Michael Dymek
December 12, 2019

Final Project

Dr. Amy Anderson
Math 342

Abstract

I will be using WHO data to analyze the affect of 13 different predictor variables on the life expectancy of a given country in the year 2015. With this dataset, I will be asnwering the following questions.

1. Do the various predicting factors chosen initially really affect life expectancy? What predicting variables are really important?
2. Is there a specific disease immunization that is significantly better at predicting lifespan than others? Is there one that is not a good predictor?
3. Do countries with larger populations tend to have higher or lower life expectancy?
4. Do countries with more schooling tend to have higher or lower life expectancy?
5. Do countries with more income from resources tend to have higher or lower life expectancy?

The results that I found were that of the 13 predictor variables that we have started with, the variables that are the most important for a model are the Adult Mortality, $\log(\text{HIV/AIDS})$, and Income from Resources. My tests also suggest that both population and schooling are not significant predictors, while income from resources is a significant predictor. Finally, I've found that the Polio and Diphtheria vaccines are the most significant predictors of the vaccine variables.

Introduction

Health is an incredibly complex subject, and it's very important that we understand as much as we can as a global community. The healthier we are, the more prosperous we are. On top of that, there is the very human argument that we should do everything we can to preserve human lives. This dataset, which has been released by the World Health Organization, contains 15 years of global health, economic, and social data, from which a total of 16 usable predictor variables have been recorded.

I'll be specifically looking at the year of 2015, as it has the most usable data, and will remove the correlation between data from the same country. All predictor variables are summarized below. The sample size of this database is in the thousands, which correspond to annual data from different countries, and I'll be using approximately 180 data points, which is all the data from just 2015. This is above the 130 required for a large enough sample size.

Predictor Variables			
Developing Status (Binary)	Adult Mortality (per 1000)	Infant Mortality (per 1000)	Under-5 Deaths (per 1000)
Population	BMI	Hepatitis B Immunization (per 100)	GDP (Dollars)
Schooling (Years)	Polio Immunization (per 100)	Diphtheria Immunization (per 100)	HIV/AIDS Deaths (per 1000)
Income Composition from Resources (Percentage)			

With this dataset, I will be answering the following questions.

1. Do the various predicting factors chosen initially really affect life expectancy? What predicting variables are really important?

2. Is there a specific disease immunization that is significantly better at predicting lifespan than others? Is there one that is not a good predictor?

3. Do countries with larger populations tend to have higher or lower life expectancy?

4. Do countries with more schooling tend to have higher or lower life expectancy?

5. Do countries with more income from resources tend to have higher or lower life expectancy?

Methods

Before I can try to answer any of these questions, I need to see if any of these variables need to be transformed in any way to get the best fit. To do this, I regressed each variable against life expectancy on it's own, and looked at the fit of the line and the diagnostic plots. This allowed me to see which variables needed to be transformed to have a more linear fit. In the supplementary figures section, I've attached each plot that I transformed and their transformations. Here, I'll just summarize the variables that I will need to transform for this regression.

$$\text{HIV AIDS} \longrightarrow \log(\text{HIV AIDS}) \qquad \text{GDP} \longrightarrow \log(\text{GDP})$$

I'll start my analysis by looking at the first question. Do the various predicting factors chosen appear to affect life expectancy. This boils down to a simple hypothesis test of whether $\beta_i = 0$ or not. I will perform these tests at type 1 error rate of $\alpha = 0.05$. Outputs from R are attached in the supplementary figures section.

$$H_0\colon \beta_i = 0 \qquad H_a\colon \beta_i \neq 0 \qquad \alpha = 0.05$$

The value of the test statistic and the p-value are given in the R output for the full model regression. The table below summarizes the results of this test on each variable.

