Effect of economic advancement on deforestation*

Submitted by:

Arkajyoti Bhattacharjee †
Rezoanoor Rahman †

Supervised by:

Dr. Christopher Hans †

THE OHIO STATE UNIVERSITY

COLLEGE OF ARTS AND SCIENCES

Submitted on:

25^{th} April, 2023

*This report has been prepared towards the partial fulfillment of the requirements of the course *STAT6950: Applied Statistics II.*

†Department of Statistics, The Ohio State University.

Contents

1 Introduction

Deforestation is a significant concern because of its role in climate change, greenhouse gas emissions, and many more. We suspect that one primary reason behind deforestation is aggressive economic growth. However, developed countries tend to have relatively stable economies, so they do not require aggressive economic activities to maintain their natural economic growth. Moreover, we suspect that countries with lower corruption and better governance quality may be less vulnerable to the problem of deforestation. As a result, we want to analyze whether economic and governance quality indices affect the net forest conservation rate ('NFCR') for the year 2019. Additionally, as deforestation has been a significant topic of discussion and eco-friendly production is becoming more and more popular, are the effects of the year 2019 statistically different than the effects of the year 2014? We aim to explore these questions based on a subset of our primary data obtained from the Harvard Dataverse(HDB).
A description of the dataset along with its source is given in Table 1

Table 1: Data Description

Variable Name	Description	Source
NFCR	Net forest conversion rate	Food and Agricultural Organization (FAO)
CPI	Corruption Perception Index	Transparency International
AML	Anti-money-laundering Index	Basel Institute on Governance
ARR	Inbound Tourism - Arrivals	World Tourism Organization
GDP	Gross domestic product per capita	World Bank
WES	Wood export share	World Bank
GE	Government Effectiveness	Economist Intelligence Unit
PV	Political Stability and Absence of Violence/Terrorism	Economist Intelligence Unit
RQ	Regulatory Quality	Economist Intelligence Unit
RL	Rule of Law	Economist Intelligence Unit
VA	Voice and Accountability	Economist Intelligence Unit
RGDP	Rate of change in GDP. $RGDP_t = \frac{GDP_t - GDP_{t-1}}{GDP_{t-1}}$	Created

Furthermore, a simple comparison between variables from the year 2014 and 2019 are given in Table 1.
Clearly, 'NFCR' is our response variable and we choose regressors so as to make our inferences useful. We first begin with some exploratory data analysis in Section 2 to make sense of the data by understanding empirical relationships across variables, to detect possible multicollinearity in the predictors, and to detect possible transformation of predictors for linearity in regressors. In Section 3, we take two approaches to the model building - a parametric 3.1 and a non-parametric 3.2 approach. In Section 5, we discuss our findings and summarize our findings. In Section 7, we discuss future goals, what possibly went wrong, and what changes we would like to make to amend them. Finally, Section 9 provides some additional plots and a little theory behind some methods used in this analysis.

2 Exploratory Data Analysis

All our variables are continuous data. After cleaning up the data (changing GDP to relative GDP, selecting data pertaining to the year 2019, and formatting variable types), we look at the heatmap of the correlation matrix of the data (Figure 1) to detect the presence of multicollinearity in the data. Clearly, there seems to be a high positive linear correlation between 'RQ2019' & 'RL2019' (0.931), 'GE2019' & 'RQ2019'

Summary of variables for year 2014 and 2019

Variable	Mean		SD		Min		Max	
Name	2014	2019	2014	2019	2014	2019	2014	2019
NFCR	0.00	0.00	0.01	0.01	-0.05	-0.04	0.02	0.02
CPI	46.01	48.07	19.44	18.78	11.00	18.00	92.00	87.00
AML	5.79	5.31	1.22	1.10	2.51	2.68	8.56	8.22
$ARR \times 10^4$	1.30	1.90	2.97	3.56	0.00	0.01	20.66	21.79
$GDP \times 10^4$	1.61	1.84	2.11	2.24	0.04	0.05	11.88	11.47
WES	2.85	2.90	1.01	0.96	0.48	0.79	6.04	5.77
GE	0.14	0.27	0.93	0.90	-2.04	-2.01	2.18	2.22
PV	-0.02	0.05	0.83	0.78	-2.36	-1.92	1.47	1.64
RQ	0.15	0.31	0.92	0.88	-1.90	-1.46	2.23	2.16
RL	0.11	0.20	0.97	0.94	-1.92	-1.26	2.13	2.06
VA	0.10	0.19	0.92	0.91	-1.91	-1.80	1.68	1.66
RGDP	0.02	0.01	0.06	0.06	-0.23	-0.15	0.29	0.19

(0.946), 'GE2019' & 'RL2019' (0.94), 'GE2019' & 'CPI2019' (0.927), 'RQ2019' & 'CPI2019' (0.901), and'RL2019' & 'CPI2019'(0.967). Most of these variables are reflecting *Governance Quality*, viz., 'GE', 'RL', and 'RQ'. So, dropping one or two of them seems reasonable in our later model building.

