Analyzing Global Energy Consumption: Trends, Economic Factors, and Statistical Insights

Alex Marcek, Ahmad Khan & Tara Draper

2025-03-27

Introduction

For our project, we examined a data set containing information about the energy consumption of 294 different countries. This data set contains many variables that tell us about the types of energy consumption, as well as variables that provide insights into trends in energy consumption for a particular country.

This data set is particularly interesting to examine because we can learn a lot about trends in energy consumption over time for particular countries, or on a global scale. We can also examine the relationships between variables, such as whether trends in GDP or population can account for trends in specific types of energy consumption.

All of these countries have a date range for which information about energy consumption is recorded, the maximum of which spans 124 years. A few countries have this amount of data, while the minimum date range is 11 years. The overall shape of this raw data frame is 21,812 observations of 129 variables. As we can see, we have a relatively lengthy data frame with many variables. Furthermore, many columns (relative to a specific country) are completely underpopulated with data, as some countries do not collect data on specific types of energy consumption.

Objectives

The sheer size of this dataset required special attention to which variables are useful to us. We will also need to pay special attention to missing data. If we were to simply remove all NA values, we would be left with around 150 rows left. This is a data loss of around 99.3% which is unacceptable for our purposes. Our first objective is to clean the data, or come up with a repeatable method to clean our data to fit our needs. We also aim to run exploratory data analysis and fit some regression models in order to answer the research questions detailed below. To aid in our inference, we will also be constructing confidence intervals and detailed graphs.

Research Questions

We considered five key research questions during our analysis:

1. How has global energy consumption changed over time, and what factors have contributed to these changes?
2. Is there a relationship between a country's population and its energy consumption patterns?
3. Does higher GDP always mean higher energy consumption?
4. Do countries that exhibit rapid population growth also show significant increases in overall energy demand?
5. How have primary energy consumption trends changed over time for top GDP countries?

Methods

Data Cleaning

The source data set exhibited significant sparsity due to incomplete data recorded by certain countries or within specific date ranges. Some columns contained only NA values for specific countries, while entire rows consisted of NA values for certain date ranges. For example, the table below highlights the number of missing values in each variable we deemed potentially relevant to our research questions.

Column Name	Number of Missing Values
fossil_fuel_consumption	16754
gas_consumption	16499
coal_consumption	16293
renewables_electricity	13703
solar_electricity	13610
wind_electricity	13578
hydro_electricity	12775
gdp	10037
primary_energy_consumption	9634
iso_code	5000
population	3365
country	0
year	0

To address issues with data loss, we created filter functions that allow users to specify which countries to include in their analysis. These filters automatically remove rows with NA values and exclude columns that contain no data for the selected countries. This approach makes the regression analysis more manageable and ensures the data remains as complete as possible.

Exploratory Data Analysis (EDA)

In the EDA phase, we used grouping and filtering methods to learn more about the dataset. Visualizations were constructed to reveal trends in the data and specific variables. Log transformations were particularly useful for addressing heavy skews in variable distributions and improving linearity.

We also answered two of our research questions through higher-level EDA and constructed visualizations to understand time trends in energy consumption.

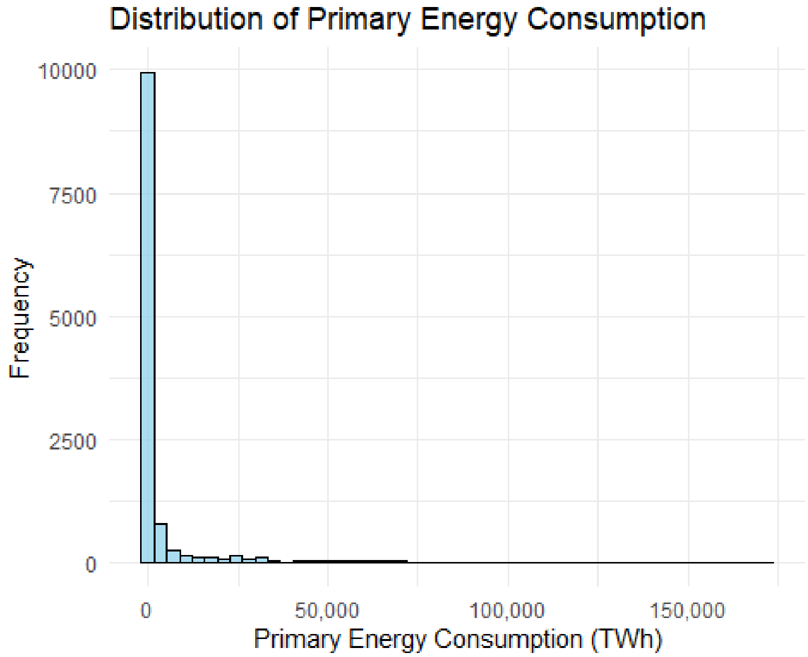
High-Level Summary of Dataset

Summary of the Dataset

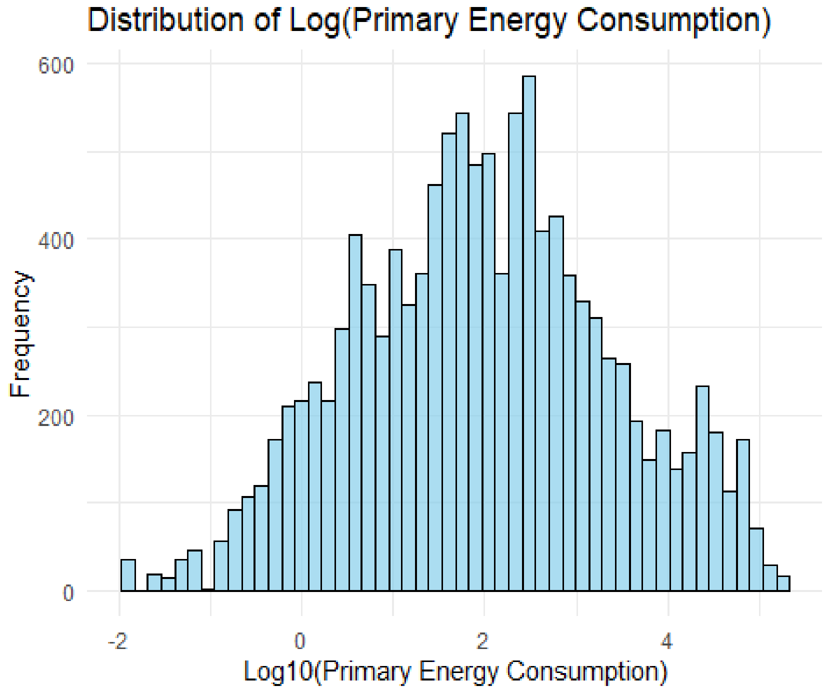
	V1
rows	21812
columns	129
discrete_columns	2
continuous_columns	127
all_missing_columns	0
total_missing_values	1892712
complete_rows	138
total_observations	2813748
memory_usage	22571640

Exploring Distributions

We plot the distribution of differnt variables to underdstand the data like primary energy consumption below.