Result of Hypothesis Test					
Name of Variable	**p-value**	**Result**	**Name of Variable**	**p-value**	**Result**
Developing Status	0.3579	Do Not Reject	Hepatitis B Immunization	0.1197	Do Not Reject
Adult Mortality	≈ 0	Reject H_0	Diphtheria Immunization	0.5908	Do Not Reject
Infant Mortality	0.2166	Do Not Reject	Polio Immunization	0.1905	Do Not Reject
Under-5 Deaths	0.1886	Do Not Reject	log(HIV/AIDS) Deaths	0.0003	Reject H_0
Population	0.8545	Do Not Reject	Income Composition	≈ 0	Reject H_0
BMI	0.7387	Do Not Reject	Schooling	0.7820	Do Not Reject
log(GDP)	0.2405	Do Not Reject			

From the results given in the table, I must remove every variable which failed to reject the null hypothesis. After doing this, I have a new model where life expectancy is only a function of the 3 variable left. I've written the function and the ANOVA table of this regression below.

$$\text{Life Expectancy} \sim \text{Adult Mortality} + \log(\text{HIV/AIDS Deaths}) + \text{Income Composition}$$

Source of Variation	**df**	**Sun of Squares**	**Mean Squares**	**F-statistic**	**p-value**
Adult Mortality	1	6249.5	6249.5	912.91	≈ 0
log(HIV/AIDS)	1	1494.5	1494.5	218.31	≈ 0
Income from Resources	1	1924.8	1924.8	281.16	≈ 0
Residuals	169	1156.9	6.8		

I feel okay with removing these variables because what I look at the diagnostic plots (in figures section), I can see that the residulal plot is still a very good fit, so the removed variables were not significantly affecting the fit. This is a very surprising result, that out of a total of 13 possible prediction variables, only 3 are significant. Looking at those variables, it possibly makes some sense, as two of the three variables are specifically correlated with deaths, so it makes sense that a metric measuring death is the best at predicting how long it will be until people die. For the last one about income from natural resources, a lot of the other variables are also general metrics for wealth, so it's possible that all of the wealth variables are wrapped into this one.

Using these results, I can also answer a lot of the other questions that I had about specific variables. Any question that was phrased in the form of "Is variable A a significant predictor of Y?" is simply a test of whether $\beta = 0$.

Do countries with larger populations tend to have higher or lower life expectancy? My analysis shows that population is not a significant predictor of life expectancy. This is because I was not able to reject H_0. The β value is not large enough to be considered different from 0. I hypothesize that this may be because of a few very large countries that have lower life expediencies and a few small countries with high life expectancy will be messing with any relationship that there may be. To investigate if this was the case, I would need to run a test on the total population dataset, and the same test with a reduced data set in which I remove the influential points.

Does schooling have an effect on life expectancy? My analysis shows that the average schooling level is not a significant predictor of life expectancy. This is because I was not able to reject H_0. As I stated above, I imagine that that is because the schooling data is well approximated using the income from resources data.

Do countries with more income from resources tend to have higher or lower life expectancy? My analysis suggests that income from population is a significant predictor of life expectancy. The correlation tends to be positive, meaning that the more income that a country gets from their natural resources, then the more likely it will be for them to have a high life expectancy.

Now that I've answered these questions, I can move on to finding if some variables are better at predicting than others. First, I want to know if there is a specific disease immunization that is the best at predicting life expectancy, and if there is one that is the worst. To do this, I ran a test regressing life expectancy as a function of only the disease immunization variables. This will isolate things so that I can look at just the immunization. Running this regression, I now want to test if any of the $\beta_i = 0$ or not, and I will use the p-values to compare the significance of each variable. The results of this test are summarized below.

Result of Hypothesis Test					
Name of Variable	**p-value**	**Result**	**Name of Variable**	**p-value**	**Result**
Diphtheria Immunization	0.0071	Reject H_0	Hepatitis B Immunization	0.2282	Do not Reject H_0
Polio Immunization	0.0002	Reject H_0			

Looking at the results above, I can see that both the diphtheria immunization, and the polio immunization both seem to be significant predictors of life expectancy. This means that if I were advising a country one what immunizations that they should prioritize, I would say that the diphtheria and the polio vaccinations are the most important, and of those two the polio one is the most important, as it has the smallest p-value.