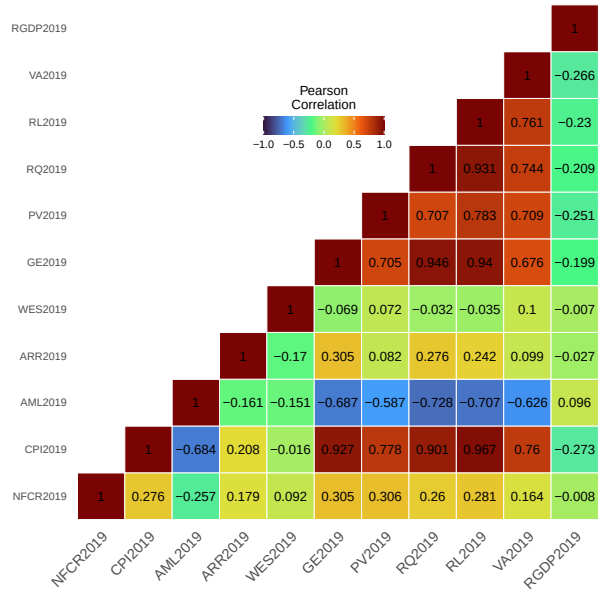

Figure 1: Heatmap of the correlation matrix of the data.

We look at the scatterplot matrix (Figure 2), after dropping 'GE2019' and 'RL2019', to understand the empirical relationship between the response and other variables. Note that since there are 5 variables for *Governance Quality*, dropping two seems reasonable given that they are strongly correlated with some other variables. 'ARR2019' seems to be skewed to the right (see Figure 3) and a log transformation seems reasonable. There is slight skewness in 'CPI2019' as well (see Figure 3) and may require some transformation. In general, some transformations in predictors, denoted as a whole by X, are needed here so that the graphs in the scatterplot matrix of X have a straight-line mean function (see Weisberg (2005), Chapter 8). None of the correlation values seem to be very high, which is what we expected from dropping the two variables.

However, we note that there is a moderately strong positive correlation between ‘PV2019’ & ‘CPI2019‘, ‘RQ2019’ & ‘CPI2019‘, ‘VA2019’ & ‘CPI2019‘, and ‘RQ2019’ & ‘VA2019‘. Note that all of these represent more or less the same thing and hence, the correlation is intuitive. The response ‘NFCR2019’ does not seem to be linearly correlated much with any of the predictors, suggesting non-linear transformations may help. It also has a heavy-tailed distribution, that is more evidently visible from the histogram in Figure 3. This is likely to affect the normality assumption and hence, cause problems in doing inferences. A Box-Cox transformation(see Box and Cox (1964)) or other similar transformations may help. Based on the boxplots (see Figure 4) of the *Governance Quality* variables, it seems ‘GE2019’, ‘PV2019’, & ‘RL2019’ have similar distributions and so does ‘VA2019’ & ‘RQ2019’, which is again intuitive from the data description as to what the variables represent. This suggests that we need to judiciously choose regressors from this set of 5 predictors.
Intuitively, ‘ARR2019’ and ‘WES2019’ seem important towards ‘NFCR2019’. We are also interested in modeling how ‘RGDP2019’ affects ‘NFCR2019‘. So, we will keep these three in all of our models.

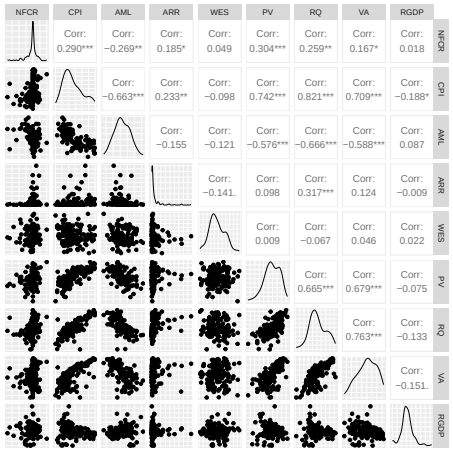

Figure 2: Scatterplot matrix of the data after dropping ‘GE2019’ and ‘RL2019‘.

