Do different Variables influence the Poverty Rate of the World?

Uriel S. Pacheco-Guerrero, Donovan Lin, Marianne

2025-12-05

Abstract:

The idea of this study is to narrow down which of the 6 predictors selected for this study has considerably influenced the Poverty Headcount Ratio at $3.00 per day. The lucky predictors are Inflation, Population Growth, Unemployment, Education, Death rate, and GDP. All but GDP are noted down by countries' annual percent amounts, if that will be by percentages of GDP, total labor force, or consumer price. By using two versions of the linear regression model, we analyze the world's Poverty rate using data spanning 270 countries from 1960 to 2024.

By running a multi-variable linear regression model, where model 1 will be the variables themselves with no transformation, and model 2 will contain two transformations of the square root and logarithmic. The results demonstrate that heavy-weight predictors ended up being population growth, Death rate, GDP, and Education since they ended up being more significant than Inflation and Unemployment. We applied the square root to the predicted variable (Poverty rate) to prevent skewness from extremely low or high points influencing our results. Knowing how critical to be informed about the inner workings of what affects the poverty rate could potentially help raise the standard of living of most folks living right now. Additionally, it could help influential individuals like world leaders, legislators, and organizations like the United Nations to start enforcing policies. That can allow for more agencies to focus more on homelessness, literacy levels, and free education for those willing to learn.

Introduction:

The goal of the analysis is to figure out if the chosen variables do, in fact, influence the Poverty Rate. There are two questions we intend to answer, the first being: Are the selected variables statistically significant to be able to influence the Poverty rate of the World? The second question: Is it possible to forecast the Poverty Headcount ratio for the foreseeable future accurately?

The results of the analysis are meant to provide those in power, like legislators, who have the ability to relocate funds/pass laws to reduce the negative effects that are associated with the Poverty rate. Additionally, by making these results public, they will allow the general public to be aware of what's influencing the poverty rate for most countries. To hopefully allow them to make their respective governments take more accountability, or even be more hands-on with the causes themselves.

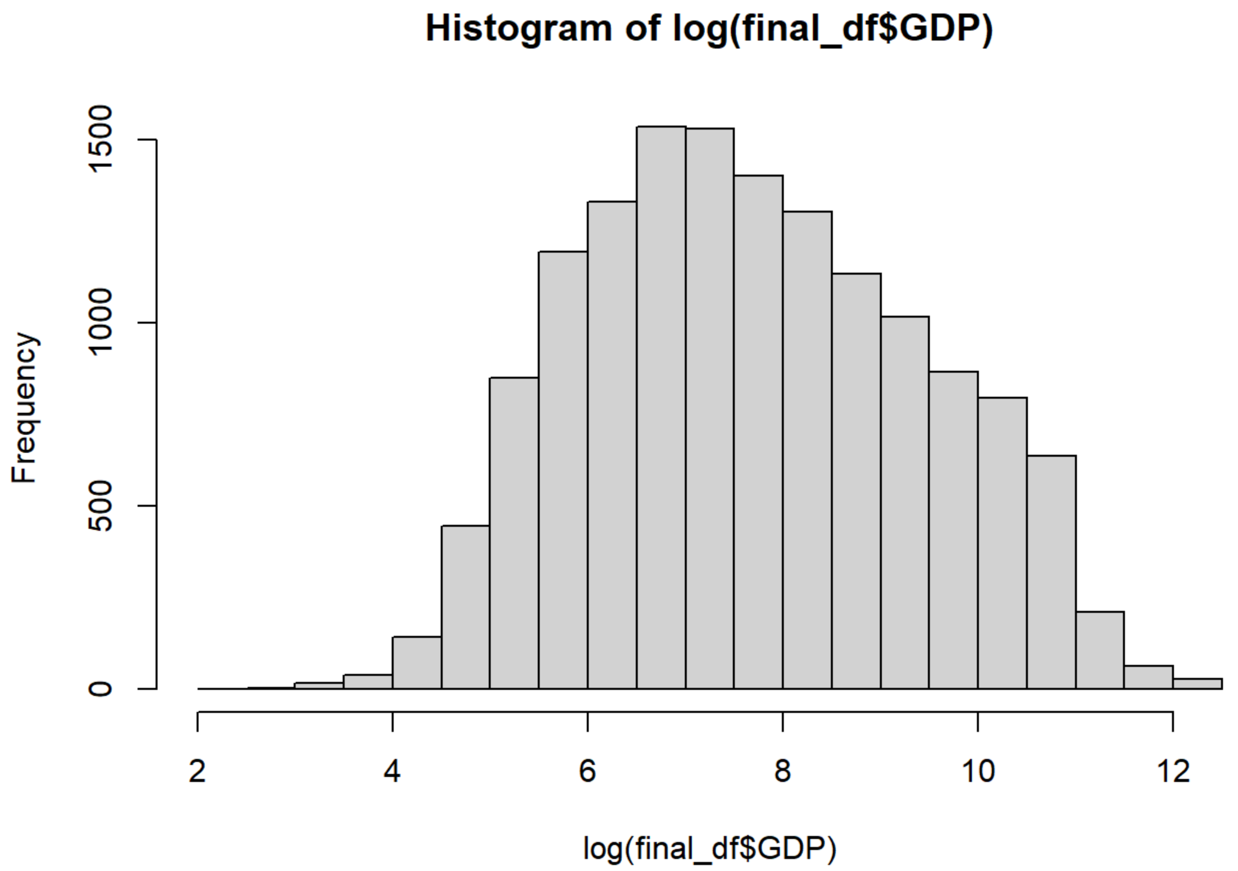
#Data: We are using 6 databases, all of which are from the World Bank Group database. The predicted variable is currently the Poverty Headcount ratio at $3.00 a day (Poverty rate). The predictor variables will be Population Growth (annual %), Inflation consumer price (annual %), Education(% of GDP), Unemployment rate (% of total workforce), GDP (in US dollars), and Death rate (crude % by [actual death/1000]).

Data Cleaning:

For all the data sets being presented here, all were downloaded in a wide form for an excel spread sheet.The rows were the 270 countries, and each column starting at column 5 were years of 1960 to 2024. However, actual data from any of the 7 data sets accumulated to be 17290 observation for each one.

So to clean the data it was necessity to convert each data set into long form by using pivot_longer() of the tidyverse package. We started at 5 because we removed columns 2 through 4 because they were indicators from the original makers, and we did this by manually removing the columns such as ND[-2:-4] into a new data frame called hot. Additionally we skipped the first 4 rows of the data set to better set up the conversion function.

Then we converted the data frame by setting into a long form by using pivot_longer() of the tidyverse package. Were we would associated each data frame with its own long form such as df_pv for Poverty rate or df_INf for Inflation rate and so on. As so not to cause unintentional skewness, many entries have NA as their entry not because they were deleted or removed but because the given data set did not have an entry for them.



#Visualization: Before trying to see the results of the linear regression line, it is critical to see the relationship each of the variables have with the Poverty rate. Since the majority of the data was collect based of annual or total percentages, applying a log transformation will provided to be useless. Even though the majority of the predictors for the potential linear regression model are all skewed to the left, since there might be countries like first world countries like Finland, United States and others that have for example a high GDP but a relativity low Poverty Rate. In this case, they might skew the data not because they are unique data. Instead because a vast amount of countries typically do not have an incredible amount of GDP in the first but also do not have a high poverty rate as well. Which is where they become clustered together with those who have low GDP but a high poverty rate, that results in getting an overall skewed graph. This is why, we intend to apply log transformation to GDP, since the units when recorded were not based off percentages. Additionally, we will apply a square root transformation to the dependent Poverty rate in hopes for it can potentially handle the skewness and result in a more normally distributed graphs.

```
##           Poverty Rate
## Inflation   0.08918017
```

```
##                   Poverty Rate
## Population Growth    0.5846786
```

```
##            Poverty Rate
## Death Rate   0.05183212
```

```
##              Poverty Rate
## Unempolyment   -0.1255289
```

```
##           Poverty Rate
## Edcuation   -0.4425305
```

```
##     Poverty Rate
## GDP   -0.8210587
```

Based off the histogram of a log-transformed GDP proves that a linear regression is possible, though not in the way that one may think. By looking at the plots for the predictors, it is possible to assume that a majority of the variables do not really have much of a correlation to the data itself. Out of all the variables, a log-transformed GDP is showed to have the strongest correlation since it is currently standing at a -.8210587. Which implies that has GDP increases, Poverty rate is expected to fall a considerable amount. In the case for the rest, plots of Population growth, and Education it is quite apparent that they have a moderately strong correlation to Poverty rate. Since Population Growth is currently standing at a .5846786 correlation, implying as for every percent increased for population growth the poverty rate is predicted to rise along side of it. While for Education is at a -.4425305, which is considerably weaker, but it implies as for every percent increase for Education the Poverty is expected to drop from it. However, in the case for Inflation, Death Rate, and Unemployment they have have from a weak to extremely weak correlation to Poverty rate since none of them pass -0.20 or +0.20.

