ISTA 321 Midterm Project Report

Claire Baker and Sherif Shawashen

Introduction

For our Midterm Project, we were assigned to analyze and apply data mining techniques to the dataset, “Life Expectancy (WHO)”, a statistical analysis of factors influencing life expectancy. The dataset was compiled from the World Health Organization and contains 2938 observations of 22 variables. In order to clean this dataset, we aimed to make the variable names more cohesive and to omit “NA” values. To maintain clarity, we changed all of the variable names to lowercase and changed “.” to “_” in the variable names to access them. We also used the R function “na.omit” to clean the dataset of any “NA” values. Also, we created functions in order to perform our data analysis more efficiently. We created functions for visualizing the residual plots, visualizing the linear regression, and visualizing the confidence interval to maintain cohesion within our output. It is also important to note that our dataset contains a column for “year”, so this could affect some of our data. There could be trends or changes over time that are not accounted for in our models that could affect the relationships between our variables.

Research Question #1

The first research question that we used to build a linear regression model was “What is the relationship between consumption of alcohol and life expectancy?”. We assumed that there would be a positive relationship before testing out the assumption since increased alcohol consumption is bad for your health. We created a linear regression model comparing the variables “life expectancy” and “alcohol” and calculated the summary statistics for our data. Based on our output, the intercept showed us that when the alcohol consumption is 0, life expectancy is 64.76 years. We found that for each unit increase in alcohol consumption, life expectancy increases by 0.87925 years, holding all other factors are constant. The p-values for both the intercept and alcohol are less than 0.05, which indicates a relationship between alcohol consumption and life expectancy and we can reject the null hypothesis. Also, both the intercept and slope have very small standard errors, indicating the coefficient estimates are precise. The residual standard error is 8.054, which indicates that the regression model did not fit the dataset very well. The model's multiple r-squared is 0.1622, which means that the model explains 16.22% of the variance in life expectancy based on the independent variable of alcohol. In Figure 1, we visualized the linear regression for life expectancy versus alcohol. As you can see from the visualization, the blue regression line has a positive slope, indicating that there is a positive correlation between alcohol consumption and life expectancy in this dataset. However, this does not necessarily imply causation, because the data points are widely scattered around the regression line. While the visualization shows a positive relationship between the variables, the wide scatter of points implies that there is a weaker relationship between the two, and other factors are more significant determinants of life expectancy. The next visualization shown in Figure 2 shows that the distribution of residuals is widely spread and varies greatly from the regression line. This suggests that the model is not the best fit for this data and suggests heteroscedasticity from the uneven distribution. We also made a model to visualize the confidence intervals for our model, shown in Figures 3 and 4. We can see that in Figure 3, the confidence interval for the intercept, life expectancy, is between 64.7308102 and 65.9022146 which means that we are 95% confident that all of life expectancy years lie in this range. The same for the confidence interval for alcohol which is between 0.7826617 and 0.9758293, meaning that we are 95% confident that the alcohol consumption lies in this range. The confidence interval range for both life expectancy and alcohol does not contain a zero, which implies that there is a statistically significant relationship between alcohol consumption and life expectancy. In Figure 4, the graph shows a green dashed line, the confidence interval, which tells us that 95% of the model’s predictions fall within this range and that the predictions are reasonably reliable. The model diagnostics for this model can give us more information on whether it fits the linear assumptions (Shown in Figure 13). The Q-Q,

residuals vs. fitted, scale-location, and Cook's distance plots indicate that our model meets the assumptions of linear regression and represents a normal distribution.

