**Performing a Simple and Multiple Linear Regression in R**

**Introduction**

For many people, the world we live in is connected in many ways. To conceptualize this connection, we have invented different models to explain the ways and extents to which certain things might affect another. With these models, we have predicted the impacts and occurrences of various phenomenon. Although our predictions have not always been accurate, our understanding of the world has improved greatly from say a century ago. This has been of immense benefit to the modern world, and accordingly, we strive to understand the world better by developing more effective tools for analysis.

Amongst these methods of analyses are the different types of regression analyses. A regression analysis is when we fit a model to predict the relationship between associated variables. For a linear regression, we assume that the relationship between a response variable and its explanatory variables are linear and additive (Prabhakaran, 2016). A simple linear regression assumes that a response variable is explained by one other explanatory variable. It also assumes that the influences not captured by the model are captured in the error term (Hanck *et al.*,2019). This error term is assumed to be uncorrelated with the regressor. A multiple linear regression goes a step further to find other explanatory variables which vary with the regressor and determine our response variable. They explain our response more holistically. Apart from a linear regression, one can move further to model relationships that might not be additive in nature, using other methods such as a logistic regression or a time series analysis. For this project however, I focused on understanding a simple and multiple linear regression and how they can be effectively carried out in R.

In the initial stages of the project, I attempted to understand the connection between financial integration and indebtedness in Sub- Saharan Africa. Out of 39 countries under the the Highly Indebted Poor Countries Initiative, 32 are within the sub-continent of Sub-Saharan Africa (World Bank,2019). The aforementioned fact communicates that macroeconomic debt is a problem within the subcontinent and affects and is reproduced by other conditions within these countries. As a result, a series of scholars have sought to understand what causes debt in the region; and how to reduce indebtedness and alleviate its effects. Some scholars promote financial integration as an effective way to deal with this problem. Financial integration simply describes a country’s linkages to international capital markets (Prasad *et al.,*2003). Theoretically, it augments a country’s savings, it allows the transfer of technology and fosters the development of the financial sector amongst many (Prasad *et* al., 2003). In efforts to empirically capture the relationship between financial integration and debt on the sub-continent, I attempted to perform a linear regression on data collected by The World Bank (“International Debt Statistics”, n.d). The data I collected however, was ridden with missing data. For the values I collected, there was no correlation between my proxy of financial integration and debt-stock in these countries. This made any further analysis futile and inconclusive. As a result, I simulated data using R, and explored carrying out a simple and multiple linear regression on the simulated data. Most methods described in this paper are gotten from Hanck *et al.,* 2019, and Prabhakaran, 2016.

**Methodology**

Simulation of Data

Since the original data for the research proved incompatible with the methods of simple and multiple linear regression, I simulated data with which to explore these two methods of analyses. The regressors were randomly generated but with a pre-set covariance. This makes it such that our regressors vary together. This is one of the conditions that need to be met for a multiple regression. Error terms were randomly generated and he response variable was predetermined by the equation

, where Y is the response variable, X1 and X2 are regressors, and u is a randomly generated error term. This equation can be considered as the true model that the subsequent regression analyses will attempt to proxy.

Simple Linear Regression

As iterated in the introduction, a simple linear regression assumes that the relationship between one response variable and one explanatory variable is linear and additive. This shows itself in the form of:

, where y is the response variable, x is the explanatory variable, b is the intercept and m is the coefficient of x

For the parts of the response variable that is not explained by the model, a simple linear model assumes to be simply due to errors. To explore a simple linear regression, the relationship between Y and X1 from the simulated data was assessed. To familiarize myself with the two variables, different graphical analyses were carried out.

To begin with, a scatter plot that explored the relationship between both variables was created (*Figure 1)*. This showed a generally positive and linear relationship. To check for outliers, box plots of both variables were created as well (*Figure 2)*. For a small dataset, outliers can influence the model and being aware if any of your variables contain them is necessary in some cases. To check if both variables were normally distributed, a density plot of both variables were created, with the skewness estimated (*Figure 3)*. Before generating a linear model, the correlation of both variables was calculated. This was done to estimate the level of linear dependence between the 2 variables. Being satisfied with the results of the graphical analyses and the correlation value presented, I proceeded to allow R to estimate the linear model with no changes to the simulated data.

However, to truly know if a simple linear model would satisfactorily predict Y, the data was split into two parts. The first part which contained 80% of the data is known as the training data. A simple linear model of the training data was generated. This linear model was then used to estimate the remaining 20% of the data known as the test data. This process is done to assess the ability of a specific type of regression analysis in predicting what it is one might want to predict. A table of the actual and predicted y values was created (*Table 1)* and a correlation test was carried out to estimate how well our linear model predicted the actual Y values. Further tools used to evaluate the accuracy of the model was minimum-maximum accuracy and the mean absolute percentage error. Pushing this a step further, a 10-fold cross validation was used. This iswhen the data is randomly parsed out into 10 mutually exclusive portions. In turns, one portion is kept as test data, and the other parts are used to predict this portion. This is done until all the portions have gone through this process. When plotted, the linear models predicted should be close to each other and parallel. This shows that the model’s prediction accuracy is not varying too much for any sample. A plot of the 10-fold cross validation of a linear model for the simulated data was generated as well (*Figure 4).* To test the performance, the mean squared error of the results was calculated as well.

Assumptions

­ To test the first linear model created for the simulated data, I run it through the assumptions of a simple linear regression. These are:

* The mean of the residuals of the model is zero. The residuals of a model are the differences between the observed and predicted values. For a linear model to be considered good enough, the residuals of the model should be close to zero.
* The residuals should be homoscedastic. This means, it should not vary much as the predictor variable changes.
* The residuals should not be autocorrelated. This means, they should not depend on each other.
* The residuals should be normally distributed
* And the residuals of the linear model and the regressor should be uncorrelated

Through various visualizations and tests, these assumptions were tested for our simple linear model (linearMod).

Multiple Linear Regression.

A multiple linear regression is like a simple linear regression in that it assumes a linear and additive relationship between associated variables. However, a multiple linear regression, considers more than one explanatory variable. What happens when our response variable can be explained by yet another variable that varies, to some extent, with our explanatory variable? In such cases, when the additional explanatory variable is not included in our model, we tend to overestimate the coefficient of the available regressor. This is what is termed as the Omitted Variable Bias. To avoid this, we take into account any additional regressor. However, adding too many regressors to one’s model has its downside. The variance of the coefficient estimators tends to inflate when more regressors are added to a model. Choosing which of the two burdens to bear is what is known as the bias-variance trade-off. For our simulated data, our additional regressor, X2, was taken into account, and a multiple regression model was generated. To make sure that both regressors were not perfectly correlated or too highly correlated, we plotted a correlation of X1 and X2 (*Figure 7).* The variance inflation factors of both coefficients in the model was calculated as well.

