**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 great 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.*,2020). 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.

First, 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”, 2020). 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.

**Methods**

Simulated Data

With the help of the R-package ‘mvtnorm’, the data used to carry out subsequent analyses was simulated. The package mvtnorm allows one to randomly derive multivariate vectors. It allows one to predetermine the means of these vectors as well as to indicate a matrix of the covariance between the vectors it develops. Since for a multiple regression analysis, the regressors are expected to vary together, the rmvnorm function was used to develop two regressor variables (X1 and X2). The base r function rnorm, was used to generate error terms (u) with a standard deviation of 5. The response variable (Y) was predetermined as a variable with an additive relationship determined by the regressors and the randomly derived error terms. The pre-set equation (*Equation 1)* counted as the true population model that the regression analyses would then attempt to proxy. The regressors and response variable were then written as a comma separated values file and saved in the clean data folder.

The first type of analysis carried out on the data was a simple linear regression that only considered X1 and Y. To begin with, the data was analysed graphically. First, the relationship between both variables was visualized with a smoothed scatter plot (*Figure 1)*. To visualize if there were any outliers within the data that affected the relationship, a box plot of both variables was plotted (*Figure 2)*. This showed no outliers for either variables. To check if both variables were normally distributed, a density plot for each variable was generated and the r-package ‘e1071’ was used to estimate the skewness (*Figure 3)*. Following that, the correlation of both variables was estimated, and a simple linear model (linearMod) as created. The f-statistic and p-values and other measures of the model was calculated as well. Following that, a linear model (lmMod) of 80% of the data (Training data) was used to test the other 20% of the data. The accuracy of this prediction was estimated by assessing the correlation between the actual and predicted values. The minimum-maximum accuracy and mean absolute percentage error of the model was calculated as well. For further analysis, using a 10-fold cross validation, the model’s performance was tested repeatedly. This meant, a randomly chosen 80% of the data was used to predict the other 20% of the data, ten times. This was computed and plotted with the help of the ‘lattice’ and ‘DAAG’ R-packages (*Figure 4)*. After, the mean squared error of the 10 linear models was extracted from the plots.

Secondly, the linear model and simulated data was tested according to the assumptions of a linear model. First, the oversized calculator in R was used to check if the mean of the residuals of the linear model was close to zero. By plotting the linear Model with R (*Figure 5)*, the homoscedasticity and normality of the residuals of the linear model was visually analyzed. To check if the residuals were autocorrelated, they were plotted using the acf function in ‘ggplot2’ (*Figure 6)*, and a Durbin Watson test was carried out on the linear model with the help of the R-package ‘lmtest’. How to rectify autocorrelated residuals was shortly discussed in the R script as well. To make sure that the regressor and the residuals of the model were not correlated, a correlation test was carried out as well. With the Rpackage gvlma , the assumptions of the linear model was automatically tested.

Following that, a multiple regression model that considered both regressors was estimated. This model was then extensively compared to the simple linear model to assses which of them estimated the response variable, Y, better. The measures of fit and coefficients were compared. The assumption of no perfect multicollinearity was explored with the multiple regression model R developed. First, the vif function from the DAAG package was used to estimate the variance inflation factors of both regressors. To visualise the correlation between both regressors, the Rpackage ‘corrplot’ was used (*Figure 7).* R was then used to randomly develop 10000 other datasets with the same parameters as the original simulation of this project. Multiple regression models of these datasets were created, and their coefficients were saved into the object coefs. Using kde2d from the R-package ‘MASS’, the density estimates of the coefficients were stored in the object kde. These density estimates were then plotted using the persp function (*Figure 8)*. This plot showed the distribution of the coefficients estimated by R.

By doing all this, I was able to explore a simple and multiple linear regression in R.

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 % of GNI, Total debt service as a % of export and financial integration of the countries for the years 2009, 2014 and 2019 was compiled into one data frame. This data frame was written as a comma separated values file and saved into the clean data folder as “debt.clean.csv”. A data frame of only complete rows from “debt.clean.csv” was also saved as “debt.complete.csv”.

To analyse the data, first subsets of the data according years was created. Then, debt stock was plotted against financial integration (*Figure 9).* Box plots of both variables were created which showed many outliers for both variables (*Figure 10)*. Density plots of both variables were created, and skewness estimated (*Figure 11).* A correlation test of moth variables was calculated as well. A linear model and polynomial model were created 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.

**Results**

Simulated Data

The Scatter plot of Y against X1 showed a generally positive trend. There were no outliers in X1 and Y shown in the boxplot (*Figure 1)*. The density plot of both variables also showed normal distribution with X1 having a skewness of .01 and Y with a skewness .18 (*Figure 2)*. The Pearson’s r of X1 and Y was 0.775. The simple linear model (linearMod) estimated an intercept of 105.308 and an X1 coefficient of 4.436. The p-values of both the coefficient estimates and the model were lower than 0.05, denoting a statistically significant model. The Adjusted R-squared of the model was 0.5962. The correlation accuracy of the lmMod for the other 20% of the data was 92.001%. The minimum-maximum accuracy and the mean absolute percentage error of lmMod were 98.31% and 1.73% respectively. The mean squared error of the 10-fold cross validation was 53.15313.

The mean of the residuals of the linear model was -1.o6e-16 which is close to zero. The residuals were normally distributed and homoscedastic (*Figure 5)*. The acf plot showed that the residuals were not autocorrelated (*Figure 6)*. The results of the Durbin Watson test of linearMod had a p-value of 0.6. The Pearson’s r of the residuals and the regressor recorded a p-value of 1. Gvlma outputted that all assumptions of linearMod were acceptable.

The multiple regression model (mult.mod) estimated an intercept of -20.424, an X1 coefficient of 3.451 and an X2 coefficient of 3.009. The adjusted R-squared of mult.mod was 0.8661. The variance inflation factors of mult.mod was 1.1114 for both regressors. The correlation plot showed very little correlation between X1 and X2.