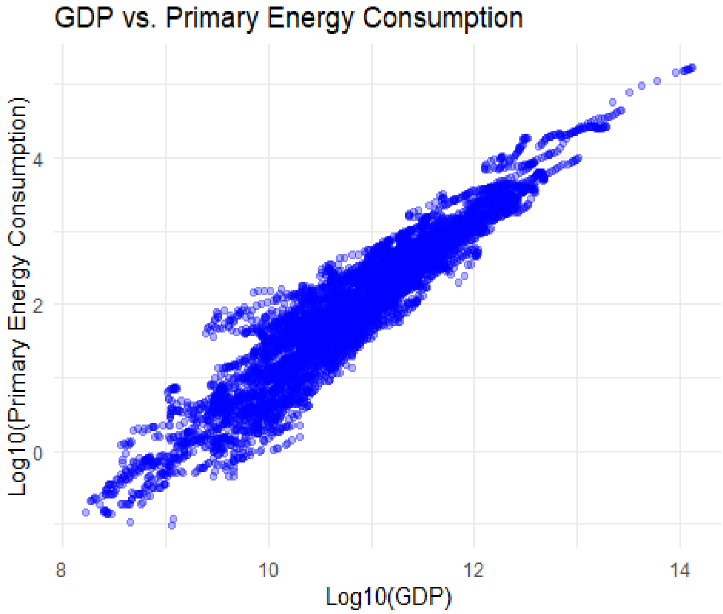


However, the scale is not optimal so a better way would be to use log-scale



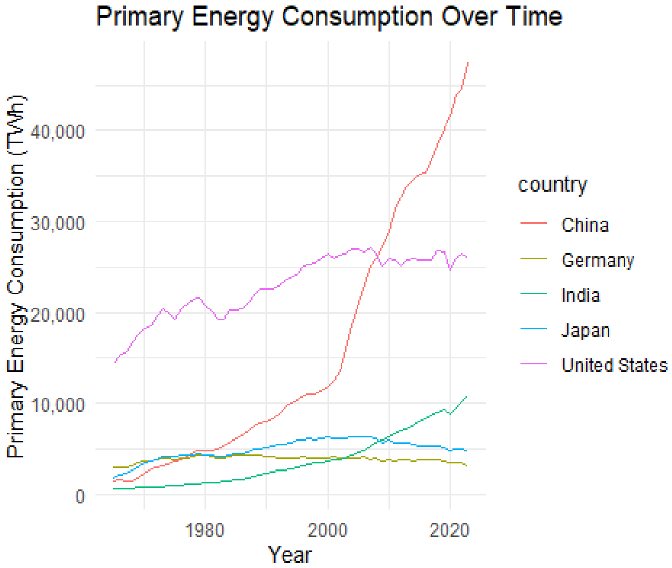
Checking Correlations

We get an idea about the data if we check the correlation between different variables like GDP and Primary Energy Consumption (Again we use a log-transformed graph since that's better suited).

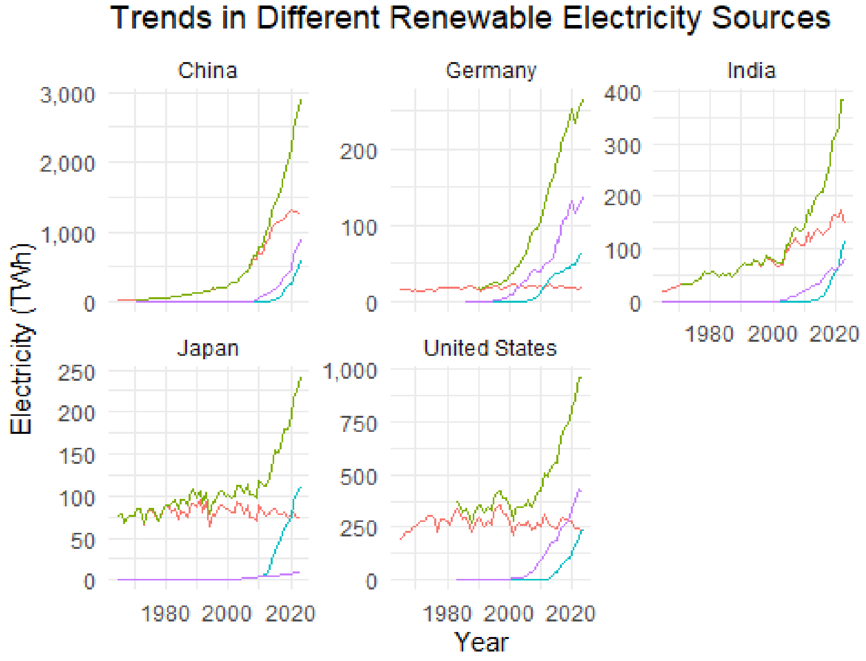


Time Trends

We look at how primary energy consumption changes over time for the top gdp countries in the world. This help us answer one of our research questions RQ1: *How has global energy consumption changed over time, and what factors have contributed to these changes?*



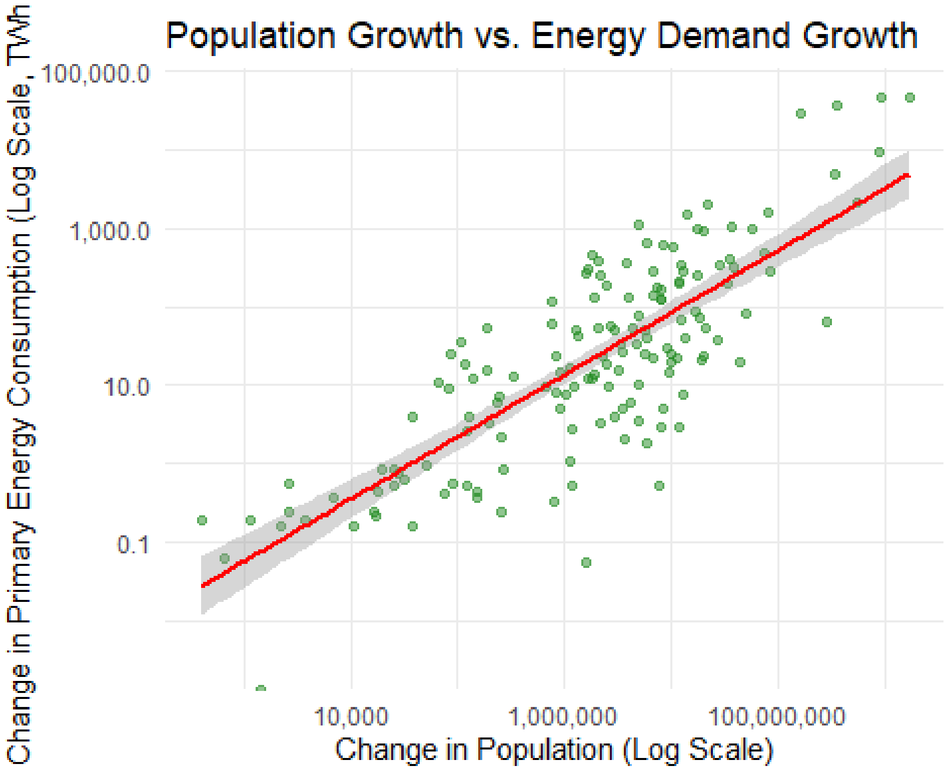
We can also explore Renewable (e.g., wind, solar, hydro) Over Time for each of the selected country, answering question 5, RQ5: *How have primary energy consumption trends changed over time for top GDP countries?*