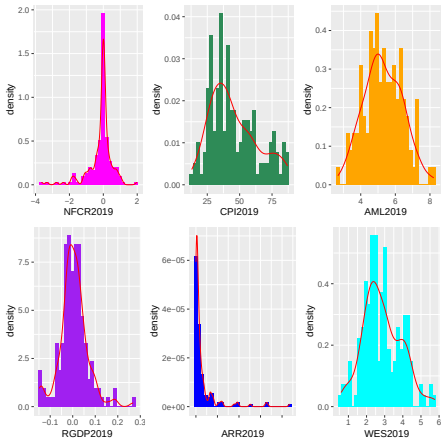

Figure 3: Histogram of the response and some other variables.

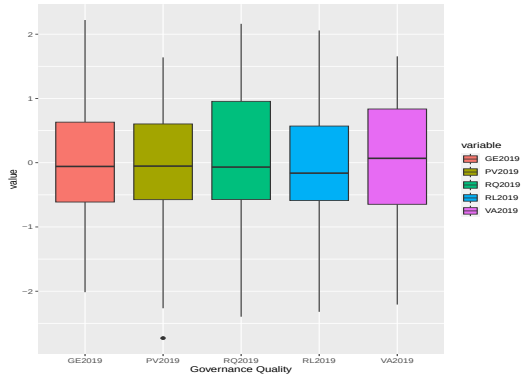

Figure 4: Boxplots of the variables representing Governance Quality.

3 Methods and Model Building

We explore a parametric as well as a non-parametric approach for modeling and making inferences.

3.1 Parametric Approach

3.1.1 Dealing with multicollinearity

First, we started with a naive model wherein we log-transformed 'ARR2019' and fit a linear model, regressing 'NFCR2019' with all other variables. As made obvious in the EDA stage, there is multicollinearity in the data. This is evidenced by the Variance Inflation Factor (VIF) for the model, as shown in Table 3.1.1. Clearly, we see that VIFs corresponding to 'CPI2019', 'GE2019', 'RQ2019', and 'RL2019' are all greater than 10.

CPI2019	AML2019	log(ARR2019)	WES2019	GE2019
19.278	2.395	1.962	1.135	18.443
PV2019	RQ2019	RL2019	VA2019	RGDP2019
3.077	13.762	23.836	3.263	1.147

Next, as suggested by the EDA stage, we drop 'RL2019'. Note that there are other *Government Quality* variables with which it is highly correlated; so, the drop makes sense and is reasonable. The VIFs for this model are shown in Table 3.1.1

CPI2019	AML2019	log(ARR2019)	WES2019	GE2019
11.697	2.389	1.961	1.125	18.006
PV2019	RQ2019	VA2019	RGDP2019	
2.935	12.668	3.249	1.134	

We still observe that the VIFs corresponding to 'CPI2019', 'GE2019', and 'RQ2019' are all greater than 10. As done in EDA stage, we drop 'GE2019', which is reasonable given that there are other similar *Governance Quality* variables. The VIFs for this model are all less than 10, as shown in Table 3.1.1, and we choose this as our baseline model.

CPI2019	AML2019	log(ARR2019)	WES2019
7.108541	2.387777	1.698668	1.121906
PV2019	RQ2019	VA2019	RGDP2019
2.918924	8.205519	3.067421	1.124691

3.1.2 Transformation of predictors

Once we have our baseline model, we improve upon it. First, we use variable transformation in X to get a linear mean function in the Xs. Note, here X comprises the columns of the data as given in Table 3.1.1. We use the Yeo-Johnson family of transformations (see Yeo and Johnson (2000), Chapter 8 of Weisberg (2005)) to achieve this. The suggested transformations are given in Table 2

Table 2: Transformation of the Independent Variables using Yeo-Johnson family of transformation

Predictor	Rounder Power	Lower Bound	Upper Bound
CPI2019	0.50	0.0591	0.7656
AML2019	1.00	-0.3670	1.1341
log(ARR2019)	1.00	0.3442	2.1085
WES2019	1.00	0.1340	1.2750
PV2019	1.32	1.0404	1.5912
VA2019	1.20	1.0509	1.3518
RGDP2019	1.00	-1.4781	2.8589

From the suggested transformations, variables raised 1.32 and 1.20 are not interpretable. So, we only transform 'CPI2019' to '$\sqrt{CPI2019}$'. We note that for the sake of exploration, we did perform all the model transformations as suggested by Table 2. However, after variable selection, we saw that the simple, interpretable model performed better in terms of adjusted R^2.

