MSCI 609 Deliverable 3

Group 9

1. Group Members & ID

Anson Ma, #20916612
Tahmid Bari, #20864394

2. Introduction and Objective

The objective of the paper is to examine the relationship between carbon dioxide (CO_2) emissions, gross domestic product (GDP) and natural factors through the statistical method of regression. Natural factors include the differences in average climate temperatures, the proportion of urban land areas, and the availability of renewable and fossil fuel resources. While many studies have shown the empirical relationship between CO_2 emissions and income, the question to what extent natural factor determines cross-country differences in CO_2 emissions has been somehow neglected, which will be addressed in this paper. The analysis would be mainly based on and compared against four countries of different income groups, namely Ethiopia (low-income), Vietnam (lower-middle-income), Thailand (middle-income), and Japan (high-income).

3. Proposed Model

The basic model would be

$$Y_i = a_i + b_{i1}x_{i1} + b_{i2}{x_{i1}}^2 + b_{i3}x_{i3} + b_{i4}x_{i4} + b_{i5}x_{i5} + b_{i6}x_{i6} + b_{i7}x_{i7} + e_i$$

Where

$Y = CO_2\ emission\ per\ capita$

$x_1 = GDP\ per\ capita$
${x_1}^2 = squared\ GDP\ per\ capita$
$x_3 = lowest\ monthly\ average\ temperature$
$x_4 = highest\ monthly\ average\ temperature$
$x_5 = percentage\ of\ urban\ land\ in\ total\ land\ area$
$x_6 = percentage\ of\ renewable\ energy\ use\ in\ total\ energy\ use$
$x_7 = percentage\ of\ fossil\ fuel\ consumption\ in\ total\ energy\ use$
$\varepsilon\ = error\ term$

$i = 1: Ethiopia;\ 2: Vietnam, 3: Thailand, 4: Japan$

4. Data Table

Variable Name	Description	Type	Source	Website
country	countries under analysis	categorical	–	–
country_code	country code	categorical	–	–
year	calendar year	integer	–	–
co2_emission	carbon dioxide emission per capita	continuous	The World Bank	https://data.worldbank.org/indicator/EN.ATM.CO2E.PC
gdp_capita	US$ gross domestic product per capita	continuous	The World Bank	https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
gdp_capita2	squared US$ gross domestic product per capita	continuous	The World Bank	https://data.worldbank.org/indicator/NY.GDP.PCAP.CD
lowest_temp	lowest monthly average temperature	continuous	The World Bank	https://climateknowledgeportal.worldbank.org/download-data
highest_temp	highest monthly average temperature	continuous	The World Bank	https://climateknowledgeportal.worldbank.org/download-data
urban_area	percentage of urban land in total land area	continuous	The World Bank	https://data.worldbank.org/indicator/AG.LND.TOTL.UR.K2
renewable_energy	percentage of renewable energy use in total energy use	continuous	The World Bank	https://data.worldbank.org/indicator/EG.FEC.RNEW.ZS
fossil_fuel	percentage of fossil fuel consumption in total energy use	continuous	The World Bank	https://data.worldbank.org/indicator/EG.USE.COMM.FO.ZS

Descriptive Statistics

Variable	# Fitted	Mean	Median	Std.Dev.	Root.M.Sqr.	Std.Err.Mean	Minimum	Maximum
co2_emission	100	3.440	1.727	3.740	5.068	0.374	0.041	9.881
gdp_capita	100	10,764	5,138	12,139	16,179	1,214	346.584	40,396
gdp_capita2	100	261,744,033	26,407,275	430,488,940	501,973,564	43,048,894	120,120	1,631,856,465
highest_temp	100	25.807	25.170	2.497	25.926	0.250	20.688	30.336
lowest_temp	100	15.754	20.856	9.755	18.504	0.975	-1.908	24.544
urban_area	100	10.247	7.163	11.907	15.664	1.191	0.517	29.816
renewable_energy	100	43.679	33.327	35.481	56.162	3.548	3.568	97.740
fossil_fuel	100	54.392	69.160	32.842	63.453	3.284	2.245	94.633

5. Use of Data

All data is collected from the World Bank DataBank to maintain consistency. The raw data is consisted of a panel covering 25 years (1991-2015) on 170 countries around the globe.

CO_2 emissions per capita, forming the dependent variable, is based on the data from Carbon Dioxide Information Analysis Center (CDIAC). CO_2 emissions are those stemming from the burning of fossil fuels and the manufacture of cement. These emissions also include CO_2 produced during consumption of solid, liquid, and gas fuels and gas flaring.

Income, as one of the independent variables, is based on the data from World Bank National Accounts and OECD National Accounts. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. GDP per capita is obtained by diving GDP by mid-year population. Higher income countries would have greater emissions to sustain economic development and activities, and vice versa.

The lowest and highest monthly average temperatures are based on the data from World Bank Climate Change Knowledge Portal (CCKP). The temperatures are the average of temperatures of a month, which are then compared against each other within the year to identify the lowest and the highest of the year. It is expected that cold countries would have greater heating demands while hot countries would have greater cooling demands. Cold countries and hot countries would therefore have higher CO_2 emissions.

The percentage of urban areas is based on the data from CIESIN Urban-Rural Population and Land Area Estimates and the Food and Agriculture Organization. The urban area is computed on a combination of population counts, settlement points, and the presence of nighttime lights. The numbers are then divided by the country total land area to obtain the percentage. Countries with less urban areas are sparsely inhabited and have higher transportation demands to move goods and people over long distances. Higher transportation demands would have higher emissions, and vice versa.

The percentage of renewable energy in total energy use is based on the data from World Bank Sustainable Energy for All (SE4ALL). Renewable resources encompass hydroelectric, geothermal, solar and wind resources as well as “fuel and waste”, which comprise biomass and animal products, gas/liquids from biomass, industrial waste, and municipal waste. It is expected that countries that have access to domestic renewable energy resources would have lower emissions than countries that lack such resources.

The percentage of fossil fuel consumed is based on the data from IEA Statistics. Fossil fuel comprises coal, oil, petroleum, and natural gas products. The higher the consumption, the fewer the reserve, vice versa. Countries that have fewer fossil fuel reserves should have lower CO_2 emissions than countries that are rich in such reserves. This is for two reasons: First, because of the emissions generated in the extraction and possibly the transport and processing of such resources. Second, because of countries that lacked major domestic fossil fuel reserves have had strong incentives to develop in a less fossil fuel intensive way to cut down on energy import costs.