Figure 2.2

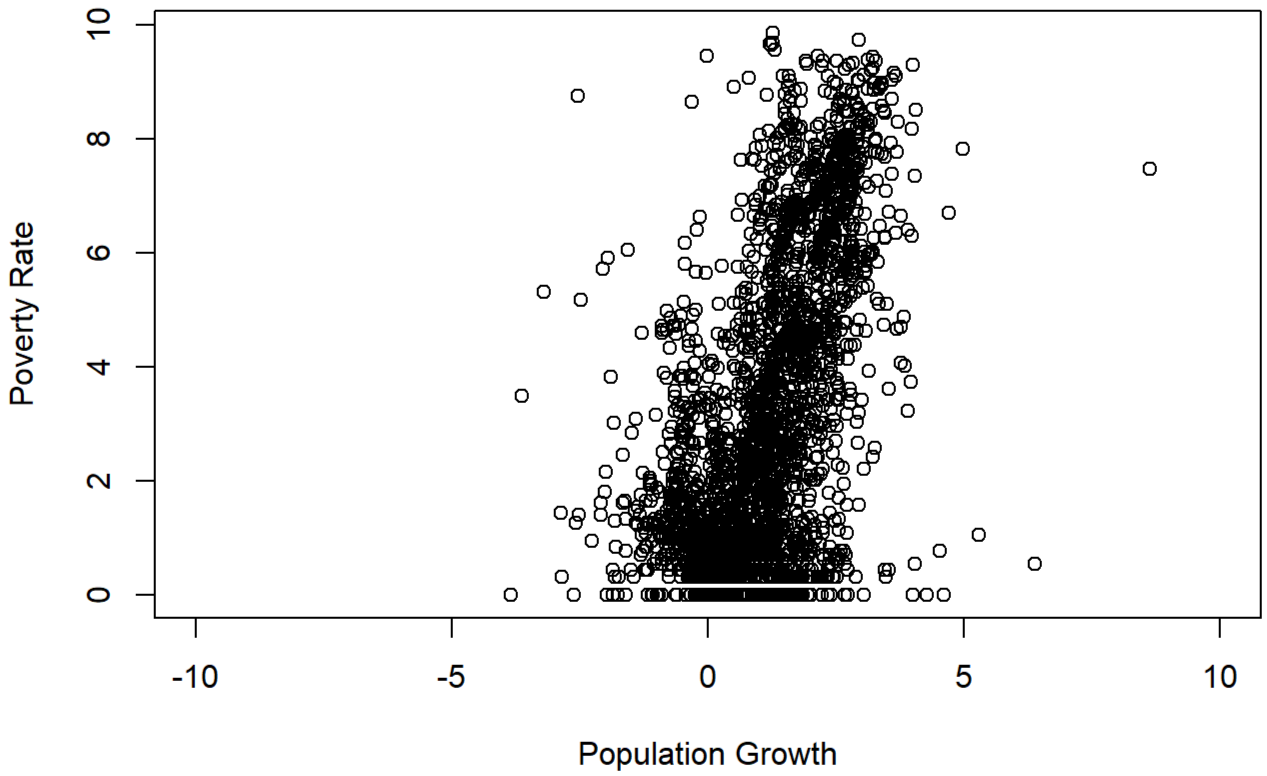


Figure 2.4

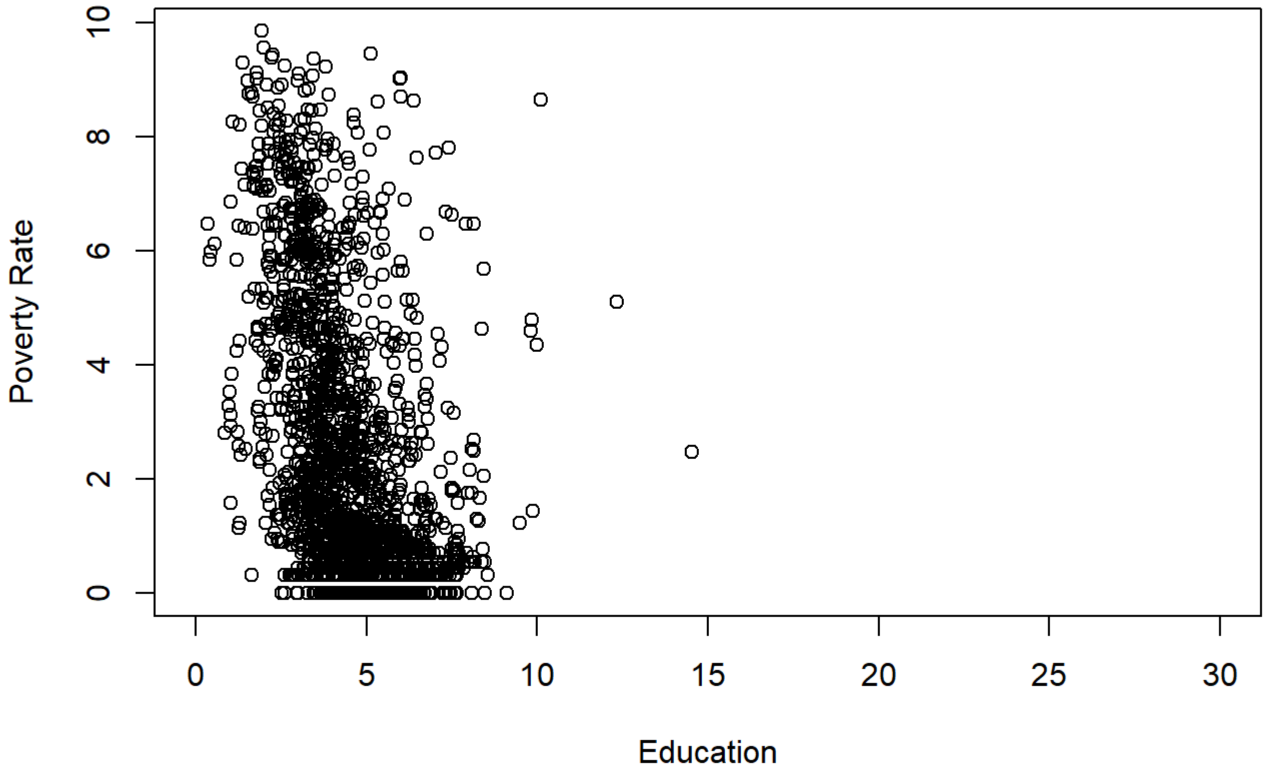
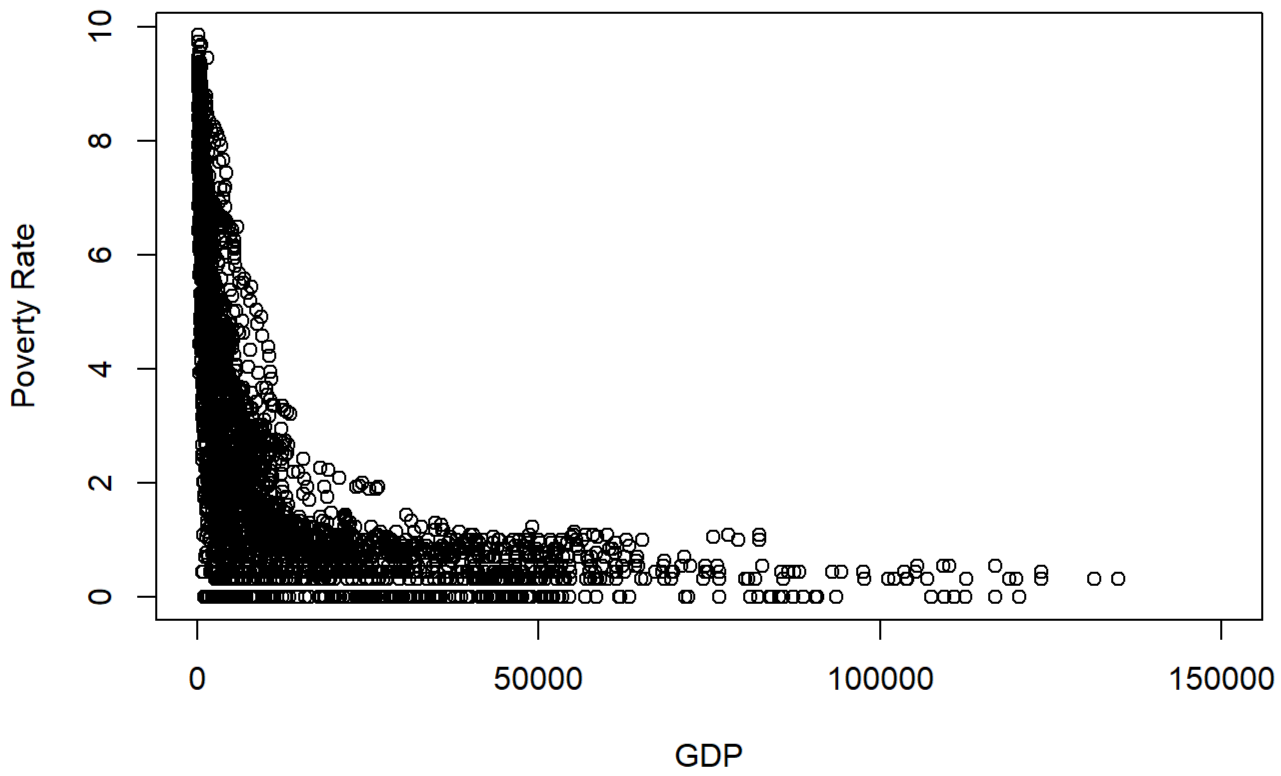


