DS6372 WHO Project

Yusi Zhang

Aurian Ghaemmaghami

Sreeni Prabhala

**Data Description**

The 2014 WHO dataset was collected from the WHO and United Nations website with the help of Deeksha Russell and Duan Wang for use in data science education. We have 183 observations with 22 explanatory variables, and it is collected from various countries around the world. We will bring up NA filtration in our “Model Build” section depicted below.

**Objective #1, Model 1**

1. **Problem Statement:** We must build a model with the main goal to identify relationships and is highly interpretable regarding our response variable of interest: Life Expectancy.

**Exploratory Data Analysis**

1. **Initial EDA**
   1. We first began our data exploration process by viewing the different types of categorical and numerical variables of interest. Variables of “Country Name” and “Year” were deemed unnecessary to include in any further analysis, so we went ahead and filtered those variables from our model.
   2. Next, we built out a correlation matrix (Appendix: 1A) to assess any relationships we see prior to doing any filtering of NA values in our data set. The reason we started with a correlation matrix is because 20/22 of our variables were either integer or numerical values. Based off a correlation cut off point of 0.5 (moderate relationships with Life Expectancy), the variables of interest we wanted to focus on were Adult Mortality, Alcohol, BMI, HIV/AIDS, Income Composition of Resources, and Schooling.
   3. We then built out a scatterplot matrix (Appendix: 1B) of our predictors of interest to view any sort of collinearity between predictors or irregular distributions. We see that both Income Composition of Resources and Schooling show high correlation with each other, as well as, correlation with other predictors. We also observe that BMI seems to have a quadratic relationship with Life Expectancy. Lastly, both Alcohol and HIV/AIDS variables show evidence of a possible log transformation to help better explain the Life Expectancy response variable.
   4. Prior to addressing the collinearity shown between Income Comp. and Schooling, we wanted to observe the individual relationship of BMI vs Life Expectancy (Appendix: 1C). We see visual evidence of a quadratic relationship and will move forward without this predictor in our model since we are trying to achieve the most interpretable model. Including quadratic terms may muddy up the interpretation of our problem of interest.
   5. Since the consideration of log transforming Alcohol and HIV seemed applicable, we first observed the correlation coefficients of our Alcohol and HIV variables as depicted in appendix: 1A (0.53,-0.62, respectively). We then log transformed both variables and saw an increase of correlation with both variables on Life Expectancy (0.56,-0.78, respectively – Appendix: 1D). With this information, we re-plotted our matrix of scatterplots (Appendix: 1E) and plotted each variable individually against Life Expectancy (Appendix: 1F/1G) to see if there is evidence of a linear relationship. Based off our graphics above, we will move forward with including both log transformed variables.
   6. Lastly, there is a categorical variable of interest named “Status” that shows the countries of being developed or developing. Our team believes this is a predictor that needs to be included to create a more intuitive regression model for these subsets of countries. More specifically, we don’t want developed countries explaining levels of variability for the response in unison with developing countries. We believe this will create a disingenuous view of the analysis and will move forward with this categorical variable in our model.

**Model Build**

Predictors of interest: Status, Adult Mortality, Log Alcohol, Log HIV/AIDS, Income Composition, and Schooling