6. Normality of Dependent Variable

The Shapiro-Wilk test is performed to check normality of the dependent variable (DV) using R programming. In this paper, CO_2 emission is the dependent variable.

```
Shapiro-Wilk normality test
data:  dv_normality
W = 0.99874, p-value = 0.7135
```

The Shapiro-Wilk test for normality is available when using the Distribution plat form to examine a continuous variable. The shapiro.test() function in R employsthe Shapiro-Wilk test on data to test whether the data are normally distributed. Use of the Shapiro-Wilk test is contingent on univariate and continuous data. The hypotheses for the test are:

$$H_0: The\ data\ are\ normally\ distributed$$
$$H_1: The\ data\ are\ not\ normally\ distributed$$

The Shapiro-Wilk test for normality is a statistical test that provides a p-value of the test statistic, W. This lab will focus on the p-value approach for statistical tests, using an α value of 0.05 as the desired significance level.

The Shapiro-Wilk test p-value is greater than α = 0.05, therefore failing to reject H_0 concluding the data are normally distributed. Again, this is expected given the `dv_normality` object was created via the `rnorm()` function.

```
> describe(d3_dv_normality$co2_emission)
vars   n mean   sd median trimmed  mad  min  max range skew kurtosis   se
X1     1  104 3.39 3.69   1.78    3.03 2.54 0.04 9.88  9.84 0.82    -0.99 0.36
```

The density plot provides a visual judgment about whether the distribution is bell shaped. Comparing the co2_emission data to a normal distribution, given μ and σ from the co2_emission data, is available via the *geom_density()* and *stat_function()* functions.

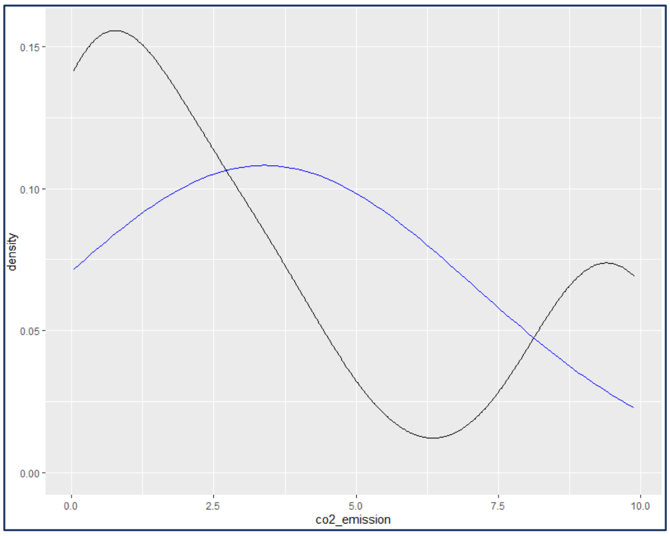


Figure 1: Density plot of the dependent variable DV = CO2 emission

The black line represents the co2_emission data, and the blue line represents the normal distribution given the μ and σ values of co2_emission. The co2_emission data appears bimodal, and does not fit the normal distribution model given the parameters calculated via the co2_emission data. The co2_emission data is further examined using QQ plots via the qqPlot() function:

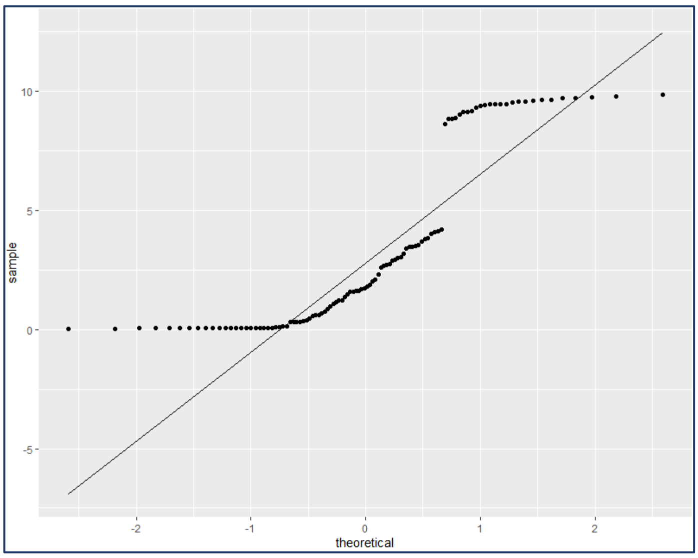


Figure 2: Normal Q-Q plot of DV = CO2 emission

Most of the points in the QQ plot fall outside the region defined by the dashed lines, further suggesting the co2_emission data is likely not normally distributed. Lastly, a Shapiro-Wilk test can confirm whether the co2_emission data is normally distributed:

```
Shapiro-Wilk normality test
data:  d3_dv_normality$co2_emission
W = 0.77506, p-value = 2.549e-11
```

The Shapiro-Wilk test p-value is less than $\alpha = 0.05$, leading to reject H0: data are normally distributed. In conclusion, the co2_emission data is not normally distributed. The visual plots are likely enough to confirm the co2_emission data are not normally distributed.

7. Comparison of Dependent Variable Means

The independent two-sample t-test is performed to compare the paired sample means among the 4 countries; Japan, Thailand, Vietnam, Ethiopia using Microsoft Excel.

Japan: mean (9.391375457) > Thailand: mean (3.103617)

t-Test: Paired Two Sample for Means

	Japan	*Thailand*
Mean	9.391375457	3.103617
Variance	0.108738244	0.568407
Observations	26	26
Pearson Correlation	0.447874747	
Hypothesized Mean Difference	0	
df	25	
t Stat	47.55965533	
P(T<=t) one-tail	2.41779E-26	
t Critical one-tail	1.708140761	
P(T<=t) two-tail	4.83558E-26	
t Critical two-tail	2.059538553	

Japan: mean (9.391375457) > Vietnam: mean (0.979067422)

t-Test: Paired Two Sample for Means

	Japan	*Vietnam*
Mean	9.391375457	0.979067422
Variance	0.108738244	0.299928883
Observations	26	26
Pearson Correlation	0.216847863	
Hypothesized Mean Difference	0	
df	25	
t Stat	74.63079834	
P(T<=t) one-tail	3.35339E-31	
t Critical one-tail	1.708140761	
P(T<=t) two-tail	6.70678E-31	
t Critical two-tail	2.059538553	

