Stats 101A Capstone Regression Model

Cassidy Sadowski

March 2024

1 Introduction

Since 2010, net global forest loss has been 4.7 million hectares per year. Up to 15 million trees are chopped down each year. Deforestation has staggering implications for wildlife conservation and climate change. Forests are natural habitats to more than half of the world's land-based animals and plants and three-quarters of all birds. Trees are also natural carbon capture facilities reducing overall carbon dioxide levels in the atmosphere and thus decreasing the impact of human activities on the environment. Deforestation is the cause of 10% of global warming. Thus, to work towards reducing human impact on the environment and reversing the effects of climate change a model must be developed to understand the drivers of deforestation.

The dataset analyzed in this project comes from the Harvard Dataverse and was constructed by Cozma et al. in 2023[1]. The response variable considered in this analysis is Net Forest Conversion Rate which measures net gain in forestland (in 1000 ha) per year as a percent, with negative gain representing loss of forest land within a given country. The predictors considered in this analysis were the following:

1. **CPI: Corruption Perception Index.** The level of perceived corruption in a country on a scale from 0 (extremely corrupt) to 100 (not corrupt).
2. **AML: Anit-Money-Laundering Index.** The vulnerability of a country to money laundering and the country's capacity to counter money laundering measured on a scale from 0 (not vulnerable) to 100 (extremely vulnerable).
3. **ARR: Inbound Tourism - Arrivals.** The number of tourists that arrive to a country per year, in thousands.
4. **GDP: Gross Domestic Product per Capita.** Measured in USD ($).
5. **WES: Wood Export Share.** The percent of a country's total exports made up by wood.
6. **GE: Government Effectiveness.** The quality of public and civil services, independence from political pressures, policy formulation and implementation, and the credibility of the government's commitment to policies measured on a scale from -2.5 (least effective) to 2.5 (most effective).
7. **PV: Political Stability and Absence of Violence and Terrorism.** The perception of political instability and politically motivated violence measured on a scale of -2.5 (weakest) to 2.5 (strongest).

[1] https://doi.org/10.7910/DVN/I8WFGF

8. **RQ: Regulatory Quality.** The perceived ability of the government to formulate and enforce regulations which promote and permit private sector development measured on a scale of -2.5 (weakest) to 2.5 (strongest).

9. **RL: Rule of Law.** The perceived quality of law enforcement, property rights, and the courts as well as the likelihood of crime and violence measured on a scale of -2.5 (weakest) to 2.5 (strongest).

10. **VA: Voice and Accountability.** The perception of how much citizens are able to select their own government in a given country and the existence of freedoms of association, expression, and media measured on a scale of -2.5 (weakest) to 2.5 (strongest).

2 The Full Linear Model

$$\begin{aligned} NFCR &= \beta_1 CPI + \beta_2 AML + \beta_3 ARR + \beta_4 GDP + \beta_5 WES \\ &\quad + \beta_6 GE + \beta_7 PV + \beta_8 RQ + \beta_9 RL + \beta_{10} VA + \beta_0 \end{aligned}$$

Table 1: Regression analysis result of the full linear model.

Factor	$\boldsymbol{\beta}$	**S.E.**	**T**	**VIF**
Intercept	$2.859 * 10^{-1}$	$3.285 * 10^{-1}$	0.871	–
CPI	$-3.144 * 10^{-3}$	$5.286 * 10^{-3}$	-0.595	15.610
AML	$-7.037 * 10^{-2}$	$3.238 * 10^{-2}$	-2.173*	2.349
ARR	$1.953 * 10^{-6}$	$9.020 * 10^{-7}$	2.165*	1.178
GDP	$-4.491 * 10^{-6}$	$2.232 * 10^{-6}$	-2.011*	3.330
WES	$4.243 * 10^{-2}$	$2.707 * 10^{-2}$	1.567	1.152
GE	$6.274 * 10^{-1}$	$1.060 * 10^{-1}$	5.918*	14.368
PV	$7.194 * 10^{-2}$	$5.055 * 10^{-2}$	1.423	2.689
RQ	$-2.831 * 10^{-1}$	$9.604 * 10^{-2}$	-2.948*	11.135
RL	$-5.057 * 10^{-2}$	$1.290 * 10^{-1}$	-0.392	22.642
VA	$1.561 * 10^{-2}$	$5.014 * 10^{-2}$	0.311	3.172
$R^2_{adj} = 0.1257$				