1. **Dealing with NA Values:**
   1. With the predictors of interest listed above, we went ahead and filtered out the countries who have NA’s listed in those columns. Countries that were filtered out are listed below:
      1. South Sudan, Somalia, Coute d’ Ivor, Democratic Republic of Congo, Democratic Republic of Korea, Czechia, Republic of Korea, Tanzania, Republic of Moldova, Great Britain and Ireland. Below is the distribution of the countries by continent:
         1. Africa – 5/54 (9.3%)
         2. Europe – 4/44 (9.1%)
         3. Asia – 2/48 (4.2%)
      2. We are confident in moving forward with the filtered data set without these countries as it is only a small percentage of each of the world’s continents.
2. **Model Build:**
   1. We built out a total of 8 models by cross validating certain metric statistics like adjusted R^2, ASE, AIC, BIC, VIF, Residual Diagnostics, and statistical significance of our predictors. After multiple trial and error (Appendix: Model Build) of dealing with correlated predictors, insignificant terms, and differing metrics, we arrived at the following model:
   2. **Life Expectancy = 55.682 - 0.018Adult.Mortality -1.963StatusDeveloping + 26.827Income.comp.of.resources -1.423LogHIVAIDS + 0.226LogAlcohol**
      1. It is important to note here that each of the predictors included in our model have slopes that are all statistically significant from an alpha = 0.05 threshold. Through multiple trial and error, we found that schooling was a redundant predictor that we ended up excluding from the model due to it making certain predictors insignificant (Appendix: Model Build). We have also included metrics of R^2, AIC, BIC, ASE and VIF (Appendix: 1I).
3. **Assumptions:** 
   1. See Appendix: 1J for plot visuals regarding assumption analysis below
   2. **Q-Q Plot of Residuals**: The Q-Q Plot of residuals provides no evidence against normality
   3. **Histogram of Residuals**: The histogram of residuals does not provide strong evidence that the residuals are not normally distributed.
   4. **Residual Plot**: A nice scattered cloud around 0 doesn’t show outliers or lack of homoscedasticity
   5. **Studentized Residual Plot**: Identifies 1 concerning point. We will explore further as this may provide some evidence against the normality assumption.
   6. **Leverage:** Shows 2 observations with high residual and high leverage that may be affecting our model. Observation 49 (Equatorial Guinea) seems to be an egregious point that may need further examination.
   7. **Cook’s D:** The Cook’s D corroborates our initial assumption above about observation 49; however, we will keep this point in our model as we do not have enough domain knowledge to believe it shouldn’t be included.
   8. **Statement:** This model is the most reasonable fit and will move forward as such. You can reference Appendix: 1I for these metrics.
      1. ASE = 9.081
      2. AIC = 881.575
      3. BIC = 903.607
      4. Adj R^2 = 0.8657
      5. Summary Statistics of each variable including count, mean, standard deviation, min, and max (Appendix: 1K)
4. **Model Cross-Validation:**
   1. We ran a 70/30 train test split to see how well the predictive ability of our model truly was. We used ASE as our metric of interest when doing this cross validation. We ran 200 iterations of different train test splits and averaged out the test ASE to see how it performed on our training set. See Appendix: 1L for code and output. Below we will list the ASE of our original model vs the average ASE of our train/test cross validation method:
      1. **Original Model:**
         1. ASE = 9.081
      2. **Cross-Validation:**
         1. ASE = 10.08 … over 200 iterations averaged out.

\*We strongly believe our model is an adequate fit for the 2014 WHO data.\*

1. **Parameters (estimates/interpretation, 95% confidence intervals)**
   1. Parameter Estimates

**Life Expectancy = 55.682 - 0.018Adult.Mortality -1.963StatusDeveloping + 26.827Income.comp.of.resources -1.423LogHIVAIDS + 0.226LogAlcohol**

**Individual Regression Equations**

* + 1. Developing Countries (When StatusDeveloping = 1)

Life Expectancy = 53.72 – 0.018Adult.Mortality + 28.827Income - 1.423LogHIVAIDS + 0.226LogAlcohol

* + 1. Developed Countries (When StatusDeveloping = 0)

Life Expectancy = 55.68 – 0.018Adult.Mortality + 28.827Income – 1.423LogHIVAIDS + 0.226LogAlcohol

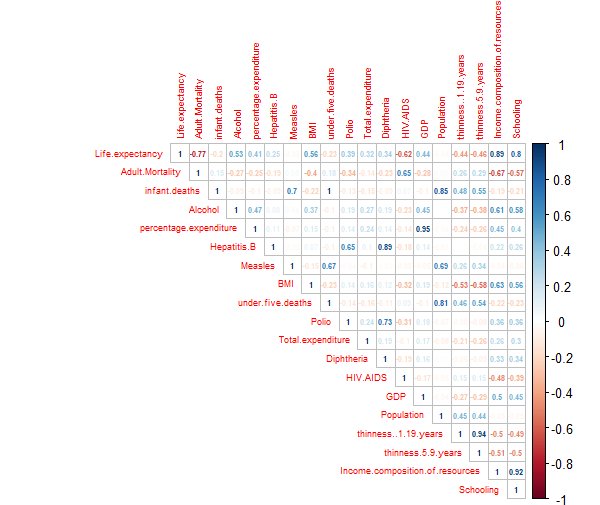
* 1. Interpretation
     1. **Developing Countries**: The associated Life Expectancy is 53.72 when holding all variables constant except the intercept (p-value: <0.0001). It is important to note this doesn’t have any practical significance. Holding all variables constant, for every unit increase in Adult Mortality, there is an associated decrease of 0.018 in Life Expectancy (p-value: <0.0001). Holding all variables constant, for every unit increase in Income composition of resources, there is an associated increase of 28.827 increase in Life Expectancy (p-value: <0.0001). Holding all variables constant, a doubling in HIV/AIDS is associated with a decrease of 0.986 (1.423\*ln(2)) in Life Expectancy (p-value: <0.0001). Holding all variables constant, a doubling in Alcohol is associated with an increase of .157 (0.226\*ln(2)) in Life Expectancy (p-value: 0.018).
     2. **Developed Countries**: All the interpretations are the same except for the intercept of 55.68 when holding all variables constant. This doesn’t hold much significance, but it is important to note that it does make practical sense that developed countries would have a higher intercept value since they have more resources and amenities that would increase live expectancy vs developing countries.
  2. Confidence Intervals
     1. The 95% confidence intervals for the coefficients are as follows:
        1. Intercept = (50.305, 61.058)
        2. Adult Mortality = (-0.025, -0.011)
        3. StatusDeveloping = (-3.481, -0.446)
        4. Income Comp. = (20.824, 32.829)
        5. LogHIVAIDS = (-1.992,-0.855)
        6. LogAlcohol = (0.039, 0.413)

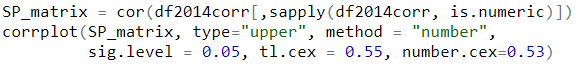
**Objective #1, Model 2**

1. **Problem Statement:** In relation to our model 1, we must now build a model with the main goal to produce the best predictions possible regarding our response variable of interest: Life Expectancy.