Japan: mean (9.391375457) > Ethiopia: mean (0.070484)

t-Test: Paired Two Sample for Means

	Japan	*Ethiopia*
Mean	9.391375457	0.070484
Variance	0.108738244	0.000553
Observations	26	26
Pearson Correlation	0.06671267	
Hypothesized Mean Difference	0	
df	25	
t Stat	144.4497966	
P(T<=t) one-tail	2.35866E-38	
t Critical one-tail	1.708140761	
P(T<=t) two-tail	4.71732E-38	
t Critical two-tail	2.059538553	

Thailand: mean (3.103616614) > Vietnam: mean (0.979067422)

t-Test: Paired Two Sample for Means

	Thailand	*Vietnam*
Mean	3.103616614	0.979067422
Variance	0.568406843	0.299928883
Observations	26	26
Pearson Correlation	0.937013138	
Hypothesized Mean Difference	0	
df	25	
t Stat	35.22874312	
P(T<=t) one-tail	3.93694E-23	
t Critical one-tail	1.708140761	
P(T<=t) two-tail	7.87387E-23	
t Critical two-tail	2.059538553	

Thailand: mean (3.103616614) > Ethiopia: mean (0.070484)

t-Test: Paired Two Sample for Means

	Thailand	*Ethiopia*
Mean	3.103616614	0.070484
Variance	0.568406843	0.000553
Observations	26	26
Pearson Correlation	0.73200966	
Hypothesized Mean Difference	0	
df	25	
t Stat	20.98822351	
P(T<=t) one-tail	1.0811E-17	
t Critical one-tail	1.708140761	
P(T<=t) two-tail	2.1622E-17	
t Critical two-tail	2.059538553	

Vietnam: mean (0.979067422) > Ethiopia: mean (0.070484)

t-Test: Paired Two Sample for Means

	Vietnam	*Ethiopia*
Mean	0.979067422	0.070484373
Variance	0.299928883	0.000552898
Observations	26	26
Pearson Correlation	0.858473926	
Hypothesized Mean Difference	0	
df	25	
t Stat	8.780893293	
P(T<=t) one-tail	2.06402E-09	
t Critical one-tail	1.708140761	
P(T<=t) two-tail	4.12804E-09	
t Critical two-tail	2.059538553	

95% confidence intervals around the largest mean and the smallest mean:

(Largest) Mean	(Smallest) Mean
9.391375457	3.103616614
9.391375457	0.979067422
9.391375457	0.070484373
3.103616614	0.979067422
3.103616614	0.070484373
0.979067422	0.070484373

(Largest) Mean	9.320891	Sample SD	5.948400032
(Smallest) Min	0.908583	Margin of Error	0.331031499
(Largest) Mean LB (Mean - E)	8.98986		
(Largest) Mean UB (Mean + E)	9.651922		
(Smallest) Min LB (Mean - E)	0.577552		
(Smallest) Min UB (Mean + E)	1.239614		

(Largest) Mean CI = (8.98986, 9.651922)

(Smallest) Mean CI = (0.577552, 1.239614)

8. Regression Analysis

Model_1 is generated using Regressit on Microsoft Excel based on the DV = CO2_emission