Figure 2.5



In which ever the case, we need to illustrate that the results of the correlation was determined by the transformation of log and square root to GDP and Poverty rate. The results showed that it is possible to run a linear regression, because the transformation may not have completely normalized the results in a way that was hope. However, we can not over look the fact that at least three of the predictors were proven to not have at least a moderate relationship to Poverty rate. Due to this, swapping those variables with others may provide a much stronger correlation to our predictor and not gathering mostly annual percentages may prove to increase the variables relation with the dependent variable.

Figure 2

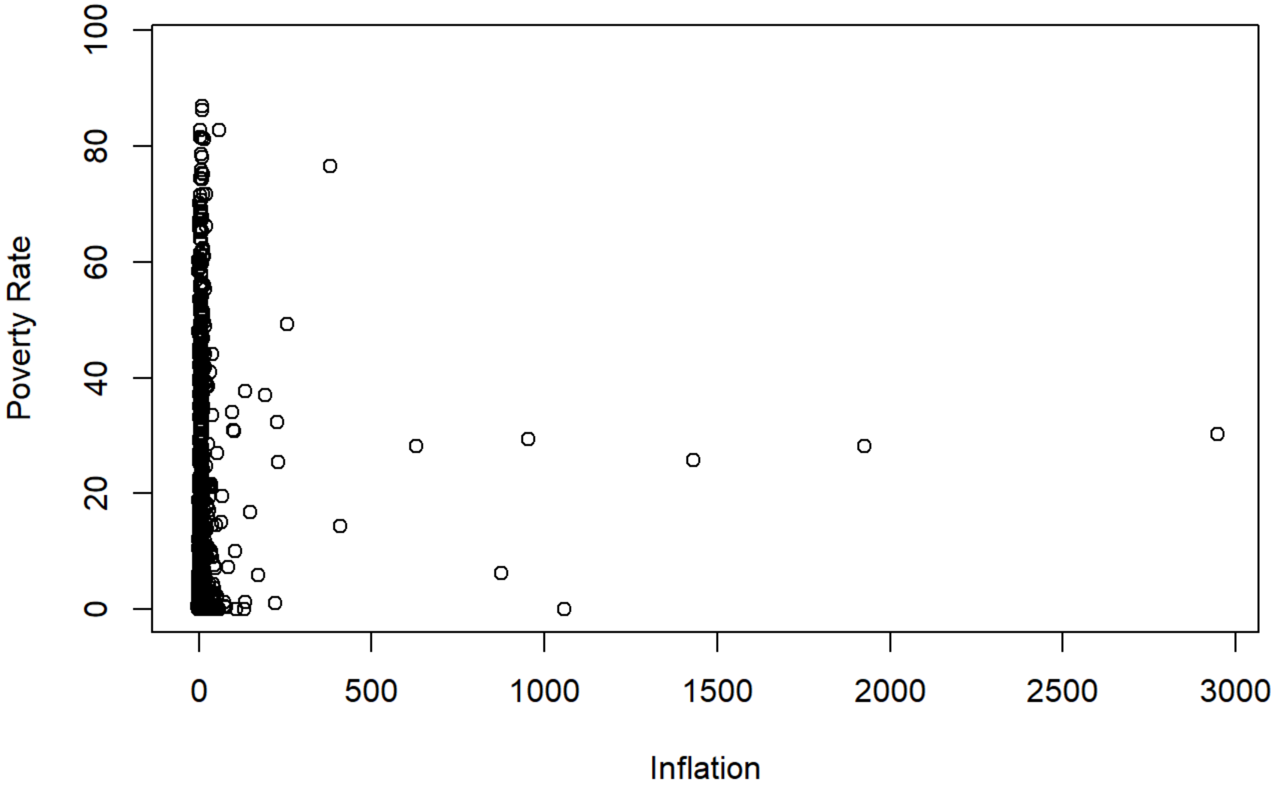


Figure 2.3

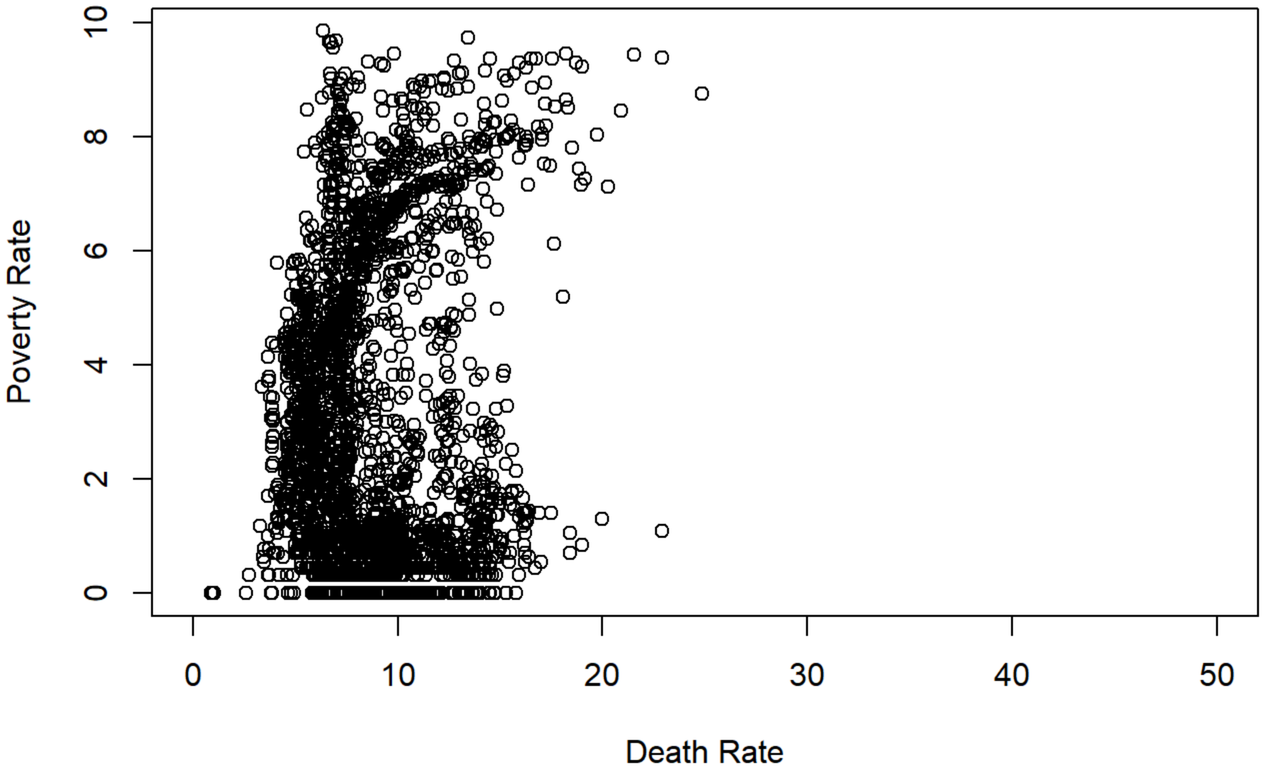
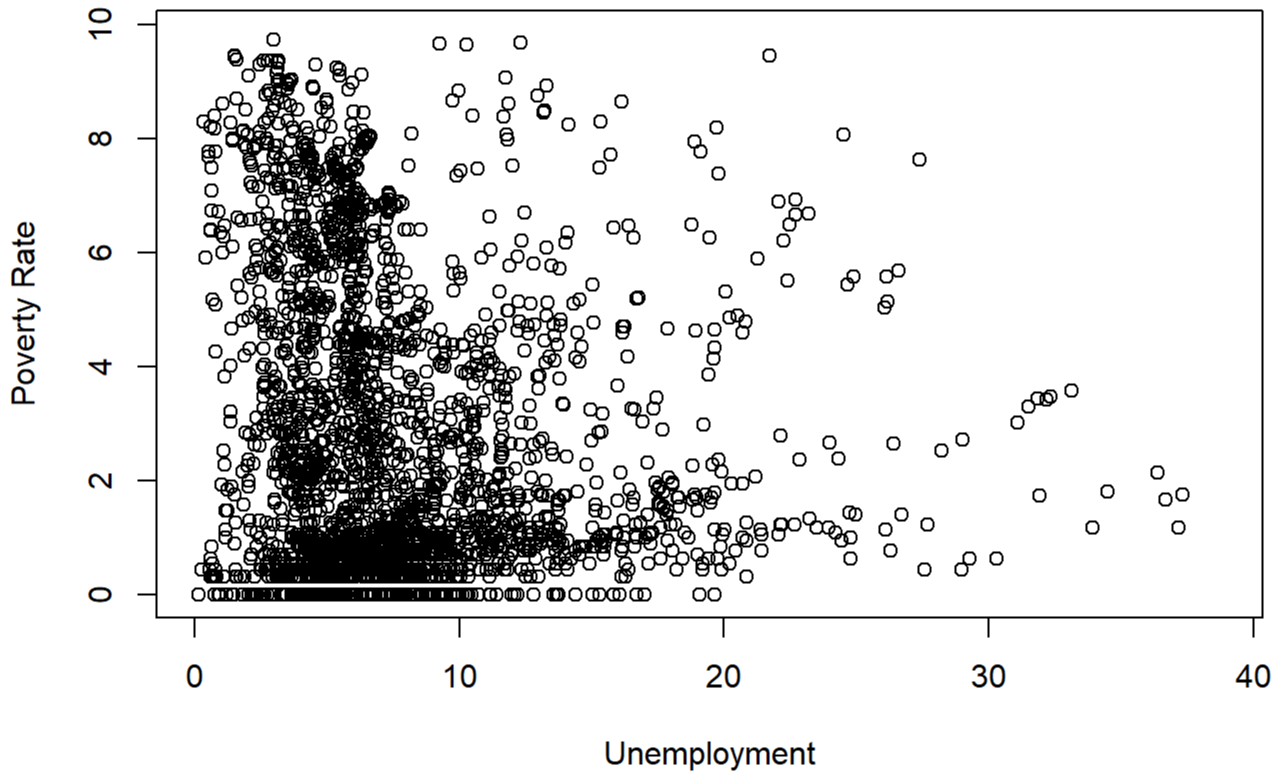
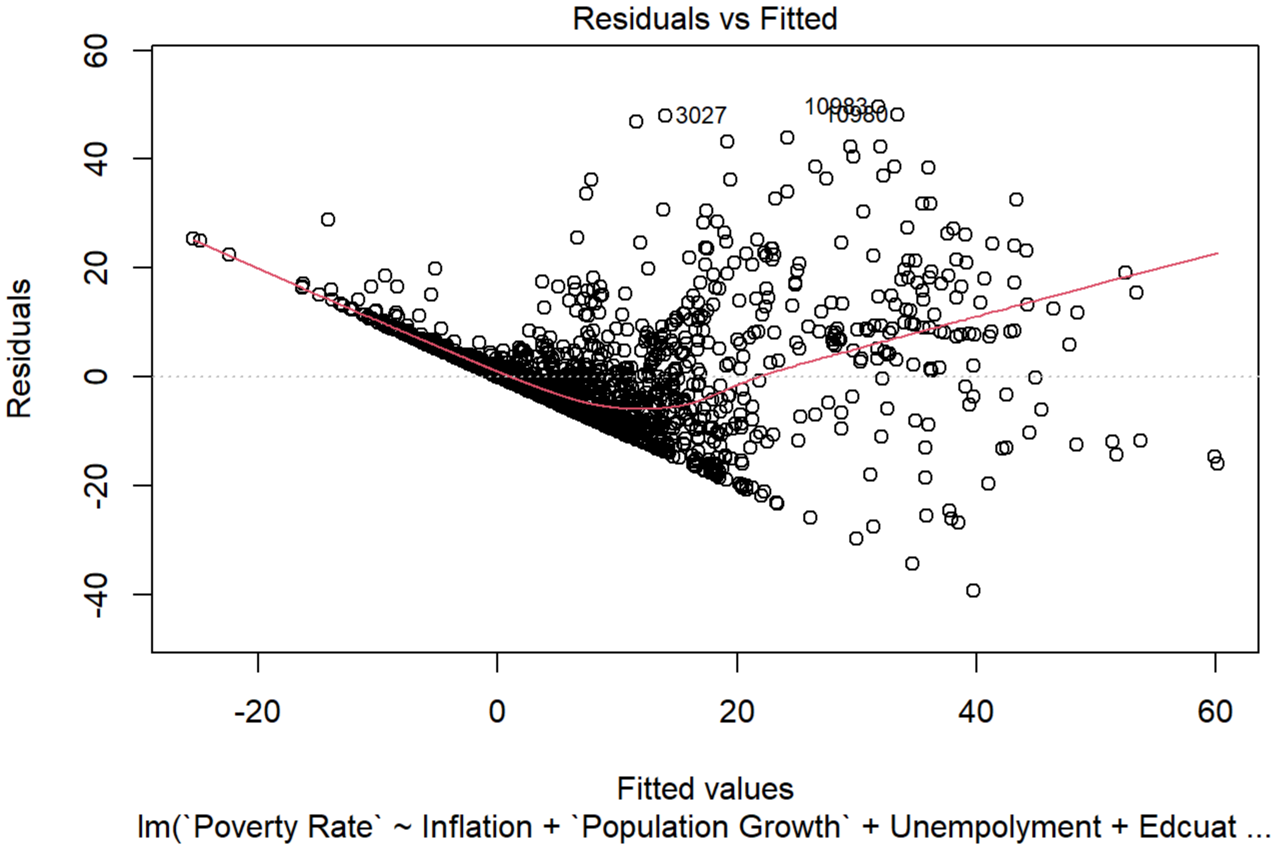