The multiple regression model and simple linear regression models were compared. This was firstly done by comparing both their adjusted R-squared values. The R-squared value tells us the proportion of variation in our response variable that is explained by our model. However, as more values are added to our X variable, the R-squared value of our model increases. The Adjusted R-squared overcomes this problem and gives an effect size that remains fairly constant as more variables are added to our regressor. Therefore, we consider this value when comparing the effect sizes of our model. Prabhakaran (2020) however suggests that the Akaike’s and Bayesian information criterion (AIC & BIC) are better ways to compare statistical models. Hence, the AIC of both models were compared. For model comparison, the model with the lowest AIC score is preferred.

Debt Data

The data used in this case was received from the World Bank. To begin with, data frames of each of the variables that could potentially be used for analysis was created. Financial Integration was then proxied by adding the export and import of goods and services, the net foreign direct investment, and the net portfolio investment by a 2.5:2.5:3:2 ratio. This was done with the understanding that financial integration occurred primarily through trade and investment. The external debt stocks as a percentage of GNI, total debt service as a percentage of export, and financial integration of the countries for the years 2009, 2014 and 2019 were compiled into one data frame which I intended to use for analysis.

To graphically analyse the two variables I was most interested in—debt stock and financial integration—a scatter plot of both variables were created (*Figure 9)*. This scatter plot showed no relationship between both variables. Further graphical analysis of the variables included box plots and density plots (*Figure 10-11)*. This was not promising of the data either. To confirm that both variables were not associated, a correlation test of both variables was calculated as well. I then proceeded to generate a linear model and polynomial model for the data. A multiple regression model that considered years was developed as well. Despite the ineffectiveness of the linear models, the effect sizes of the simple and multiple regression models were shortly compared as well as their AICs.

**Results**

The variables X1 and Y both included no outliers (*Figure 2) and* were normally distributed with a skewness of 0.01 and 0.18, respectively. The Pearson’s r for both variables was 0.775. This denoted a strong positive correlation between both variables. The linear model generated by R was:

The coefficient estimates and model p-values were both lower than 0.05. This implies that the linear model generated by R was statistically significant.

The correlation accuracy between the predicted values of the test data and the actual values was 0.92. This is quite high. The minimum maximum accuracy was 98.31% and the mean absolute percentage error was 1.73%. This means that the linear model developed for the 80% of the data predicted the other 20% quite well.

The plot of the 10-fold cross validation showed linear models that were parallel and quite close to each other (*Figure 4)*. This speaks well to the efficacy of a simple linear model in predicting the simulated data. The mean squared error of the cross validation was 53.1513.

The mean of the residuals of the linear model was -1.06e-16 which is close to zero. The plots of the models (*Figure* *5)* showed that the residuals were normally distributed (top right plot) and homoscedastic (relatively flat line in top and bottom left plots). An autocorrelation factor plot (*Figure 6)* showed that the residuals were not autocorrelated. A Pearson’s r of the correlation of the residuals and X1 showed a p-value of 1. This denoted that they were not correlated. In all, the linear model fit the assumptions it was required to.

The multiple regression model generated was:

A correlation plot of both regressors was created as well (*Figure 7)*. This showed little correlation between the two. The variance inflation factor for both regressors was 1.114 which shows a satisfactory variance inflation. A variance inflation factor greater than 4 would have however been of concern. In such cases, when one has a list of regressors, removing one out of the most correlated regressors help.

In comparing the two, I found that the omitted variable bias was evident in the simple linear regression. The coefficient of X1 was 4.44 in the simple linear model whereas it was 3.45 when X2 was considered. The adjusted R-squared of both models also showed that whilst the simple linear model accounted for 59.62% of the data, the multiple regression model accounted for 86.1 % of the data. The AIC of the multiple regression model was lower than that of the simple linear model as well. This establishes the multiple regression model as better at explaining our data than the simple linear model.

For the debt data, both variables had a lot of outliers (*Figure 10)* and were highly skewed with debt stock recording a skewness of 2.69 and financial integration recording a skewness of 3.69 (*Figure 11)*. The linear, polynomial, and multiple regression models were as follows:

, where D is debt stock, and F is financial integration.

The multiple linear regression in this case had the highest Adjusted R-squared of 35.2% which was a stretch of an improvement from that of the simple linear model, which was 2.1%.

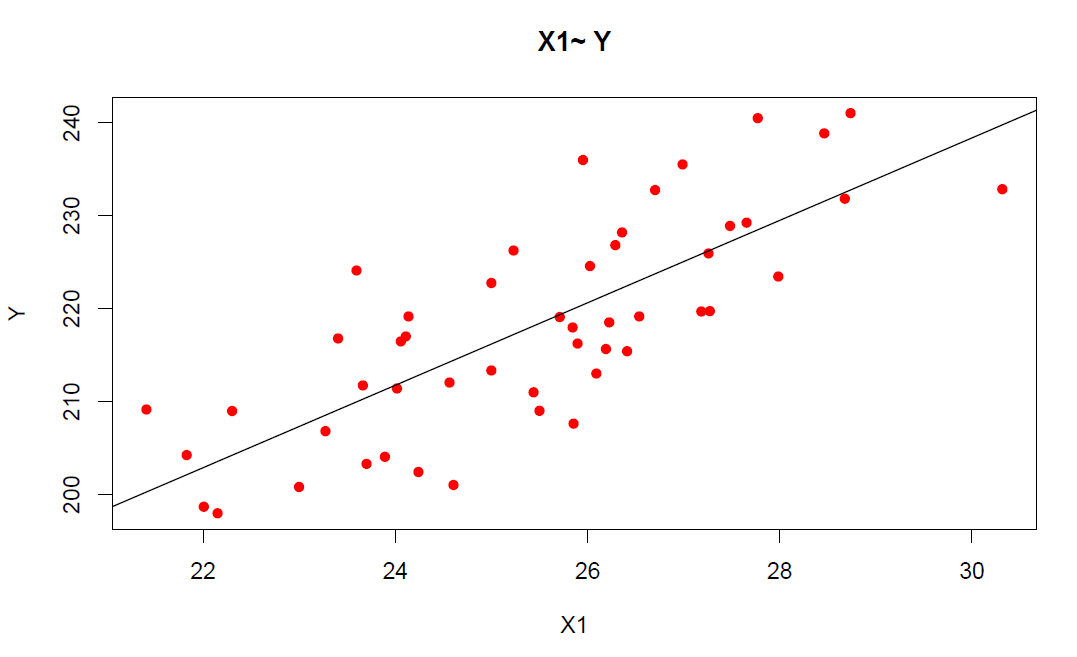
**Concluding Remarks**

Using R, I was able to employ a simple and multiple linear regression to simulated data in R and investigate the efficacy of my models. These methods focus simply on relationships that are additive in nature. However, the truth is, for most things in nature, the association between two variables might go beyond what can be explained with these two models. That is when other models such as a logistic regression comes into play. Time also usually plays a role and using time series analyses allows one to take it into account. This project however, allowed me to investigate the two types of regression analyses and walk my audience through how they are conducted in R. For the debt data, a lot of data was missing. My proxy of financial integration was quite lacking as well. This made it hard to carry out much of an analysis on it. Further research might be to look at several other features that might be related to debt stock, and carefully select a model that explains this relationship accordingly. Going beyond Sub-Saharan Africa to inspect developing countries, might also give a much larger sample from which better conclusions can be drawn.