Debt Data

The plot of debt stock against financial integration showed no trend (*Figure 9)*. Both variables had a lot of outliers and were both highly skewed (*Figures 10&11).* Debt stock had a skewness of 2.69 and financial integration figures had a skewness of 3.69. The Pearson’s r of both variables was -0.18. The linear model estimated an intercept of 5.4e+1 and a financial integration coefficient of -7.218e-10. The polynomial fit estimated an intercept of 6.166e+1 and regressor coefficient of -3.668e-09. Overlaying both models on a strip chart of the data showed a straight line. The multiple regression model estimated an intercept of 9.2e+1; and a 2014, 2019, and financial integration intercept of -5.9e1, -4.99e1 and -2.6e-10 respectively. The Adjusted R-squared values of the linear fit, polynomial fit, and multiple regression model were 0.021, 0.064, and 0.352 respectively.

**Discussion**

The Simulated data was clean, hence, carrying out model estimations was smooth. The linear model proved to meet all the assumptions it needed to. However, the multiple regression model was an improvement on the simple linear model. In the simple linear model, the coefficient of X1 was higher than that in the multiple regression model. This demonstrates what is called the Omitted Variable Bias. This is when, due to an omitted variable, the effect that the included variables have on the response are inflated. In real life, this can lead to inaccurate conclusions. However, there is a trade-off included in adding so many regressors to one’s model. With many regressors, that all vary together, the variance of our coefficient estimators is inflated greatly. This is known as the bias-variance trade-off and is a much bigger problem for smaller samples. Luckily, for our model that had 2 regressors only, the variance inflation was not that bad. A variance inflation factor of 4, is however cause for concern. In such cases, one can choose to remove one of the two most correlated regressors.

The adjusted R-squared values of the multiple linear models of both the simulated and debt data were better than that of the simple linear models in both cases. 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 constant as more variables are added to our regressor. Therefore, we consider this value when comparing the effect sizes of our model. The linear model accounted for just 59.62% of the variation in Y whereas the multiple regression model accounted for 86.61% of the variation. For the debt data, despite the absurd coefficient values given by all 3 models, the effect size of the multiple regression model was 35.17% which was a stretch of an improvement from the 2.1% effect size of the simple linear model.

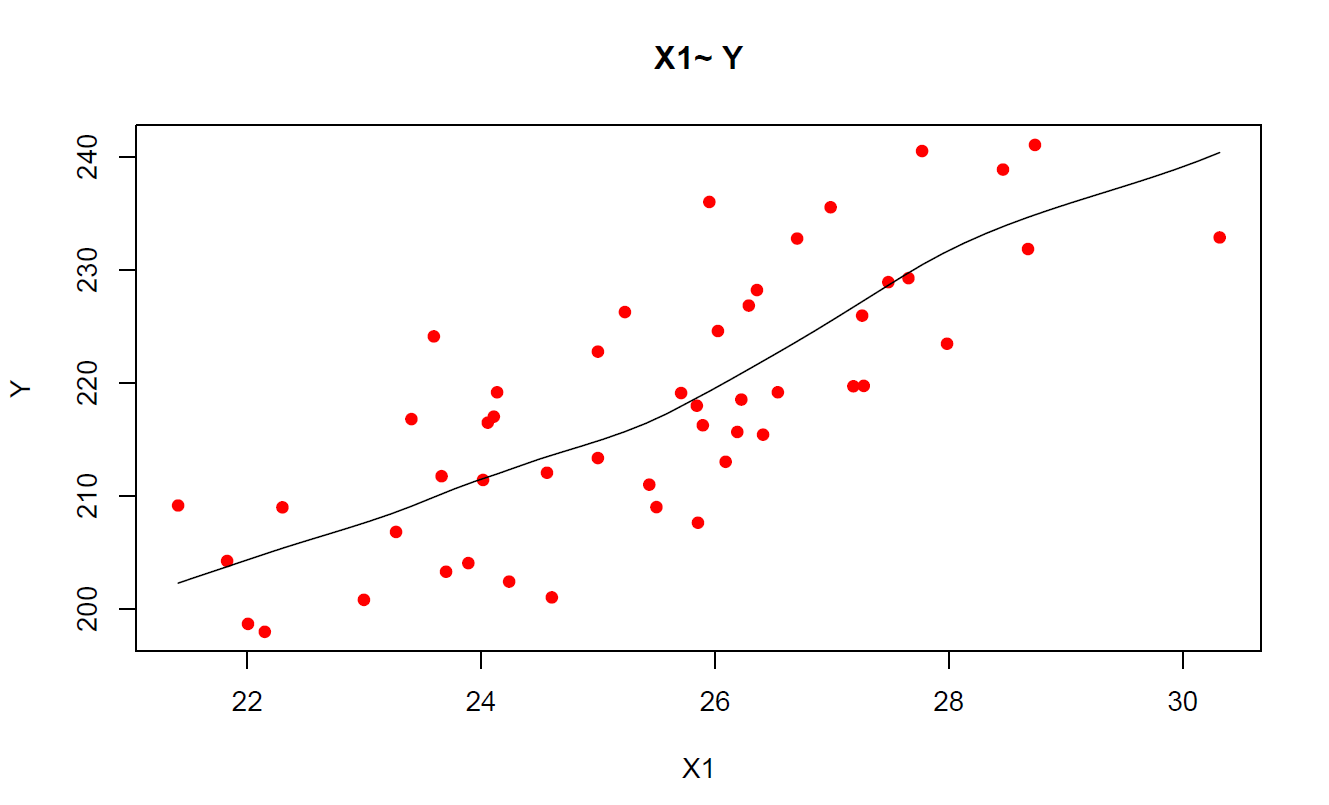
**Concluding Remarks**

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.

