Finding the strongest determinants of economic growth

Minh Nguyen

4/13/2022

Introduction

In a landmark paper in 1996, Robert J. Barro conducted an empirical study on 100 countries between 1960-1990 to find the strongest determinants of economic growth among the given sample. He concluded that for a fixed value of GDP per capita (which is GDP/total population of the country), growth rate is increased by "higher initial schooling, life expectancy, lower fertility, lower government consumption, better maintenance of the rule of law, lower inflation, and improvements in the terms of trade" [1]. It's no surprise then, that economic growth is an incredibly important topic for policy-makers, especially those outside the West. A strong economic policy plan can often improve the living standards of millions. On the contrary, bad policymaking can plunge an entire nation into poverty for multiple generations.
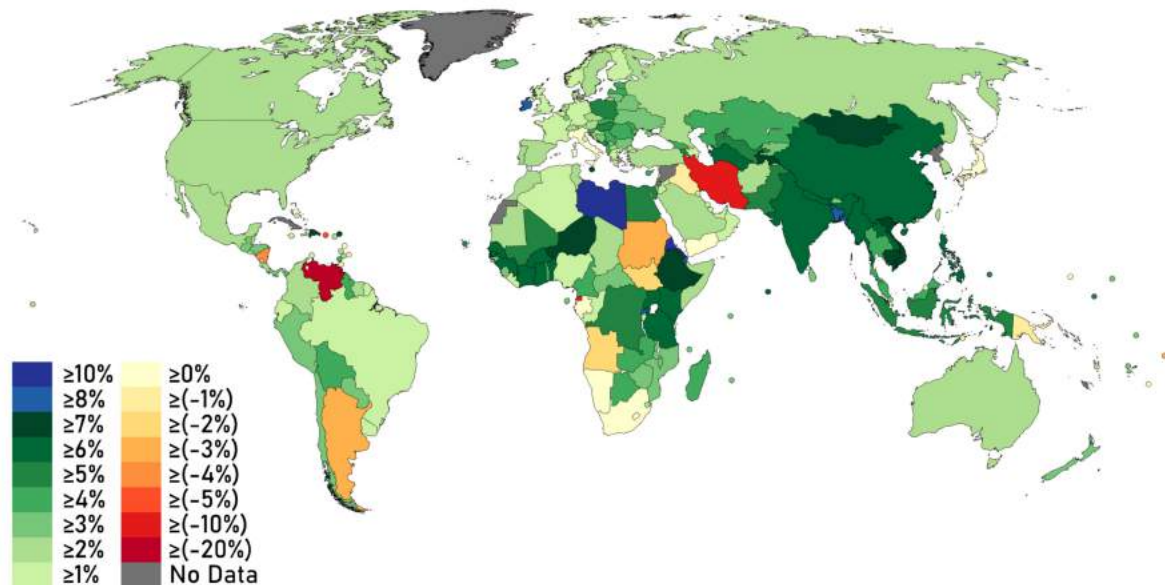

Figure 1: A depiction of countries by GDP growth rate in 2018

Generally, growth rate is measured by the percentage change of *Gross Domestic Product (GDP), which is the total value of all goods and services produced by a country.* Using data from the World Bank Development [3] and Governance Indicators [4], we attempt to conduct a similar study using Barro's criteria of economic growth. Thus, *we aim to build a linear regression model that measures the rate of GDP Growth among countries using predictors similar to Barro's, Our goal is to strike a balance between simplicity, interpretability, and accuracy.* .

Methods

Model Violations and Diagnostics

We will be using a linear regression model with p number of predictors and sample size n. Thus, the general form of our model is defined as:

$$\hat{Y}_{GDPGrowth} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon$$

The model can be broken down as follows:

- $Y_{GDPGrowth}$: The response value for our model
- $\sum_{i=1}^{n} \beta_i x_i$: The p number of predictors and their corresponding coefficients in our model
- $\hat{\beta}_0$: Our intercept value
- ϵ: Our error term

Assumptions of Linear Regression Models

Linear Regression Models require 4 assumptions to be satisfied to ensure that it makes sense to model our sample. These assumptions are as follows:

1. **Linearity**: The relationship between each of our predictors x_i and our response $\hat{Y}$ appears linear.
2. **Homoscedasticity**: All our sample observations should be spread out
3. **Independence**: All our sample observations are independent/uncorrelated with each other
4. **Normality**: All our sample observations should be normally distributed

In addition to the 4 assumptions, 2 conditions are also checked to help us understand any limitations of our model

5. **Condition 1**: The expected value of the mean of the response is a single function of a linear combination of the predictors
6. **Condition 2**: The expected value of predictor A conditional on some predictor B is a linear function of predictor A

Checking for Assumption Violations using Residual Plots

We can check for violations in each assumption using Residual plots of each predictor. Generally, we verify Conditions 1 and 2 first before we check for violations in Assumptions 1-4. If either one of the Conditions are violated and any one of the four Assumptions are violated, all we can say is that a wrong model has been fit and nothing more. However, if no Assumptions are violated, then Condition 1 is likely satisfied, and violations of Condition 2 (if there are any) will be considered a non-problematic limitation (this is not a hard rule).

When we look for assumption violations in Residuals, we look for the following patterns:

1. **Linearity**: Any systematic pattern in the residuals
2. **Homoscedasticity**: Any systematic pattern appears, especially a fanning pattern where residuals become more/less spread out
3. **Independence**: Two or more clusters of residuals that are clearly separate from each other
4. **Normality**: Usually checked with a special type of Residual plot called a QQ Plot. Violations appear when the QQ plot exhibits any non-linear abnormalities (i.e. points trailing upwards/downwards at the ends, deviations within the middle of the line)

The additional conditions are not checked using residuals, but rather with specific plots.

5. **Condition 1**: Checked by plotting the response Y against the Fitted values of our model. Violations appear when there is non-random scatter around non-linear functions (i.e. scatter around a curve, fanning out pattern).
6. **Condition 2**: Checked by plotting each variable in our dataset against every other variable. Potential violations appear if there is any distinct relationship between one predictor, or another predictor, or the response.