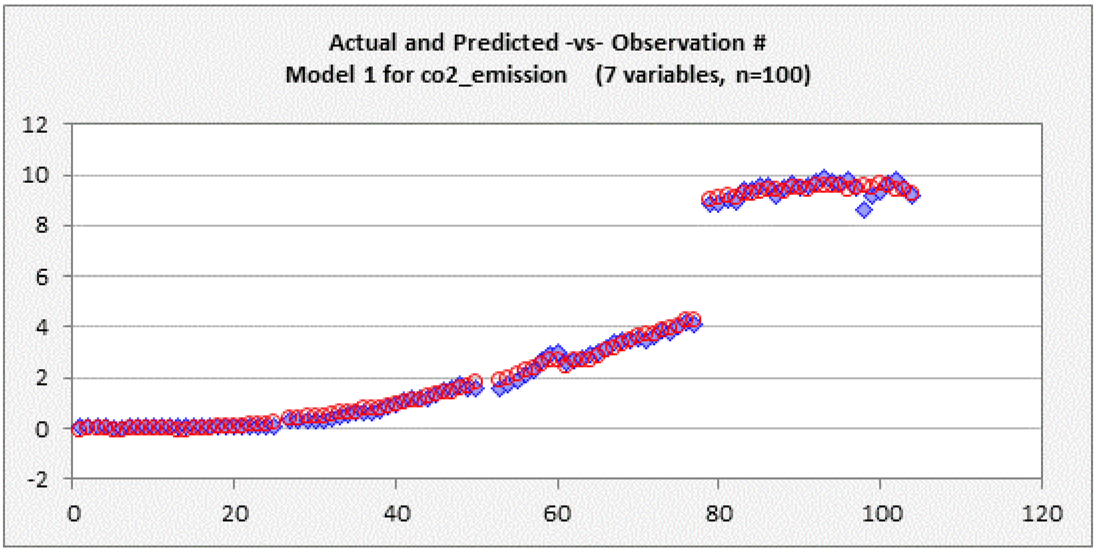
Model: Model 1

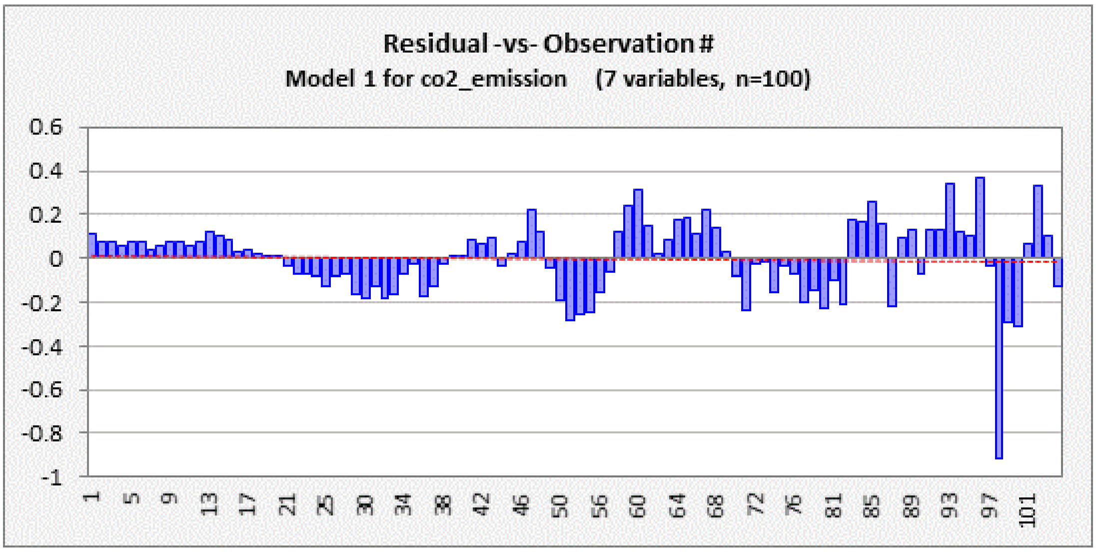
Dependent Variable: co2_emission

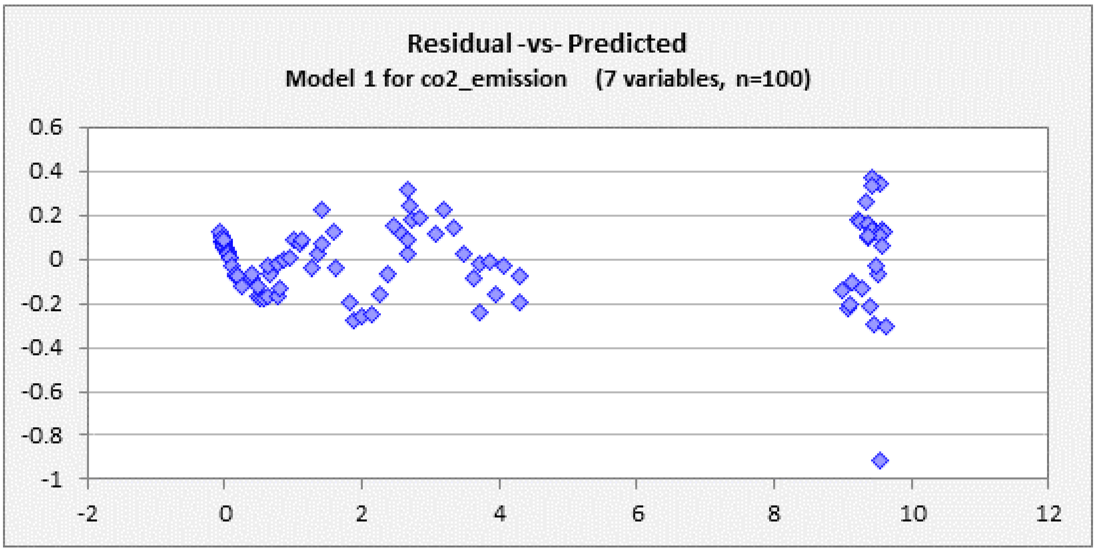
	R-Squared	Adj.R-Sqr.	Std.Err.Reg.	Std.Dep.Var.	# Fitted	# Missing	Critical t	Confidence
	0.998	0.998	0.181	3.740	100	4	1.986	95.0%

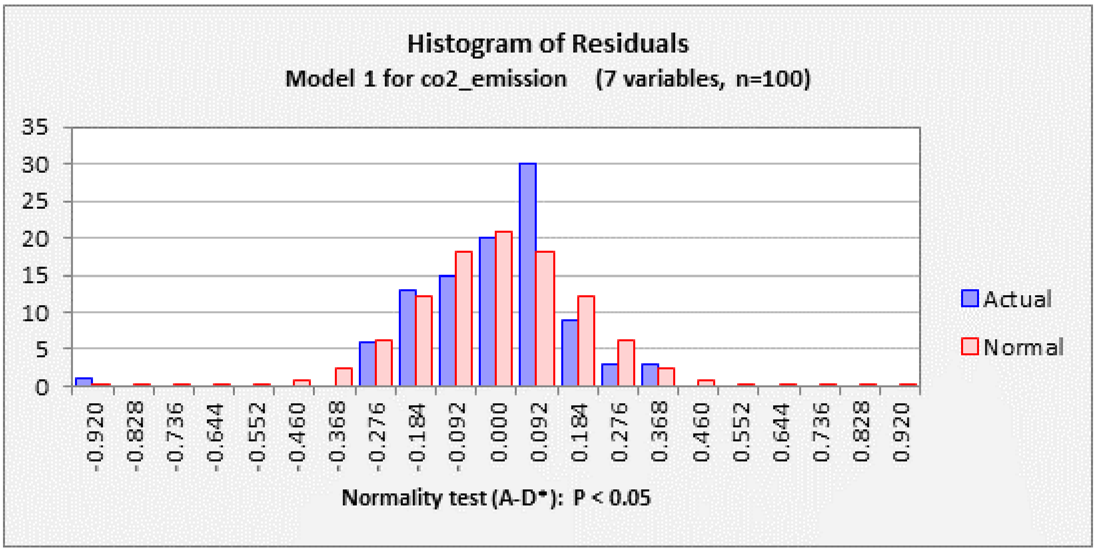
Variable	Coefficient	Std.Err.	t-Statistic	P-value	Lower95%	Upper95%	VIF	Std. Coeff.
Constant	0.009380	1.095	0.009	0.993	-2.166	2.184	0.000	0.000
fossil_fuel	0.015	0.009280	1.660	0.100	-0.003029	0.034	280.701	0.135
gdp_capita	0.000295	0.000015	19.538	0.000	0.000265	0.000325	101.212	0.956
gdp_capita2	-4.766E-09	2.981E-10	-15.986	0.000	-5.358E-09	-4.174E-09	49.776	-0.549
highest_temp	-0.034	0.026	-1.337	0.185	-0.085	0.017	12.302	-0.023
lowest_temp	-0.035	0.011	-3.037	0.003	-0.057	-0.012	37.162	-0.090
renewable_energy	0.013	0.010	1.289	0.201	-0.007050	0.033	389.579	0.124
urban_area	0.149	0.016	9.195	0.000	0.117	0.181	112.320	0.474