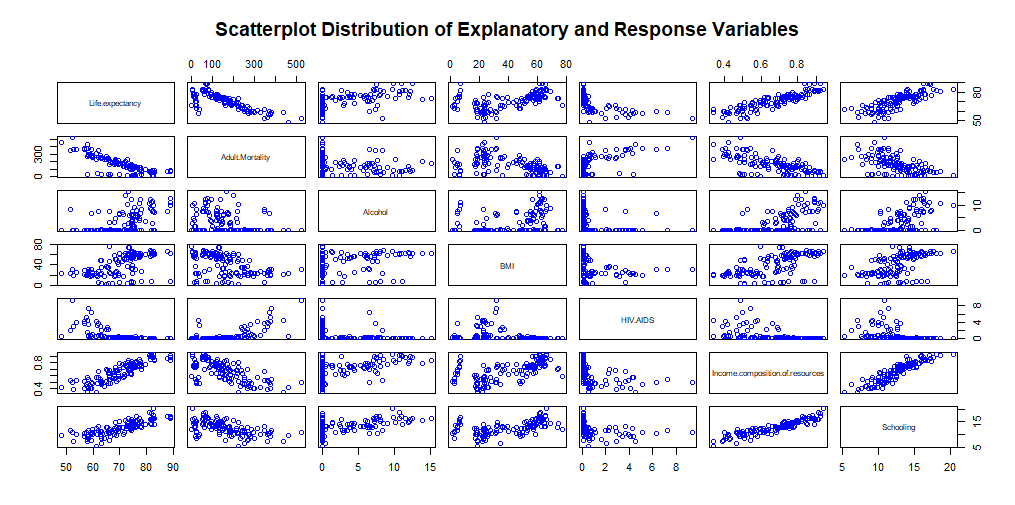
**Appendix:**

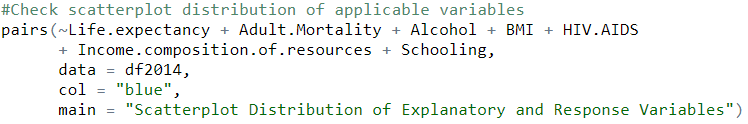
1A)



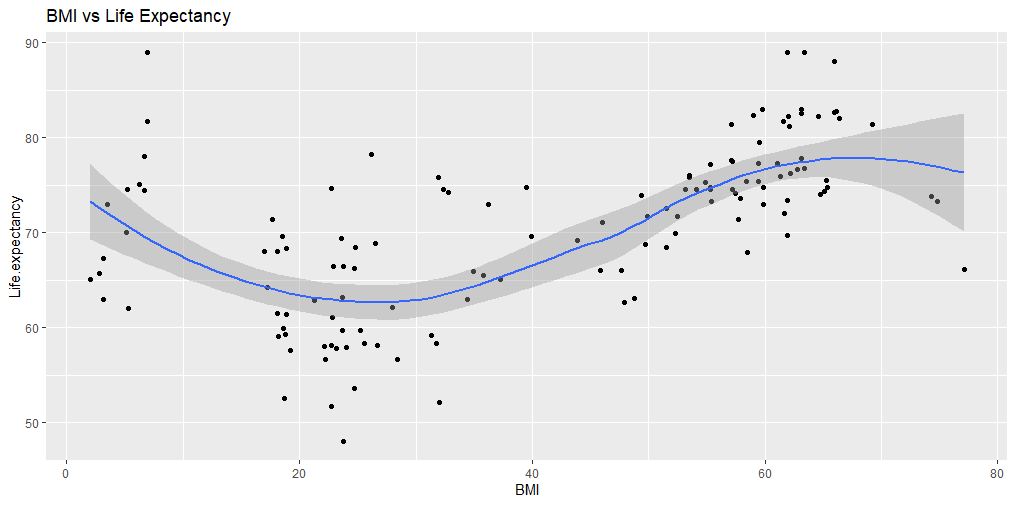


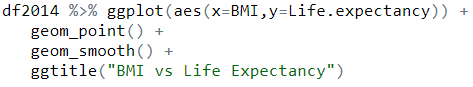
1B)