Fixing non-Homoscedasticity

Violations in Homoscedasticity can often be fixed using *Variance Stabilizing Transformations (VST)*. VST's are functions that can be applied to the response of a linear model, which usually "removes the dependence of error variance on the model's predictor values" [2]. In practice, this usually just means running a linear model with $log(Y)$ or $\sqrt{Y}$ as our response

Fixing non-Linearity and non-Normality

Violations in Linearity and Normality can occasionally be fixed using *Box Cox Transformations*. A Box Cox Transformation is a type of power transformation that can be performed either on the response or the predictors of a model. For most practical uses, this involves either taking the *log* of our response and predictors or raising them to a power:

Box Cox Transformation for a response Y:

$$\psi_M(Y, \lambda) = \begin{cases} Y^\lambda & \lambda \neq 0 \\ log(Y) & \lambda = 0 \end{cases}$$

Box Cox Transformation for some predictor x_i:

$$\psi_M(x_i, \lambda) = \begin{cases} x_i^\lambda & \lambda \neq 0 \\ log(x_i) & \lambda = 0 \end{cases}$$

In either case, our calculations will depend on λ, which is called the power parameter. This value is provided to us for each predictor and response in R. Depending on what power parameter a variable has, we will attempt to transform it in accordance with the above functions.

Note that power transformations are more of a band-aid for our model rather than a proper fix. Too many power transforms can make a model un-interpretable to the average policy-maker, hence we will only use it for serious violations.

Model Selection

Adjusted Coefficient of Determination

The Adjusted Coefficient of Determination (R^2_{adj}) is often used as a way to see whether models with different numbers of predictors can explain variation in the sample better than others [2]. It is formally represented by the equation below:

$$R^2_{adj} = 1 - \frac{\left(\frac{RSS}{n-p-1}\right)}{\left(\frac{SST}{n-1}\right)}$$

The equation can be broken down as follows:

- **RSS**: The Residual Sum of Squares for our model. This is the amount of variation left over after fitting our regression model

- **SST**: The Sum of Squares Total is the original amount of variation present in the sample.
- **n**: Sample size of our population
- **p**: Number of predictors in our model

When comparing models, a higher R^2_{adj} is generally better. Notice that as if p increases and there is no substantial change in RSS, then R^2_{adj} can go down for each additional predictor. This makes the metric very useful for us in comparing simple and complex models, as it penalizes models that have too many predictors given the amount of variations it explains.

Bayesian Information Criterion (BIC)

The BIC is a model selection criteria that attempts to balance the "goodness of fit" of a model with a penalty term for increased model complexity [2]. We will be using the BIC alongside automated model selection to find our final model.

All Possible Subsets Selection

All Possible Subsets Selection (APSS) is a type of automated selection process that helps decide the best predictors for a model [2]. For each pth predictor in a model, APSS attempts to find the "best" model containing p predictors. The "best" model is the one that has the lowest BIC and highest R^2_{adj} for each p number of predictors model.

Variance Inflation Factor (VIF)

The VIF is a measures "how inflated the standard errors/variance can be due to the relationship of one predictor to the others" [2]. This helps us see whether any of our predictors have multicollinearity. For our purposes, a $VIF \geq 5$ will indicate a sign of severe multicollinearity. We can solve this by respecifying what our model should be.

Techniques to find Influential Observations

There are three main measures that help us classify influential observations:

1. **Cook's Distance**: Used to measure the influence of a single observation on an entire regression line
2. **DFFITS**: Measures the effect of an observation on its own fitted value for every observation in the data [2]
3. **DFBETAS**: Measure's the effect that some observation i has on the estimation of it's regression coefficient $\hat{\beta}_i$. A large DFBETAS means an observation has considerable influence on $\hat{\beta}_i$, but not necessarily on other coefficients [2]

Generally, an observation will be considered influential if:

1. **Cook's Distance**: $Cook's \geq 50th$ percentile of $F(p+1, n-p-1)$
2. **DFFITS**: $|DFFITS| > 2\sqrt{\frac{p+1}{n}}$
3. **DFBETAS**: $|DFBETAS| > \frac{2}{\sqrt{n}}$

where p is the number of predictors in our model and n is the sample size

Model Validation

Our dataset has a total of 88 observations, which we will split into two datasets of 44 observations each: a *training dataset* and a *test dataset*. To validate our model, we have to first conduct all of our model selection and diagnostics work on the training dataset. Once a final model has been determined, we will run this model using the test dataset, and our goal is for the resulting model to "look similar" to the one run on the training dataset.

By "look similar", we mean that the final model on both the training and test datasets must have:

- Minimal differences in estimated regression coefficients, as this shows us that we're estimating similar relationships between training and test sets [2]
- The same predictors are statistically significant between both models ($\alpha = 0.05$)
- No new and worse model violations, which shows that the model won't provide good predictions on different datasets [2]
- Similar R^2_{adj} scores, which shows us that our model has good and consistent performance in explaining variation in the data [2]

Results

Data

We will be using 6 predictors and 1 response in total for this dataset, with 3 predictors from the World Bank Development Indicators (WDI), and 3 from the World Bank Governance Indicators (WGI). To avoid the collection of time series data, we will focus on GDP Growth Rate of all countries in the year 2018. This year was not arbitrarily chosen; 2018 happens to be the most recent year which contained the most amount of data among both datasets.

Table 1: Description of the Important Variables in WGI and WDI

Variable	Description
Government Expenses (% of GDP)	The amount of money the state expends in maintaining government operations in goods and services. This includes (but is not limited to) wages for public sector employees, social and medical benefits, subsidies, and grants.
Life Expectancy (in years)	The average life expectancy of a citizen in the country.
Secondary Education (in years)	The proportion of the total population that have attained a secondary education
Political Stability	An estimate of the likelihood of political instability and politically motivated violence. Values range from -2.5 to 2.5, with the former indicating complete instability and the latter indicating complete stability in governance
Rule of Law	An estimate of the perceptions that citizens have about the laws of their country and their willingness to abide by them. Values range from -2.5 to 2.5, with the former indicating no respect for the law and the latter indicating complete obedience to the law
Government Effectiveness	An estimate of the perceptions that citizens have about the quality of civil and public service, the quality of policy formulations and implementations, and the governments commitments to such policies

Visualizations of Variables

Plotting the scatterplot of all our variables results in Figure 2, which is shown in the following page.