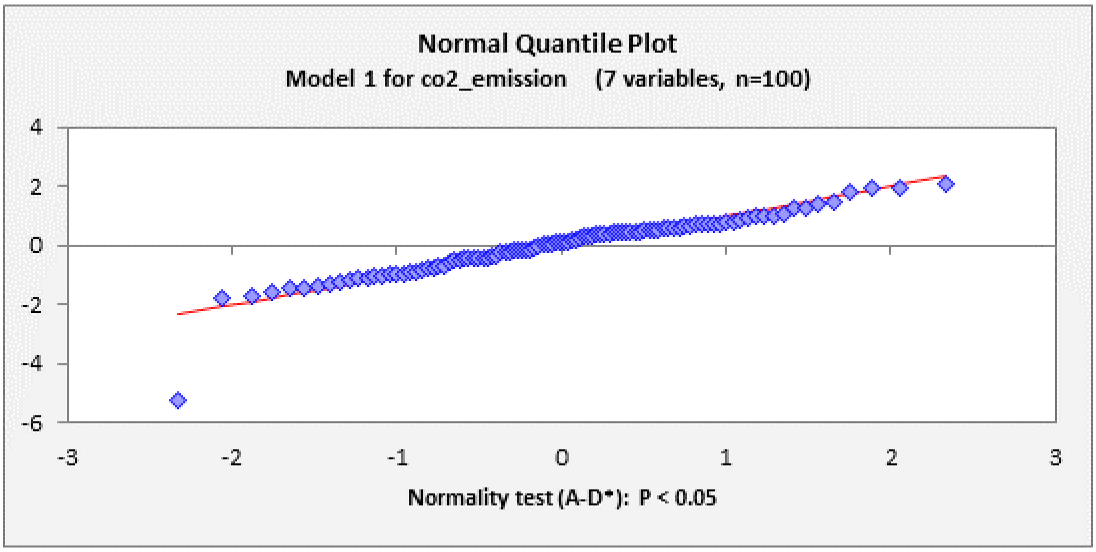
	Mean Error	RMSE	MAE	Minimum	Maximum	MAPE	A-D* stat
Fitted (n=100)	0.000	0.174	0.130	-0.919	0.366	32.1%	1.00 (P=0.012)



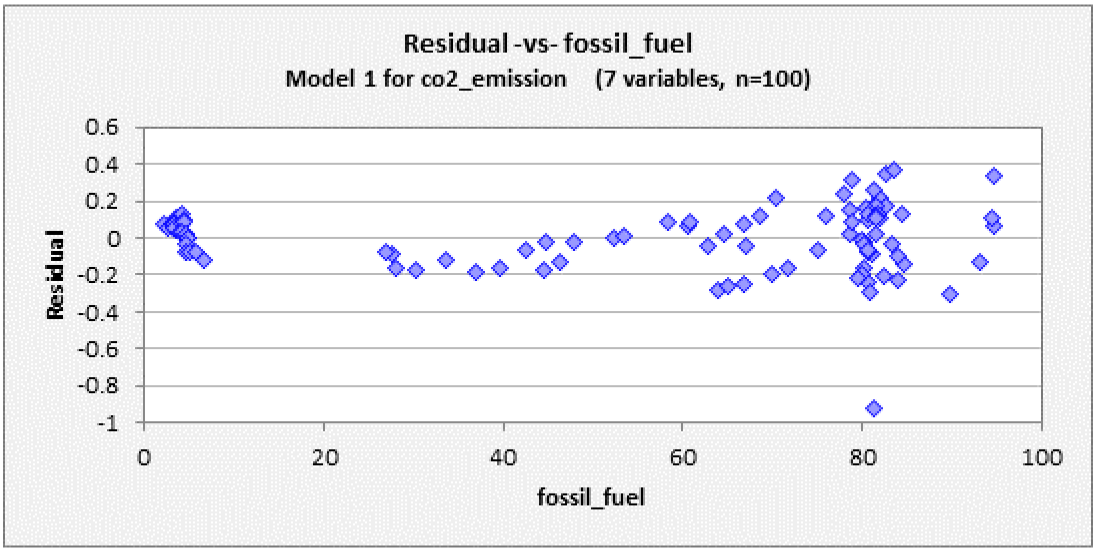

Residual -vs- Observation #
Model 1 for co2_emission (7 variables, n=100)


Residual -vs- Predicted
Model 1 for co2_emission (7 variables, n=100)

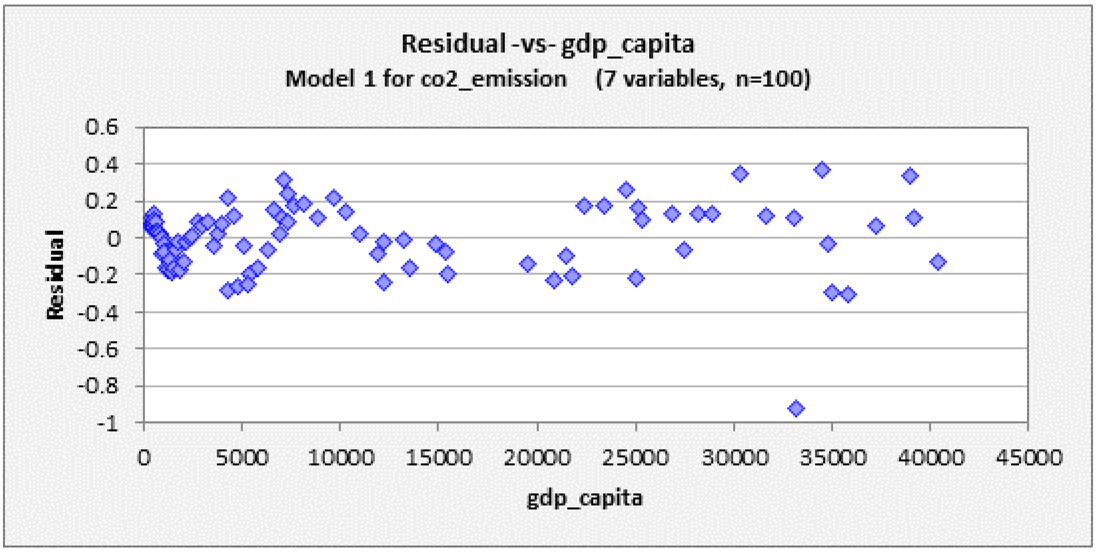

Histogram of Residuals
Model 1 for co2_emission (7 variables, n=100)
Actual
Normal
Normality test (A-D*): P < 0.05