* $p < 0.05$

The given model suggests that for each increase of one on the corruption perception index, $3.144 * 10^{-3}$% of forestland in a country is lost per year; for an increase of one on the anti-money-laundering index, $7.037 * 10^{-2}$% of forestland in a country is lost per year; for an increase of one thousand tourist arrivals per year, forestland grows by $1.953 * 10^{-6}$% each year; for an increase of one US dollar in GDP per capita $4.491 * 10^{-6}$% of forestland is lost; for an increase of one percent in wood export share forestland increases by $4.243 * 10^{-2}$%; for an increase of one in government effectiveness, forestland increases by $6.274 * 10^{-1}$% within a given country; for an increase of one in political stability and absence of violence and terrorism, forestland increases by $7.194 * 10^{-2}$% in a given country; for an increase of one in regulatory quality, $2.831 * 10^{-1}$% of forestland is lost within a given country; for an increase of one in percieved quality of rule of law, $5.057 * 10^{-2}$% of

forestland is lost within a given country; and finally for an increase of one in strength of voice and accountability, forestland increases by $1.561 * 10^{-2}\%$.

However, the t-test shows that many of the predictors fed to the full model are not significant. The significant predictors of the model are AML, ARR, GDP, GE, and RQ. Additionally, the variance inflation factor reveals that there exists high multicolinearity among predictors which also suggests the model needs to be reduced to better predict the response variable.

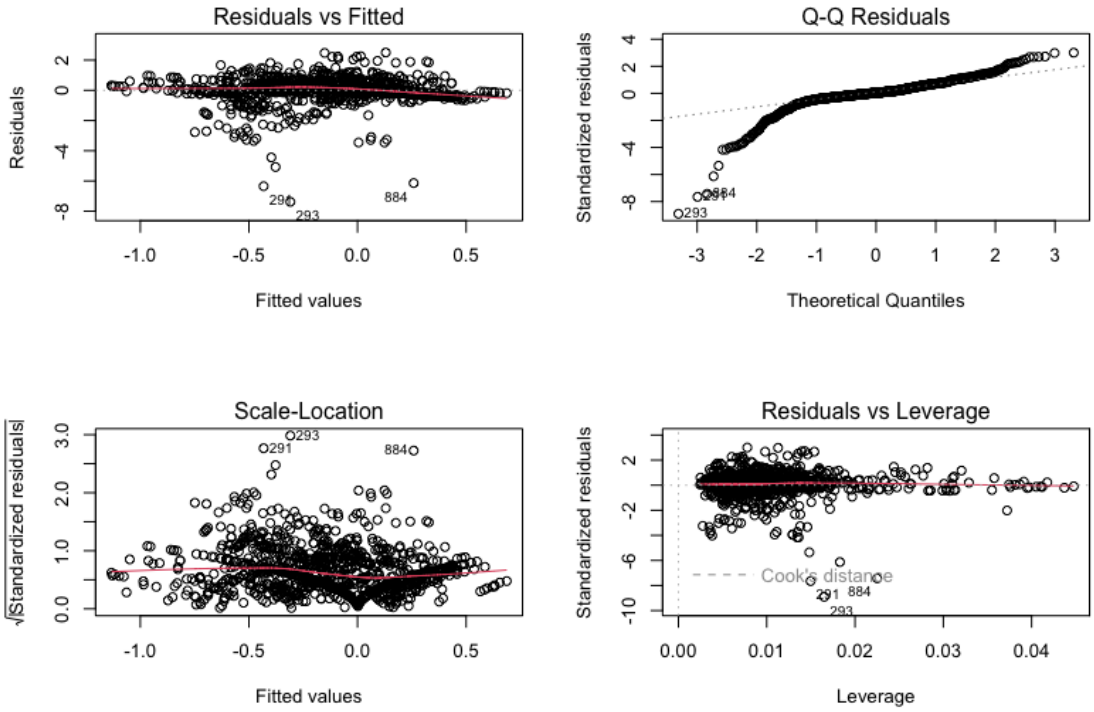

Figure 1: Full model diagnostic plots.