3.1.3 Variable Selection

We used stepwise regression (see Chapter 10 of Weisberg (2005)) to obtain the optimal number of regressors and what they are. The optimal number of regressors is 4 and the variables selected are 'log(ARR2019)', 'WES2019', 'PV2019', and 'VA2019'. Intuitively, keeping ARR and WES makes absolute sense; and, PV and VA represent some aspects of governance quality/corruption. We are interested to keep 'RGDP2019' in the model and we note that it improves the model, in terms of adjusted R^2 than when it is not included.

3.1.4 Transformation of the response

We now check if a transformation is suggested by the Yeo-Johnson family of transformation for the response so that we can get closer to normality. Table 3.1.4 suggests raising the response to the power 1.61, which is not interpretable. So, we keep the response as it is. However, we note that the model with the transformed Y did not perform better than the non-transformed model in terms of adjusted R^2.

Response	Rounder Power	Lower Bound	Upper Bound
NFCR2019	1.61	1.394	1.834

3.2 Non-parametric Approach

Here, we explore a non-parametric regression method called Alternating Conditional Expectations (ACE; see Breiman and Friedman (1985)). It is an iterative algorithm that aims to find nonlinear optimal transformations that produce the best-fitting additive model. ACE transforms the response variable Y and the predictor variables, X_i to minimize the fraction of variance not explained.
We fit a model using the ACE algorithm keeping all the variables. The non-linear transformations for the variables are given in Figure 14. Clearly, the transformations are non-trivial.

4 Regression Diagnostics

At this stage, we look at whether the assumptions of the Gauss-Markov theorem are met or not.

4.1 Parametric Model

4.1.1 Outlier Analysis

We used Cook's Distance (see Cook (2000), Cook (1979), Chapter 9 of Weisberg (2005)), leverage statistics (see Chapter 9 of Weisberg (2005)), and studentized residuals to find out potential outliers. We took the cutoff for each as 0.5, $\frac{3p}{n}$, and 3 respectively, where p is the number of predictor variables including intercept and n is the number of observations. We used a combination of all these statistics and deleted a few of the observations which seemed like potential outliers. The outcome of the outlier analysis is shown in Figure 5.

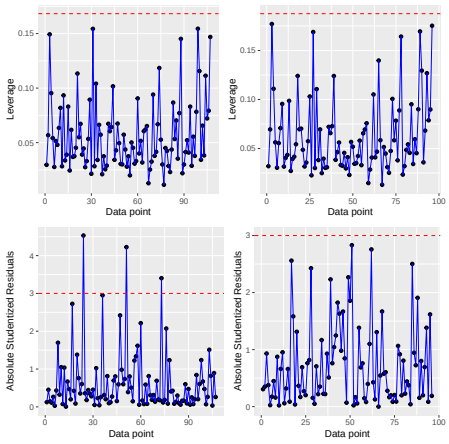

Figure 5: Leverage before (topleft) & after(topright) and absolute studentized residuals before(bottomleft) & after(bottomright) outlier analysis.

4.1.2 Homoscedasticity of errors

We check if the errors are homoscedastic (constant variance) or heteroscedastic. We look at the plot of the residuals vs. fitted values in Figure 6. The plot seems to be fairly random and centered at 0. There is a slight pattern, although negligible. The variance apparently looks constant from the plot.
To confirm, we perform the Breusch-Pagan test (see Breusch and Pagan (1979)). The result is presented in Table 4.1.2, wherein we see that we fail to reject the null hypothesis of homoscedasticity at %5 level of significance.

χ^2	Df	$p-value$
2.399	1	0.121

4.1.3 Multicollinearity

We already dealt with multicollinearity for a bigger model. For the current model, our VIFs are all less than 10, as shown in table 4.1.3.

ARR	WES2019	PV2019	VA2019	RGDP2019
1.078	1.083	2.081	2.095	1.101

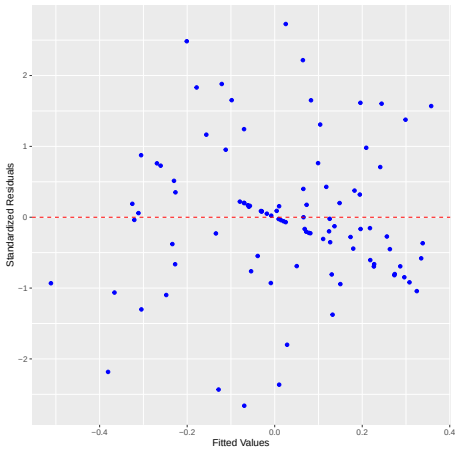

Figure 6: Leverage before (topleft) & after(topright) and absolute studentized residuals before(bottomleft) & after(bottomright) outlier analysis.