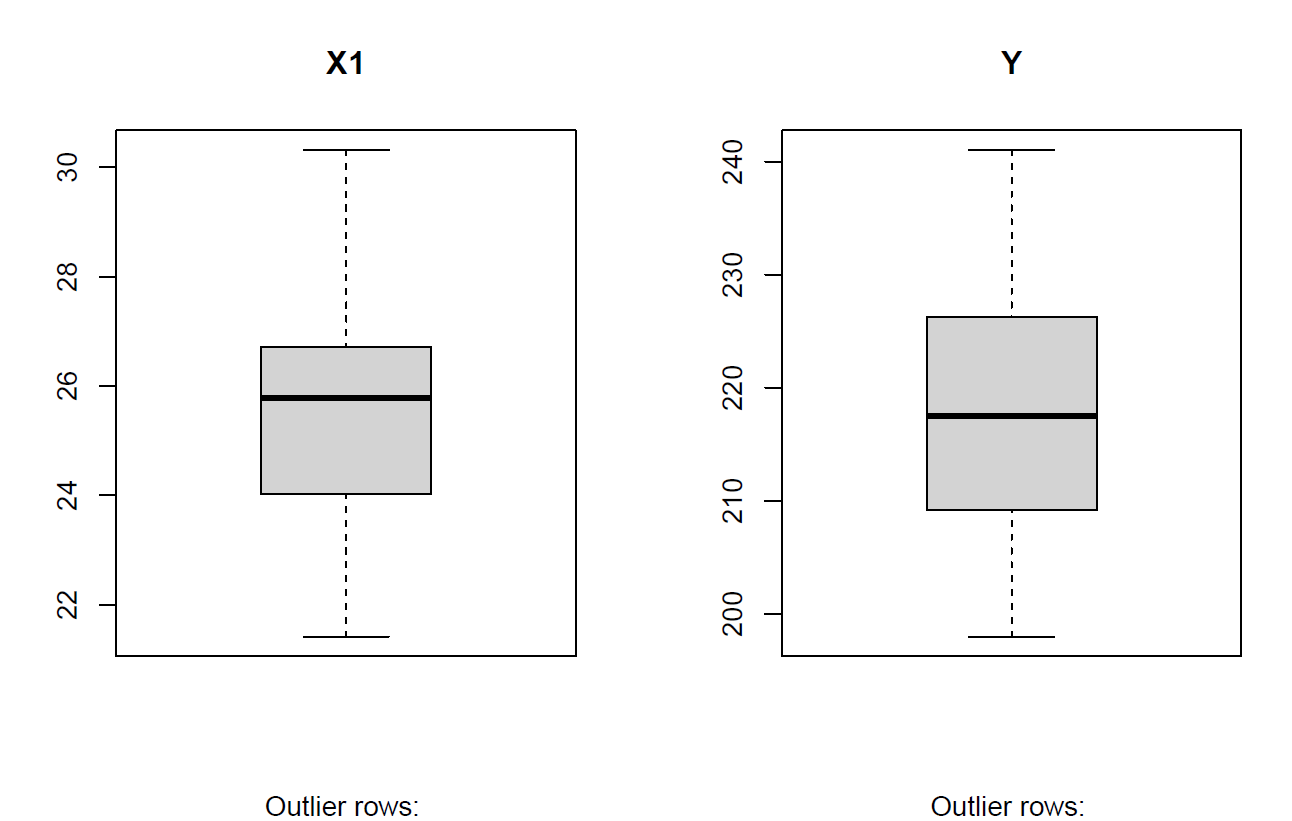
**Progress Report**

Tuesday went as planned. However, I decided to merge Wednesday’s goal with Thursday’s since they both had to do with analysis. I spent Tuesday and Wednesday cleaning up my data. Wednesday and Thursday proved very challenging since my data was not cooperating and clad with struggles. At the end of the day, I gave up on carrying out a regression and started making maps. After a meeting with Thor, I was told to stick to one thing, so despite the progress in making maps, that was put aside. On Friday, with frustration from uncooperative data, I decided to focus on understanding a linear regression instead. Friday was spent understanding a simple linear regression model and the assumptions associated with it. Saturday was spent studying a multiple linear regression. Saturday was also spent figuring out how to simulate data, simulating data and applying what I had learnt to the simulated data. Sunday was spent doing a write up of what I had done so far.