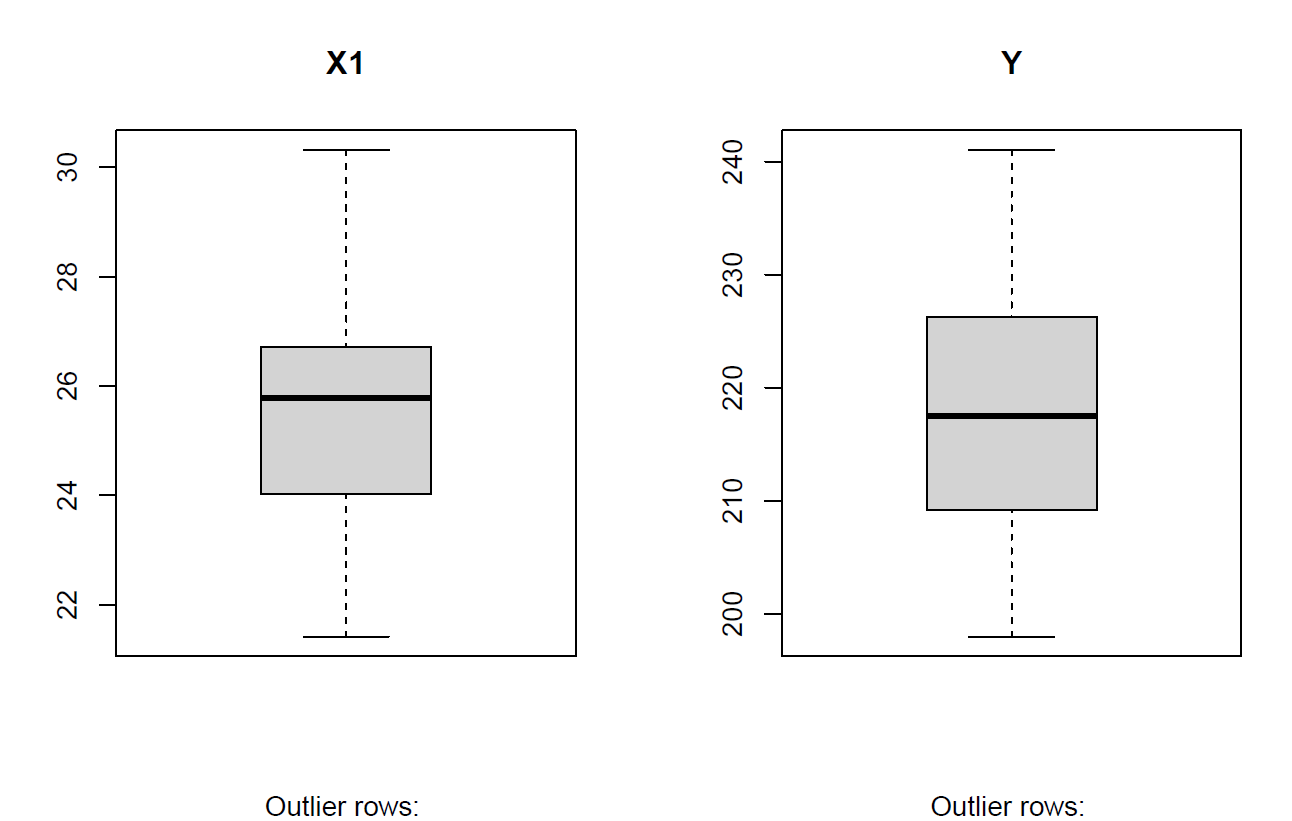
**Feedback incorporation**

I displayed a table of the predictions of the training and test data in the results as suggested. I also showed equations of the models instead of describing their coefficients. I changed the method section from explaining the code to explaining the substance of what I did. I fixed the grammatical errors and carefully explained a 10-fold cross- validation. I used the figures to tell my story more clearly, however, I did not change their descriptions much. I removed all explanations of the 3d graph since I did not find it exactly relevant to this paper. It would have counted as extraneous information. I changed the first figure to having the linear model overlaid instead of an R scatter smooth line thingy since it seemed confusing.