The residuals vs fitted plot does not show any discernible pattern, the variance of the residuals appears constant, and the average residuals are zero shown by the line highlighted in red thus no model assumptions are violated. The q-q residuals plot is light tailed and shows a slight deviation between theoretical quantiles -3 and -2, but generally shows that the errors are mostly consistent with the normal distribution. The Scale-Location plot shows a generally equal random spread of points suggesting that homoscedasticity is not violated. Finally, the Residuals vs Leverage plot does not show any points that are influential against the regression line. With all this in mind, we can be confident that the model assumptions are not violated and we can move forward and reduce the linear model with step-wise regression.

3 Step-wise Regression

A backwards step-wise regression of the full model with regard to AIC yields the following reduced model:

$$NFCR = \beta_1 AML + \beta_2 ARR + \beta_3 GDP + \beta_4 GE + \beta_5 RQ + \beta_0$$

Table 2: Regression analysis result of the backwards step-wise regression model.

Factor	β	**S.E.**	**T**	**VIF**
Intercept	$3.844 * 10^{-1}$	$1.792 * 10^{-1}$	2.145*	–
AML	$-8.838 * 10^{-2}$	$3.078 * 10^{-2}$	-2.871*	2.120
ARR	$1.674 * 10^{-6}$	$8.755 * 10^{-7}$	1.912*	1.109
GDP	$-4.983 * 10^{-6}$	$2.124 * 10^{-6}$	-2.346*	3.010
GE	$5.900 * 10^{-1}$	$8.685 * 10^{-2}$	6.794*	9.637
RQ	$-3.040 * 10^{-1}$	$8.698 * 10^{-2}$	-3.495*	9.127
$R^2_{adj} = 0.1251$				

* $p < 0.05$

Following the utilization of backwards step-wise regression to optimize our linear model, all predictors remaining in the model are significant in predicting NFCR. The β values can be interpreted in the same manner as presented in the full model. It is of note that the R^2_{adj} did not improve in response to this model reduction which means that 12.5% of the variation in y is accounted for by the variation in the predictors of both the full model and the reduced model. Since variance inflation factor is still high for both GE and RQ it important to check that no model assumptions are violated under our reduced model.