Research Question #2

The second research question that we chose to build a linear regression model was, “What is the relationship between the number of years in schooling and GDP per capita?”. We wanted to find out if there was a relationship between the average number of years of schooling that a country’s population has and its gross domestic product. In order to test this, we created a linear regression model comparing the variables “gdp” and “schooling”. From our summary statistics, the intercept shows that when schooling is 0, GDP is -$17717.00 dollars, meaning that when a country has no years of schooling, its GDP level is lower. The other statistic for schooling is 1921.1, meaning that for every unit increase in schooling, GDP increases by $1921.10. The p-values for both GDP and schooling are less than 0.05, which indicates a relationship between schooling and GDP, and we can reject the null hypothesis. Both the intercept and slope have large standard errors, especially the intercept, which is 1111.9, indicating the coefficient estimates are not precise. The residual standard error is 10140, which indicates to us that the residual spread is very large. The model's multiple r-squared is 0.219, which means that the model only explains 21.9% of the variance in GDP. So, there are other factors that could explain GDP better than years of schooling. In Figure 5, we visualized the relationship between GDP and schooling. This model shows us that while schooling and GDP have a positive linear relationship, the regression line is almost flat and there are a lot of observations that do not fall close to the regression line. We can also see that from the years of schooling from 0-10 years, there is little increase in the country’s GDP, and the GDP only increases by a significant amount after 10 years of schooling and on. The next visualization in Figure 6 shows that the distribution of residuals is widely spread and varies greatly from the regression line. Many of the residuals towards the right side of the plot appear to be large, indicating that the linear model might not be the best representation of the data. For lower levels of schooling, the GDP values are relatively low, and the residuals are smaller. As the years of schooling increase, the GDP values become more spread out, and the residuals increase significantly, suggesting that GDP becomes more variable for higher education levels. To visualize the confidence intervals for our model, we created two plots shown in Figures 7 and 8. In Figure 7, we can see that the confidence interval for the intercept ranges from -19897.975 to -15536.084. This means that we can be 95% confident that the true value of GDP lies within this range. The confidence interval for schooling is between 1745.715 and 2096.409, indicating that we can be 95% confident that the true effect of schooling on GDP falls within this interval. Since the confidence intervals for both the intercept and schooling do not include zero, we can reject the null hypothesis and there is a statistically significant relationship between schooling and GDP. Figure 8 visualizes the confidence interval. The confidence interval tells us that 95% of the model’s predictions for the relationship between GDP and years of schooling fall within this range and that the predictions are reasonably reliable. The green dashed line is fairly narrow to the regression line, indicating that the estimates have a higher certainty. Similarly to our other research question, we created model diagnostics plots to test our assumptions (Shown in Figure 14). The Q-Q, residuals vs. fitted, scale-location, and Cook's distance plots indicate that there is a violation of normality and heteroskedasticity, meaning that the model does not meet the assumptions of linear regression.

Research Question #3

Our final research question for our dataset is, “Does a country’s GDP, years of schooling, status, or total expenditure affect life expectancy?”. First, we created a multiple regression model to evaluate the summary statistics for our predictor variables. The baseline for our model was that the country has a status of “Developed”. For the intercept, the model predicts that when all variables, GDP, schooling, and total expenditure are zero and status is Developed (baseline), life expectancy equals 45.67 years. Also, the intercept’s p-value is less than 0.05, which indicates that this estimate is statistically significant. All of the other coefficients (GDP, statusDeveloping, schooling) had a statistically significant p-value and we were able to prove that they had a relationship with life expectancy. The only coefficient that did not have a statistically significant p-value was total expenditure, which means that this predictor does not imply a relationship with