4.1.4 Normality of errors

Although normality is not a required assumption of the Gauss-Markov model, we do need it for inference purposes. We look at the normal Q-Q plot in Figure 7. Although the middle part is fairly linear, the tails indicate clear non-normality; it indicates a heavy-tailed distribution. We confirm non-normality using the Shapiro-Wilks test(see Shapiro and Wilk (1965)), whose result is presented in Table 4.1.4.

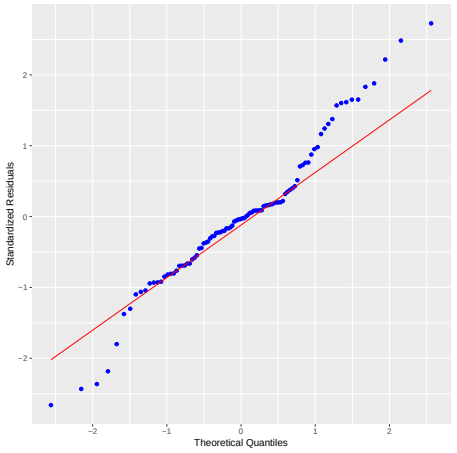

Figure 7: Normal Q-Q plot.

$W-statistic$	$p-value$
0.96792	0.01873

4.2 Non-parametric Model

4.2.1 Outlier Analysis

Similar to the parametric approach, we used Cook's Distance, leverage statistics, and studentized residuals to find out potential outliers. We took the cutoff for each as 0.5, $\frac{3p}{n}$, and 3 respectively, where p is the number

of predictor variables including intercept and n is the number of observations. We used a combination of all these statistics and deleted a few of the observations which seemed like potential outliers. The outcome of the outlier analysis is shown in Figure 8.

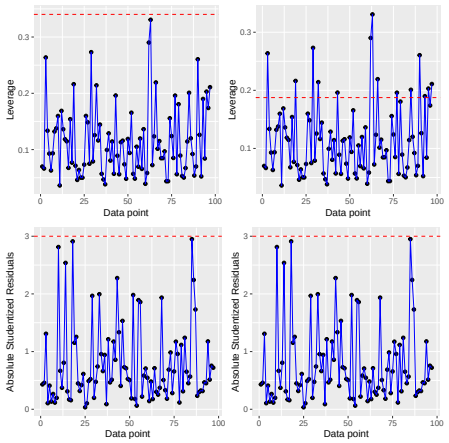

Figure 8: Leverage before (topleft) & after(topright) and absolute studentized residuals before(bottomleft) & after(bottomright) outlier analysis.

4.2.2 Homoscedasticity of errors

We check if the errors are homoscedastic (constant variance) or heteroscedastic. We look at the plot of the residuals vs. fitted values in Figure 9. The plot seems to be fairly random and centered at 0. There is a slight pattern, although negligible. The variance apparently looks constant from the plot.

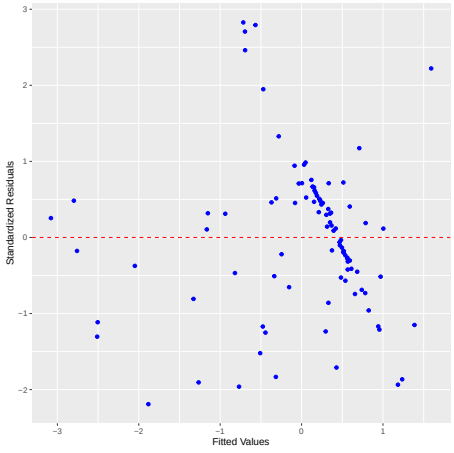

Figure 9: Leverage before (topleft) & after(topright) and absolute studentized residuals before(bottomleft) & after(bottomright) outlier analysis.

To confirm, we perform the Breusch-Pagan test (see Breusch and Pagan (1979)). The result is presented in Table 4.2.2, wherein we see that we fail to reject the null hypothesis of homoscedasticity at %5 level of significance.

χ^2	Df	$p-value$
3.365463	1	0.066577

4.2.3 Multicollinearity

For the current model, our VIFs for all the transformed predictors are all less than 10, as shown in table 4.2.3.

CPI2019	AML2019	ARR2019	WES2019	GE2019
6.980	1.368	1.169	1.098	3.720
PV2019	RQ2019	RL2019	VA2019	RGDP2019
2.794	6.362	3.945	1.207	1.189

4.2.4 Normality of errors

Although normality is not a required assumption of the Gauss-Markov model, we do need it for inference purposes. We look at the normal Q-Q plot in Figure 10. Although the middle part is fairly linear, the tails indicate clear non-normality; it indicates a heavy-tailed distribution. We confirm non-normality using the Shapiro-Wilks test(see Shapiro and Wilk (1965)), whose result is presented in Table 4.2.4.