While our response seems normally distributed, Figure 2 suggests to us that a lot of our predictors do not have a strong linear relationship with our response. Individually, there seems to be a lot of random scatter around the center of the plots, with the exception of Government Expenses. In particular, Government Expenses may present itself as a violation of Linearity and Homoscedasticity.

Verifying Assumptions

Our Residuals vs. Fitted plot (see Appendix 2a) also exhibits random scatter around a linear function, which suggests to us that Condition 1 is satisfied. However, when we plot every predictor in our dataset against the other (see Appendix 1a), aside from the categorical variables, a very strong linear relationship appears
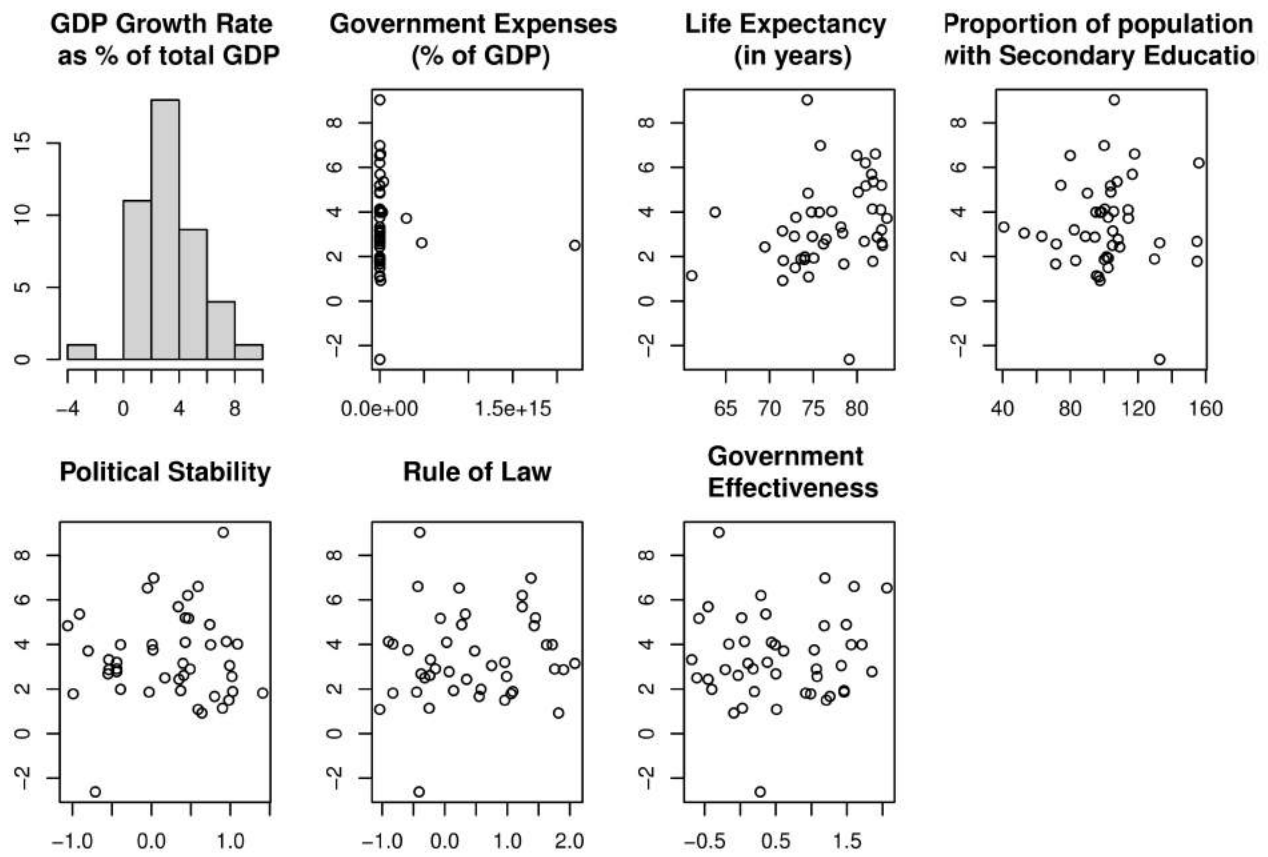
Figure 2: Scatterplot of every predictor plotted against the response

between Government Expenses and every other predictor, again suggesting a violation of Linearity

Indeed, our Residuals (see Appendix 3a) show us that while most predictors seem to have random scatter, the residuals of Government Expenses seem to be heavily influenced by a handful of observations. Given that Government Expenses is numerical, it would seem like it would benefit from a Box Cox Transform.

After applying the power transform on Government Expenses, we verify the residuals (Appendix 3b), Fitted plot (Appendix 2b), and Pairs plots (Appendix 1b) again. These results suggest to us that our model with a transformed Government Expenses predictor verifies the assumptions.

Checking for multicollinearity, we see that all our predictors have a VIF score below 2. This is a good sign for us, hence we can move on

Identifying and addressing Influential Observations

We will utilize the tools mentioned in our Methods section to see whether identify any leverage points that are influential on our model. After constructing a model containing our 6 main predictors, Figure 3 represents the results of running Cook's Distance, DFFITS, and DFBETAS on our model.