Time Trends 2

To find answer one of our research question RQ4: *Do countries that exhibit rapid population growth also show significant increases in overall energy demand?* we use the following EDA metho: 1. The dataset is filtered for the years 2000 and 2020, selecting relevant columns and removing rows with missing population or energy consumption data. 2. The data is reshaped into a wide format to compare values between 2000 and 2020, and changes in population and energy consumption are calculated. 3. A correlation test is performed to assess the relationship between population growth and energy consumption growth. 4. Scatter plots with regression lines (linear and log-transformed scales) are created to visualize the relationship and trends.

```
##  Pearson's product-moment correlation
## 
## data:  energy_growth_wide$pop_change and energy_growth_wide$energy_change
## t = 19.894, df = 222, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7477820 0.8430391
## sample estimates:
##       cor 
## 0.8004077
```



Regression Analysis

To address the research questions, we fitted linear regression models to analyze the relationships between variables such as GDP, population, and energy consumption. The log transformations applied during EDA were also incorporated into the regression to mitigate skewness and improve the model fit.

SIMPLE LINEAR REGRESSION CASE

We first try to model Primary Energy Consumption against GDP. These are the results of the SLR.

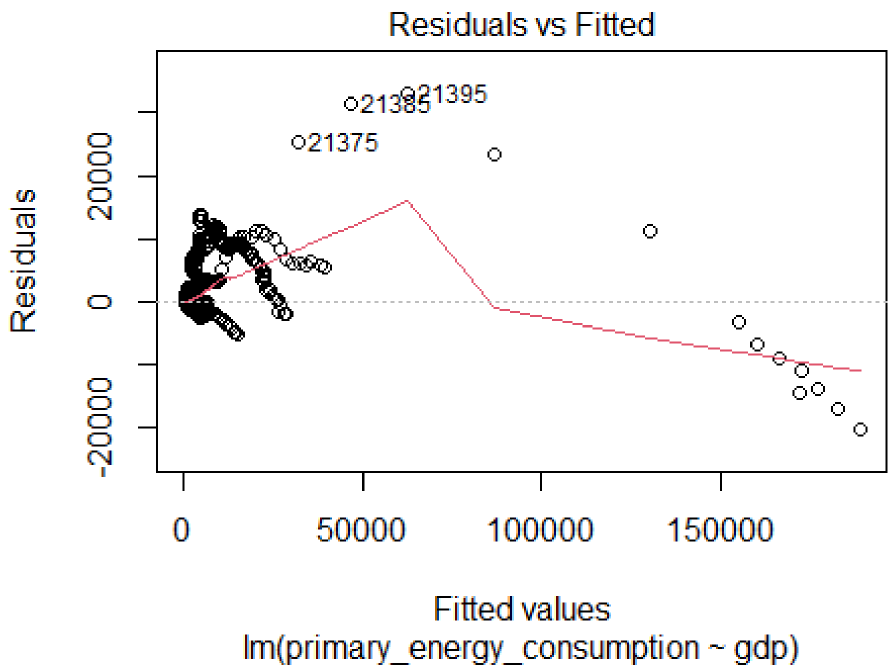
Coefficient Estimates

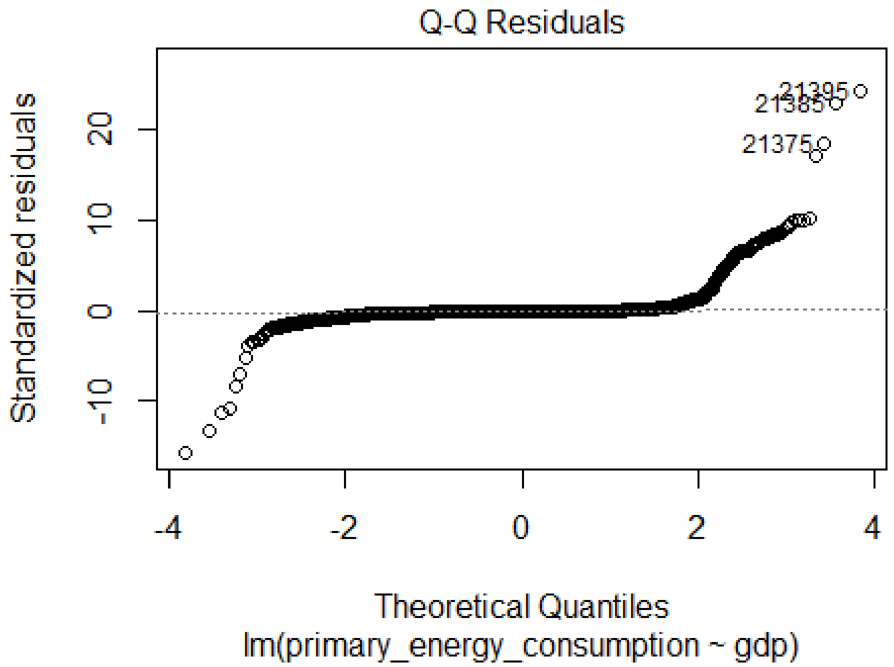
term	estimate	std.error	statistic	p.value
(Intercept)	160.657769	15.616031	10.288003	< 2e-16
gdp	0.000000	0.000000	402.678110	< 2e-16