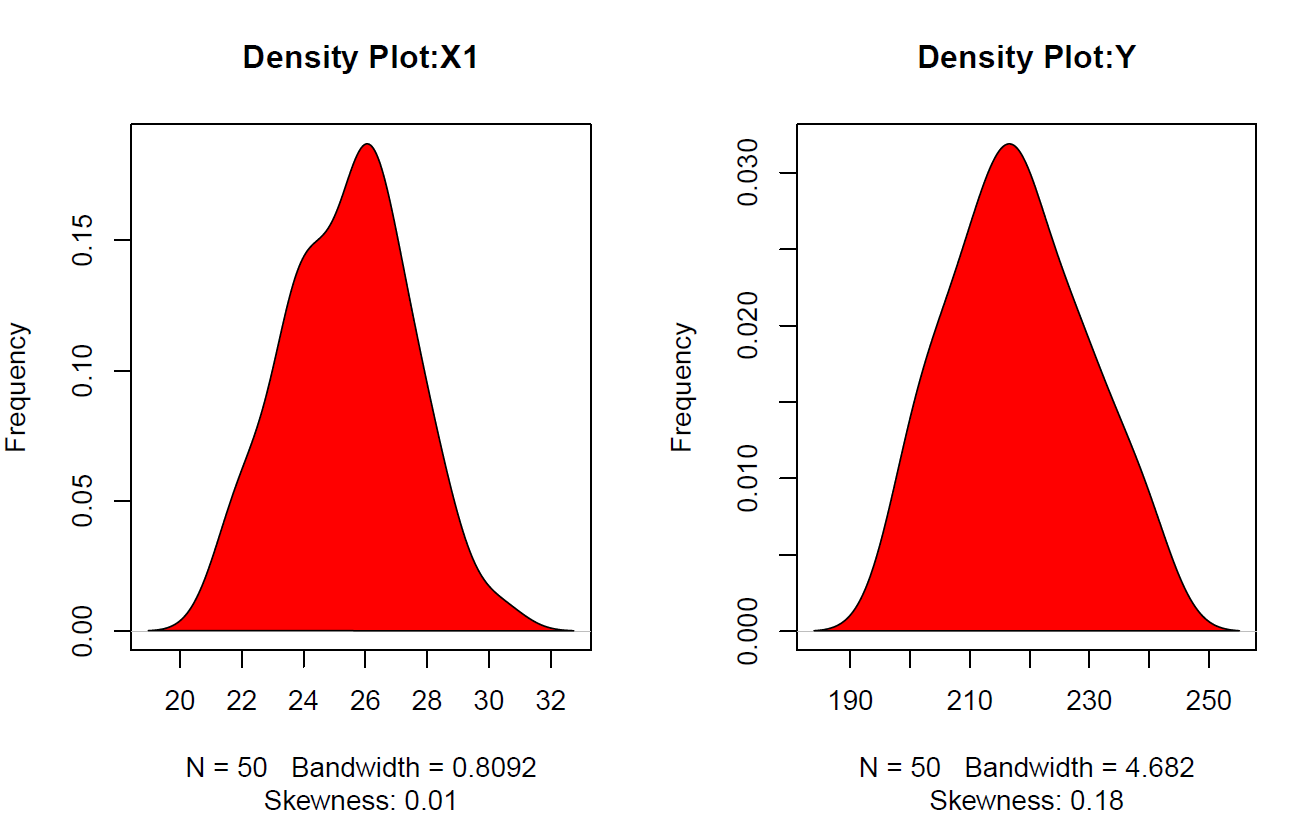
**APPENDIX A: FIGURES**



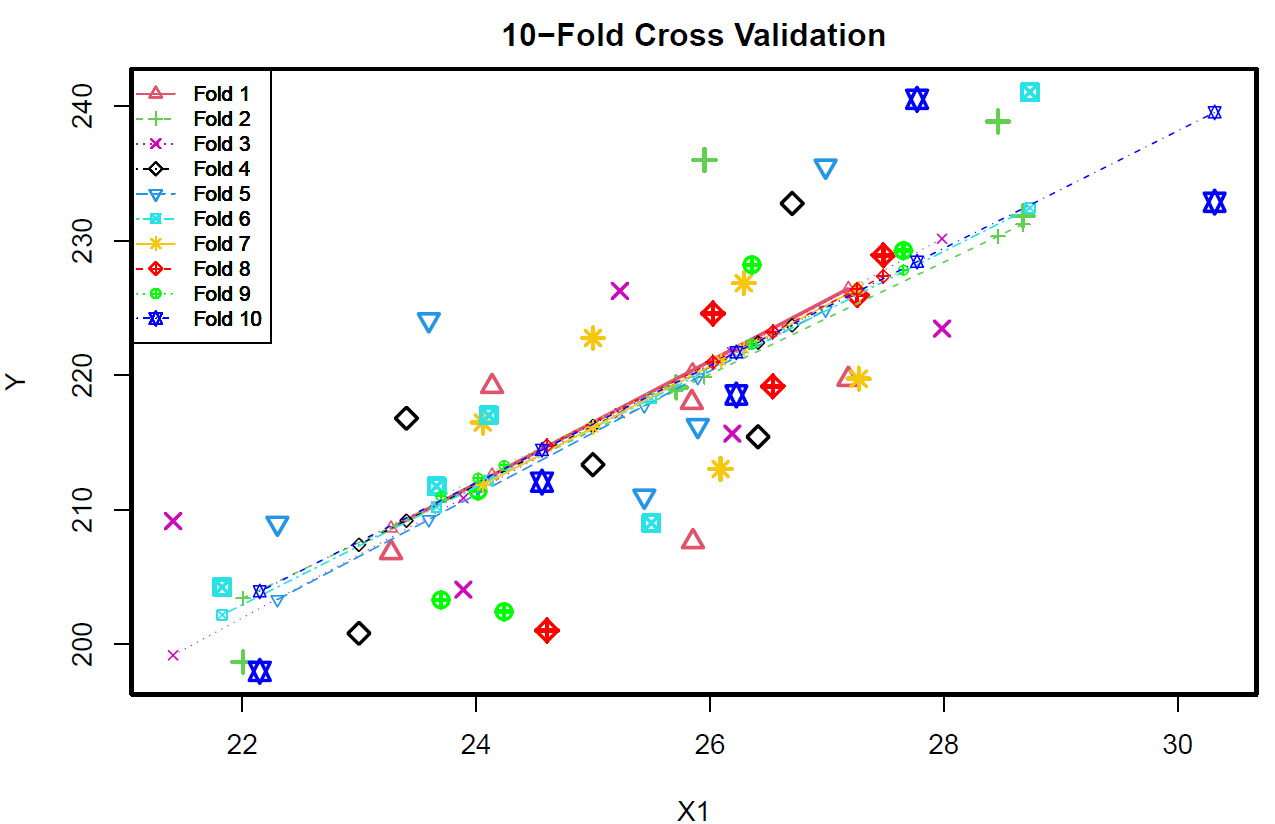
*Figure 1: A smoothed scatter plot of Y against X1*