We can see that observation 18, 24, and 40 appear in at least 3 of the 8 plots, with the rest of the observations appearing only once. While not a hard rule, this is convincing evidence for us to remove them.
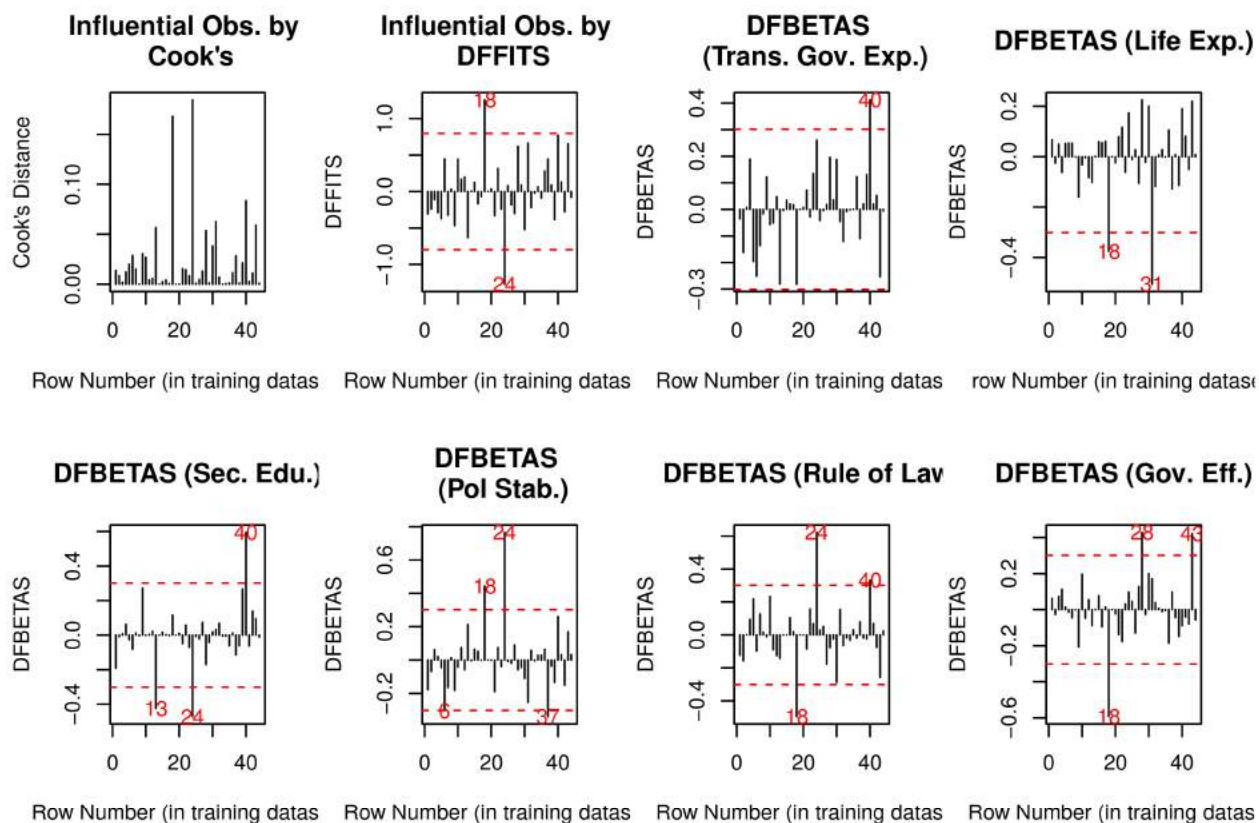

Figure 3: Influential observations by Cook's Distance, DFFITS and DFBETAS

Final Model and Model Validation

Once the influential observations are removed, we will run All Possible Subsets Selection on the aforementioned model. Among all possible subsets for each fixed number of predictors, the *two predictor model containing Life Expectancy and Government Effectiveness* has the highest R^2_{adj}. However, the BIC suggest to us that the one predictor model containing Life Expectancy is the best one. Due to the oversimplicity of the one predictor model, we've elected to keep the two predictor model instead.

The Pairs plots, Fitted plot, and Residuals plots can be found in Appendix 1c, 2c, and 3c respectively. Our final model does not seem to violate any assumptions.

To validate the model, we run through the same process as we did before, but for the test dataset. The results of this process, in comparison with the model built off our training dataset, is summarized in the table below:

Table 2: Summary table of our models

Characteristic	First Model (Train)	First Model (Test)	Final Model (Train)	Final Model (Test)
Largest VIF value	1.1446223	1.1583473	1.0147182	1.0139333
# Cook's D	0	0	0	0
# DFFITS	2	4	4	3
Adj. R^2	-0.030	-0.005	0.106	0.034
Violations	slight linearity	slight linearity	none	none
Intercept	-6.646 ± 10.862	-0.338 ± 6.902	-6.739 ± 6.748	0.396 ± 6.606
Gov. Exp.	3.621 ± 48.094	-8.232 ± 47.942	-	-
Life Exp.	0.13 ± 0.13 (.)	0.045 ± 0.092	0.128 ± 0.086 (*)	0.038 ± 0.086
Sec. Edu.	-0.005 ± 0.028	0.007 ± 0.022	-	-
Pol. Stab.	0.429 ± 1.008	-0.483 ± 0.67	-	-
Rule of Law	0.335 ± 0.726	-0.135 ± 0.61	-	-
Gov. Eff.	0.19 ± 0.838	0.498 ± 0.626	0.42 ± 0.59	0.519 ± 0.596 (.)

In short, it would seem that neither of our models look completely similar to how it performed with the training dataset. While there are minimal distances within the estimated regression coefficients, we encounter fairly different R^2_{adj} scores, in particular between the Final Model Training with Final Model Test.

Discussion and Limitations

There are several key takeaways from this report. If we interpret the results of the final model, we will see that for every additional percent in GDP Growth Rate, Average Life Expectancy is expected to increase by 0.038-0.128 years, or 4-6 weeks. Similarly, we also see a positive increase in Government Efficiency, indicating that a government that drafts and enforces effective policies will usually see enhanced economic growth. In general, there is a positive correlation between GDP Growth Rate, Life Expectancy, and Government Efficiency. Our model is also incredibly simple and interpretable, and policymakers will have no trouble understanding what they need to do to achieve their growth goals.

