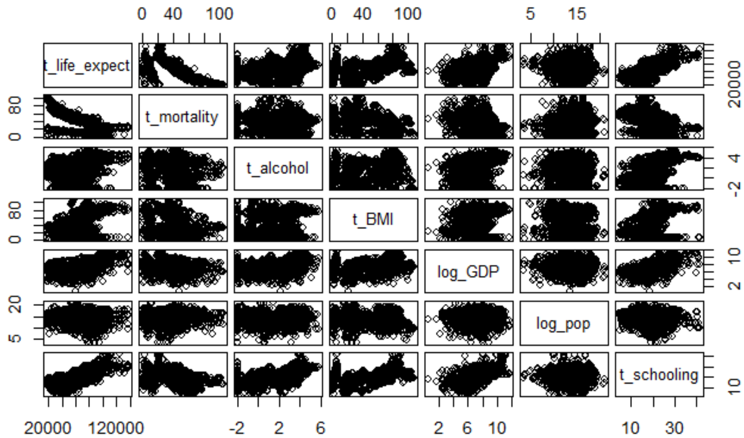

Figure 4: The transformed model and the corresponding scatterplot matrix.

The transformed model meets all five model assumptions. Furthermore, the F-test shows the model is statistically significant with a p-value less than 2.2e-16. However, predictors t_alcohol and log_pop are statistically insignificant with a p-value above alpha 0.05. This means variable selection is needed to determine if these predators could be dropped from the transformed model.

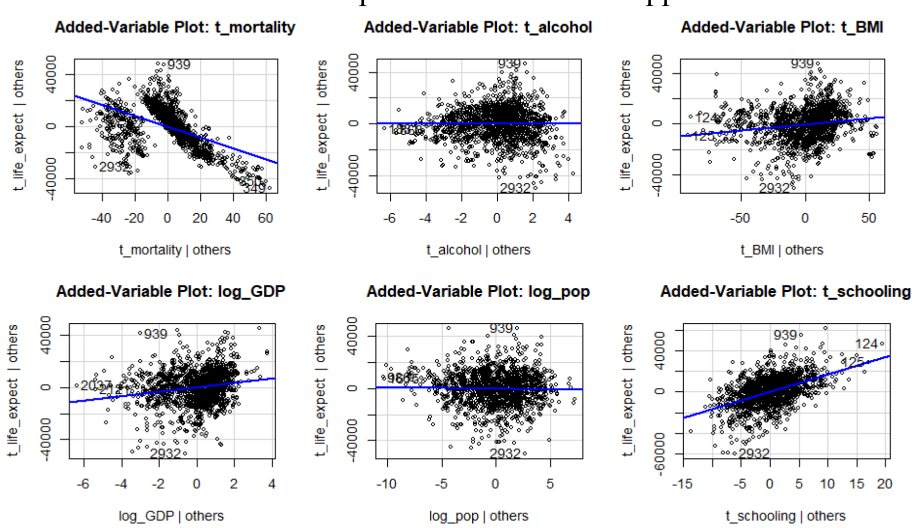

Figure 5: Added-Variable Plots examine the effect of a given predictor on life expectancy.

The Added-Variable Plots show t_alcohol and log_pop have almost no effect on life expectancy without the effects of other predictors, further supporting that they are statistically insignificant.

VARIABLE SELECTION

The first variable selection method used is to consider all possible subsets. This method suggests that based upon the adjusted R-squared value, the subset with p = 5 to be the best. This subset uses the predictors t_mortality, t_BMI, log_GDP, log_pop, and t_schooling. When considering AIC corrected and BIC values, the method suggests the subset with p = 4 to be the best, which uses the predictors t_mortality, t_BMI, log_GDP, and t_schooling.

```
Residuals:
   Min     1Q Median     3Q    Max
-49976  -7487    638   7919  46411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33664.22    1755.29  19.179  < 2e-16 ***
t_mortality  -421.08      17.30 -24.346  < 2e-16 ***
t_BMI          96.13      13.47   7.134 1.45e-12 ***
log_GDP      1691.85     221.31   7.645 3.54e-14 ***
t_schooling  1702.52      69.66  24.440  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 12450 on 1644 degrees of freedom
Multiple R-squared:  0.7112, Adjusted R-squared:  0.7105
F-statistic:  1012 on 4 and 1644 DF,  p-value: < 2.2e-16
```

```
Residuals:
   Min     1Q Median     3Q    Max
-49531  -7348    777   7921  46590

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 35685.67    2306.91  15.469  < 2e-16 ***
t_mortality  -418.96      17.36 -24.130  < 2e-16 ***
t_BMI          95.65      13.47   7.098 1.87e-12 ***
log_GDP      1698.10     221.30   7.673 2.86e-14 ***
log_pop      -151.60     112.30  -1.350    0.177
t_schooling  1701.74      69.64  24.435  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 12450 on 1643 degrees of freedom
Multiple R-squared:  0.7116, Adjusted R-squared:  0.7107
F-statistic: 810.6 on 5 and 1643 DF,  p-value: < 2.2e-16
```