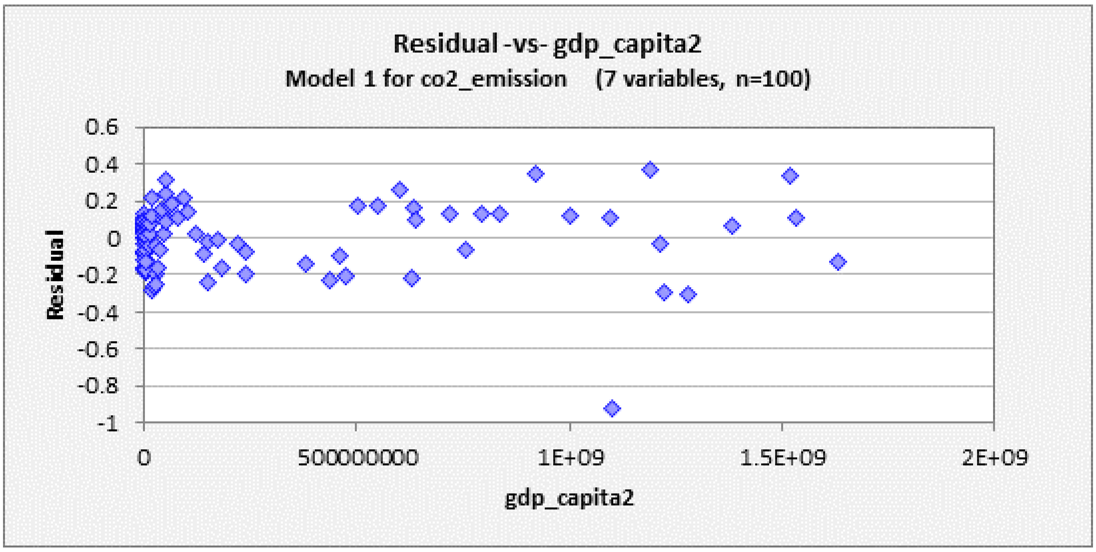
Residual -vs- fossil_fuel



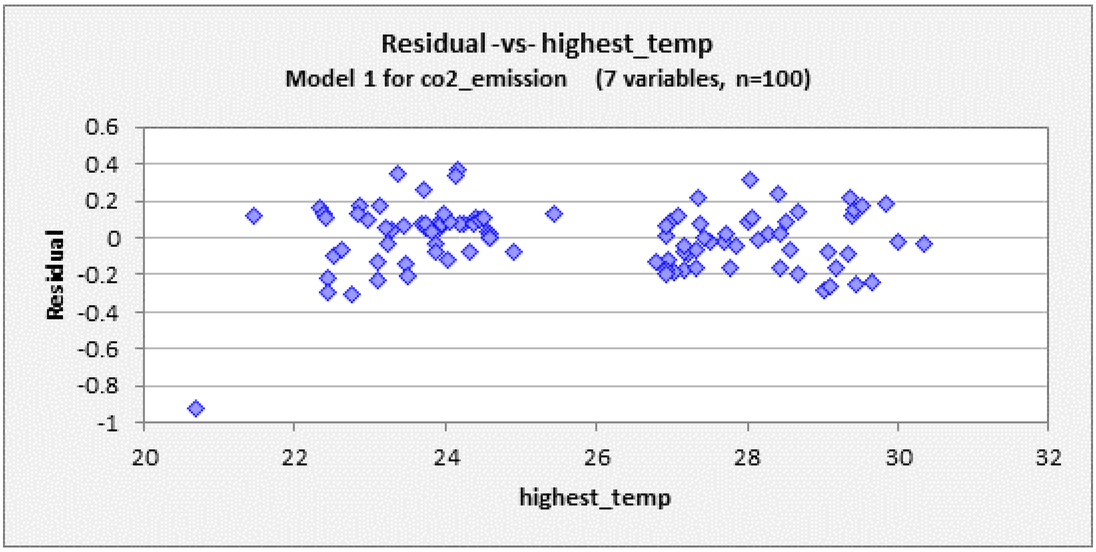
Residual -vs- gdp_capita



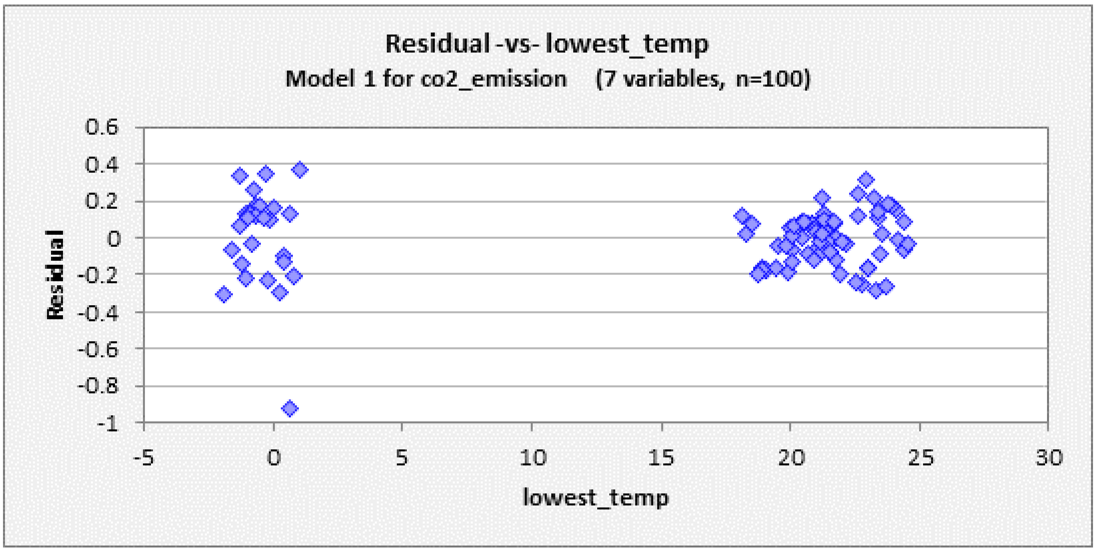
Residual -vs- gdp_capita2



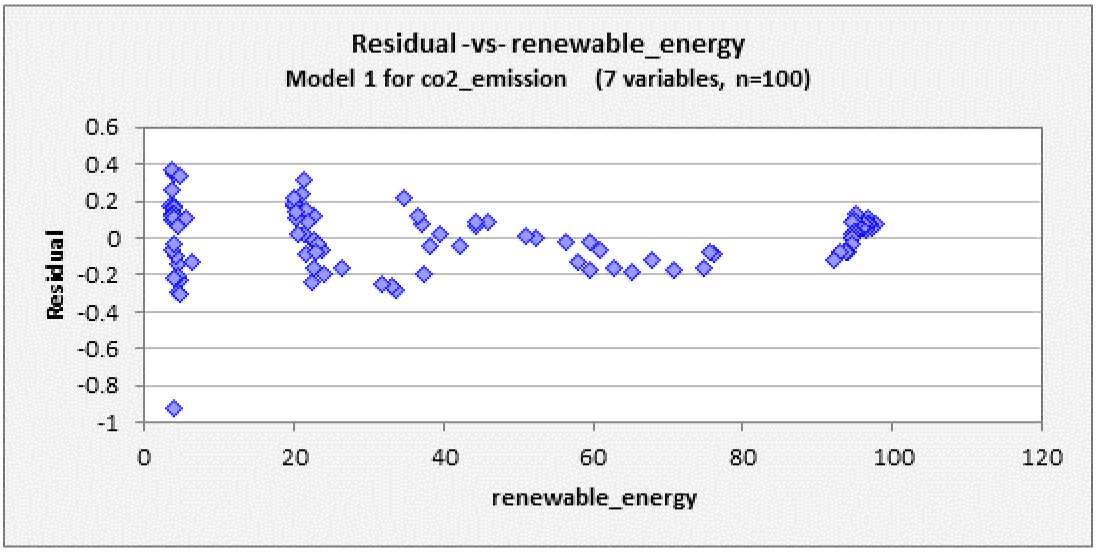
Residual -vs- highest_temp



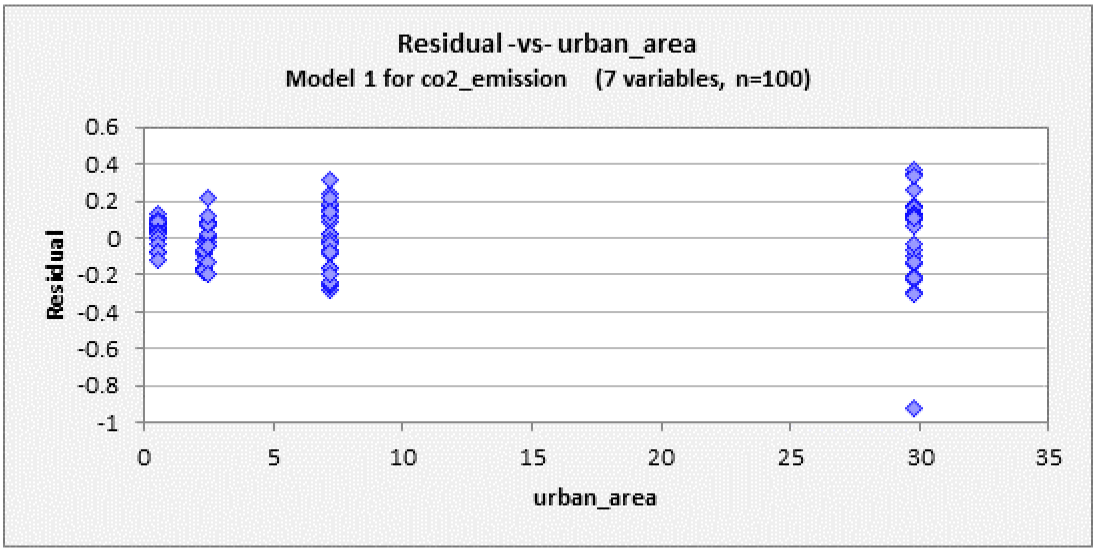
Residual -vs- lowest_temp



Residual -vs- renewable_energy



Residual -vs- urban_area



9. Analysis of Residuals

```
Shapiro-Wilk normality test
data:  d3_residual$residual
W = 0.91241, p-value = 5.696e-06
```

describe(d3_residual$residual)