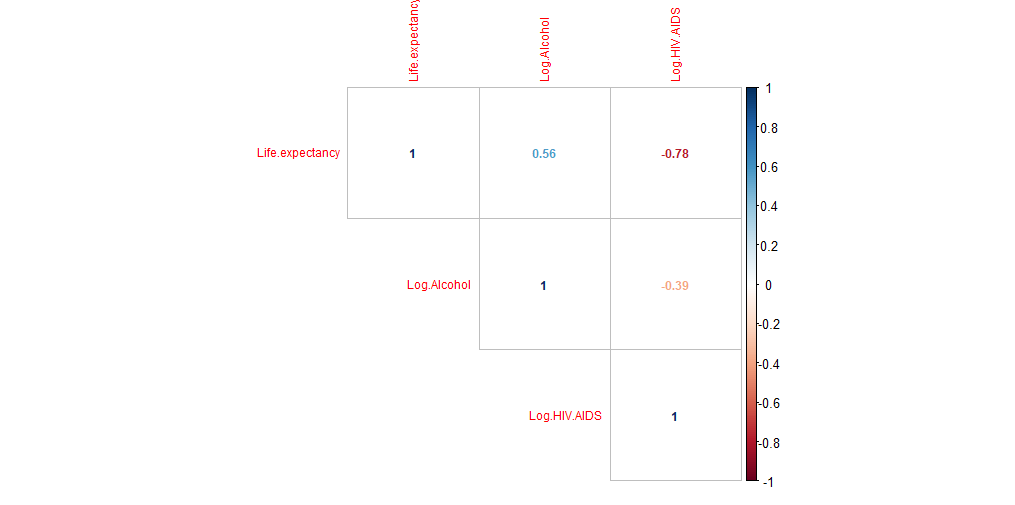


1C)



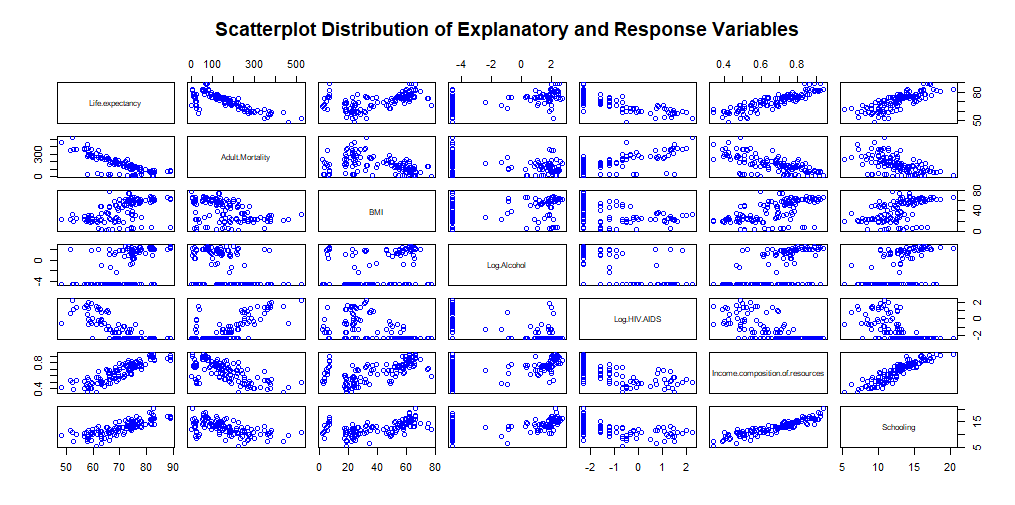


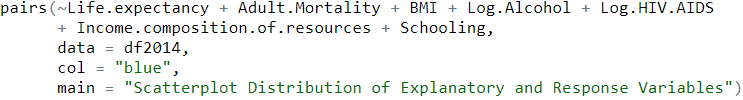
1D)