life expectancy.The multiple r-squared statistic is 0.5449, which means that about 54.49% of the variance in life expectancy is explained by our predictor variables. This is a moderate estimate, showing us that there are other factors that could predict life expectancy. The residual standard error is 5.945, and since this is a measure of the variability of the residuals, it means that the residuals are not spread out in the model. Our p-value is less than 2.2e-16, meaning that the model is statistically significant. As we can see from these statistics, schooling has the most significant impact on life expectancy, and total expenditure is the least significant out of our predictors. We ran a correlation test to see if any of our predictors were correlated and would influence our results and saw that schooling and status could be correlated, so we removed them from our predictor variables and ran another test to see how the summary statistics changed. After removing schooling and status, we used a regression model to analyze the interaction between GDP and total expenditure to predict life expectancy. The most important data that we found from this test was that the interaction term was -1.795e-05, which means that as GDP increases, the positive effect of total expenditure on life expectancy slightly decreases. This is statistically significant, as the p-value is less than 0.05. Increasing the total expenditure in countries with a higher GDP does not lead to as large an increase in life expectancy as it would in countries with lower GDP. We ran an ANOVA test to evaluate our hypotheses and found that both total expenditure and GDP have significant effects on life expectancy. There is a significant interaction effect between total expenditure and GDP, meaning that the impact of total expenditure on life expectancy varies depending on the level of GDP. In Figure 9, we visualized this interaction and divided the total expenditure levels into five groups to better understand how the regression lines change. As shown, all of the regression lines have positive slopes, indicating that as GDP increases, at all levels of total expenditure, life expectancy also increases. Most of the observations are clustered around the left side of the plot, showing that our observations mostly contain values of GDP from 0-25000. In Figure 10, we explored this relationship in terms of whether a country was developed vs. developing. Figures 11 and 12 show the confidence intervals for this research question, which told us that the confidence for the intercept, which is life expectancy, was slightly small, and the confidence intervals for total expenditure and GDP were very precisely estimated.We also conducted a Cohen’s F test to find out our effect sizes, and found that GDP had a much larger effect size (0.49) than total expenditure (0.11) meaning that it has a higher impact on life expectancy. In order to check our assumptions for non-linearity, we ran model diagnostics on our model for total expenditure, GDP, and life expectancy (Shown in Figure 15). The Q-Q, residuals vs. fitted, scale-location, and Cook's distance plots show that there could be a presence of non-linearity and heteroscedasticity, failing the assumptions of linear regression.

Conclusion

In conclusion, this report depicts our findings using data mining techniques to better understand the “Life Expectancy (WHO)” dataset with three guiding research questions. The first research question “What is the relationship between consumption of alcohol and life expectancy?”, showed us that there is a positive relationship between alcohol consumption and life expectancy and that the relationship is statistically significant. However, our model diagnostics showed us that the model could not be the best fit for predicting life expectancy because of the spread of residuals and low r-squared value. For our second research question, “What is the relationship between the number of years in schooling and GDP per capita?”, the linear regression model showed a statistically significant relationship between years of schooling and GDP per capita. As the number of years of schooling increased, the GDP per capita also increased. Similarly to our first question, we found that the presence of heteroscedasticity suggested that schooling is not the strongest predictor of GDP. The third research question, “Does a country’s GDP, years of schooling, status, or total expenditure affect life expectancy?”, was altered to “Does a country’s GDP or total expenditure affect life expectancy?” after finding the correlation between some of the predictor variables. We found that both GDP and total expenditure have a statistically significant effect on life expectancy and that as GDP increases, the positive effect of total expenditure on life expectancy slightly decreases. This report gave us knowledge of important factors that influence the variables of life expectancy, GDP, years of schooling, alcohol consumption, and total expenditure. In the future, we will further test more predictors for life expectancy to find more predictable models.

Addendum of Figures

Figure 1:

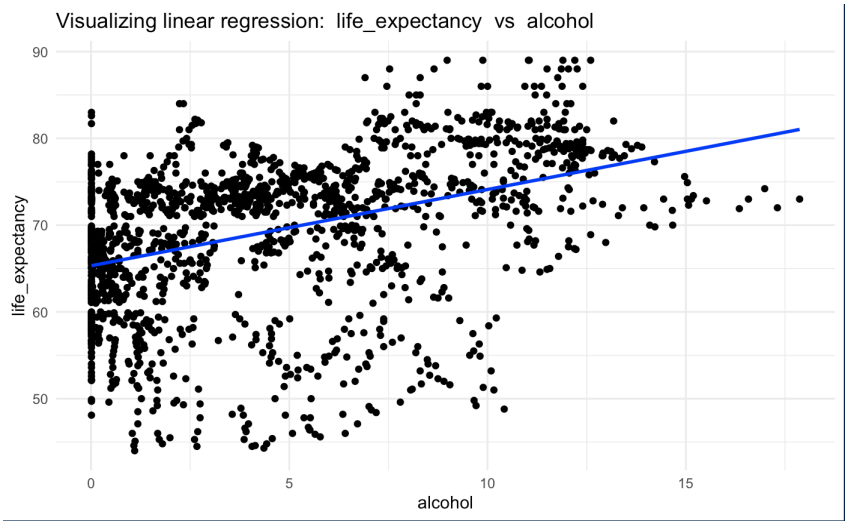


Figure 2:

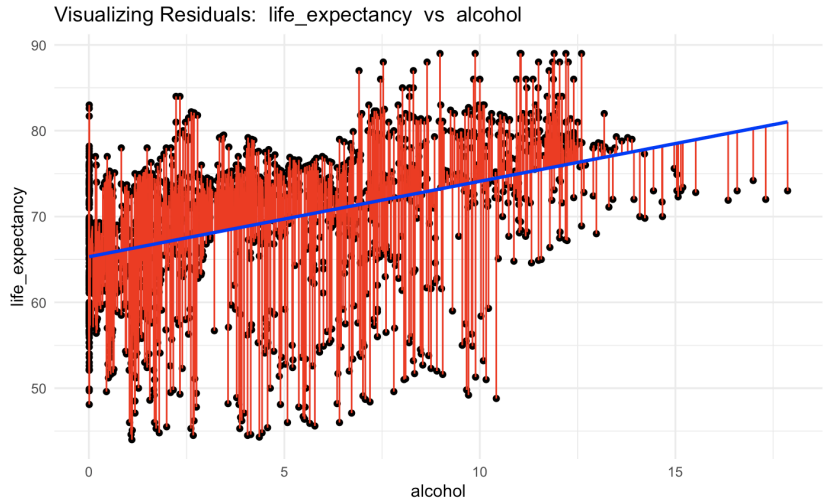


Figure 3:

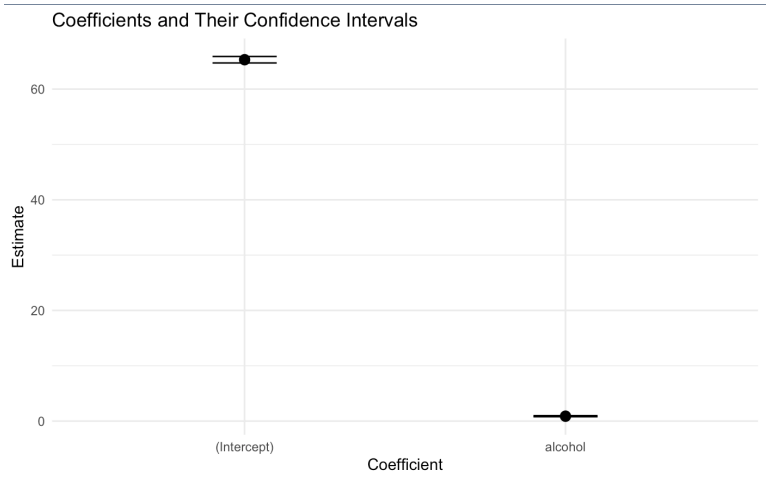


Figure 4:

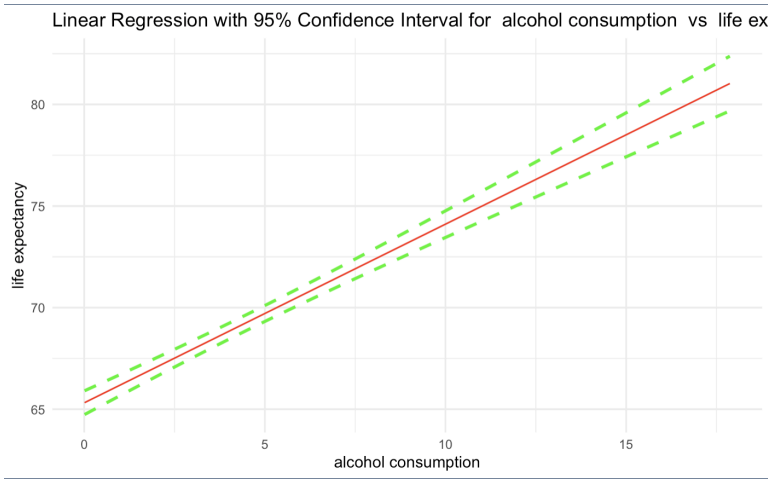


Figure 5:

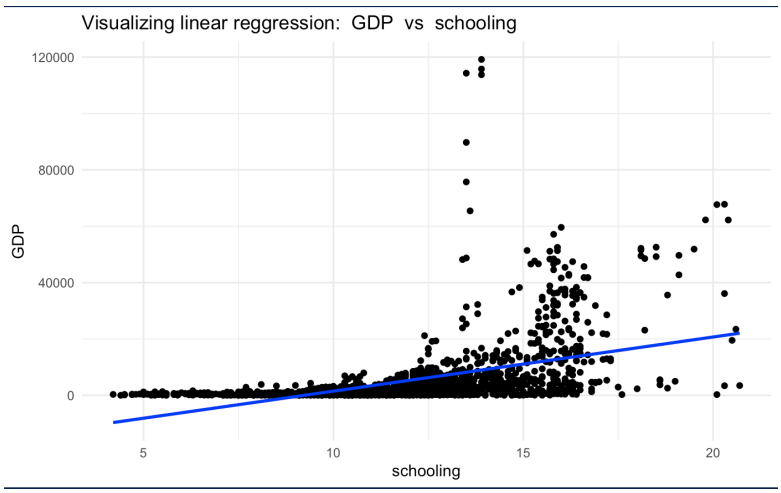


Figure 6:

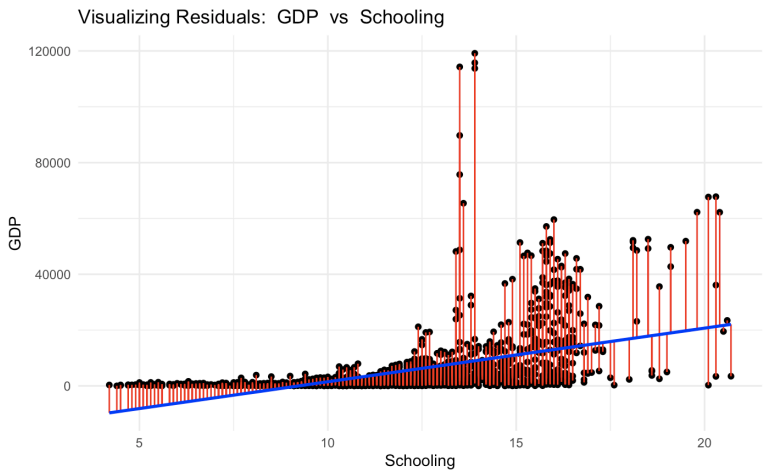


Figure 7:

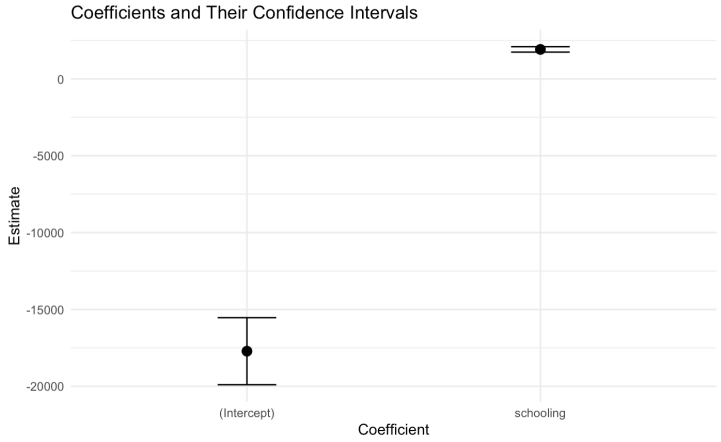


Figure 8:

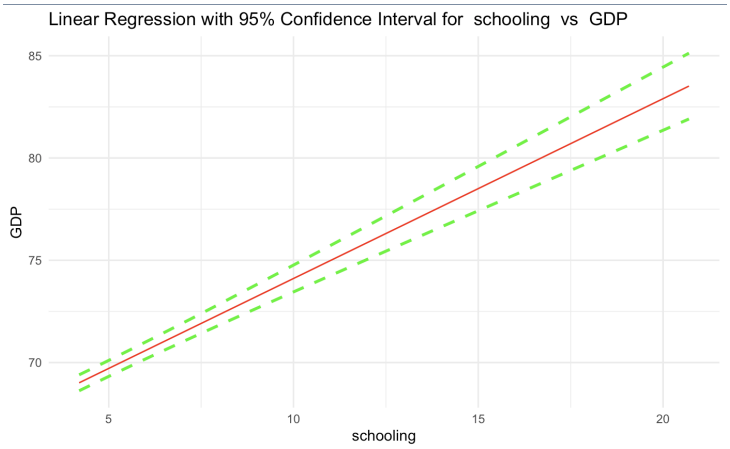


Figure 9:

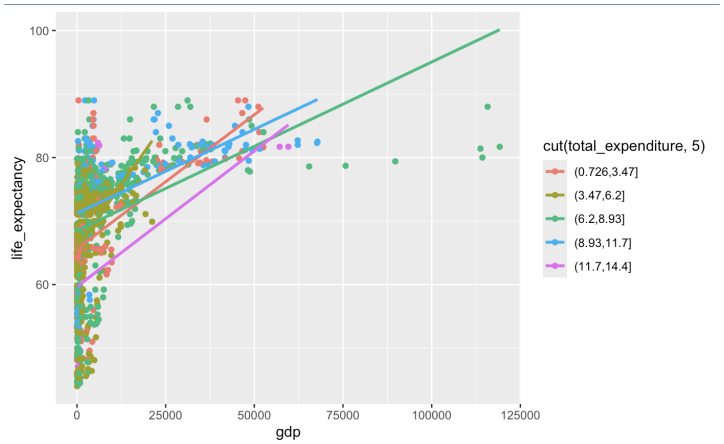


Figure 10:

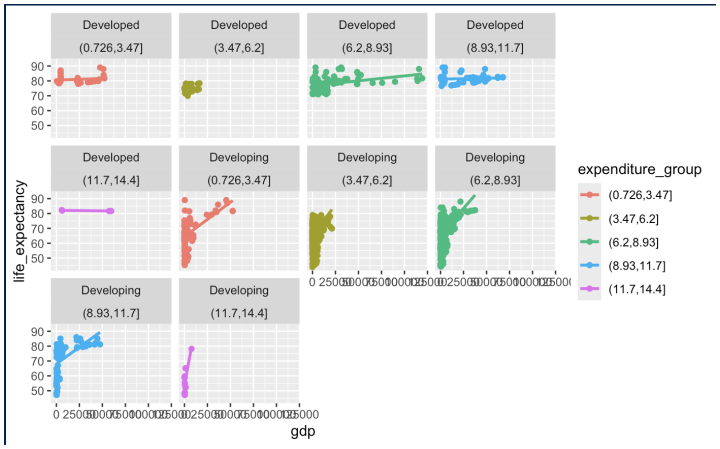


Figure 11:

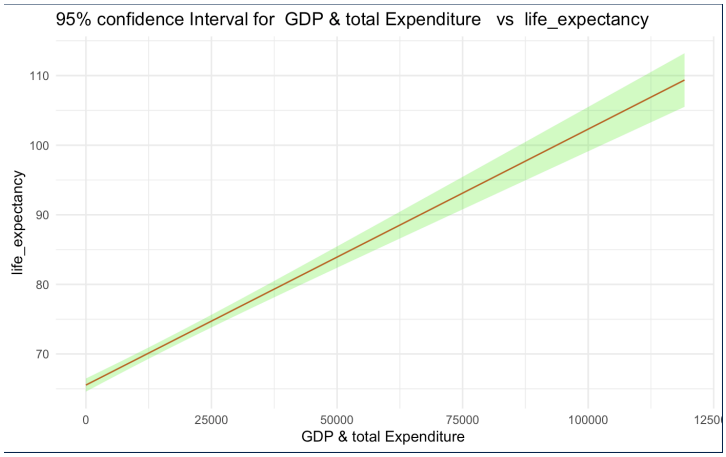


Figure 12:

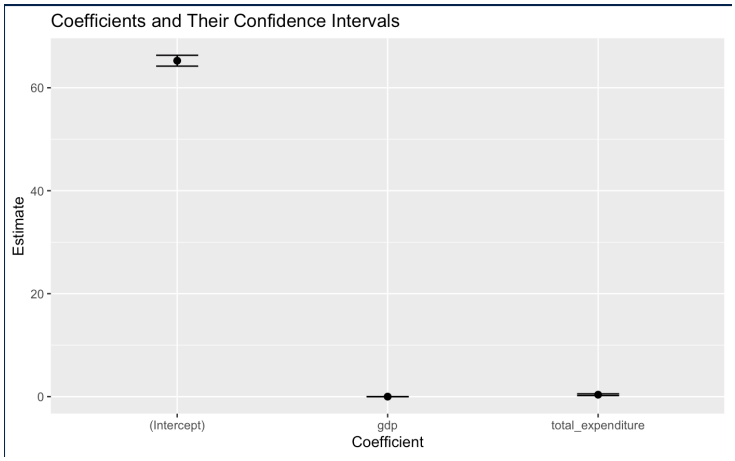


Figure 13:

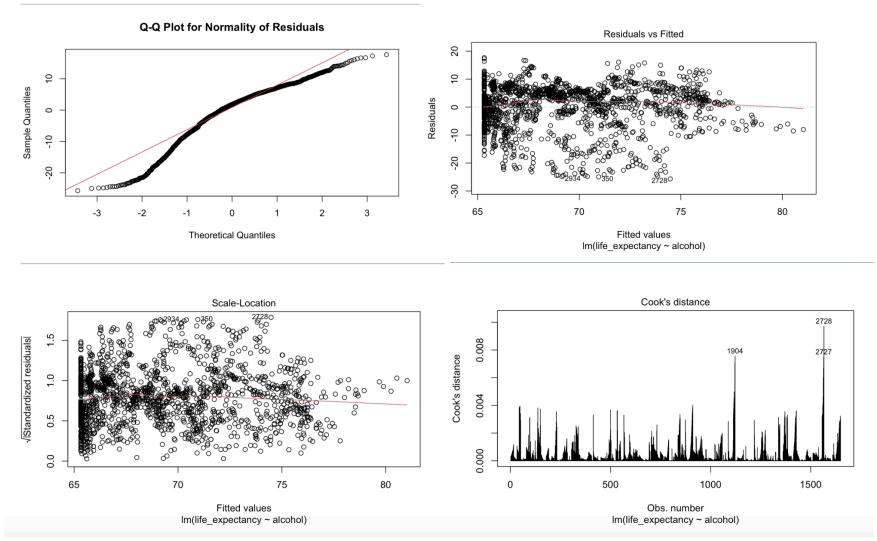


Figure 14:

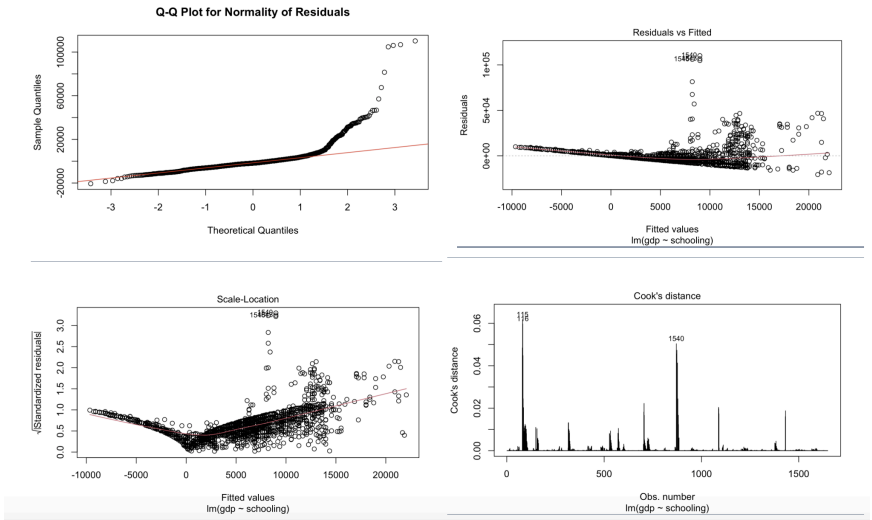
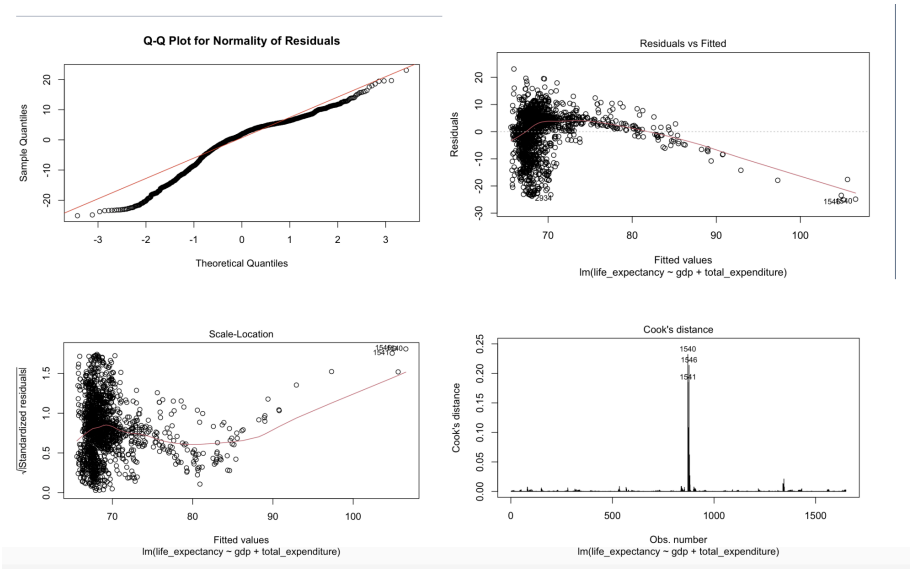


Figure 15:


Q-Q Plot for Normality of Residuals
Sample Quantiles
Theoretical Quantiles
Residuals vs Fitted
Residuals
Fitted values
lm(life_expectancy ~ gdp + total_expenditure)
Scale-Location
√|Standardized residuals|
Fitted values
lm(life_expectancy ~ gdp + total_expenditure)
Cook's distance
Cook's distance
Obs. number
lm(life_expectancy ~ gdp + total_expenditure)