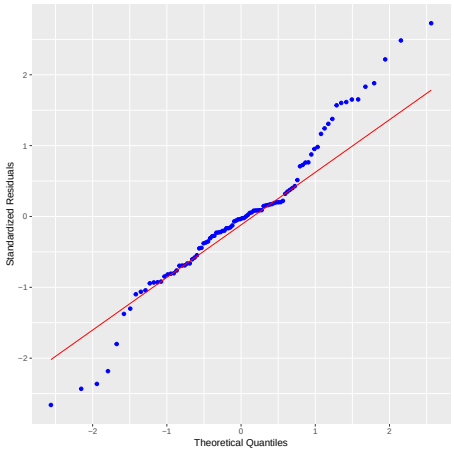

Figure 10: Normal Q-Q plot of residuals.

$W-Statistic$	$p-value$
0.9615	0.006126

5 Models, Interpretations, and Inference

Here, we summarize all our results, try to interpret our models and look into inferences.

5.1 Models, Summaries, and Interpretation

5.1.1 Models

Based on the parametric approach, our final model is given by Equation 5.1.1

$$E[NFCR|X] = \beta_0 + \beta_1 \log(ARR) + \beta_2 WES + \beta_3 PV + \beta_4 VA + \beta_5 RGDP, \quad (1)$$

Table 3: Estimated coefficients and their corresponding t and p-values

Parameter	Estimate	SE	$t-value$	$p-value$
β_0	-0.859	0.259	-3.320	0.001
β_1	0.082	0.023	3.568	0.001
β_2	0.052	0.040	1.288	0.201
β_3	0.213	0.071	2.984	0.004
β_4	-0.079	0.060	-1.311	0.193
β_5	1.319	0.705	1.871	0.065

wherein X represents all the regressors present in the equation. The estimated coefficients are given in Table 3.
Based on the non-parametric approach, our final model is given by Equation 5.1.1

$$E[tNFCR|tX] = \alpha_0 + \alpha_1 tCPI + \alpha_2 tAML + \alpha_3 tARR + \alpha_4 tWES + \alpha_5 tGE + \alpha_6 tPV + \alpha_7 tRQ + \alpha_8 tRL + \alpha_9 tVA + \alpha_{10} tRGDP, \tag{2}$$

wherein tX represents all the non-linearly transformed regressors present in the equation. The estimated coefficients are given in Table 4.

Table 4: Estimated coefficients and their corresponding t and p-values

Parameter	Estimate	SE	$t-value$	$p-value$
α_0	0.00	0.05	0.00	1.00
α_1	1.12	0.43	2.58	0.01
α_2	1.21	0.30	4.07	0.00
α_3	1.36	0.37	3.68	0.00
α_4	1.22	0.32	3.86	0.00
α_5	0.92	0.28	3.32	0.00
α_6	1.07	0.34	3.14	0.00
α_7	1.52	0.63	2.42	0.02
α_8	1.16	0.31	3.80	0.00
α_9	1.11	0.25	4.45	0.00
α_{10}	1.07	0.49	2.19	0.03

5.1.2 Interpretation

β_js in Equation 5.1.1 have the usual interpretation as in any linear regression setting: expected change in response for a unit change in the j^{th} variable, keeping the other variables fixed. Here, we focus on the sign of the coefficients, i.e., the direction of the relationship between the response and the variables. That is, based on the signs, we can say that the rate of deforestation increases as Inbound Tourism-Arrivals (ARR) increases, as wood Export Share (WES) increases, as stability and absence of Terrorism (PV) increases, as Voice and Accountability (VA) decrease, and as GDP increases.
However, the α_js in Equation 5.1.1 are not interpretable due to non-linear transformations in both X and Y and because the transformations are not one-one and monotone as evident from Figure 14.

5.1.3 Model Adequacy

The R^2 and adjusted R^2 (R^2_{adj}) values of the baseline, parametric and non-parametric models are shown in Table 5.
Although our linear model shows an improvement over the naive model, the non-parametric model is the clear winner. This is obvious from the fact that the ACE algorithm aims to maximize R^2.

Table 5: R^2 and R^2_{adj} values for the three models.