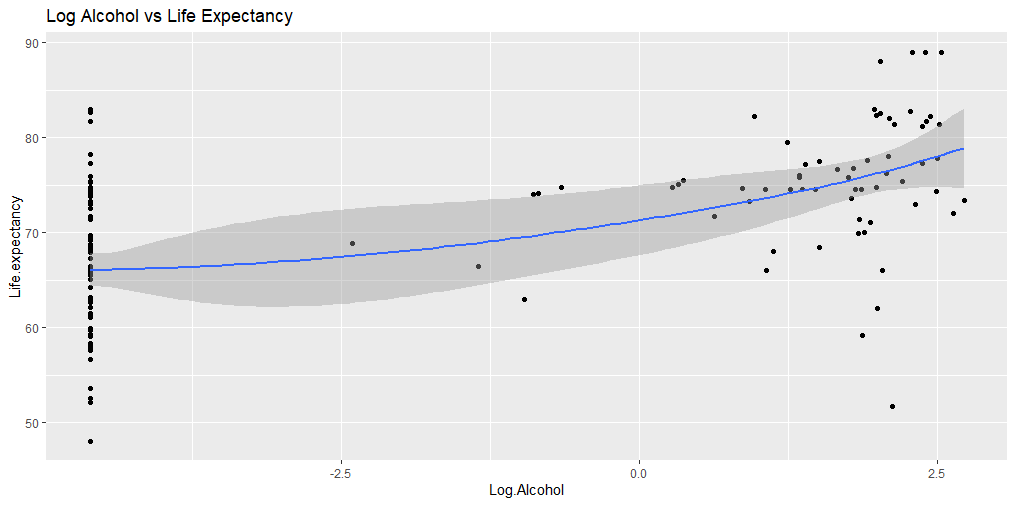


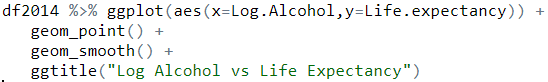
1E)

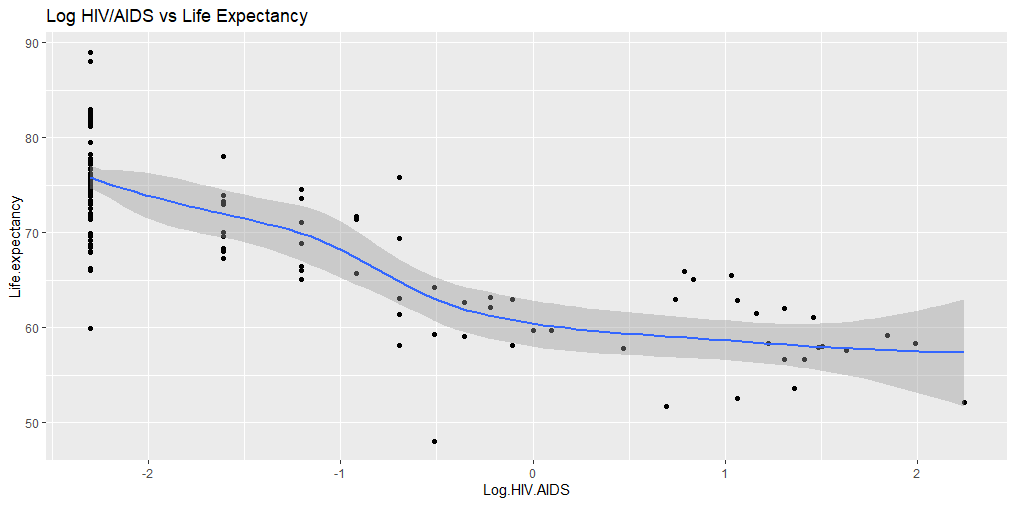


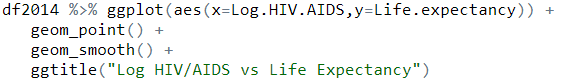


1F)

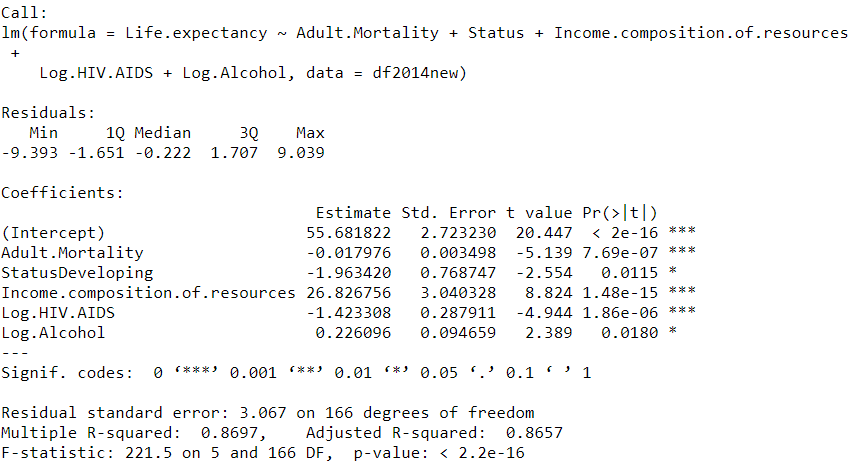




1G)

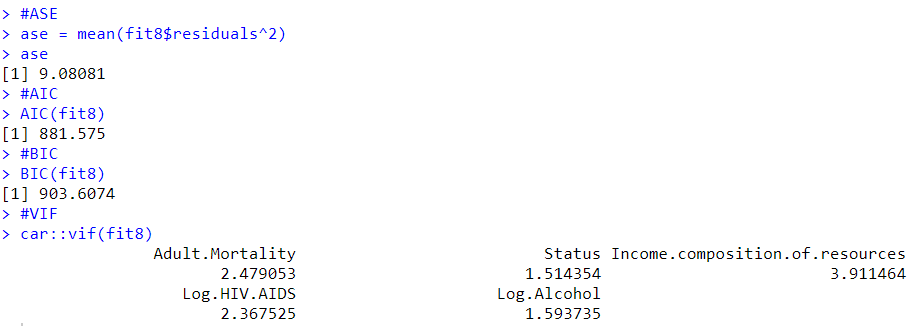


1H)