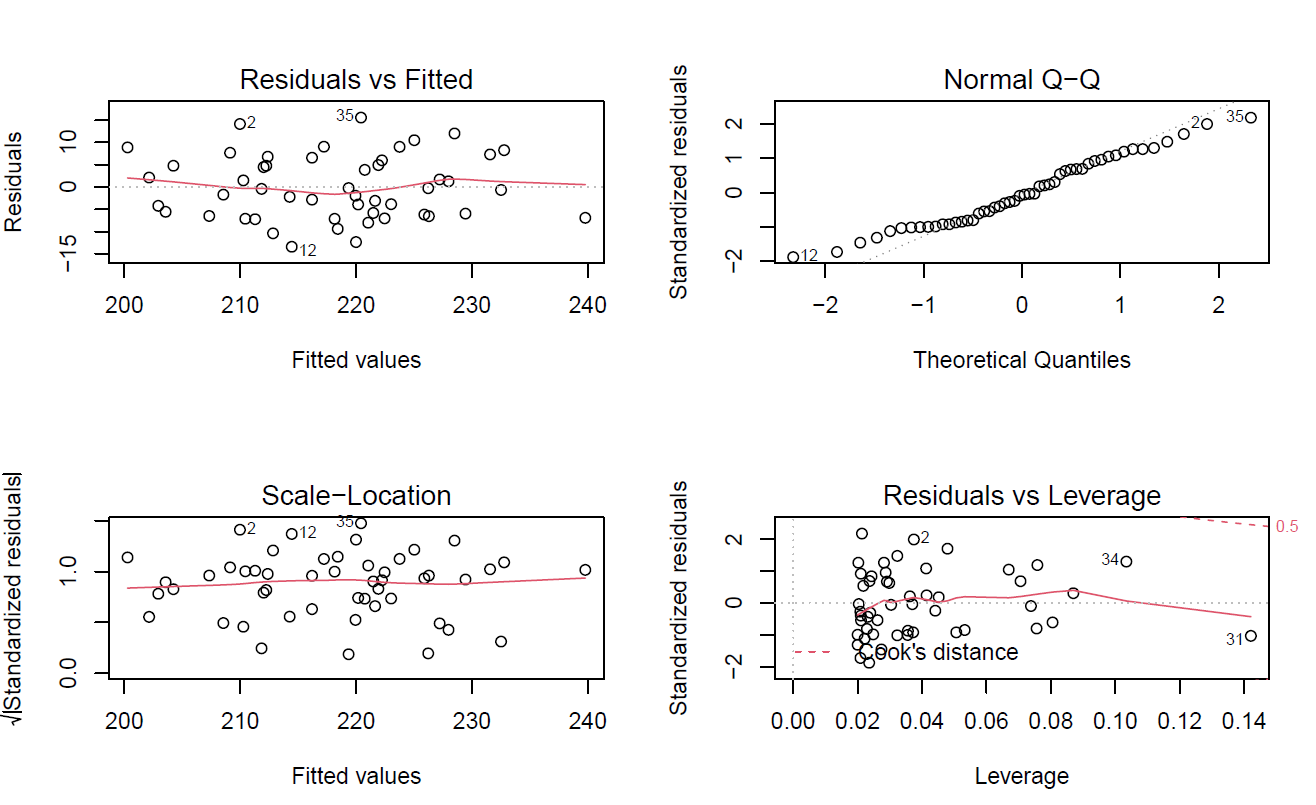
*Figure 2: Box plots of X1 and Y variables with outlier rows listed below (there are no outliers)*



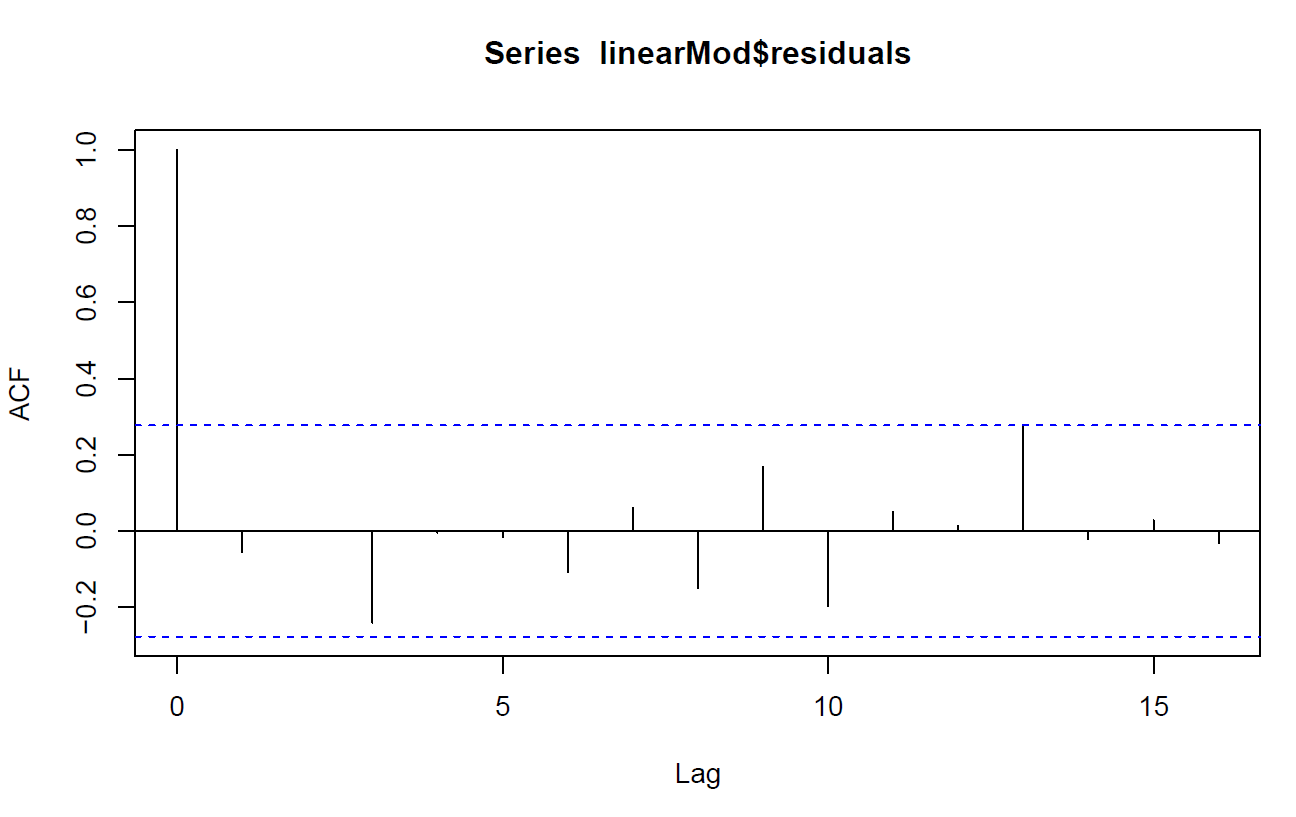
*Figure 3: Density plots of X1 and Y1 with sample size (N), bandwidth, and skewness indicated below .*