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X1	1	100	0	0.17	0.02	0.01	0.14	-0.92	0.37	1.29	-1.35	6.04	0.02

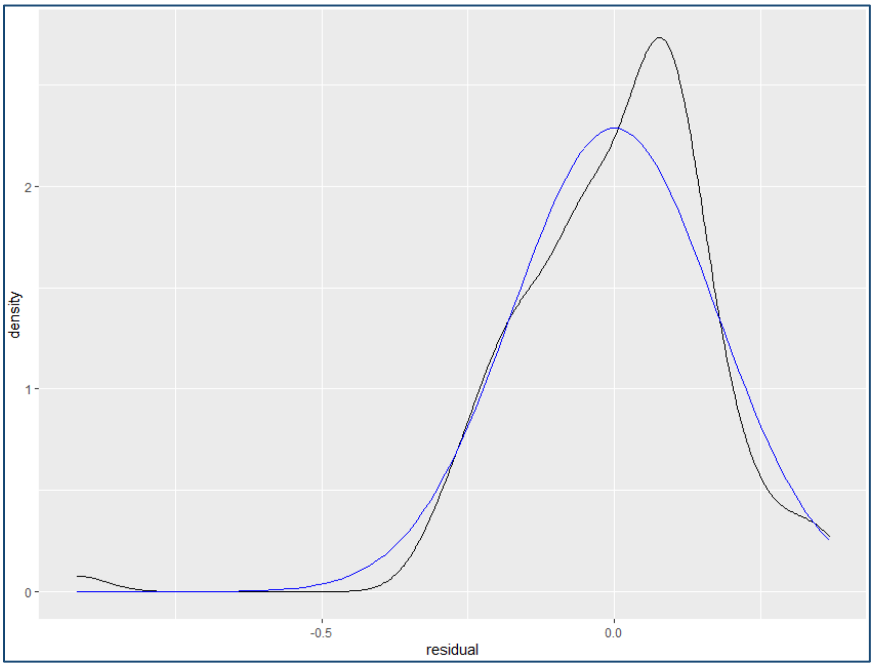


Figure 3: Density plot of the residuals

The density plot provides a visual judgment about whether the distribution is bell shaped. The black line represents the residual data, and the blue line represents the normal distribution given the μ and σ values of residual. **Note:** The residual data appears as bell shaped, and does fit the normal distribution model given the parameters calculated via the residual data.