Figure 6: Summary results for p = 4 (left) and p = 5 (right) subsets of predictors.

Both subsets are statistically significant based on the p-value from the F-statistic. In the four-variable subset (p = 4), the R^2 value is smaller than the five-variable subset (p = 5)'s R^2 value, but all four predictors are statistically significant. Meanwhile, one predictor, log_pop, is not statistically significant for the five-variable subset, indicating that the five-variable model overfits the data. Based upon the result, we choose the four-variable model.

```
Step:  AIC=31103.29
t_life_expect ~ t_mortality + t_BMI + log_GDP + t_schooling

              Df  Sum of Sq        RSS   AIC
<none>                      2.5481e+11 31103
- t_BMI        1 7.8892e+09 2.6270e+11 31152
- log_GDP      1 9.0583e+09 2.6386e+11 31159
- t_mortality  1 9.1866e+10 3.4667e+11 31609
- t_schooling  1 9.2582e+10 3.4739e+11 31612
```

Figure 7: Backward Stepwise regression result for the transformed model.

Additionally, Stepwise regression Backward stepwise also recommends the four-variable subset model. Overall, because the Added-Variable Plots show that t_alcohol and log_pop have little effects on life expectancy, and both subset methods recommend the four-variable subset model, which eliminates t_alcohol and log_pop, we confidently conclude that the four-variable subset model is the best model and the aforementioned predictors must be removed.

MODEL INTERPRETATION

```
lm(formula = t_life_expect ~ t_mortality + t_BMI + log_GDP +
    t_schooling, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-49976  -7487    638   7919  46411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33664.22    1755.29  19.179  < 2e-16 ***
t_mortality  -421.08      17.30 -24.346  < 2e-16 ***
t_BMI          96.13      13.47   7.134 1.45e-12 ***
log_GDP      1691.85     221.31   7.645 3.54e-14 ***
t_schooling  1702.52      69.66  24.440  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12450 on 1644 degrees of freedom
Multiple R-squared:  0.7112,    Adjusted R-squared:  0.7105
F-statistic:  1012 on 4 and 1644 DF,  p-value: < 2.2e-16
```

Figure 8: Final regression model.

With an R^2 value of 0.7105, the four predictors, t_mortality, t_BMI, log_GDP, and t_schooling can explain 71.05% of the variation in life expectancy. Furthermore, all four predictors are statistically significant with p-values less than alpha 0.05. Lastly, the F-statistic also has a p-value less than 2.2e-16, revealing the strength of the model. The interpretations for estimate coefficients from the model are as follows:

- The intercept is 33664.22, which is the estimated life expectancy for a given country when all predictors are 0. This interpretation is not meaningful in the real world context because predictors like mortality rate cannot be zero in a population.
- For every one-unit increase in the transformed mortality rate, transformed life expectancy decreases by 421.08 units.
- For every one-unit increase in transformed BMI score, transformed life expectancy increases by 96.13 units.
- For every one-percent change in GDP, transformed life expectancy increases by 1691.85 units.
- For every one-unit increase in transformed schooling rate, transformed life expectancy increases by 1702.52 units.

CONCLUSION

In conclusion, this research aimed to predict life expectancy using demographic indicators, economic performance, and health factors using a dataset containing data on 193 countries over 15 years. The author performed variable transformation and variable selection to construct a final model with four statistically significant predictors: transformed mortality rate, transformed BMI score, log GDP, and transformed schooling rate. The model demonstrated strong statistical significance, explaining 71.05% of the variation in life expectancy. Policymakers and public health officials seeking to improve life expectancy should focus efforts on decreasing mortality rate and increasing national GDP, average BMI score, and education.

Works Cited

"Life Expectancy (WHO)." Www.kaggle.com, www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data.

Worldometer. "Life Expectancy by Country and in the World (2023)." Worldometers.info, 2023, www.worldometers.info/demographics/life-expectancy/.