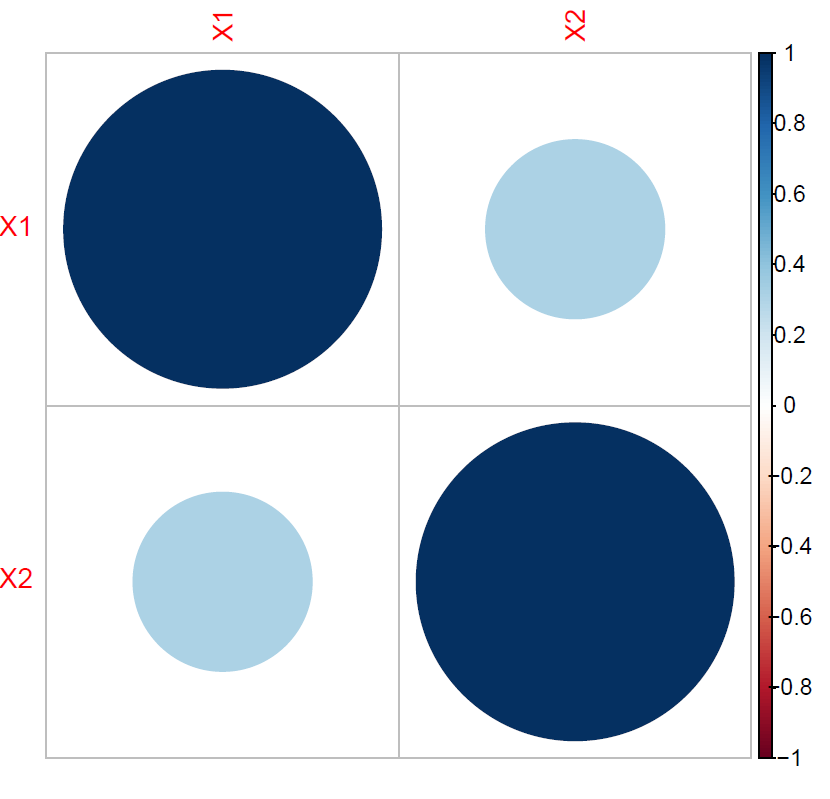
*Figure 4: A plot showing a 10-fold cross validation of the simulated data*



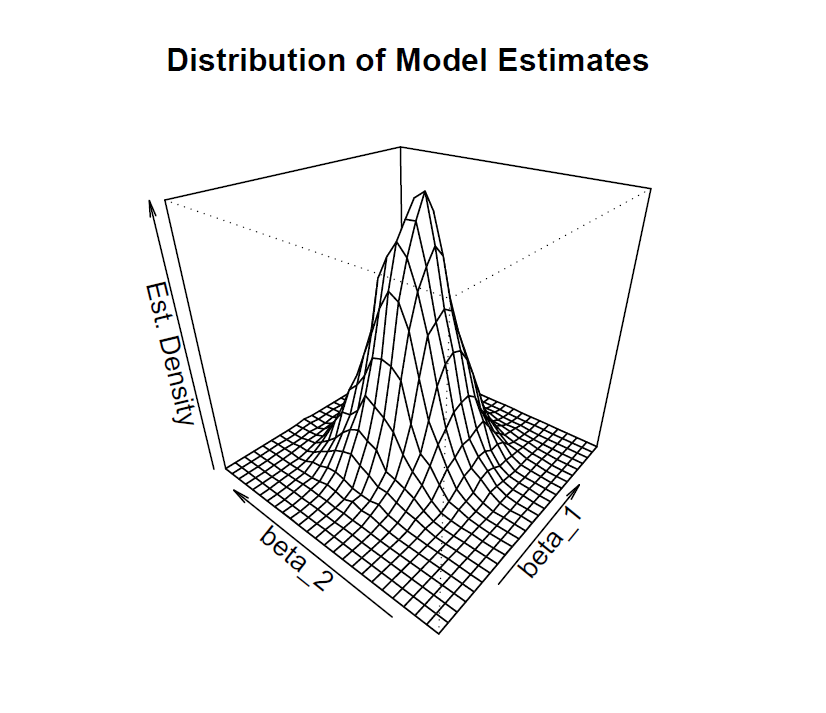
*Figure 5: Top left: A graph of the residuals of the linear model against its fitted values; Top right: A plot showing the quantile distribution of the residuals of the linear model; Bottom left: A plot of standardized residuals against fitted values; Bottom Right: A plot depicting the Cook’s distance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



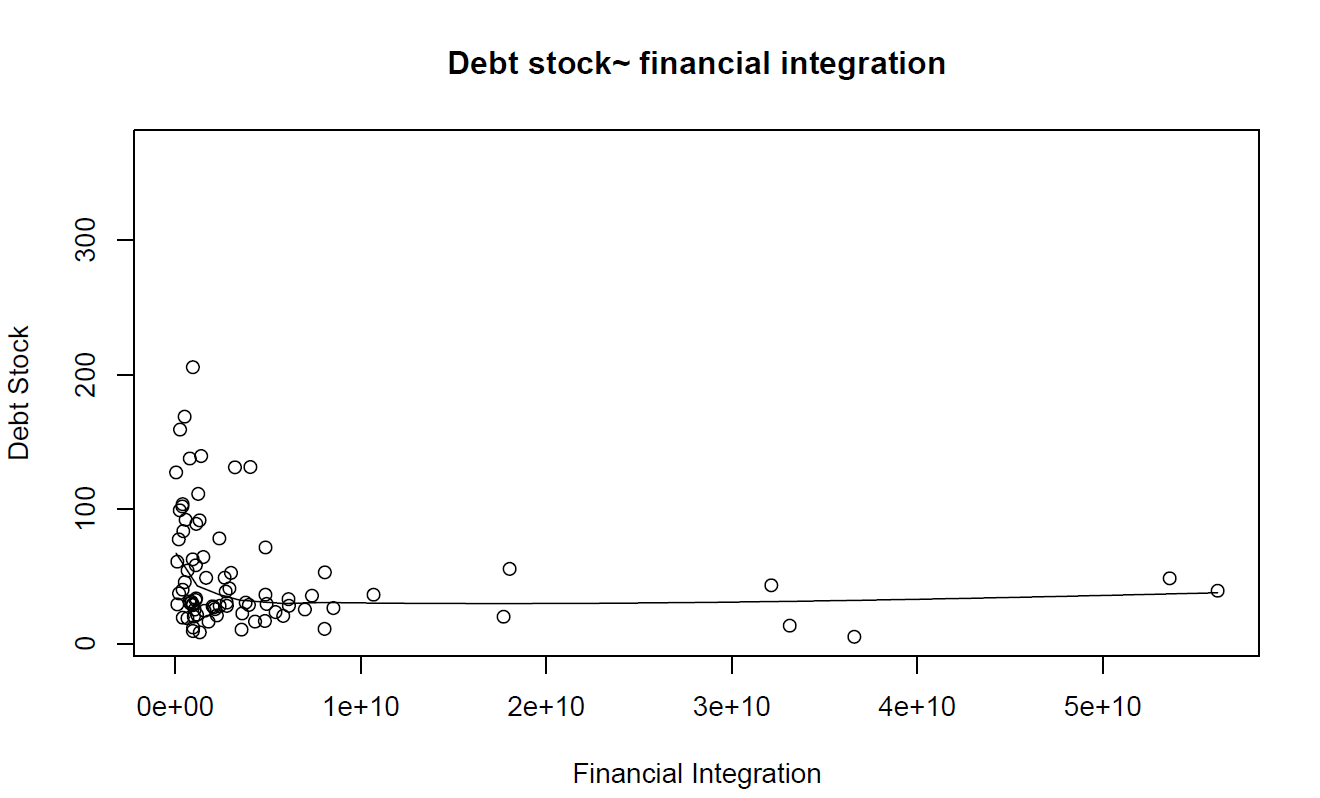
*Figure 6: A graph showing the autocorrelation factors of the residuals with respect to their lags in steps of 1.*