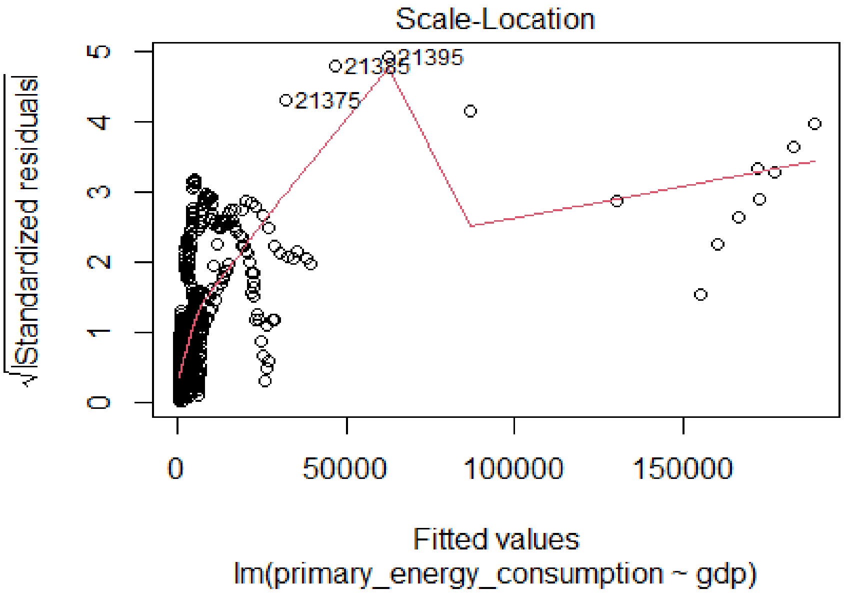
Model Statistics

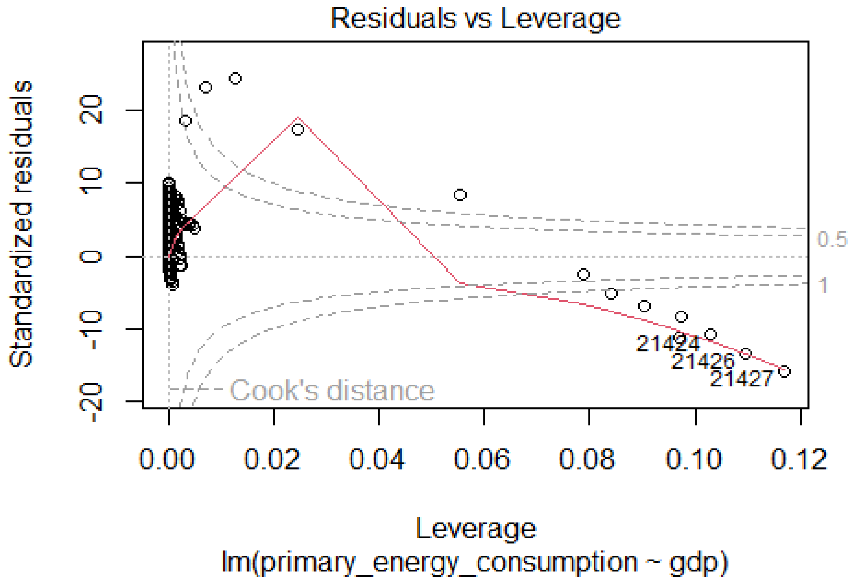
r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual
0.9542	0.9542	1,365.55	162,149.66	< 2e-16	1	7,787

However, this doesn’t seem right perticulary the high R-Squared value. Maybe its not suitable for a SLR. We can check this with the following diagnostics.



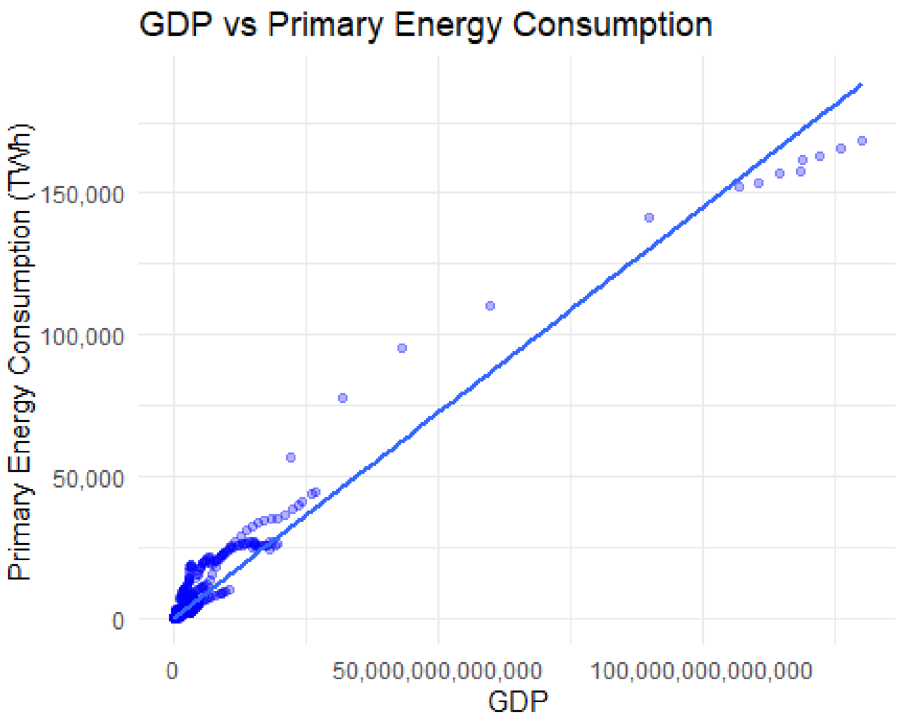




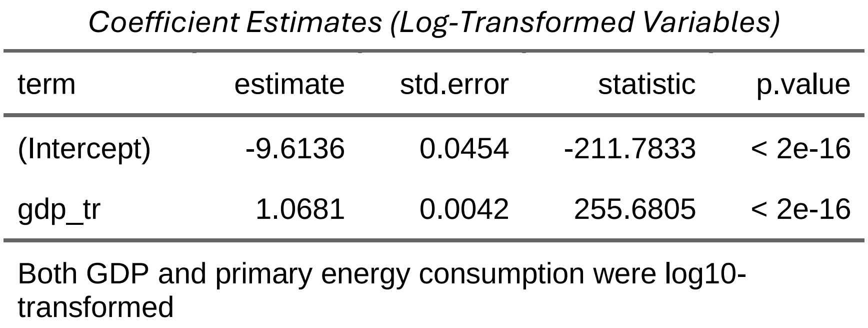


As we can see from the diagnostics run on the first linear model, there are some clear issues with our residual vs fitted plot. We can see that our data plots are clustered on one end of the scale. This is an indicator that our scale may be inappropriate, and we may need to transform the data. There are also clear issues with our qqplot, but we will address the scaling issue first.

First, lets take a look at what our un-transformed regression line looks like on the plot



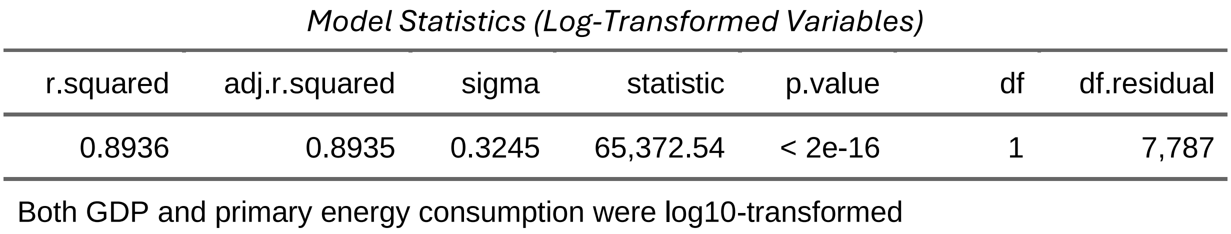
The clustering is causing clear influential points to our model, which may cause it to be much more inaccurate for lower GPDs. Lets fix the scale now and check the regression line. These are the reuslts of SLR on tranmsformed model.