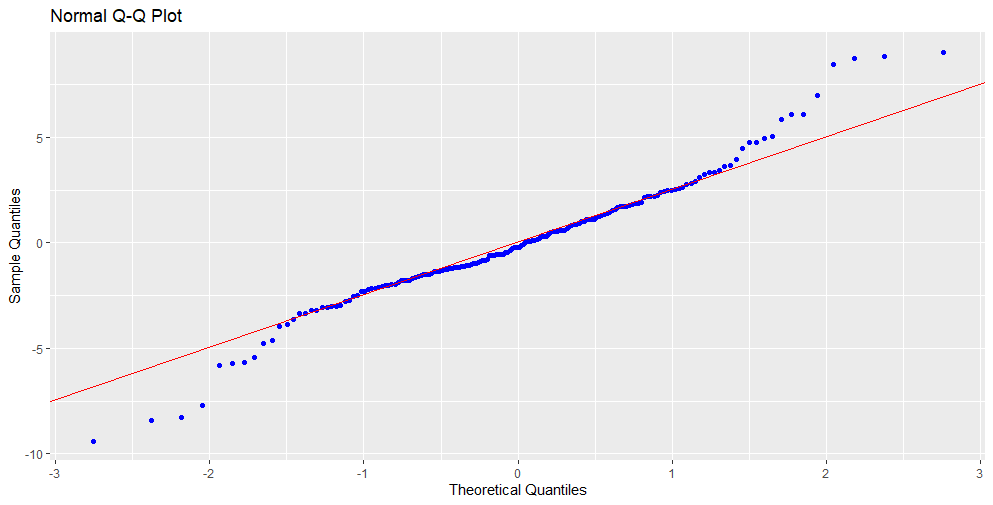


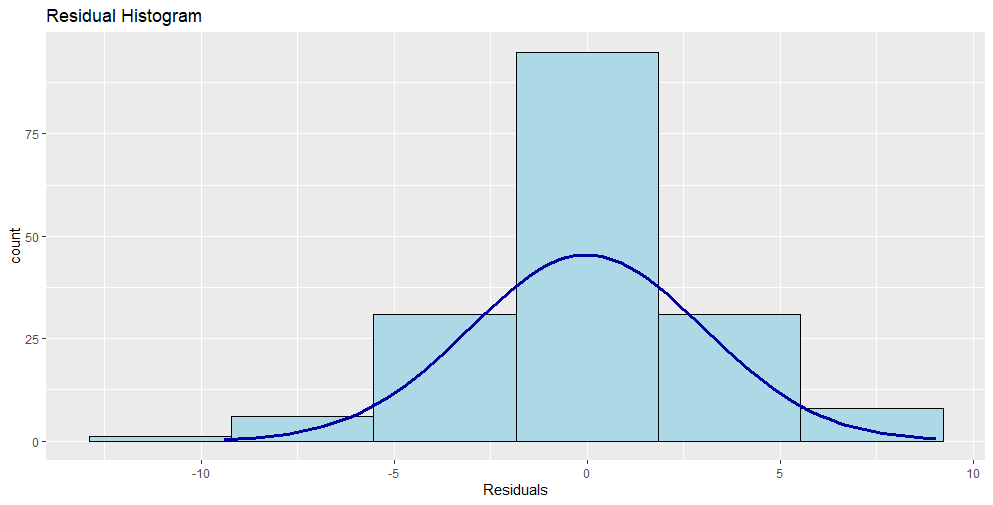


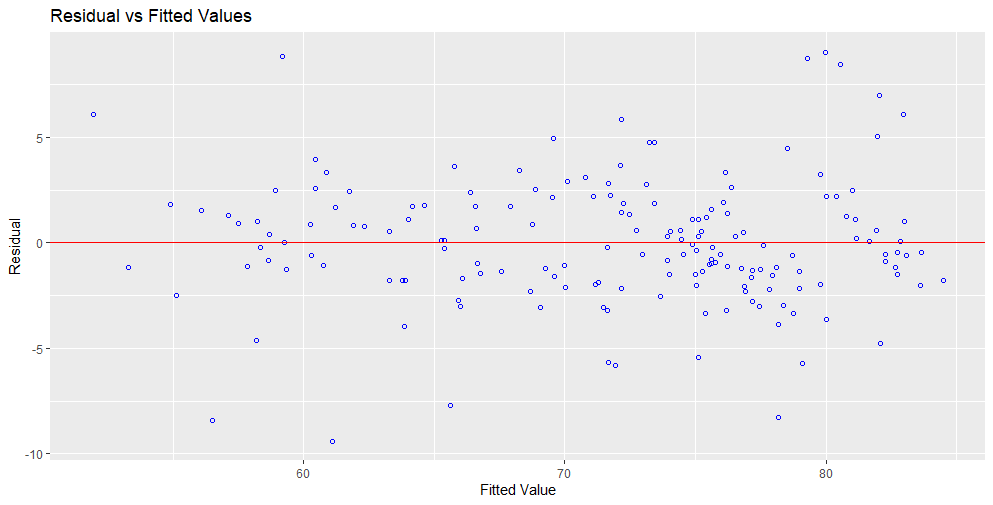
1I)