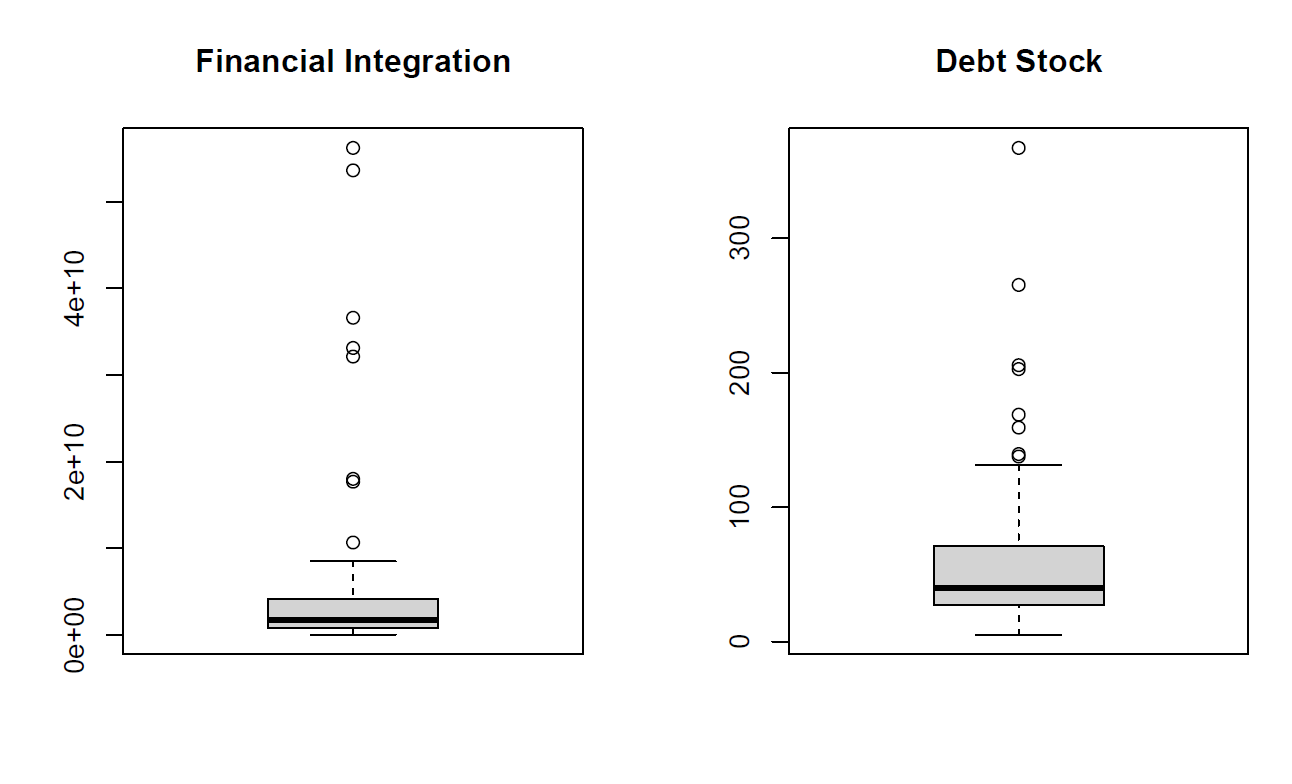
*Figure 7: A plot showing the correlation factors of the regressors X1 and X2.*