Conclusions

From these tests, I can make the following conclusion. Firstly, of the 13 predictor variables that we have started with, the variables that are the most important for a model are the Adult Mortality, log (HIV/AIDS), and Income from Resources. My tests also suggest that both population and schooling are not significant predictors, while income from resources is a significant predictor. Finally, I've found that the Polio and Diphtheria vaccines are the most significant predictors of the vaccine variables, so a country should prioritize those when starting any kind of vaccination law of public service.

Supplementary Figures

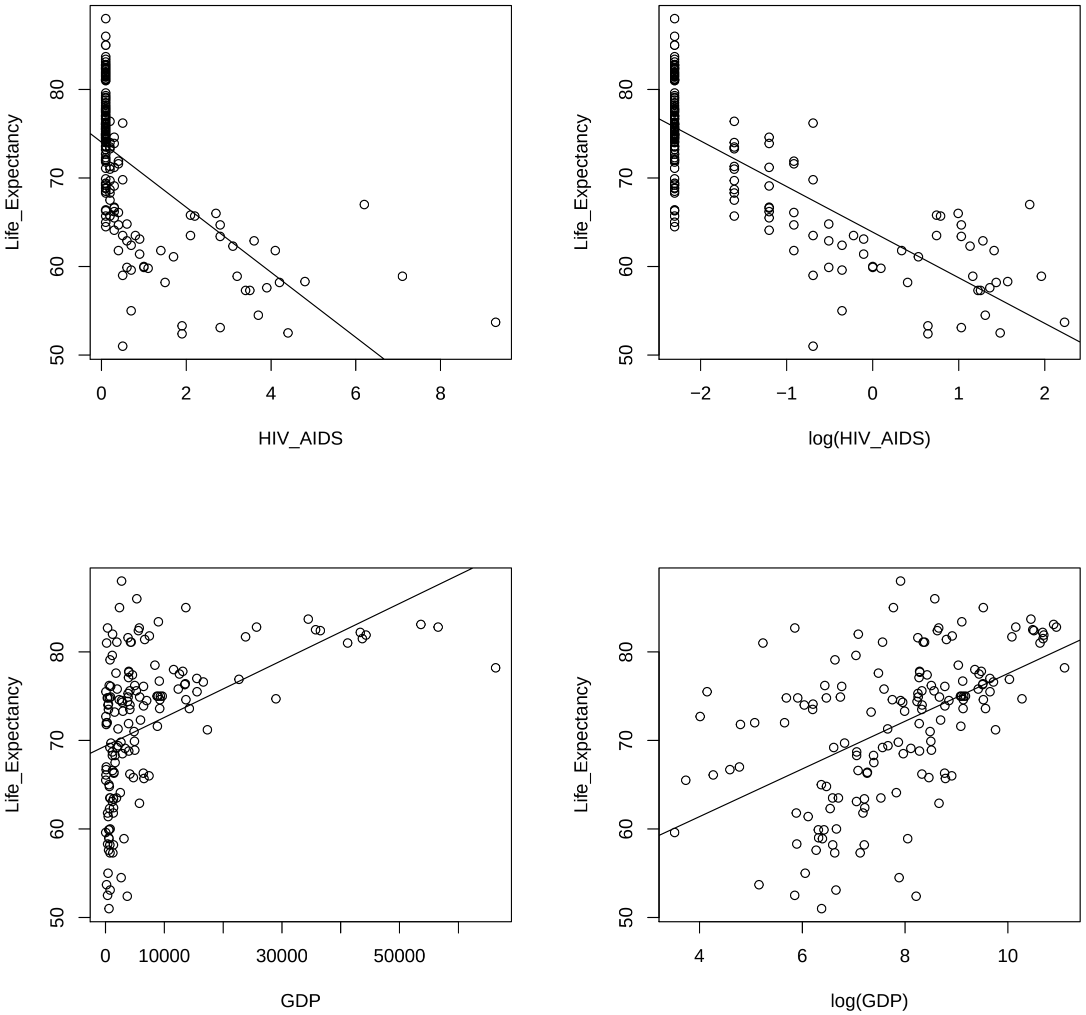
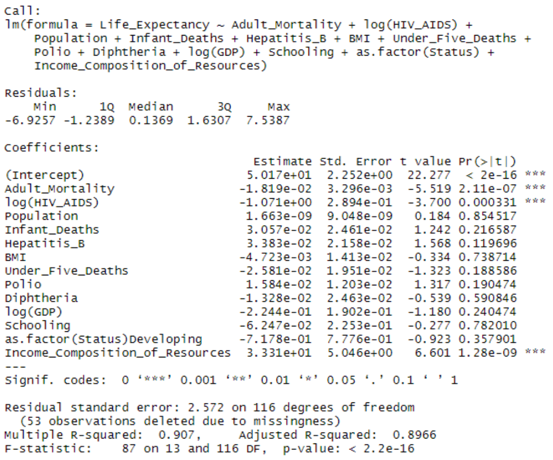
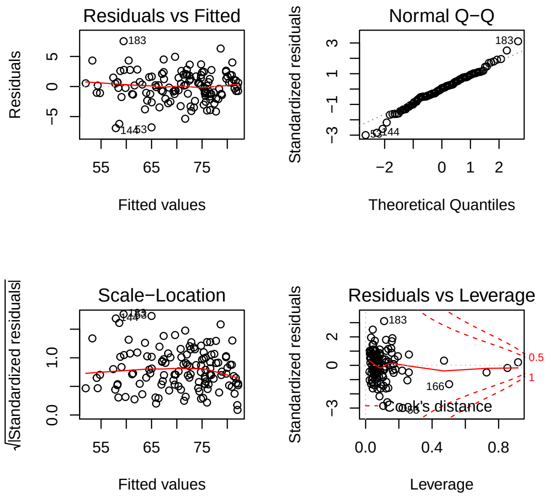
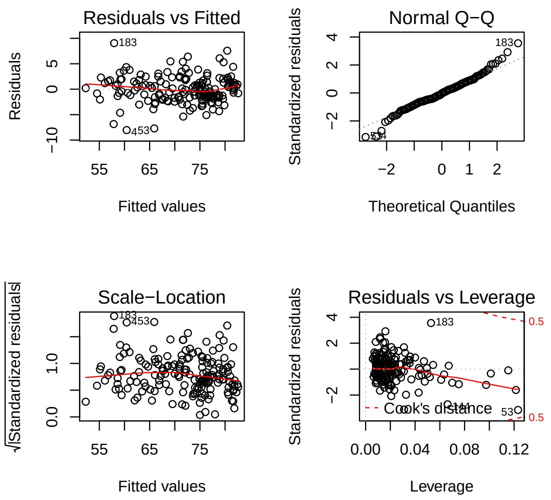