Coefficient Estimates (Log-Transformed Variables)

term	estimate	std.error	statistic	p.value
(Intercept)	-9.6136	0.0454	-211.7833	< 2e-16
gdp_tr	1.0681	0.0042	255.6805	< 2e-16

Both GDP and primary energy consumption were log10-transformed

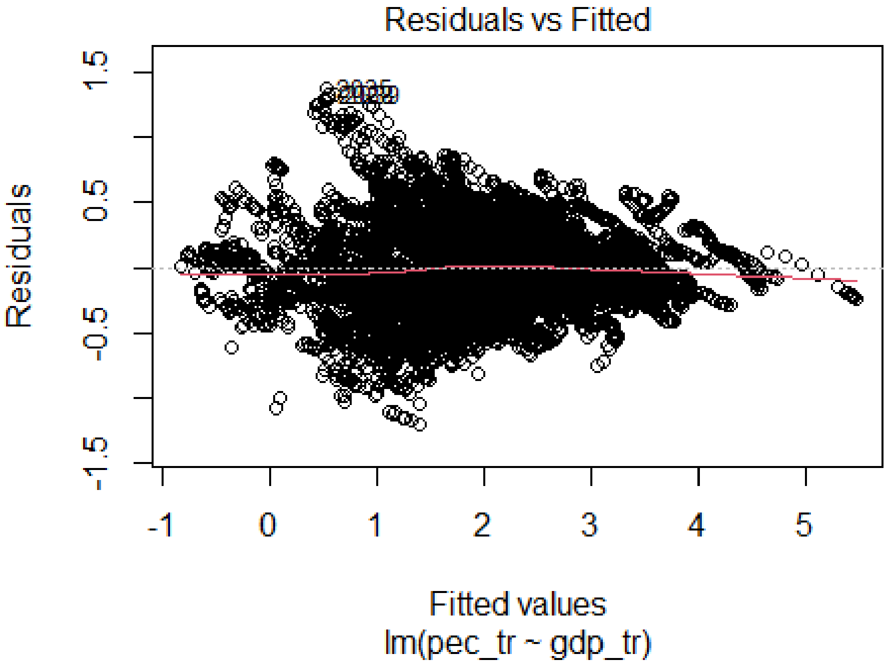


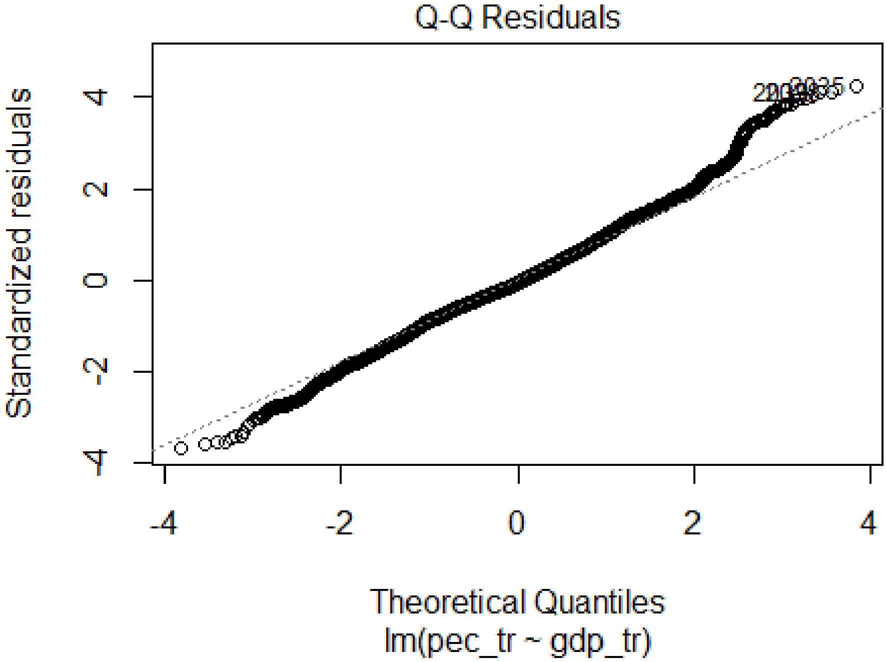
Model Statistics (Log-Transformed Variables)

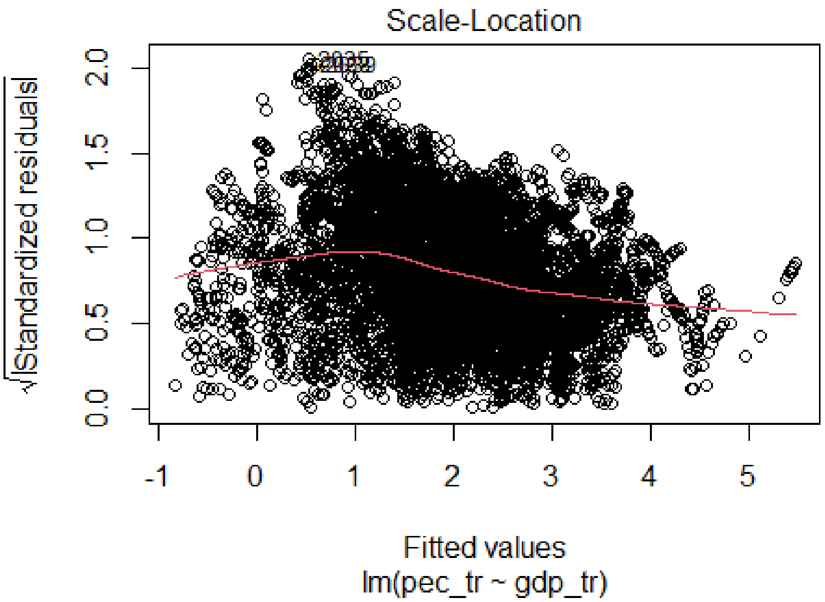
r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual
0.8936	0.8935	0.3245	65,372.54	< 2e-16	1	7,787