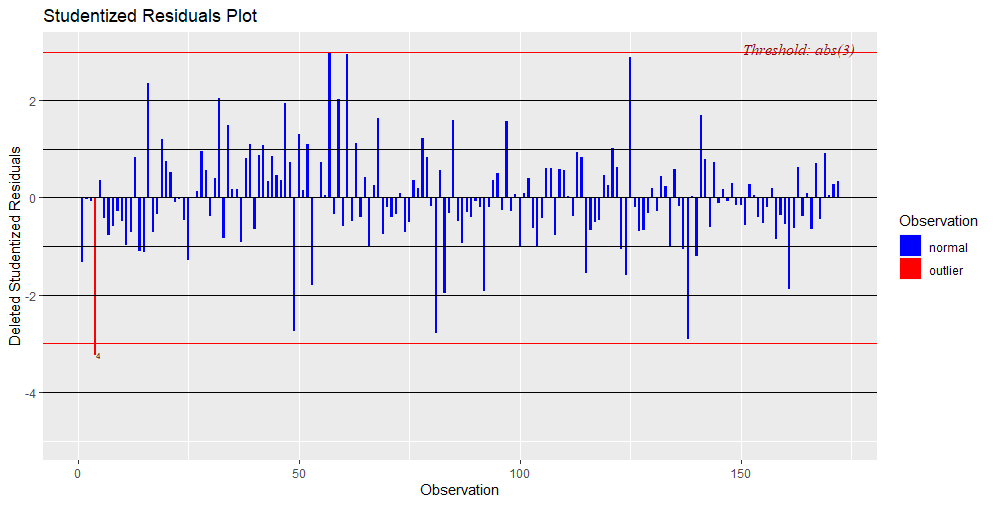


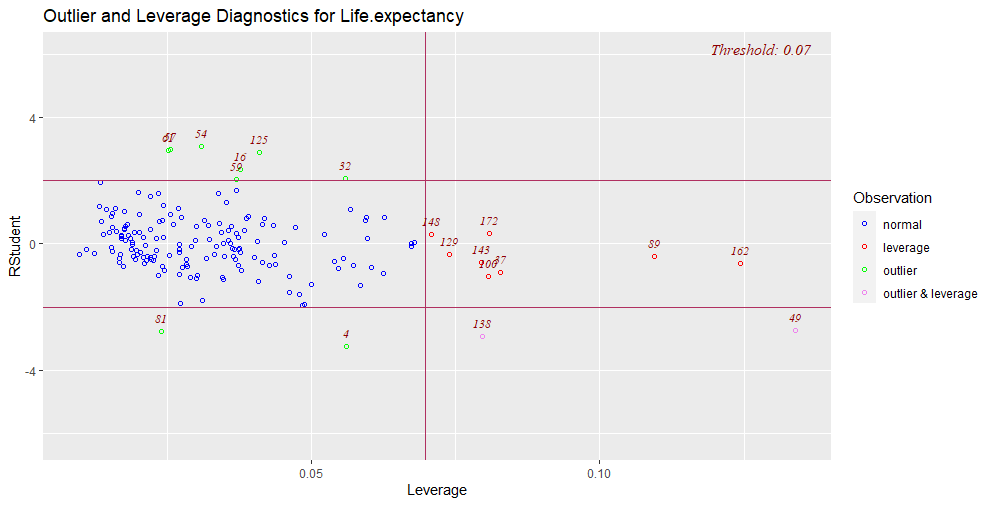
1J

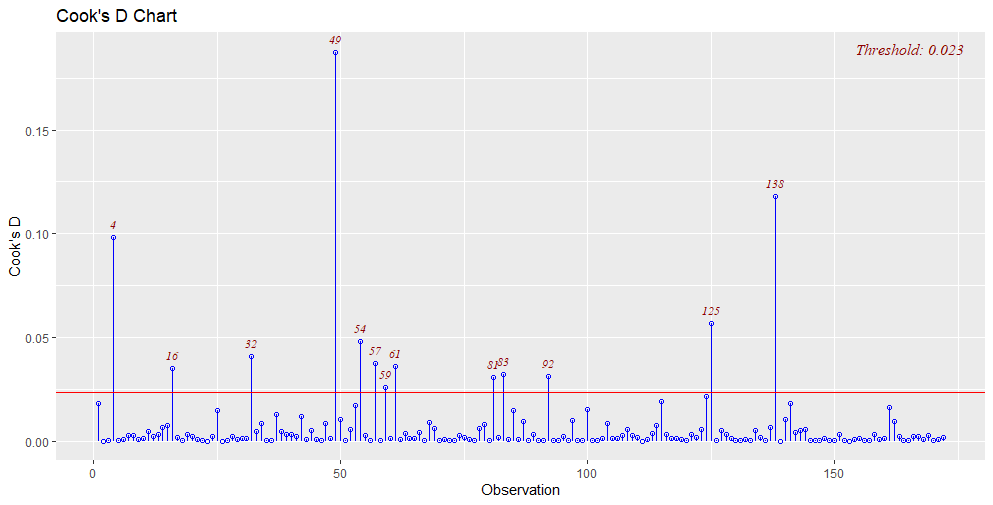