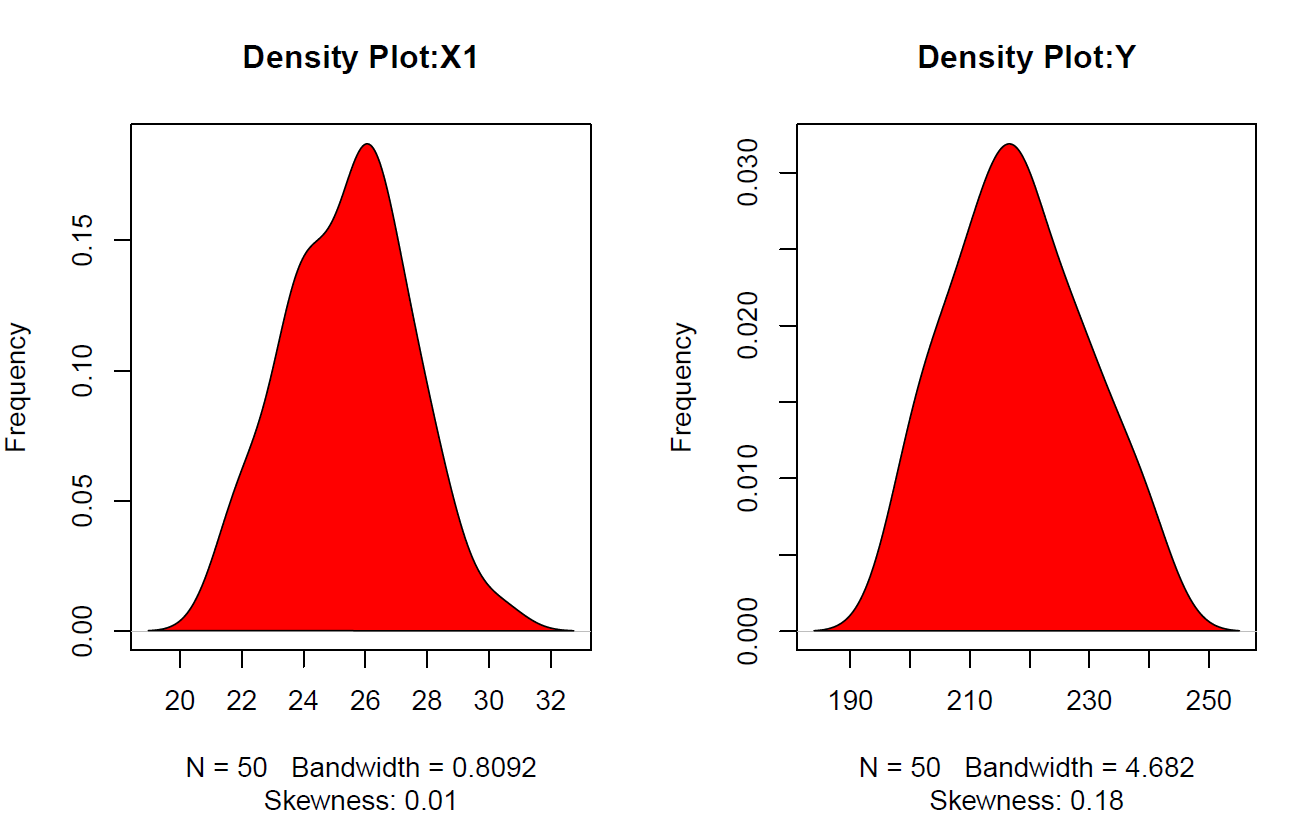
**APPENDIX A: FIGURES**



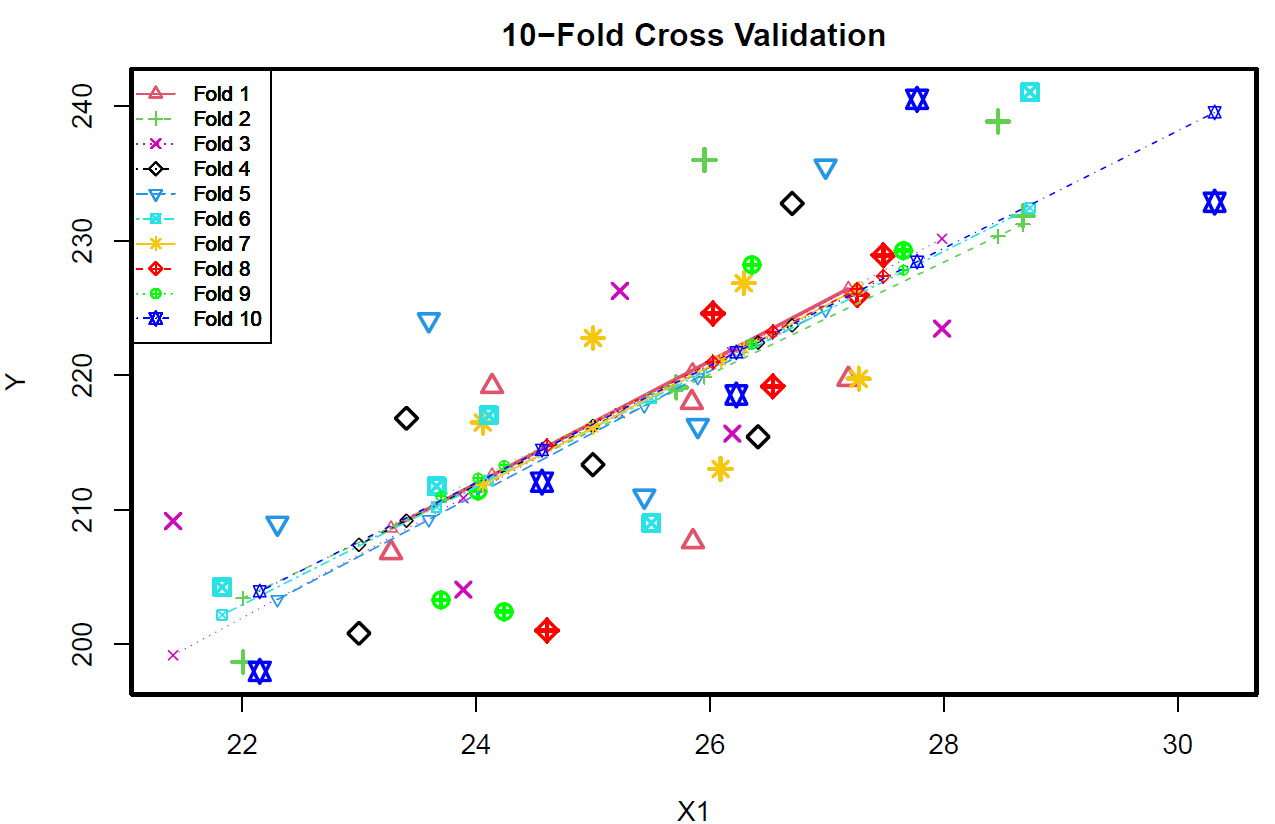
*Figure 1: A scatter plot of Y against X1 overlaid with the R generated linear model.*