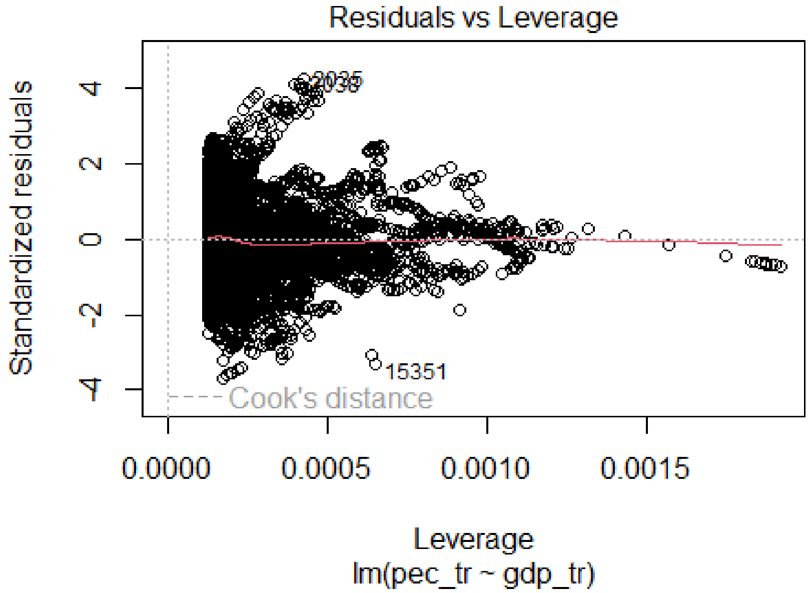
Both GDP and primary energy consumption were log10-transformed

We can check the diagnostics on this new model too

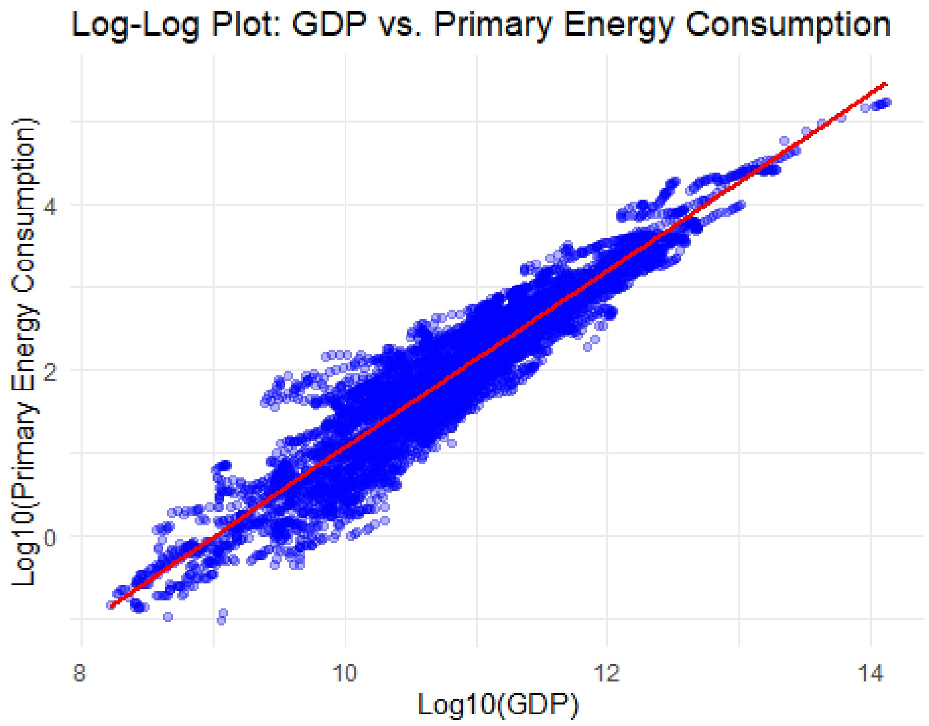








The residual vs fitted plot looks much better, but there may still be some issues with non-constant variance. *This leads to the limitation of our model: it will be worse at predicting the primary energy consumption for lower GDP countries*. Our qqplot also looks much better better than it did before. Our line is no longer shallow and our tails look good.



Confidence interval for true mean of primary energy consumption

We calculated confidence interval for true mean of primary energy consumption

Confidence Intervals of Regression Coefficients

	Coefficient	Lower_Bound	Upper_Bound
(Intercept)	(Intercept)	-9.7025	-9.5246
gdp_tr	gdp_tr	1.0599	1.0763

Cohen’s d Effect sizes

To answer the third question posed to us, RQ3: *Does higher GDP always mean higher energy consumption?* we took a look at the effect sizes of the GDP and energy consumption data for two high GDP countries.

Cohen’s d Effect Size for GDP Comparison

Measure	Value
Cohen’s d	-0.1281
Effect Size Magnitude	1.0000
Lower Bound (95% CI)	-0.4055
Upper Bound (95% CI)	0.1493

Now, looking at energy consumption

Cohen’s d Effect Size for Primary Energy Consumption Comparison

Measure	Value
Cohen’s d	-0.7217
Effect Size Magnitude	3.0000
Lower Bound (95% CI)	-1.0981
Upper Bound (95% CI)	-0.3454

As we can see from the output, there is a negligible effect size between the two countries when it comes to GDP, there is a medium effect size when it comes to the energy consumption of the two countries. This means that even though the difference in GDP between the two groups is negligible, the difference in energy consumption is not!

MULTIPLE LINEAR REGRESSION CASE

To answer the research question, RQ2: *Is there a relationship between a country’s economic population and its energy consumption patterns?* We fit a multiple linear regression model using multiple energy consumption methods to try to predict population. These are the reuslts

Coefficient Estimates

term	estimate	std.error	statistic	p.value
(Intercept)	28,845,849.7615	6,863,631.4456	4.2027	2.7e-05
primary_energy_consumption	-291,283.7785	17,382.2344	-16.7576	< 2e-16
fossil_fuel_consumption	325,526.0986	19,283.0396	16.8815	< 2e-16
coal_consumption	65,759.8515	6,874.9006	9.5652	< 2e-16
gas_consumption	55,890.2608	12,717.8436	4.3946	1.1e-05
renewables_electricity	3,365,541.8447	614,203.7394	5.4795	4.6e-08
solar_electricity	378,885.7274	618,790.3460	0.6123	5.4e-01
wind_electricity	-5,804,150.1057	855,827.3784	-6.7819	1.4e-11
hydro_electricity	-2,504,759.7947	611,711.1918	-4.0947	4.3e-05

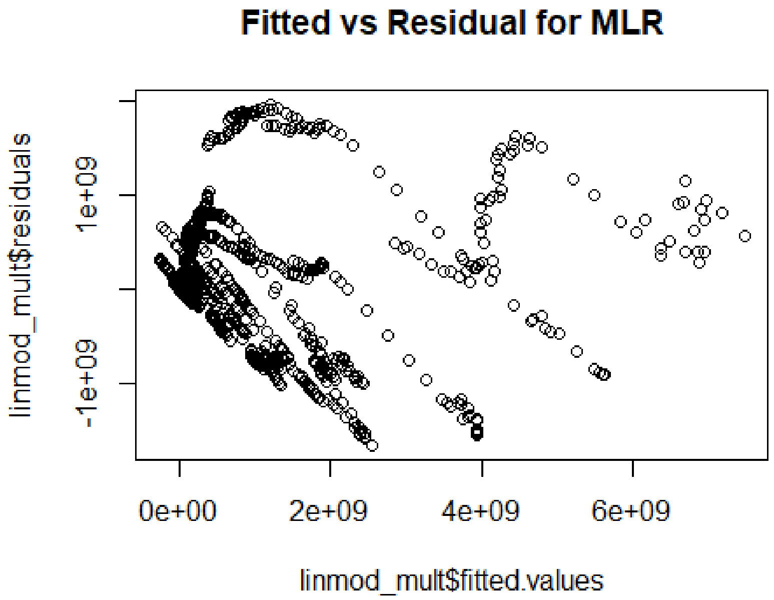
The dependent variable is population, and primary energy and other variables are predictors.