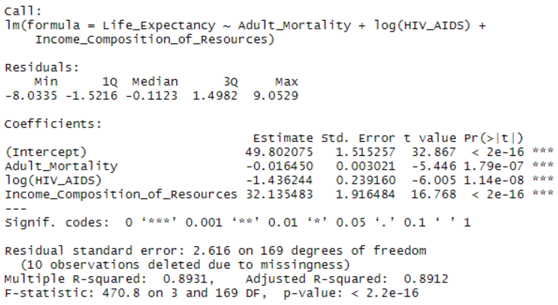
Figure 1: Before and after plots for the transformations performed on the HIV and GDP data sets.



```
Call:
lm(formula = Life_Expectancy ~ Adult_Mortality + log(HIV_AIDS) +
    Population + Infant_Deaths + Hepatitis_B + BMI + Under_Five_Deaths +
    Polio + Diphtheria + log(GDP) + Schooling + as.factor(Status) +
    Income_Composition_of_Resources)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9257 -1.2389  0.1369  1.6307  7.5387

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      5.017e+01  2.252e+00  22.277  < 2e-16 ***
Adult_Mortality                 -1.819e-02  3.296e-03  -5.519 2.11e-07 ***
log(HIV_AIDS)                   -1.071e+00  2.894e-01  -3.700 0.000331 ***
Population                       1.663e-09  9.048e-09   0.184 0.854517
Infant_Deaths                    3.057e-02  2.461e-02   1.242 0.216587
Hepatitis_B                      3.383e-02  2.158e-02   1.568 0.119696
BMI                             -4.723e-03  1.413e-02  -0.334 0.738714
Under_Five_Deaths               -2.581e-02  1.951e-02  -1.323 0.188586
Polio                            1.584e-02  1.203e-02   1.317 0.190474
Diphtheria                      -1.328e-02  2.463e-02  -0.539 0.590846
log(GDP)                        -2.244e-01  1.902e-01  -1.180 0.240474
Schooling                       -6.247e-02  2.253e-01  -0.277 0.782010
as.factor(Status)Developing     -7.178e-01  7.776e-01  -0.923 0.357901
Income_Composition_of_Resources  3.331e+01  5.046e+00   6.601 1.28e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.572 on 116 degrees of freedom
  (53 observations deleted due to missingness)
Multiple R-squared:  0.907,     Adjusted R-squared:  0.8966
F-statistic:    87 on 13 and 116 DF,  p-value: < 2.2e-16
```