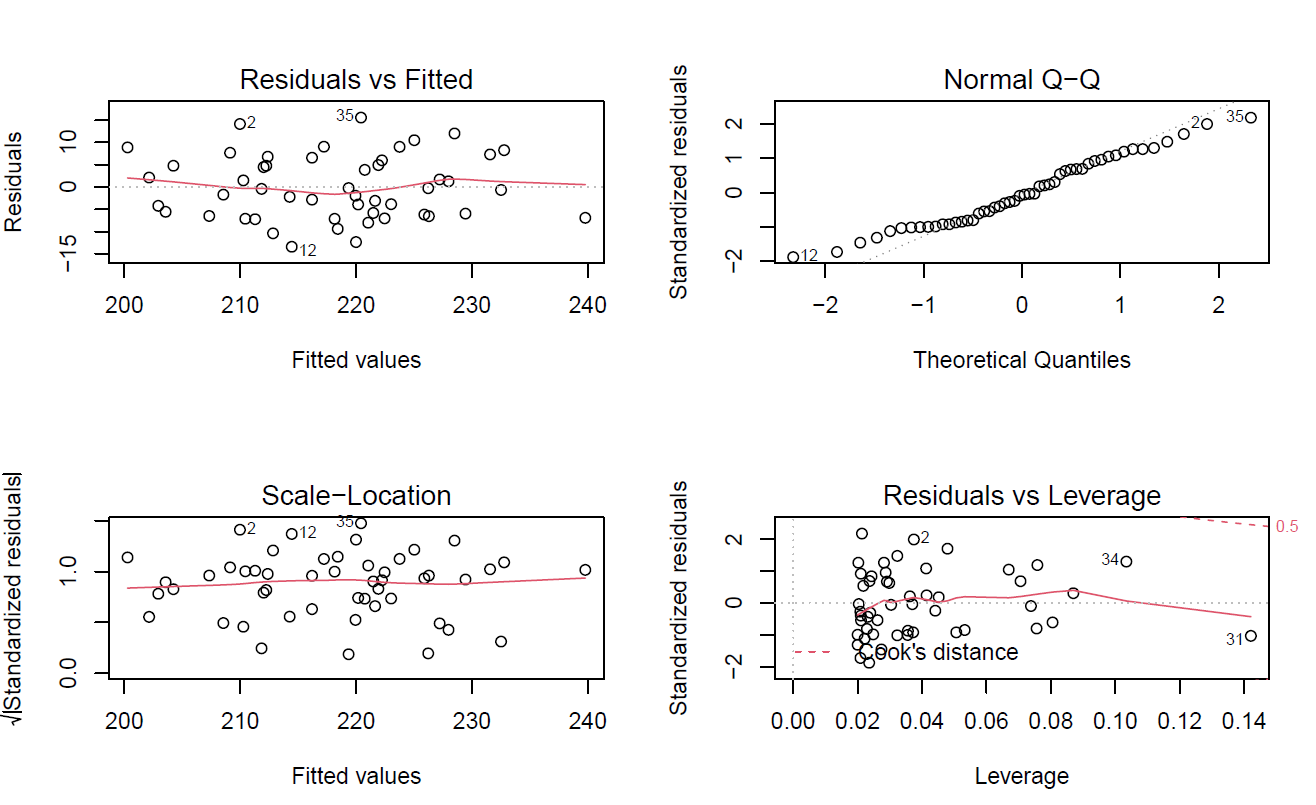
*Figure 2: Box plots of X1 and Y variables with outlier rows listed below (there are no outliers)*



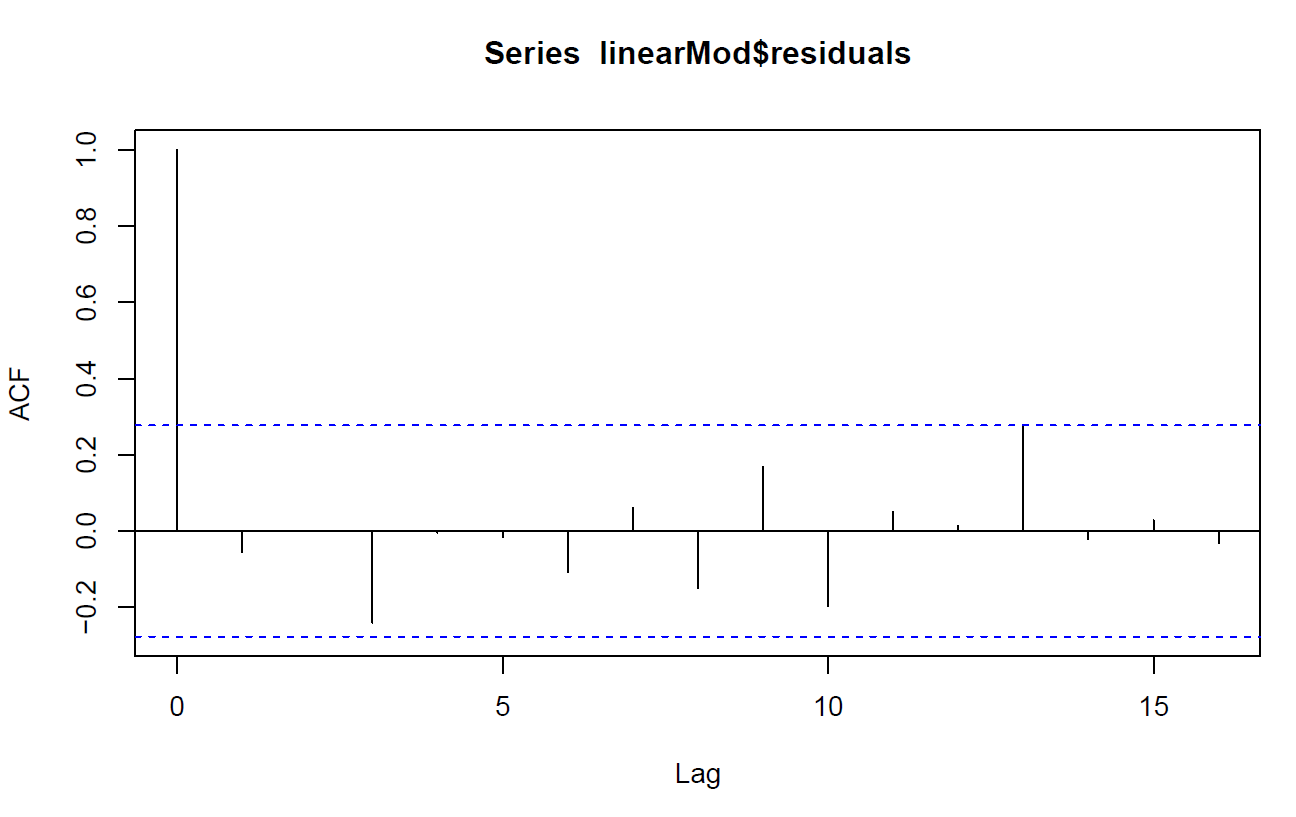
*Figure 3: Density plots of X1 and Y1 with sample size (N), bandwidth, and skewness indicated below .*