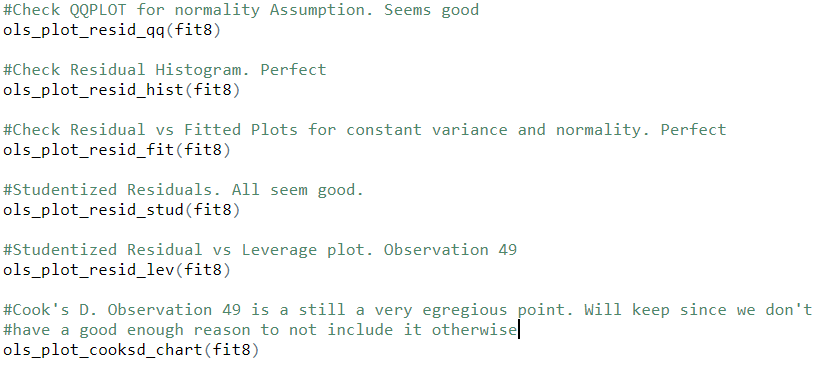




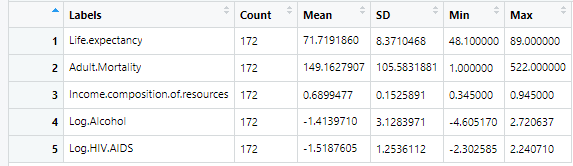


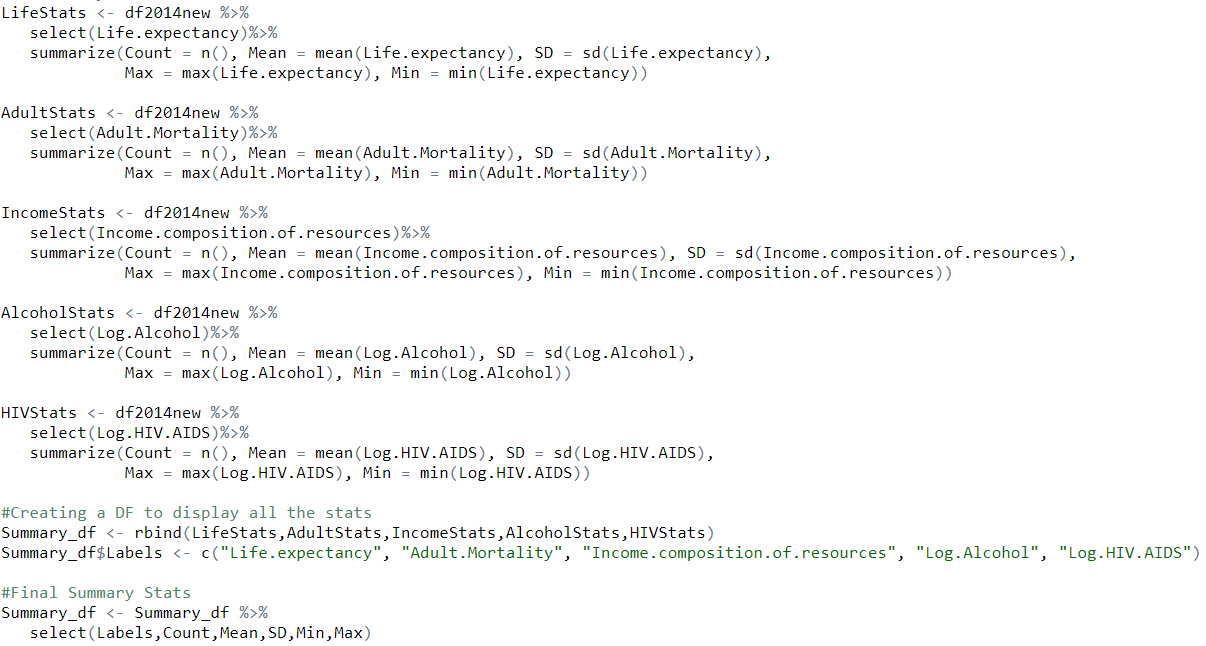






1K)





1L)