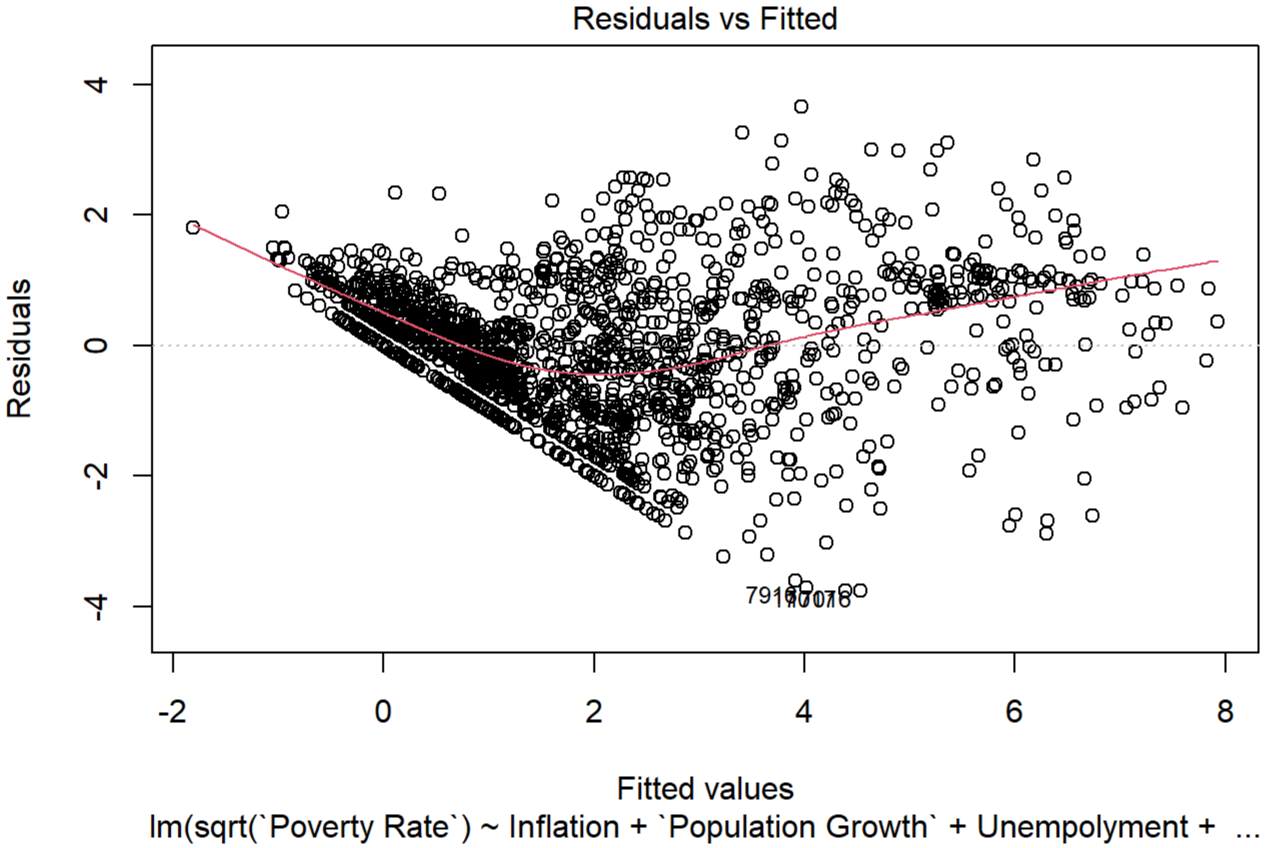
Figure 2.4



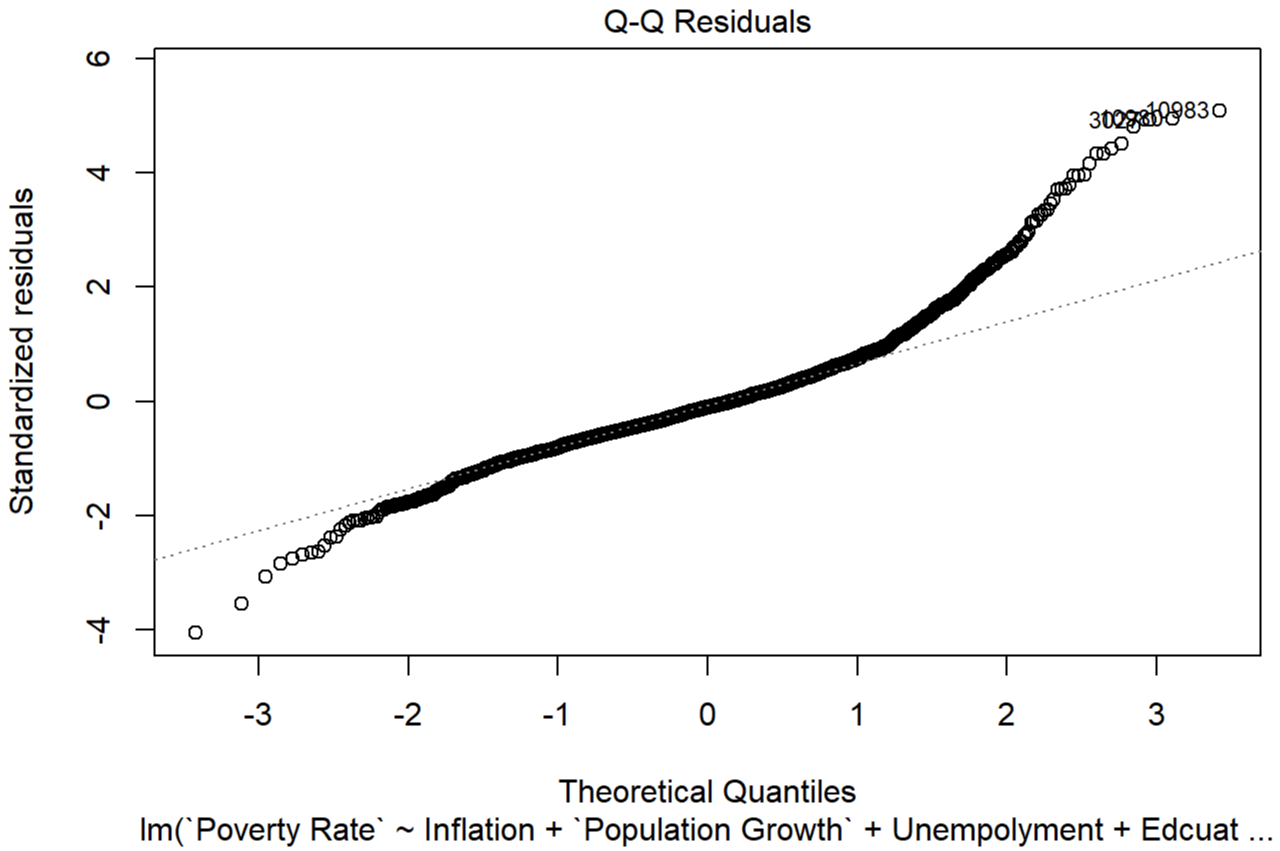
Diagnostic:

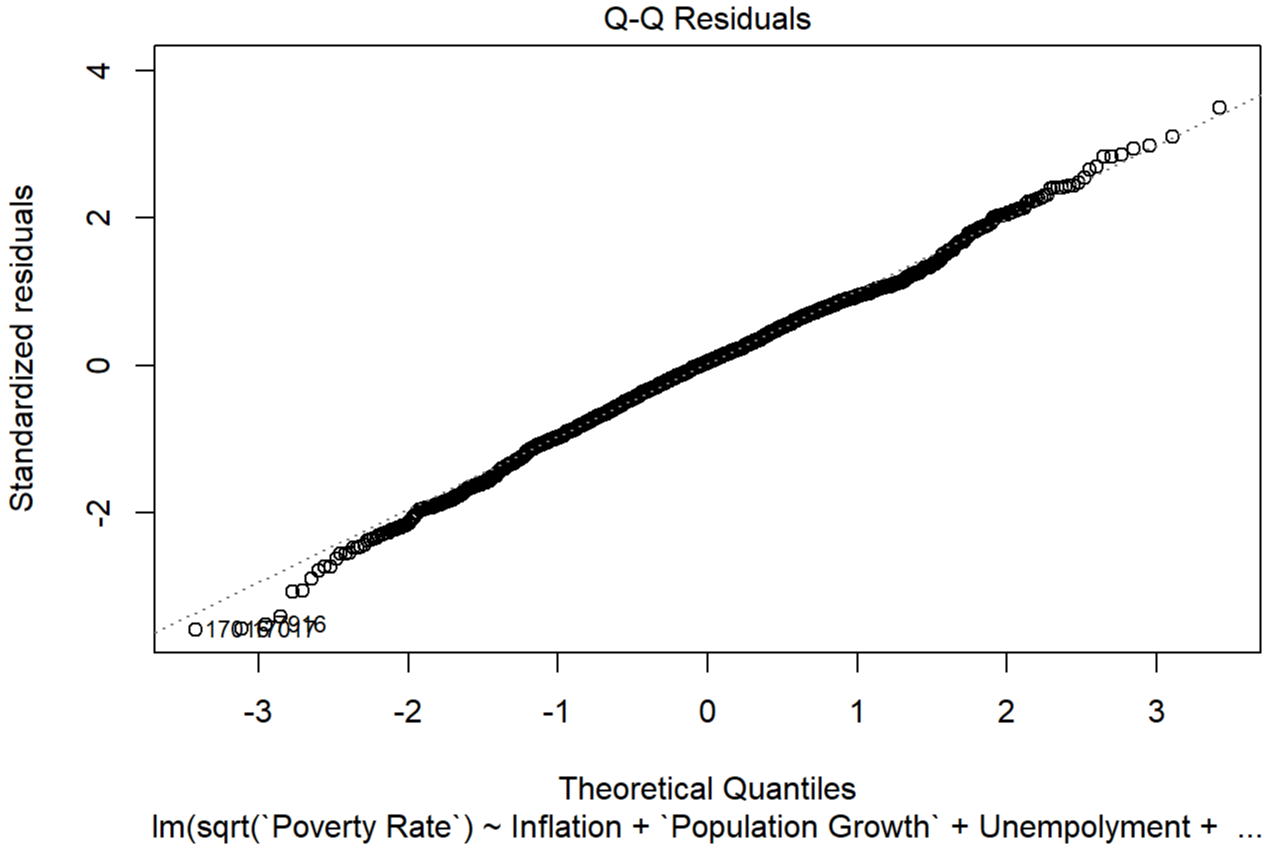
There has been a commonly recurring feature for the Residuals vs Fitted plots, that primarily they may not prove to be capturing the model properly. Even though by using the variance inflation factor function on the model, there might not be multicollinearity appear among the variables themselves. But the Res. vs. Fit for the linear regression model that does not have any transformation placed on it, shows a curve appearing for the red line. And it might not be funneling inwards, but it is defiantly funneling outwards. While for the case of the model that has the log and square root transformation placed upon it, the funnel outwards may still be visible. However, they are visible less clustered together which just proves that the transformation help the model slightly come closer to being normal.



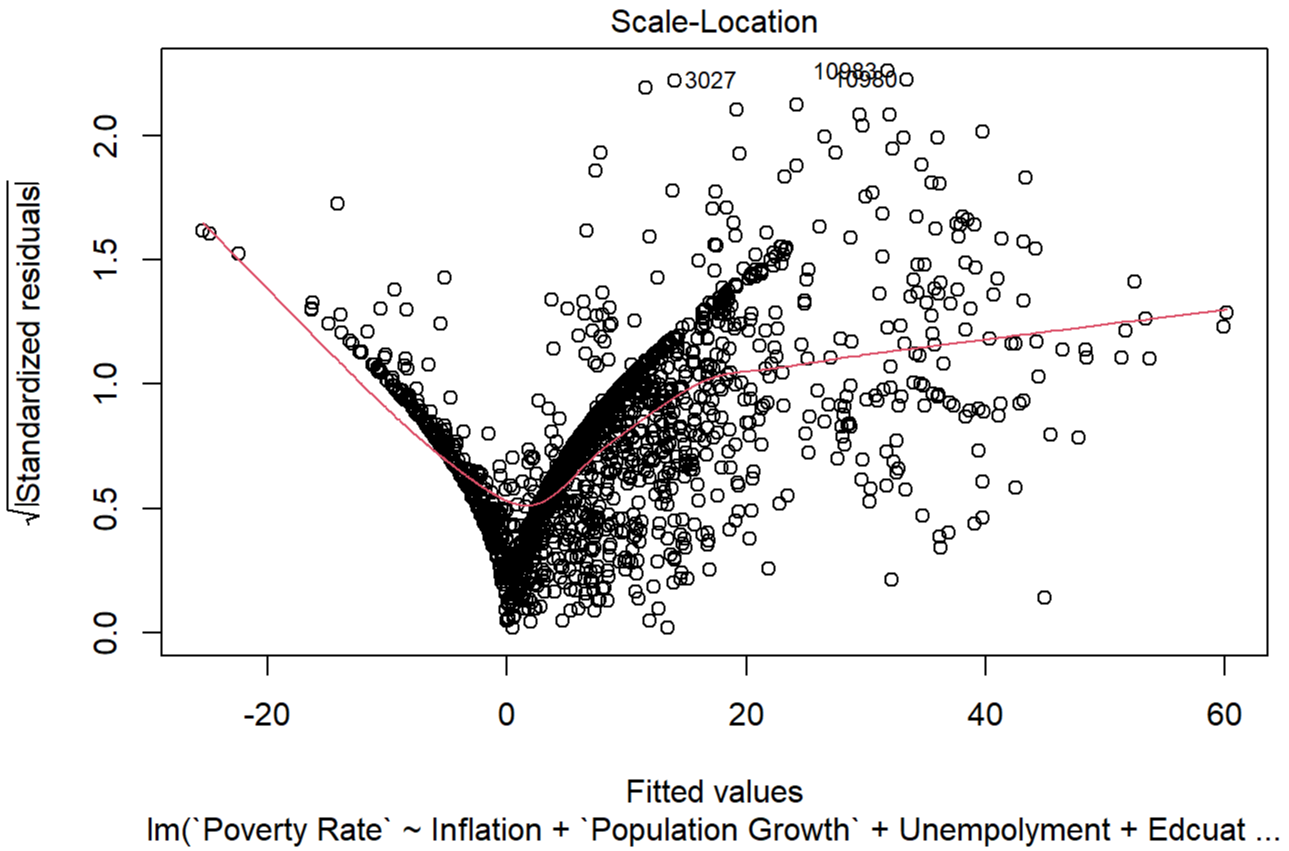


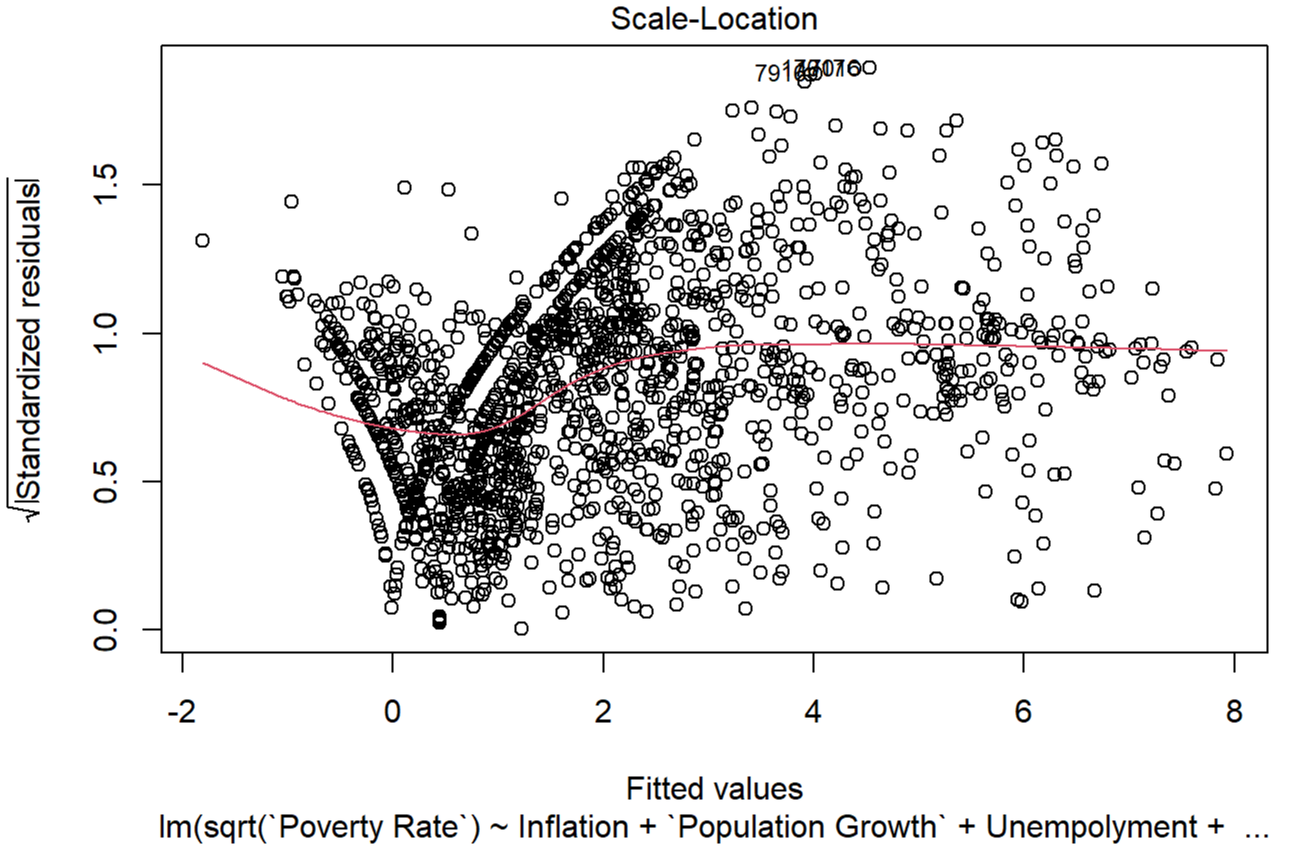
Though in the case of Q-Q Residuals, the differences between the plots are not only visible but just further shows that the transformation helped the models significantly. For the model known as life with no added transformation currently has a curve in the middle of the line and slightly deviates from both ends. Which furthers shows that without transformation, skewness from extremely high poverty rates from countries will negatively affect and cause bias among the results of the models. While in the cause of the model Life_alt, it does appear to have a slight deviation at the ends of the line in the beginning. However, besides that the Q-Q Residual plot for the model's residuals does follow a normally distribution and shows the transformation does properly handle the skewness found in the figures 2-2.5.