What we sacrificed in return, was model accuracy and legitimacy. For one, we were nowhere near the recommended amount of 120 observations, as only 88 countries/observations had all the relevant data for our analysis. Because of this, we did not have the freedom to focus on a particular region. It is important to keep in mind that there are a million and one factors that lead to economic growth for every single country in the world. Thus, it would make sense why there was so much scatter in Figure 2: there is just too much diversity in what causes economic growth. Beyond that, our model is both invalid and also just too simple, so much so that it sacrifices accuracy as demonstrated by our R^2_{adj} score. The varying R^2_{adj} scores between Final Model Training and Final Model Test suggests to us that the model is not consistent in its predictions, hence its
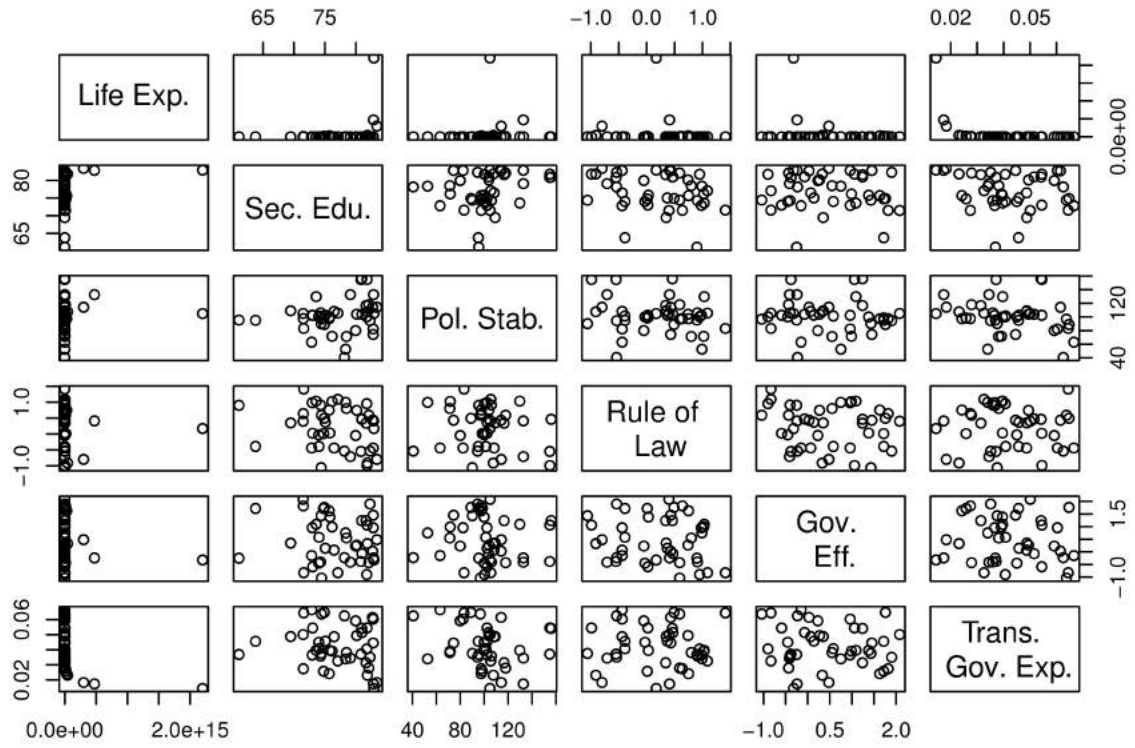
results are likely not valid.