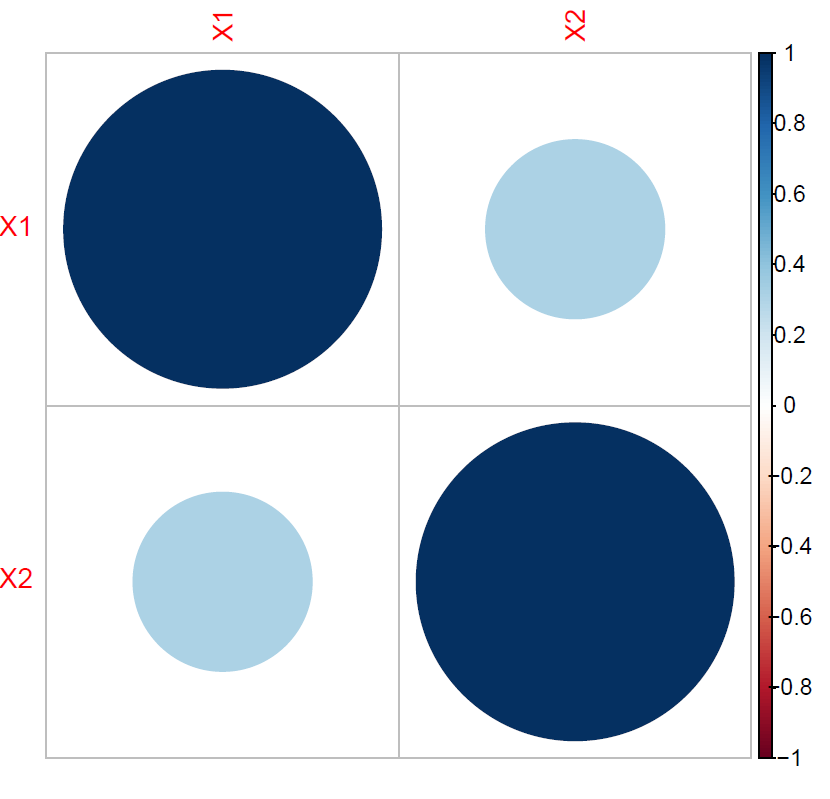
*Figure 4: A plot showing a 10-fold cross validation of the simulated data*



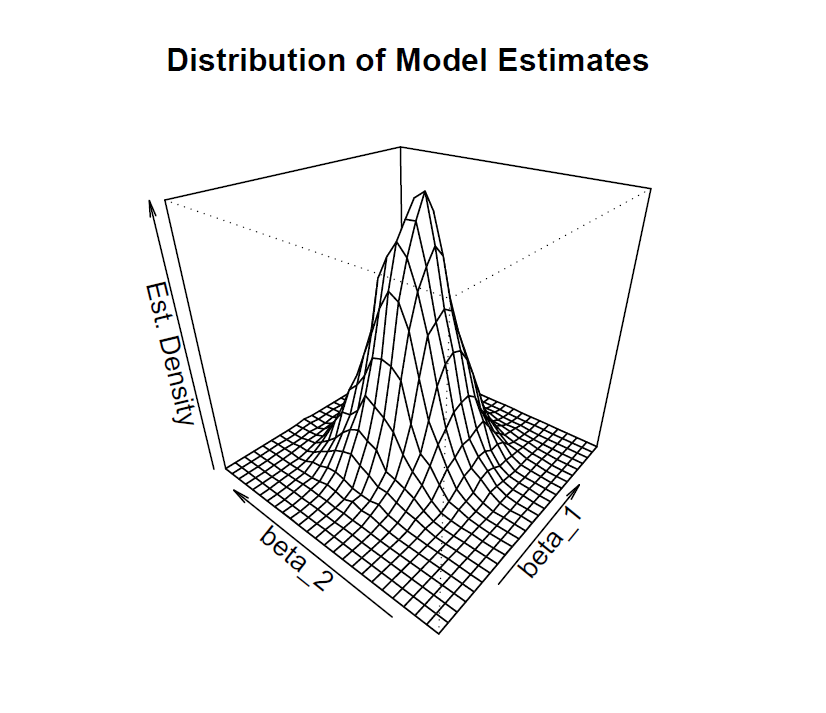
*Figure 5: Top left: A graph of the residuals of the linear model against its fitted values; Top right: A plot showing the quantile distribution of the residuals of the linear model; Bottom left: A plot of standardized residuals against fitted values; Bottom Right: A plot depicting the Cook’s distance of the residuals. Point 34 and 31 are outliers.*



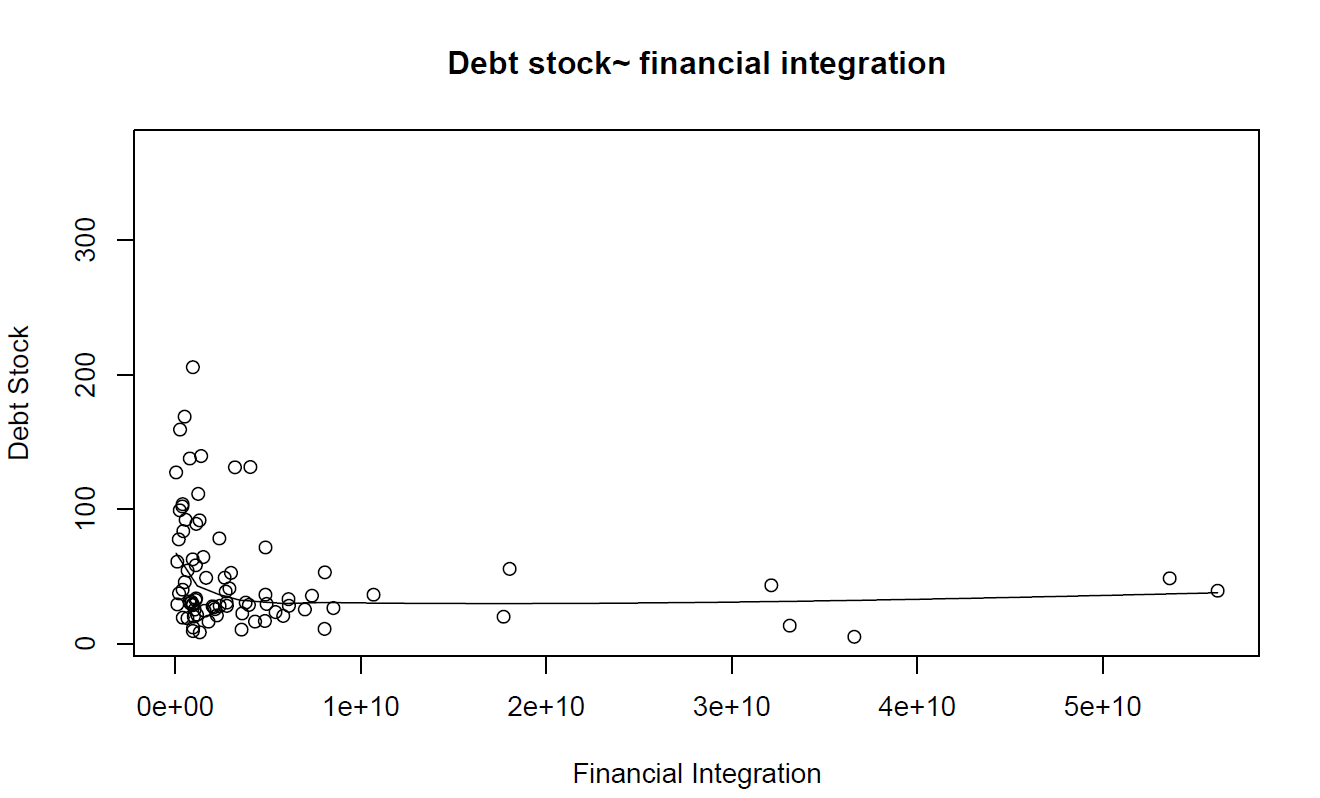
*Figure 6: A graph showing the autocorrelation factors of the residuals with respect to their lags in steps of 1.*