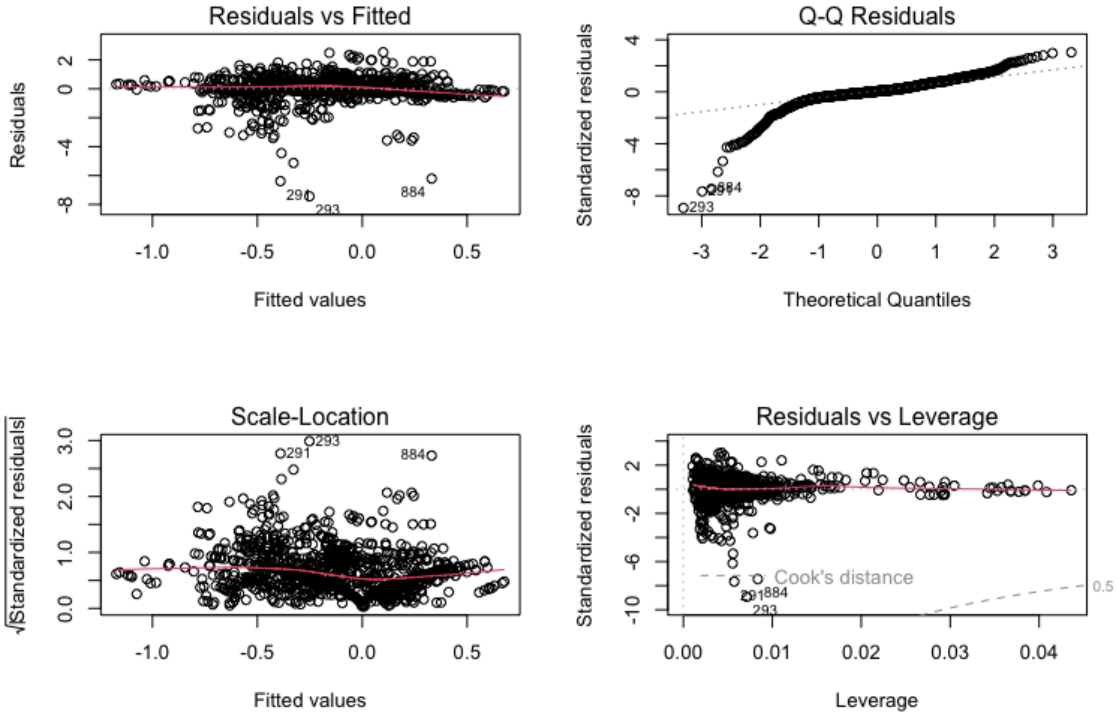

Figure 2: Reduced model diagnostic plots.

The Residuals vs Fitted, Q-Q Residuals, and Residuals vs Leverage show the same trends as visualized for the full model. The Scale-Location plot shows a slight deviation from random spread

with a dip at the fitted value of 0, but the square root of standardized residuals were fairly randomly and evenly spread which allows us to accept the model assumptions. Though both GE and RQ show a high VIF we can retain both predictors since they are both significant and no model assumptions are violated.

Finally, to determine if the reduced linear model offered by the backwards step-wise reduction in regard to AIC provides a better model than the full model a partial F-test can be utilized. The partial F-test revealed a p-value of 0.3243. From this p-value we fail to reject the null hypothesis, thus stating the reduced model is sufficient at predicting the response in NFCR.

4 Transformed Model

Though the reduced model presented a significant improvement from the full model, an improvement in R^2_{adj} was not seen. Thus a transformation of the data is likely necessary to best predict the response in NFCR. Visualization of the NFCR variable shows a strong peak around 0, with 87.3% of all observations falling between -1 and 1 and a left-skew. However, since many of the values for NFCR fall below zero, it is not possible to complete a log transformation or a box-cox transformation. Thus, in order to perform a transformation all values needed to be scaled to be positive by an addition of 8 (since the minimum value of the dataset was -7.6923).

Following this scaling, a box-cox transformation was performed revealing a 95% confidence interval for λ which did not include one. This suggests that a transformation was necessary to best predict the response variable. Finally, since the full model was utilized in the transformation a backwards step-wise reduction was calculated following the transformation to ensure only significant predictors were utilized to predict the response in the transformed NFCR. The output was as follows:

Table 3: Regression analysis result of the backwards step-wise regression model following a box-cox transformation of the scaled response variable.

Factor	$\boldsymbol{\beta}$	**S.E.**	**T**	**VIF**
Intercept	$2.498 * 10^2$	$1.644 * 10^1$	15.193*	–
AML	$-6.685 * 10^0$	$2.359 * 10^0$	-2.834*	2.340
ARR	$1.415 * 10^{-4}$	$6.430 * 10^{-5}$	2.201*	1.124
GDP	$-3.443 * 10^{-4}$	$1.550 * 10^{-4}$	-2.221*	3.014
WES	$5.209 * 10^0$	$1.948 * 10^0$	2.674*	1.120
GE	$4.556 * 10^1$	$6.337 * 10^0$	7.189*	9.641
RQ	$-2.403 * 10^1$	$6.369 * 10^0$	-3.774*	9.195
$R^2_{adj} = 0.1478$				

* $p < 0.05$

A partial F-test between the full transformed model and the step-wise reduced transformed model presented above revealed a p-value of 0.6995 which suggests the reduced transformed model is sufficient at predicting the response in transformed NFCR. While R^2_{adj} improved in comparison to the reduced linear model, it is still not large enough to demonstrate that a linear relationship exists between the predictors and the transformed NFCR.

In considering why a linear relationship is not seen between the predictors and NFCR or the transformed NFCR many limitations come to mind. First, deforestation within a country may be dependent on external variables not considered in this model such as the amount of protected land, active urban development, and the random effect of natural disasters within a country. In the future, to better understand the driving factors contributing to deforestation it may be helpful to include a time-series in the model as deforestation rates have varied significantly across decades.