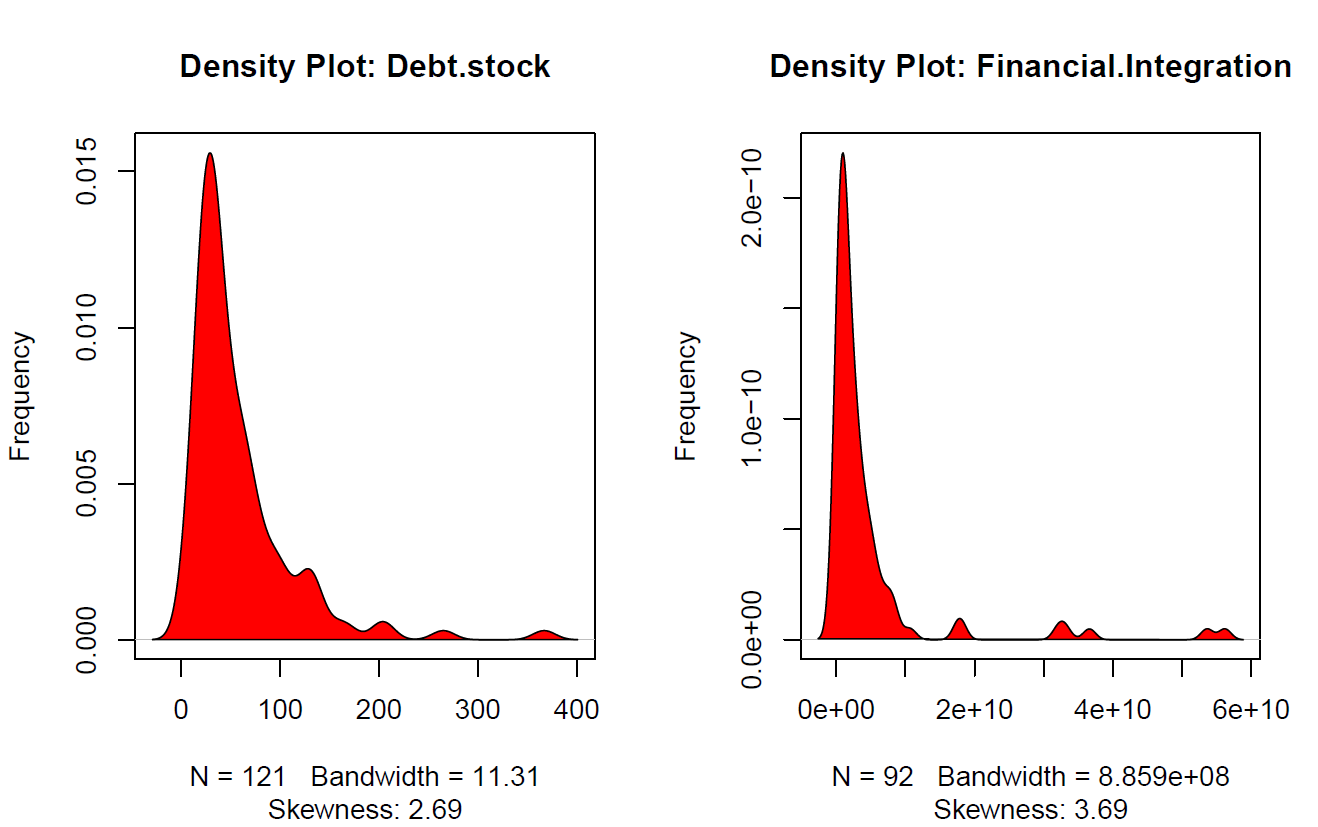
*Figure 8: A graph showing the estimated density of coefficient estimates from different multiple regression model estimators of the same true model.*



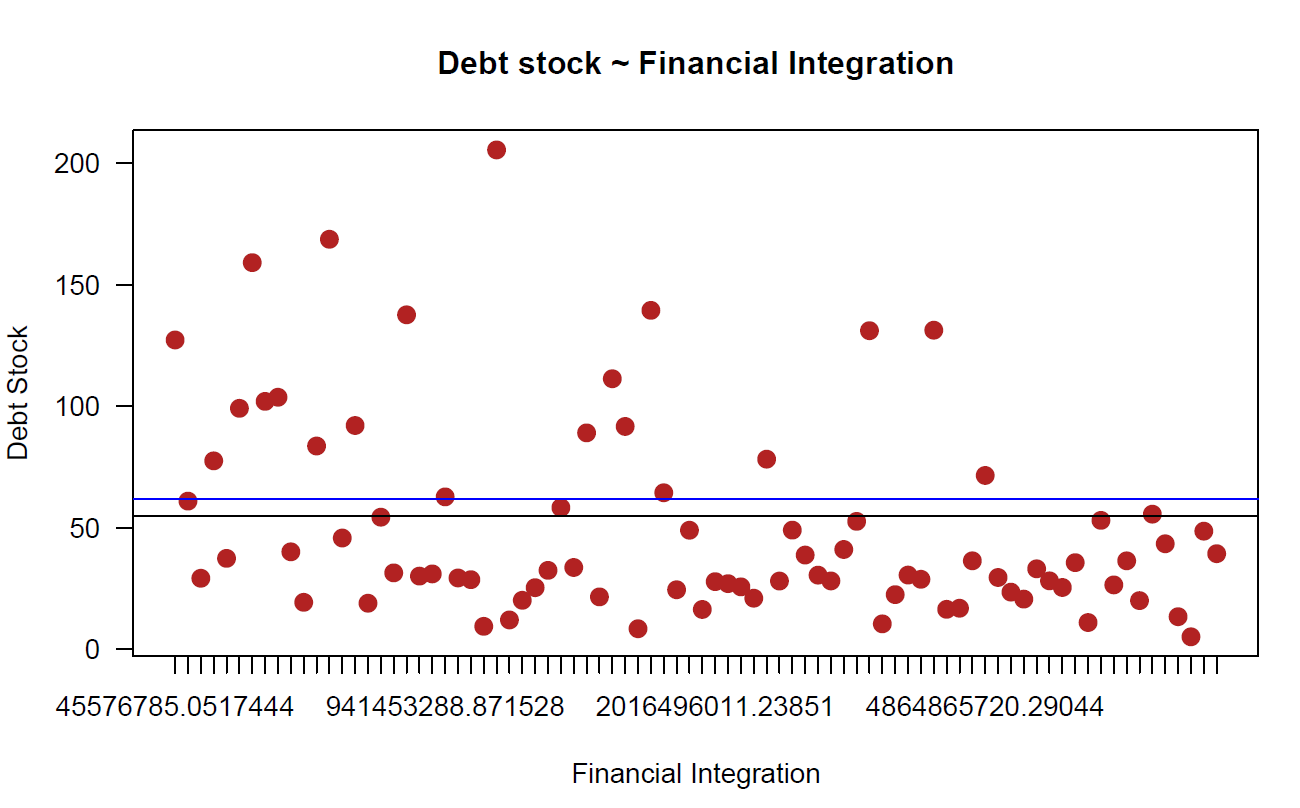
*Figure 9: A smoothed scatter plot of debt stock against financial integration values .*



*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: EQUATIONS**

*Equation 1: An equation of how the variable Y was simulated from X1, X2 and u*

**BIBLIOGRAPHY**

***To be inserted.***