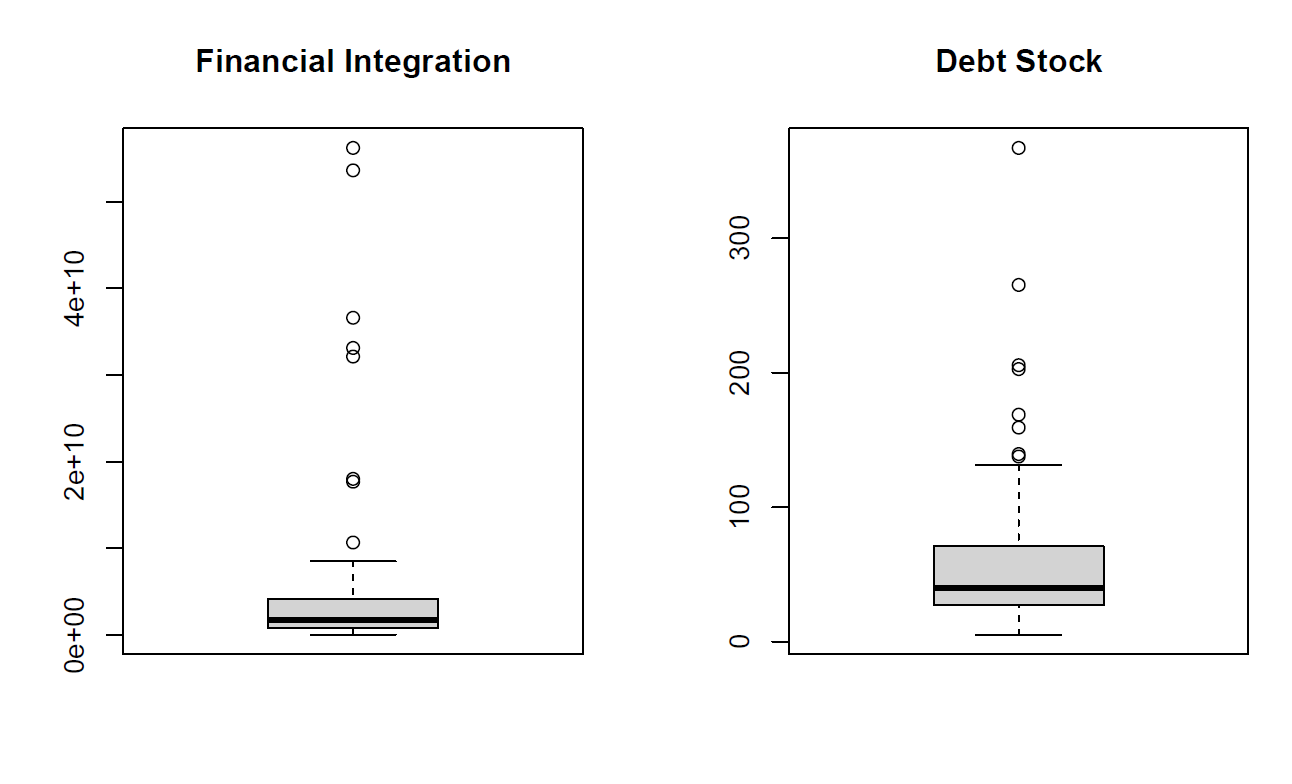
*Figure 7: A plot showing the correlation factors of the regressors X1 and X2.*



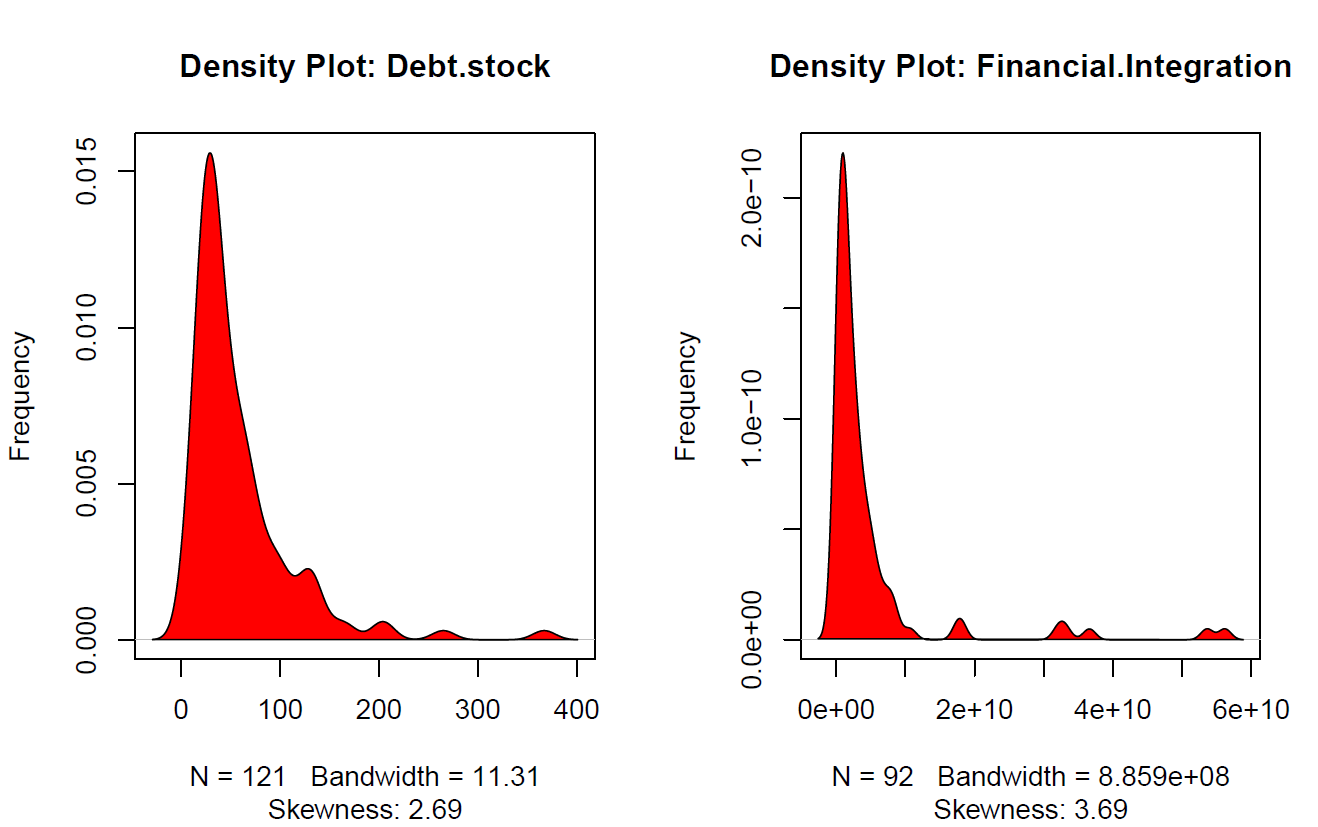
*Figure 8: A graph showing the estimated density of coefficient estimates from different multiple regression model estimators of the same true model.*