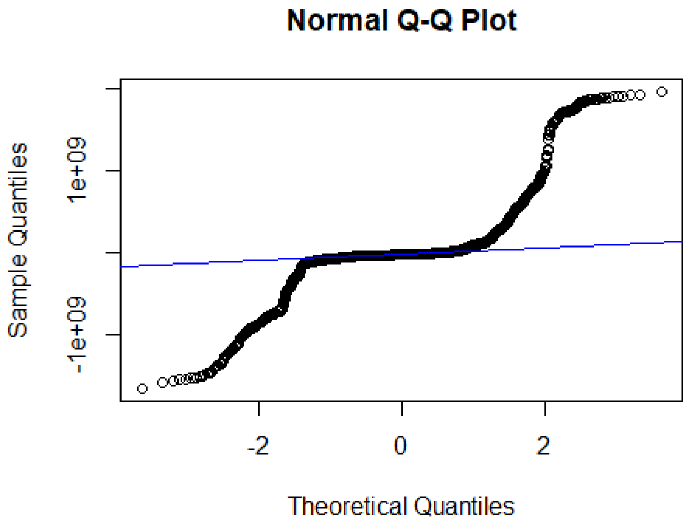
Model Statistics

r.squared	adj.r.squared	sigma	statistic	p.value	df	df.residual
0.8385	0.8381	377,987,968.1284	2,248.97	< 2e-16	8	3,466

Summary statistics for the linear regression model.

MLR CASE - Accounting for Multi Collinearity





Our model has issues, let’s test for multicollinearity

```
## primary_energy_consumption    fossil_fuel_consumption 
##                 2444.91176                 2230.79196 
##           coal_consumption            gas_consumption 
##                   35.02973                   61.80665 
##     renewables_electricity          solar_electricity 
##                 3366.00618                   33.53452 
##           wind_electricity          hydro_electricity 
##                  215.98613                 1721.61480 
```

As we can see, all values are above 10, lets remove our most concerning values that are obviously collinear due to its transparent reliance on other variables defined in the model, then re-evaluate. Those variables are primary_energy_consumption and renewables_electricity

Check again

```
## fossil_fuel_consumption        coal_consumption         gas_consumption 
##               123.92344                30.52495                57.75149 
##       solar_electricity        wind_electricity       hydro_electricity 
##                12.80104                18.01308                20.29338 
```

This helped a lot, lets remove fossil_fuel_consumption next, since it is the next highest with a still relatively high variance inflation factor of 123.923

```
##  coal_consumption   gas_consumption solar_electricity  wind_electricity 
##          8.518596         10.909338         12.368665         15.613430 
## hydro_electricity 
##         19.983011 
```

Again, with next highest which is hydro_electricity

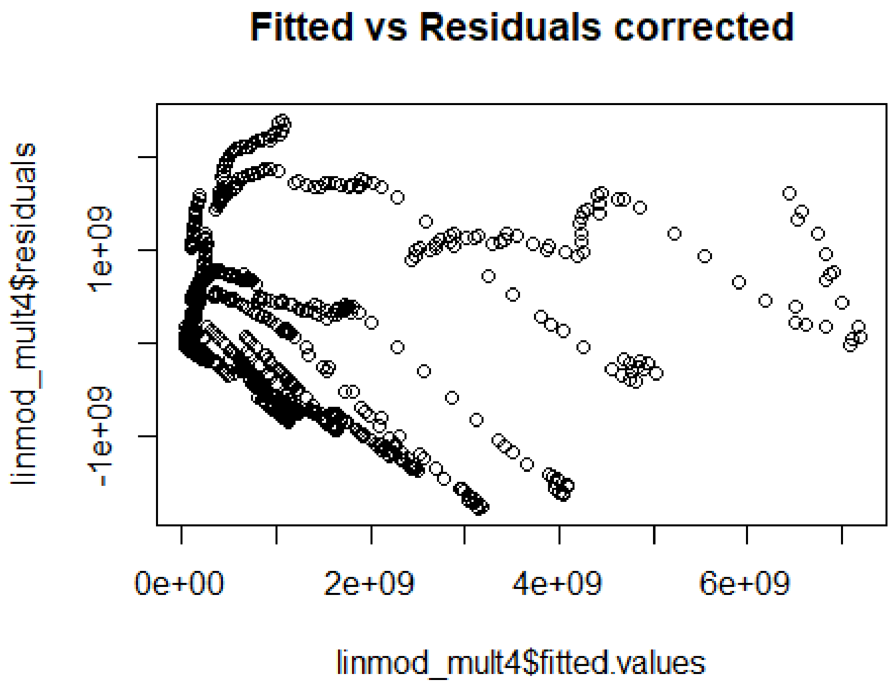
```
##  coal_consumption   gas_consumption solar_electricity  wind_electricity 
##          4.252963          5.596159         12.345037         15.513629 
```

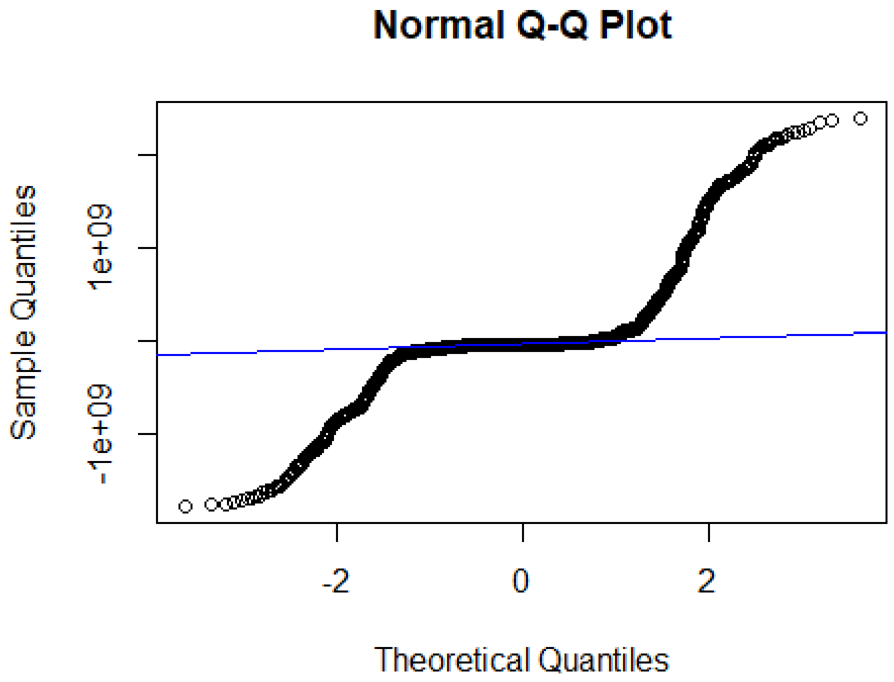
Again, with wind_electricity

```
##  coal_consumption   gas_consumption solar_electricity 
##          4.194911          4.543366          1.429500 
```