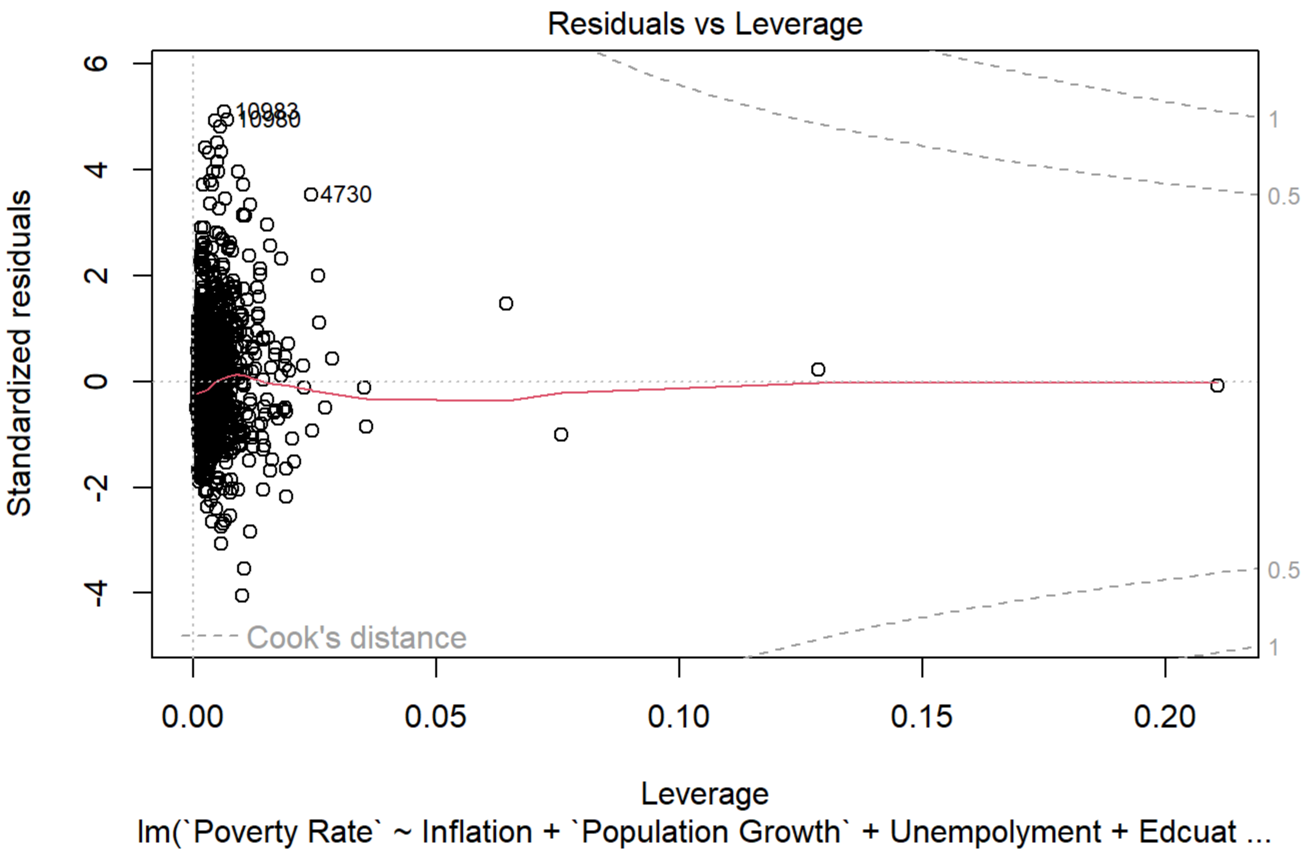


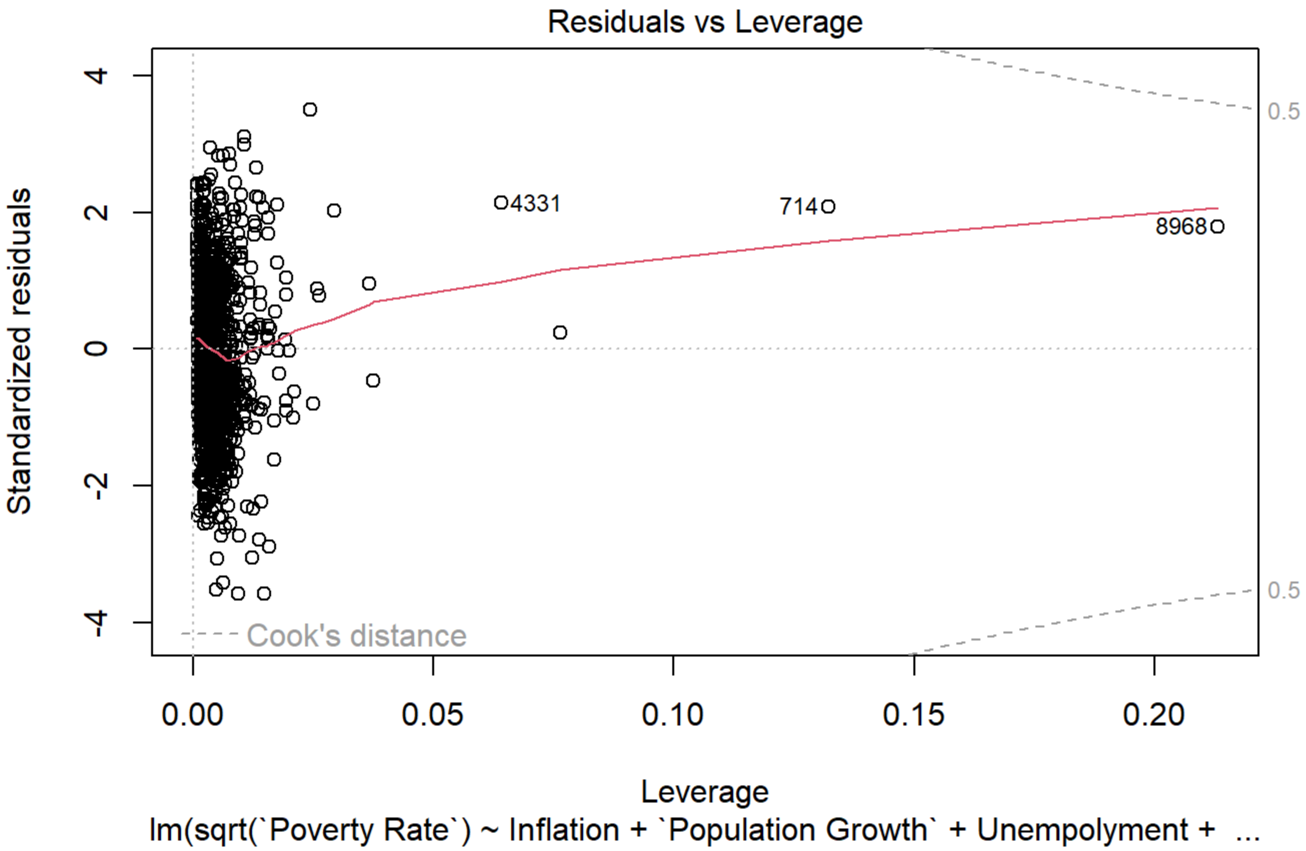
In the two plots for Scale Location, in some aspect there appears homoscedasticity in both the linear models. However it is more apparent in the model of life. Because the Scale-Location plot for it basically nailed all the points needed to be horrible plot. Because not only does the points themselves have multiple outlines that even by removing them. The plot itself will not move a single inch, they are also clustered near the interval of 0 for the plot. Where they even form a V shape that almost perfectly aligns with the red line's curve downwards curve. But in the case for the model life_alt, there is still a very mild amount of homoscedasticity. Since similar like the Scale-Location plot for model life, there is a downward curve but instead of falling nearly by 0.5 standard residuals it drops less than 0.25. Though that is where similarities conclude, because its Scale-Location losses its V shape considerably and the points themselves are significantly less cluster around form their original locations.