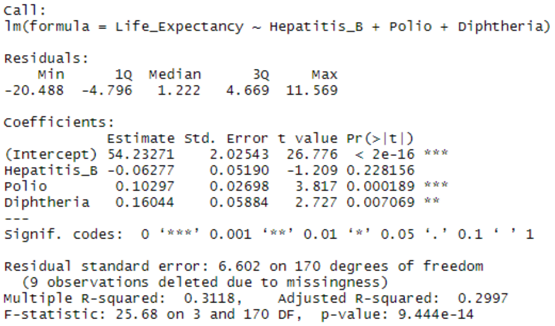






```
Call:
lm(formula = Life_Expectancy ~ Adult_Mortality + log(HIV_AIDS) +
    Income_Composition_of_Resources)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0335 -1.5216 -0.1123  1.4982  9.0529

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     49.802075   1.515257  32.867  < 2e-16 ***
Adult_Mortality                 -0.016450   0.003021  -5.446 1.79e-07 ***
log(HIV_AIDS)                   -1.436244   0.239160  -6.005 1.14e-08 ***
Income_Composition_of_Resources 32.135483   1.916484  16.768  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.616 on 169 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.8931,    Adjusted R-squared:  0.8912
F-statistic: 470.8 on 3 and 169 DF,  p-value: < 2.2e-16
```

Figure 2: Before and after plots for the regression with all of the variables considered and with only the significant variables considered.



```
Call:
lm(formula = Life_Expectancy ~ Hepatitis_B + Polio + Diphtheria)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.488  -4.796   1.222   4.669  11.569 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 54.23271    2.02543  26.776  < 2e-16 ***
Hepatitis_B -0.06277    0.05190  -1.209 0.228156    
Polio        0.10297    0.02698   3.817 0.000189 ***
Diphtheria   0.16044    0.05884   2.727 0.007069 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.602 on 170 degrees of freedom
  (9 observations deleted due to missingness)
Multiple R-squared:  0.3118,	Adjusted R-squared:  0.2997 
F-statistic: 25.68 on 3 and 170 DF,  p-value: 9.444e-14
```

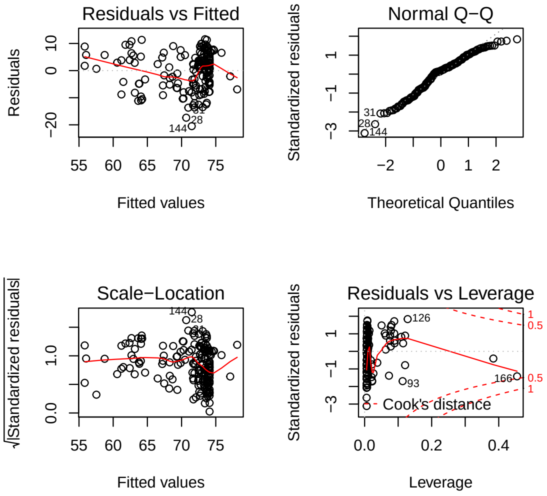


Figure 3: Regression summary and Diagnostic plots for the disease immunization tests.

The data for this analysis was obtained from https://www.kaggle.com/kumarajarshi/life-expectancy-who