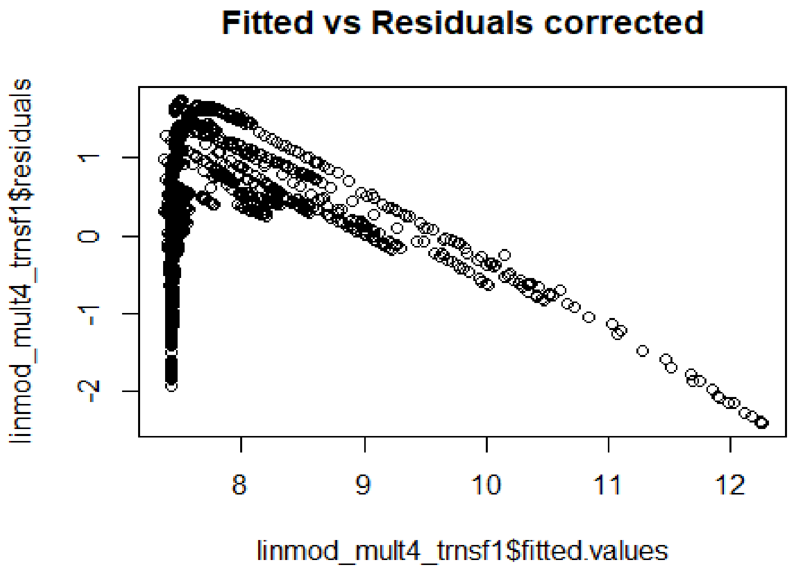
At this point, there are no major concerns for multicollinearity, though there are still some minor concerns regarding coal_consumption and gas_consumption while solar_electricity is within acceptable VIF range. We will stop removing variables here. *If multicollinearity continues to be a concern, I recommend running a ridge regression*

Lets run some diagnostics on this model.





It looks like were still having the same issues, so we will try scaling the response



Results are sub-par, try again scaling predictors as well

This cannot be completed because the vectors are different lengths, any attempts to transform will result in more data loss, efforts to remove/add data to make the vectors the same length will be poor data analysis hygiene. Perhaps random sampling our of population to obtain same length, but there may be some hidden statistical/probability theory that is being violated?

Since non-constant variance as well as clustering still appears to be an issue, I would recommend using other scales to try to un-cluster data then evaluating. If non-constant variance is still a concern, we may consider using weighted regression.

HYPOTHESIS TEST

$$H_0: all\ slope\ coefficients\ are\ equal\ to\ zero$$

$$H_a: at\ least\ one\ slope\ is\ not\ zero$$

Using the output seen above, we have that

1) F = 2249
2) The degrees of freedom are F(8, 3466)
3) p-value < 2.2e-16

Because the p-value is so small we reject H_0. In context, this means that at least one of the coefficients in our full model is nonzero and important in predicting values in population. The null model, with no predictors is not sufficient in predicting the population.

To answer the research question posed above, we see that all of our predictors that were included in this model are significant except for one. This means that there is a relation between a country's economic population and its energy consumption patterns. For example, an in the context of this model, for every one unit increase in a country's fossil fuel consumption, that country's population would be 325526 greater. Another interesting example, for every one unit increase of hydro electricity consumption, we would expect that country to have 2504760 less population.

Results - A Quick Recap

Research Question 1: Global Energy Consumption Over Time

- Using exploratory data analysis (EDA) and time trend visualizations, the study examined global energy consumption trends over time.
- The analysis revealed that global energy consumption has increased significantly, with variations based on economic and demographic factors.
- The transition from fossil fuels to renewable energy sources was also observed in some regions, though fossil fuels remain dominant in many countries.

Research Question 2: Relationship Between Population and Energy Consumption

- A multiple linear regression model was used to assess the relationship between population and different types of energy consumption.
- The results indicated a significant correlation between population size and energy consumption patterns.
- Countries with larger populations generally consume more energy, though the type of energy used varies by region and level of economic development.

Research Question 3: Higher GDP = Higher Energy consumption?

- A simple linear regression (SLR) model initially suggested a strong relationship between GDP and energy consumption.
- However, diagnostic tests revealed issues with the model, leading to a log-transformed regression for better accuracy.
- Additionally, effect size analysis (Cohen's d) indicated that while GDP and energy consumption are correlated, high GDP does not always equate to proportionally higher energy consumption.

Research Question 4: Population Growth and Energy Demand

- The dataset was filtered for the years 2000 and 2020, and correlation analysis was conducted to compare population growth with energy consumption increases.
- The findings showed a positive correlation, suggesting that countries with rapid population growth tend to experience rising energy demand.
- However, the strength of this relationship varies depending on economic and policy factors influencing energy efficiency and consumption patterns.

Research Question 5: Trends in Top GDP Countries

- Time trend visualizations focused on energy consumption in the world's top GDP countries.
- The analysis showed that while overall energy consumption has increased, some high-GDP countries have shifted towards renewable energy sources.
- Differences in energy policies and economic structures have led to varying trends in energy consumption among top economies.

Acknowledgements

This report was written by Alex Marcek, with contributions to research questions by all group members. Regarding GitHub repository contributions:

- Alex Marcek is the repository administrator and provided the preliminary analysis and data filter template, as well as additional commentary for the pre analysis process (files: `pre_analysis.rmd`, `filter_doc.rmd` and `prelim_regression.rmd`).
- Ahmad Khan authored all code and commentary for exploratory data analysis (file: `eda.rmd`) and helped in formatting report.
- Tara Draper authored all code and commentary for linear regression (file: `Energy_Consumption.rmd`) and collaborated to create the presentation

Our project has been documented with all the code available at https://github.com/AlexM866/ISTA321_Midterm1

- report.Rmd has all the code of this particular document
- eda.Rmd, pre-analysis.Rmd, prelim_regression.Rmd, filter_doc.Rmd has other of our workings

---

References

- Dataset: Our World in Data - Energy
- Variable Documentation: `energy-data/owid-energy-codebook.csv` at owid/energy-data ```