In the case for the Residuals vs Leverage plots for both models, have proven to be quite interesting. Since after apply two transformation to better handle the model and make it more appropriate to the data. The model Life with no transformation have proven to be the better residuals vs Leverage plot out of the two. Because even though the ideal form of the red line is to be completely straight, since its where residual are ideal should be centered.The plot's red line has a bit of a ditch but, other wise follows the ideal red line for this type of plot. Moreover the majority of the standardized residuals are clustered around the zero interval, and the furthest that it gets spread out is -4 to 5 standardized residuals away. Which proves that even though the cluster is around zero, imply that the majority of the data is within standard range of zero. However the points that reach past -3 or 3, should be either considered outliers, which could mean extreme values.However in the cause for the model that has transformation added to them, it mainly is quite similar to the Residual vs Leverage model's plot. However, the quite notable part of it shows that the line does not follow the ideal form. Since it looks like a exponential being charted so with even time it would increase higher and higher as the leverages increase.





Anlysis

```
##
## Poverty rate lm
## ===================================================================
##                                         Dependent variable:
##                                 -----------------------------------
##                                 `Poverty Rate` sqrt(`Poverty Rate`)
##                                      (1)               (2)
## -------------------------------------------------------------------
## Inflation                           0.002           -0.01***
##                                     (0.03)          (0.003)
##
## `Population Growth`                11.70***          0.90***
##                                     (0.33)           (0.04)
##
## Unempolyment                         0.09            0.03***
##                                     (0.06)           (0.01)
##
## Edcuation                          -1.46***         -0.17***
##                                     (0.19)           (0.02)
##
## GDP                               -0.0002***
##                                    (0.0000)
##
## log(GDP)                                             -1.00***
##                                                       (0.02)
##
## `Death Rate`                       2.14***           0.11***
##                                     (0.11)           (0.01)
##
## Constant                           -9.70***         10.12***
##                                     (1.63)           (0.29)
##
## -------------------------------------------------------------------
## Observations                        1,602             1,602
## R2                                   0.58              0.75
## Adjusted R2                          0.58              0.75
## Residual Std. Error (df = 1595)      9.76              1.06
## F Statistic (df = 6; 1595)        371.00***         791.90***
## ===================================================================
## Note:                                   *p<0.1; **p<0.05; ***p<0.01
##
## Poverty rate lm
## ====
## TRUE
## ----
##
## Poverty rate lm
## =====
## FALSE
## -----
```

Model: Life

For the variable inflation, with a coefficient of 0.002 implies that with every additional percent increase the Poverty Rate is implied to increased by 0.002 points. Sadly this variable is not statistically significant below 0.10 p-level.For variable 2 Population growth, with a coefficient of 11.7, implies that for every additional percent increase the predict value of Poverty rate will rise by 11.7 points.However, Population growth is statistically significant at the p-level below 0.01. For variable 3, unemployment with a coefficient of 0.09 implies for every additional increase in percent will lead to a rise of 0.09 points for Poverty rate. And just like inflation, it is not statistically significant at any level. For variable 4, Education with a coefficient of -1.46 implies that with every additional percent increase will lead the Poverty rate to decline by 1.46 points. Additionally, Education is statistically significant at a p-level below 0.01. For variable 5, GDP is shown to be with a coefficient of -0.0002 which implies that for every additional percent increase the poverty rate will drop by 0.0002 points. As well, this is also significant at a p-level of 0.01 and below. Lastly, Death rate with a coefficient of 2.14 does implies that for every additional percent increase will result in Poverty rate increasing by 2.14 points.

Model: Life_alt

For the variable inflation, with a coefficient of -0.01 implies that with every additional percent increase the Poverty Rate is implied to decline by 0.01 points. And this variable is statistically significant below 0.10 p-level.For variable 2 Population growth, with a coefficient of .9, implies that for every additional percent increase the predict value of Poverty rate will rise by .9 points.As with inflation, Population growth is statistically significant at the p-level below 0.01. For variable 3, unemployment with a coefficient of 0.03 implies for every additional increase in percent will lead to a rise of 0.03 points for Poverty rate. And unlike inflation, it is statistically significant at any level p-value below 0.01. For variable 4, Education with a coefficient of -0.17 implies that with every additional percent increase will lead the Poverty rate to decline by 0.17 points. Additionally, Education is statistically significant at a p-level below 0.01. For variable 5, GDP with the log transformation applied it is shown to be with a coefficient of -1 which implies that for every additional percent increase the poverty rate will drop by -1 points. As well, this is also significant at a p-level of 0.01 and below. Lastly, Death rate with a coefficient of .11% does implies that for every additional percent increase will result in Poverty rate increasing by .11 points.

Important things to note:

Even though the actually coefficients themselves may prove to be lackluster since, none of them are extremely large coefficients that for every additional percent that the Poverty rate will be greatly affect. Instead, lets take a look at the variances outputs of the models themselves.The model life that has no transformation added to them, the model itself is able explain about 58% of the variance shown from the data itself. Life model is at best, a moderate level for a model since a 58% implies at least 42% of the variance of the data can not be explained by the model. While in the case for the model life_alt, it can cover 75% of the variance of the data shown. This is significant because that is to the very least a 17% increase in variance power when compared to the original.

Conclusion:

Not only does adding log-transformation to GDP and square rooting not only handles the skewness of the data themselves, I believe it is safe to say that it almost normalizes it. In the end, the the predictors that ended up being statistically significant for the transformed model was technicality all of them. However, I believe there is so much work that can end up being done for this model, especially for the transformed altered life model. Even though the r^2 showed to be able to cover almost 3/4 of the variance from the data itself. It did end up lacking o certain aspects, when it came to the residuals vs leverage plots. What should of been the end result was model life alt should of have been the one with the red line being ideal and potential have it more spread around the zero standard residual interval.

Additionally, the plots themselves ended causing problems along the way. Since there were too many extreme data points to remove because that would end up basing the result of this research. Since every last of them were real data entries, and for all the data sets if for some reason there were no data in their row. It is believe that either for that year for that country there were no data to be collect or there were some but those who originally collect the data were not able to enter these data entries.

Instead what did happen, was that the model in its based form could be used either for business application or even predicting the future Poverty rates for the foreseeable decade. However, what limit our models was lack of experience since there were probably others ways to handle and tackle certain task and especially when it came using other means of linear regression models. We used linear regression as a potential stepping stone that could provide the perfect leverage for some else to either build off our research or even use our research and outcomes as a means of building a automatic model that can predict Poverty rates with no estimation errors.All what that was just mention could potential give this research more depth and or even a better way to tackle our question.