But this is also because we built the model off the problematic data to begin with. In short, we were limited by the lack of data and did not have a strong enough focus on what region of the world we were interested in (due to not having region-specific data as well).
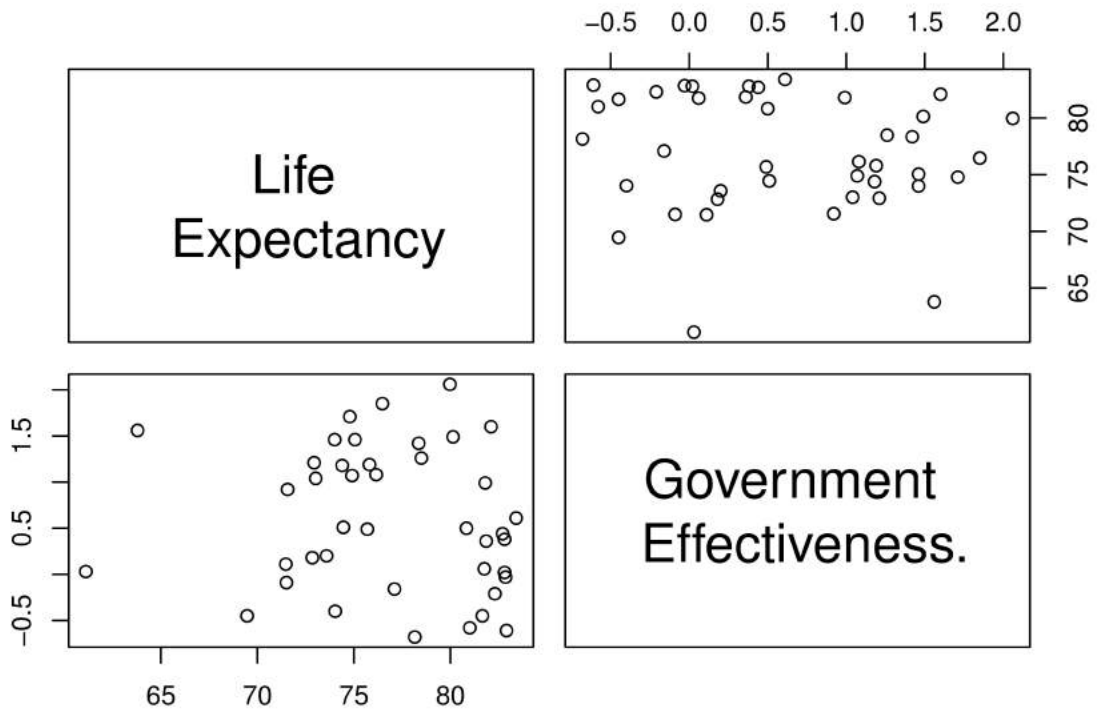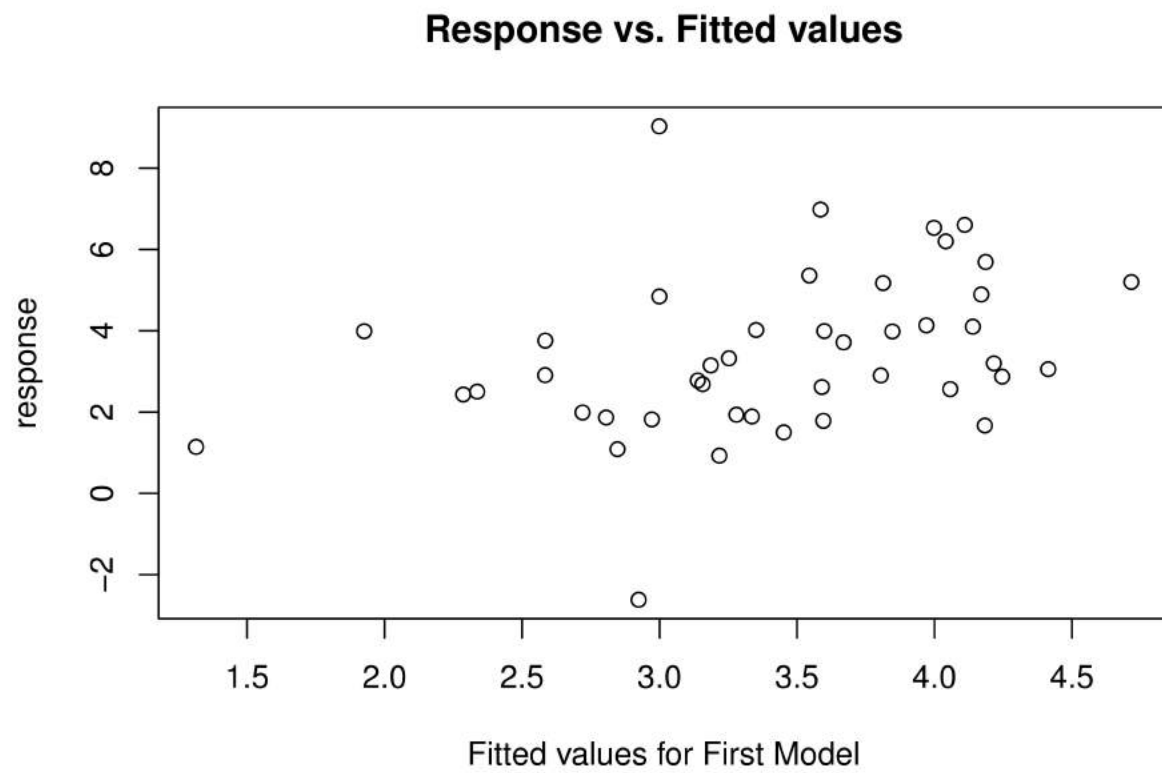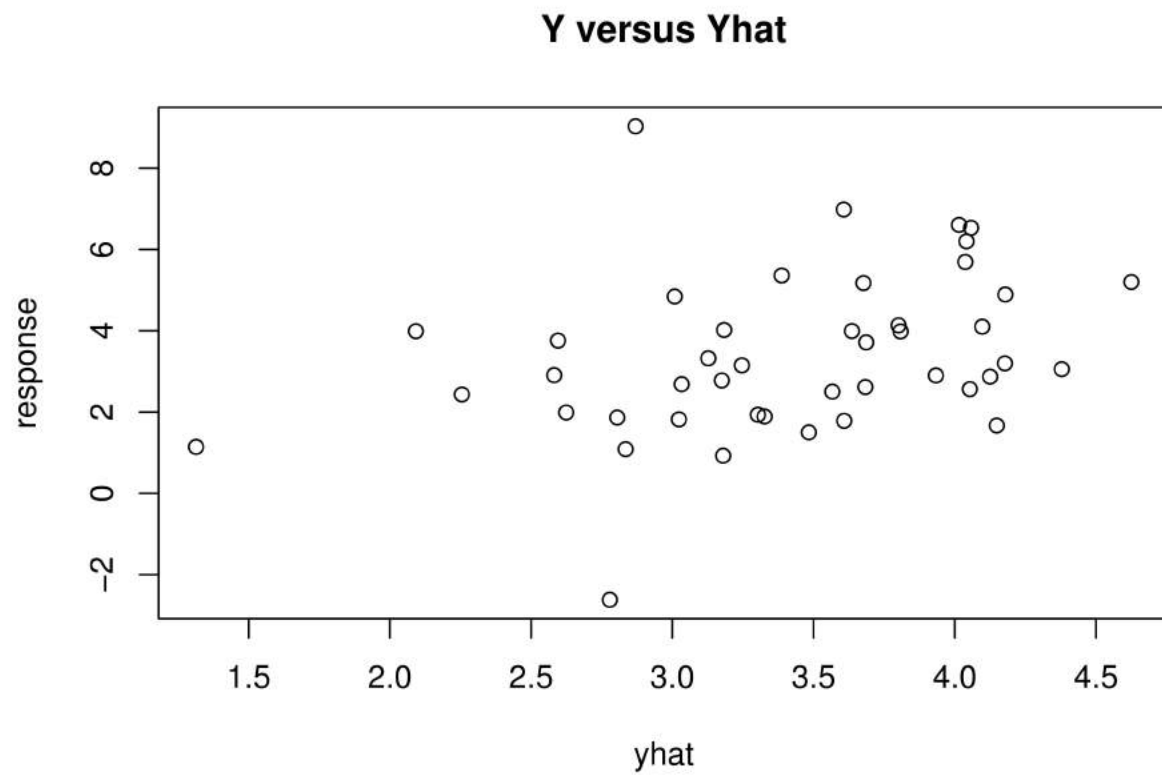
Appendix

Appendix 1a:

Appendix 1b:

Response vs. Fitted values

Fitted values for First Model

Y versus Yhat

Y versus Yhat

Appendix 3a:

Appendix 3b

Appendix 3c:

Life Avg.   **Gov. Eff.**   **Normal Q–Q Plot**
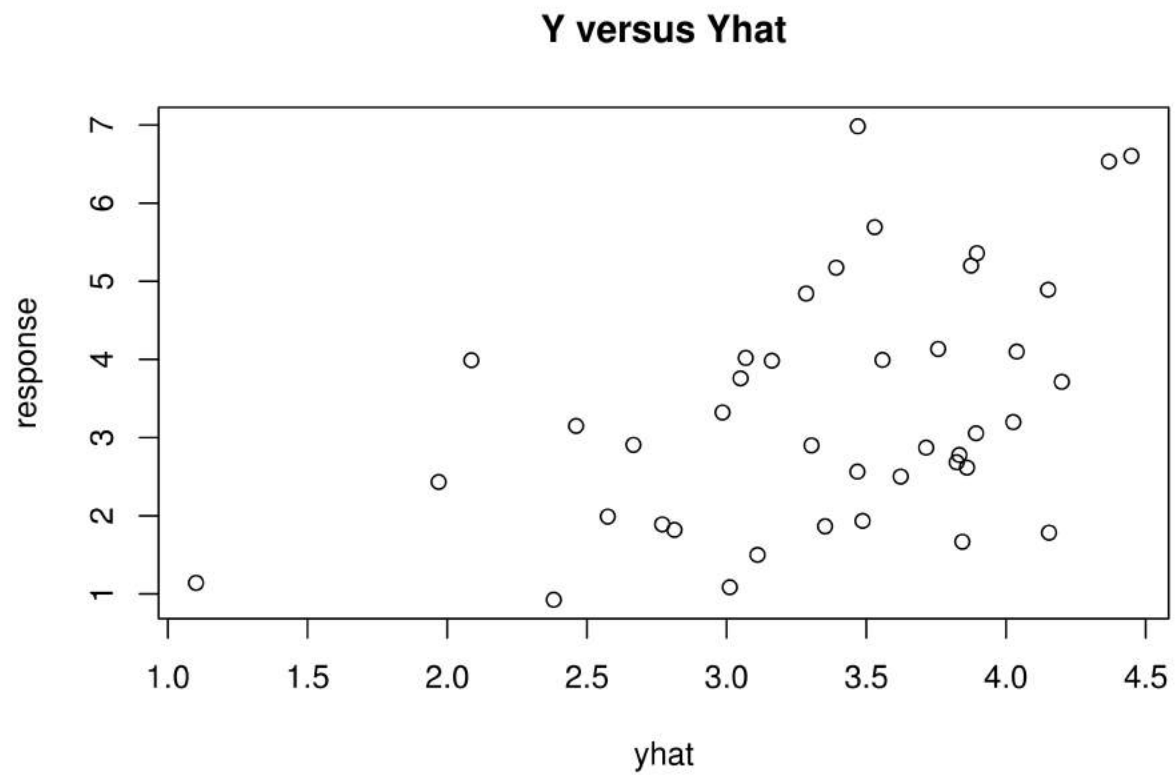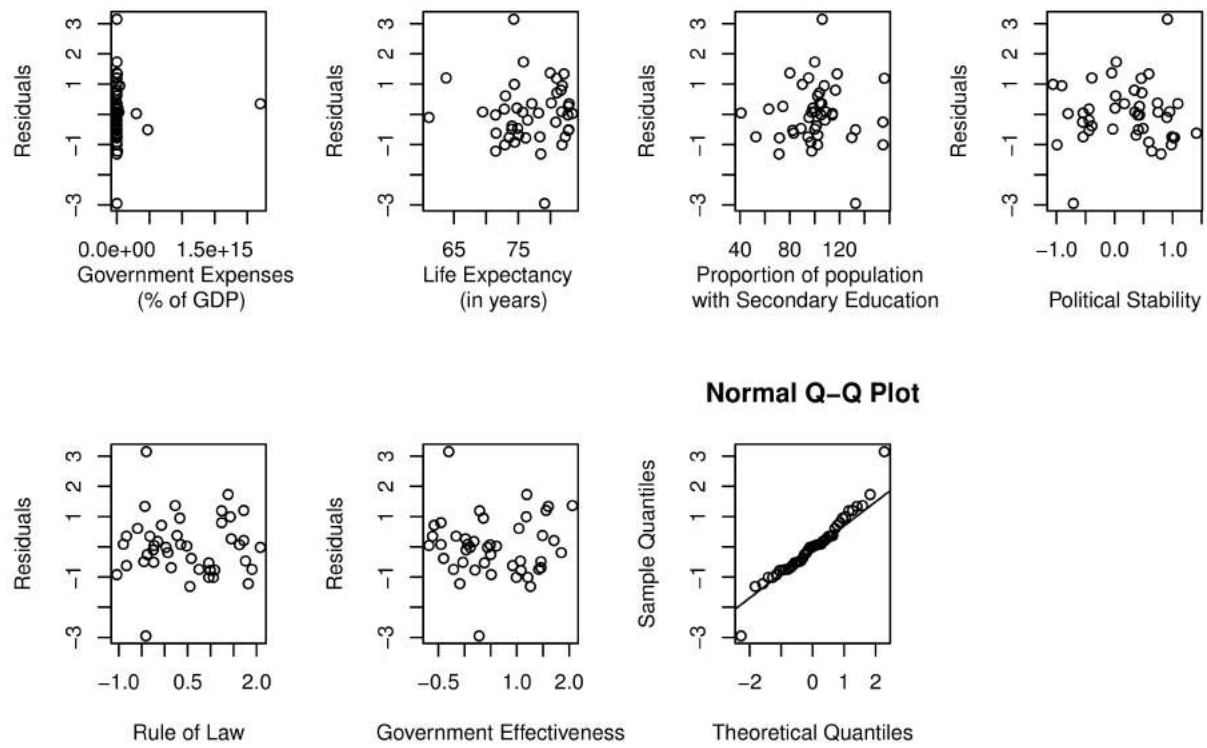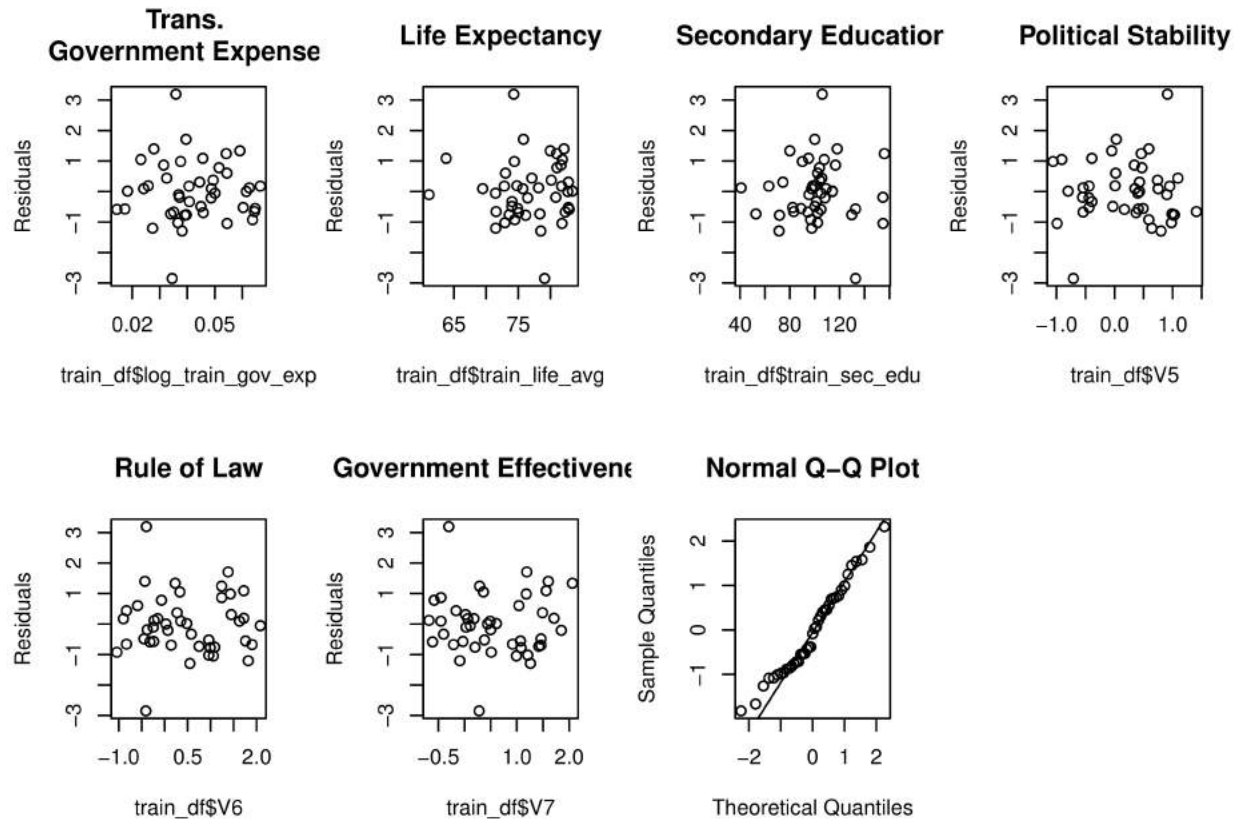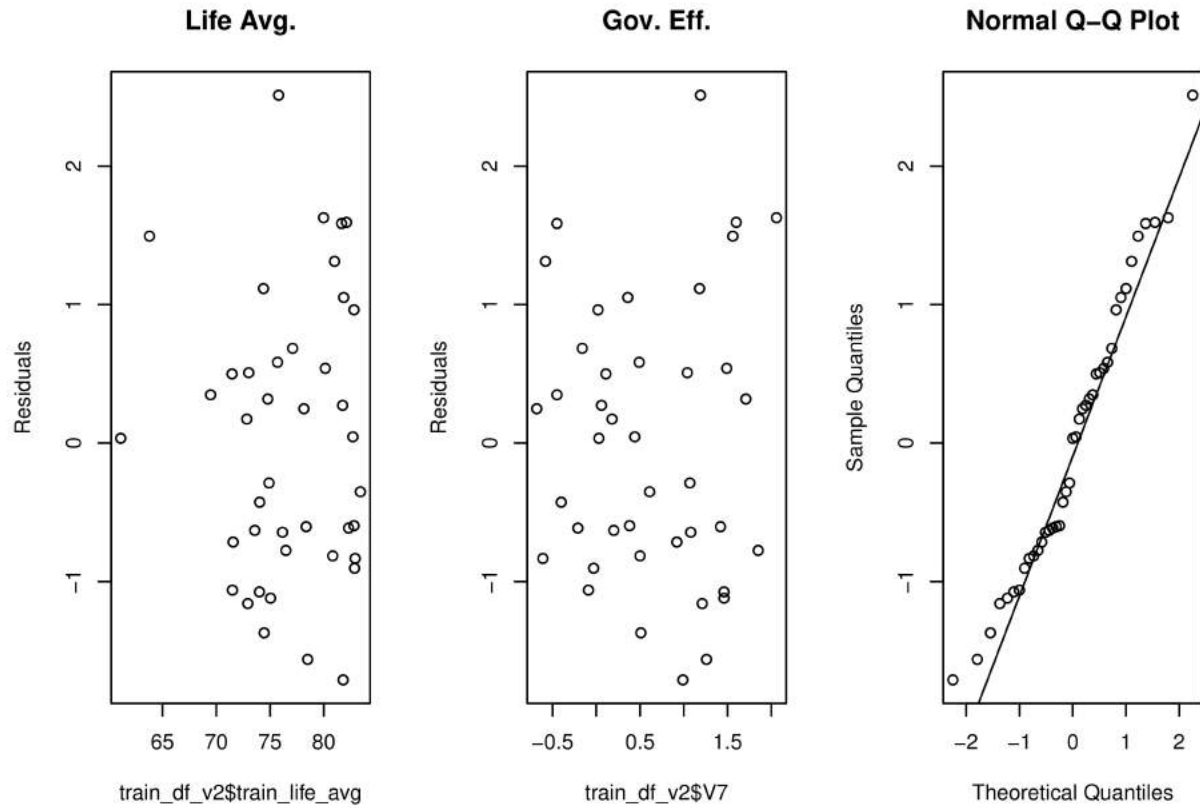
References

[1] Barro, Robert J. *Determinants of economic growth: A cross-country empirical study.* (1996). Retrieved April 15, 2022.

[2] Daignault, Katherine. *STA302 Lectures* . Retrieved April 20, 2022.

Association of Tennis Professionals, *John Isner: Overview*, ATP World Tour, Retrieved December 19, 2021, from https://www.atptour.com/en/players/john-isner/i186/overview.

[3] World Bank Group. *World Bank Development Indicators.* Retrieved April 16, 2022, from [https://datatopics.worldbank.org/world-development-indicators/] (https://datatopics.worldbank.org/world-development-indicators/).

[4] World Bank Group. *World Bank Governance Indicators.* Retrieved April 16, 2022, from http://info.world bank.org/governance/wgi/.