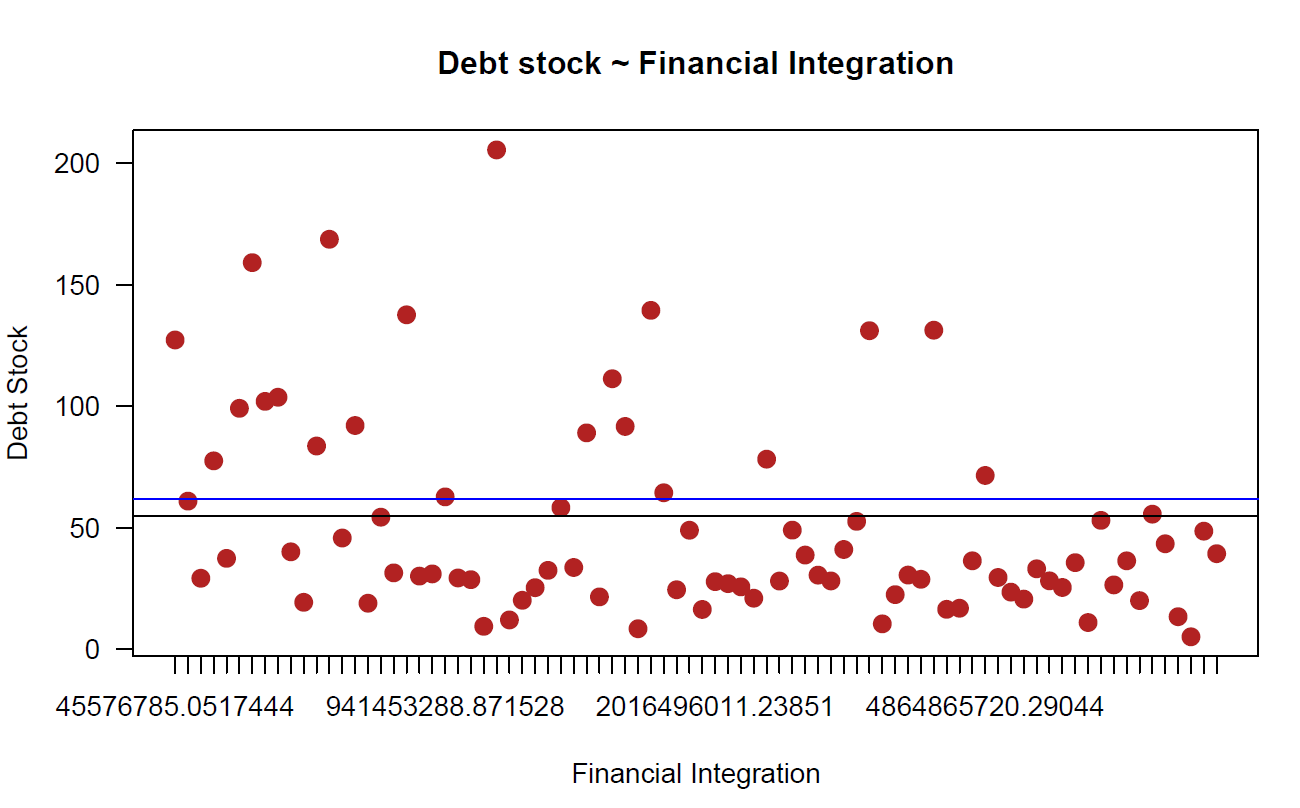
*Figure 9: A smoothed scatter plot of debt stock against financial integration values .*



*Figure 10: Box plots of Financial Integration and Debt stock.*



*Figure 11: Density plots of Debt Stock and Financial integration with sample size (N), bandwidth and skewness below of each variable below their respective density plots.*



*Figure 12: A graph showing the distribution of debt stock against financial integration values with an overlaid approximated line of best fit (black) and polynomial of best fit(blue).*

**APPENDIX B: TABLES**

|  |  |
| --- | --- |
| actuals | predicteds |
| 209.0133 | 218.704786 |
| 215.6594 | 221.65716 |
| 217.9816 | 220.178076 |
| 224.5879 | 220.948722 |
| 241.0441 | 232.526913 |
| 219.1696 | 223.137662 |
| 211.4146 | 212.373313 |
| 216.2508 | 220.396947 |
| 219.1676 | 212.885278 |
| 197.9868 | 204.401824 |

*Table 1: A table showing the actual test data values and the predictions of the test data from the linear model generated according to the training data.*

**BIBLIOGRAPHY**

Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2019). Introduction to Econometrics with R. *Obtenido de https://www. econometrics-with-r. org/ITER. pdf*.

IMF & The World Bank. (2019). Heavily Indebted Poor Countries (HIPC) Initiative and Multilateral Debt Relief Initiative (MDRI)—Statistical Update. *Policy Paper No. 19/028.* Retrieved from <https://www.imf.org/en/Publications/Policy-Papers/Issues/2019/08/06/Heavily-Indebted-Poor-Countries-HIPC-Initiative-and-Multilateral-Debt-Relief-Initiative-MDRI-48566>

Prabhakaran, S. (2016). Linear regression. *r-statistics. co*. Retrieved November 15, 2020, from <http://r-statistics.co/Linear-Regression.html>

Prasad, E., Rogoff, K., Wei, S.-J., & Kose, M. A. (2003). Effects of Financial Globalisation on Developing Countries: Some Empirical Evidence. *Economic and Political Weekly*, *38*(41), 4319–4330. JSTOR.

*International Debt Statistics | DataBank*. (n.d.). Retrieved November 15, 2020, from <https://databank.worldbank.org/source/international-debt-statistics/preview/on>