Model	R^2	R^2_{adj}
Naive	0.154	0.001
Parametric	0.225	0.182
Non-parametric	0.816	0.795

5.2 Inference

Since the errors are not normal, inferences made are not valid based on t-statistics. If they were valid, then for the parametric model the effect of log(ARR), PV, and RGDP only are significant at 10% level of significance; for the non-parametric model, all the predictors are statistically significant at 5% level of significance.
For the sake of exploration, we observed that the errors may be from a Cauchy(0, $10^{-.6}$) distribution. This is validated by Figure 11 and the Kolmogorov-Smirnov test, whose result is shown in Table 6

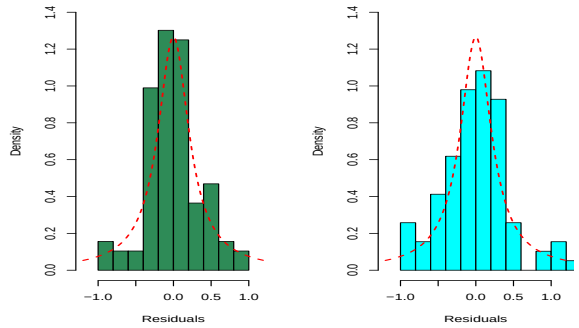

Figure 11: Comparing histogram of residuals with Cauchy(0, $10^{-0.6}$).

Table 6: Kolmogorov-Smirnov Tests.

Model	D Statistic	p-value
Parametric	0.121	0.110
Non-parametric	0.101	0.249

6 Comparison of model parameters for years 2019 and 2014

The estimated coefficients for the years 2019 and 2015 are given in Table 7. Although the other coefficients are almost the same, coefficients for *VA* and *RGDP* has different signs for two different years which suggests that there might be a significant difference between the two sets of slope parameters.

Table 7: Estimated coefficients for 2019 and 2015

Independent Variable	Estimate 2019	2014
ARR	0.082	0.086
WES	0.052	0.044
PV	0.213	0.151
VA	-0.079	0.110
RGDP	1.319	-1.175

For testing the slope parameters of years 2019 and 2014, we first build a model using the same selected independent variables and their transformations and then use a Hotelling-T^2 statistic (see Hotelling (1931)) to compare them with the fitted model in subsection 5.1.1. Here we have, the number of observations for years 2019 and 2014 as $n_{2019} = 96$ and $n_{2014} = 133$, respectively. If $\hat{\beta}_{2019}$ and $\hat{\beta}_{2014}$ be the estimated slope parameters; and $\hat{V}_{2019}$ and $\hat{V}_{2014}$ be the estimated variance-covariance matrix of them, then $T^2 = (\hat{\beta}_{2019} - \hat{\beta}_{2019})^T \left(V(1/n_{2019} + 1/n_{2014}) \right)^{-1} (\hat{\beta}_{2019} - \hat{\beta}_{2019})$, where $V = \frac{(n_{2019-1})\hat{V}_{2019} + (n_{2014-1})\hat{V}_{2014}}{n_{2019}+n_{2015}-2}$. Approximately $\frac{n-k}{k(n-1)} T^2 \sim F_{k,n-k}$. Note that, although the normality of errors assumption has not been filled, we are using this assumption based on what would be an appropriate test if it was fulfilled. The results for our test are given in Table 8. Here we have $F - statistic = 135.02$ and corresponding $p - value = 0.00$. Thus we conclude that there is a significant difference between the two sets of slope parameters at 5% level of significance.

Table 8: Test results for testing coefficients of years 2019 and 2014

n_{2019}	n_{2014}	k	T^2	$F - statistic$	$p - value$
96	133	5	687.16	135.02	0.00

7 Discussion

In this project, we explored how economic growth has affected the rate of deforestation. We saw, in general, that economic growth promotes deforestation and a causal link is intuitive here, although not formally tested. Granger causality test (see Granger (1969)) can be used to check if indeed 'RGDP' and 'NFCR' are causally related across the years 201 through 2019. We chose the year 2019 rather than 2020, simply because the latter was a "COVID" year and so, the results may have been unexpected or counter-intuitive. We think this is worth further exploration. It would be worthwhile to explore methods, parametric as well as nonparametric, that deal with the inherent non-linearity in this dataset and is amenable to inference as well. We saw a scarcity of packages in R that deals with linear regression using non-Gaussian errors. In fact, we explored a package "Heavy" (see Osorio and F. (2018)) which provides a limited amount of flexibility in defining error family parameters, with the error families being normal, Cauchy, Student's t, slash, and contaminated distributions. We would like to see more computational development in this area and is worth exploring in our opinion. The R code for the plots and analysis done here are available at https://github.com/ArkaB-DS/STAT6950Project.

8 Acknowledgements

We would like to thank Dr. Hans for the opportunity of this project and for providing us with guidance whenever we needed it.

9 Appendix

9.1 Some additional plots

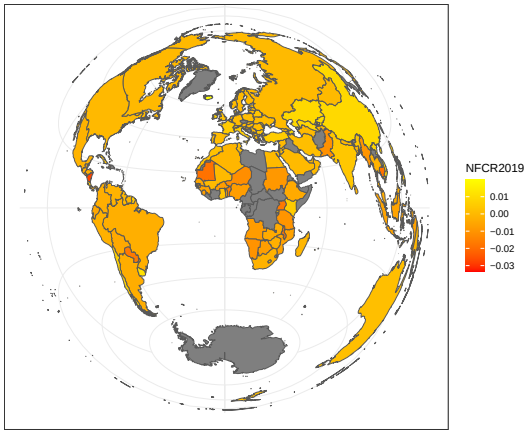

Figure 12: Deforestation in World Map

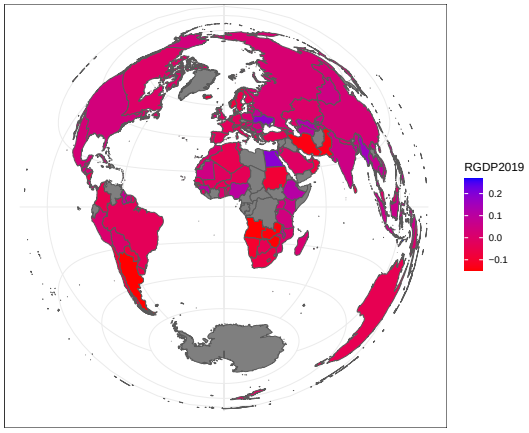

Figure 13: RGDP in World Map

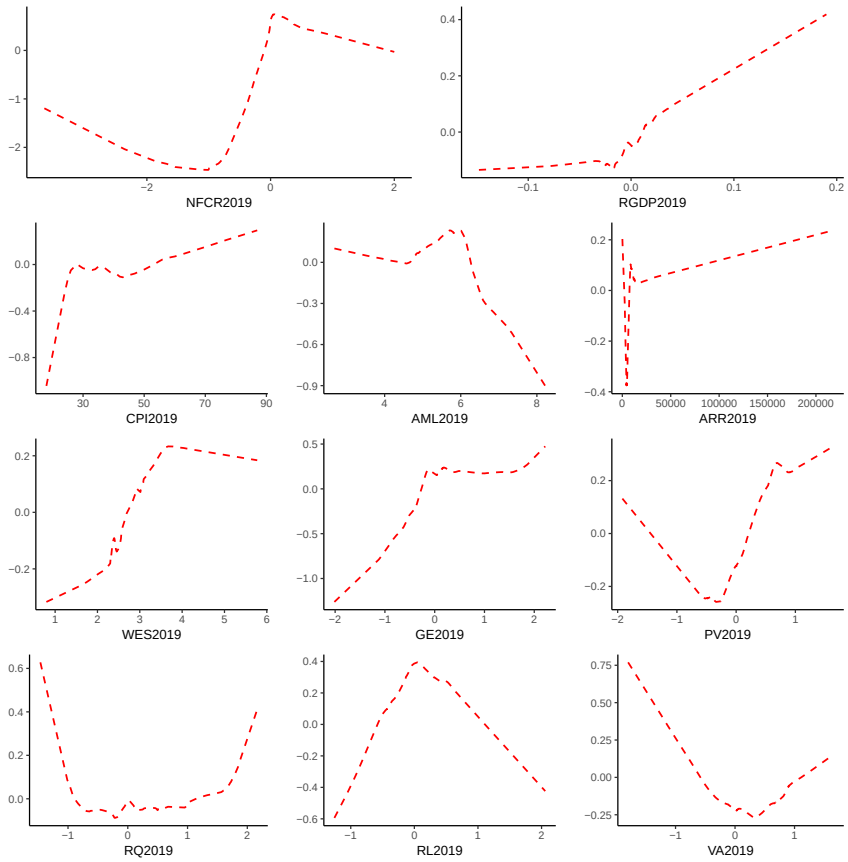

Figure 14: Non-linear optimal transformations of all variables.

References

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2):211–243.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the econometric society*, pages 1287–1294.

Cook, R. D. (1979). Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):169–174.

Cook, R. D. (2000). Detection of influential observation in linear regression. *Technometrics*, 42(1):65–68.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Hotelling, H. (1931). The Generalization of Student's Ratio. *The Annals of Mathematical Statistics*, 2(3):360 – 378.

Osorio and F. (2018). *heavy: Robust estimation using heavy-tailed distributions*. R package version 0.38.19.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Weisberg, S. (2005). *Applied linear regression*, volume 528. John Wiley & Sons.

Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.