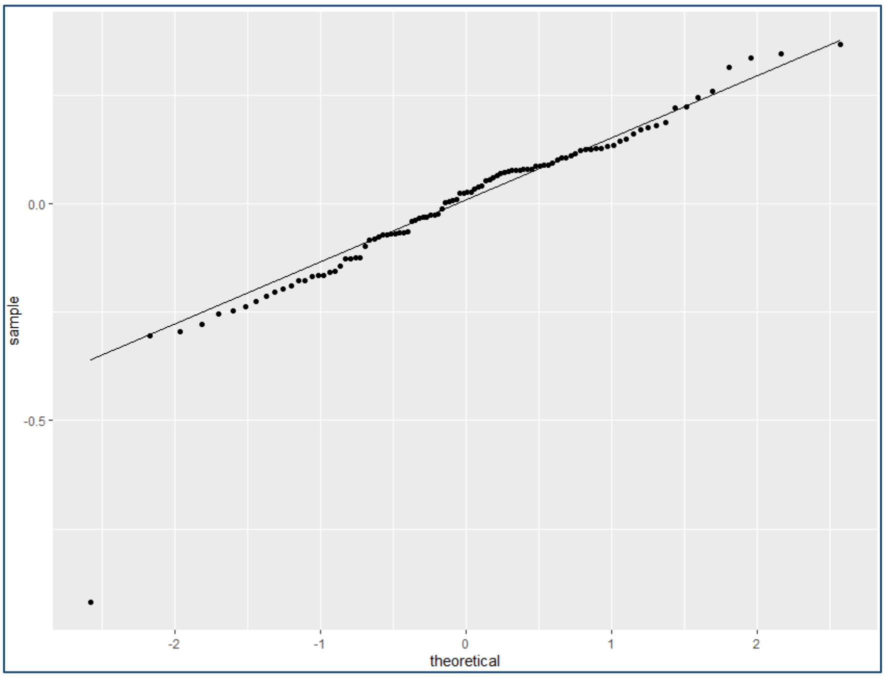


Figure 4: G-G plot of the residuals

Most of the points in the QQ plot fall inside the region defined by the dashed lines, further suggesting the residual data is likely normally distributed. Lastly, a Shapiro-Wilk test can confirm whether the residual data is normally distributed:

```
Shapiro-Wilk normality test
data:   d3_residual$residual
W = 0.91241, p-value = 5.696e-06
```

The Shapiro-Wilk test p-value is less than $\alpha = 0.05$, leading to reject H_0: data are normally distributed. In conclusion, the residual data is not normally distributed. **Note:** Though the visual plots are likely enough to confirm the residual data are normally distributed however, based on the p-value of residual data we can conclude that the data is not normal.

Check for heteroskedasticity: There seems to be no evident pattern. However, it does seem to look as if there's more variation in residuals in this sample data from the linear regression.

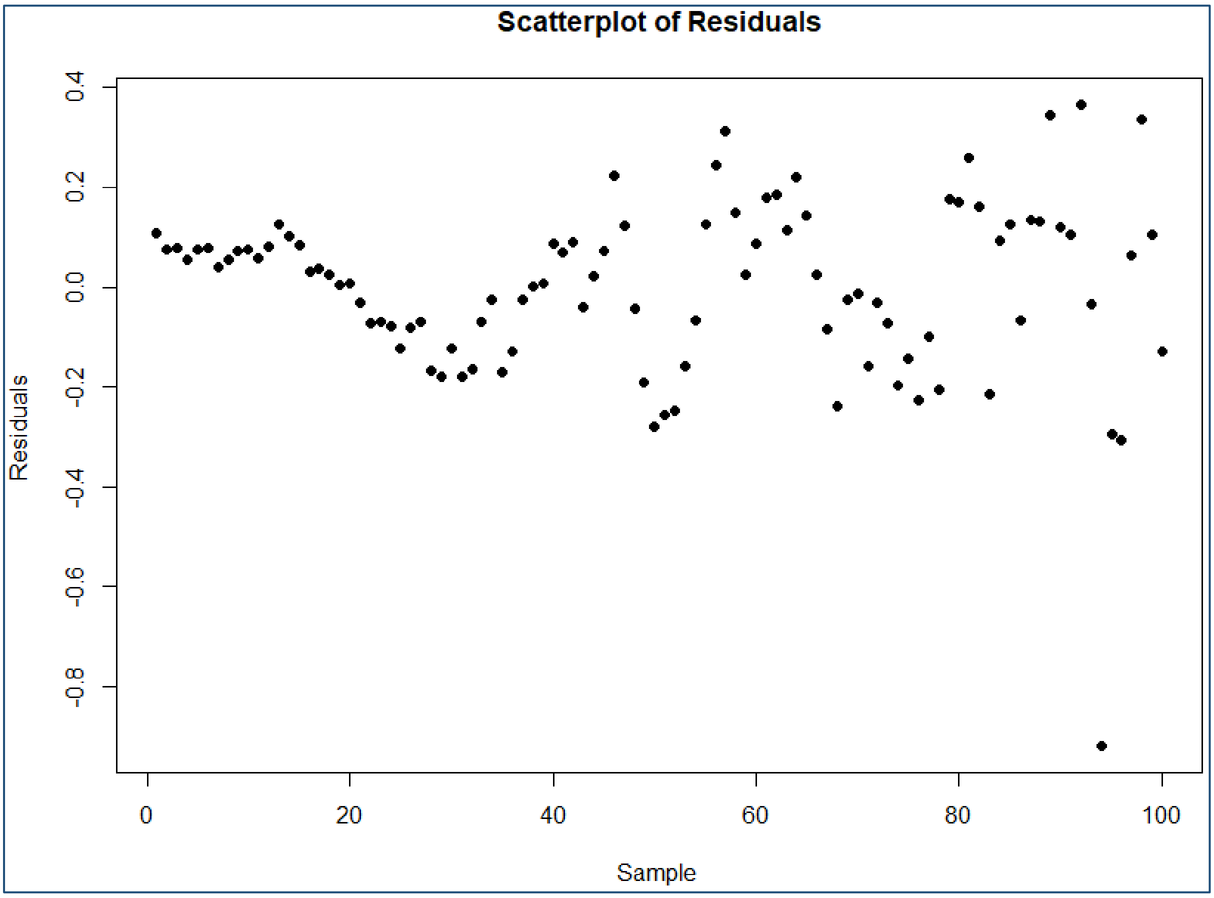


Figure 5: Scatterplot of Residuals

10. References

Neumayer, E (2002). 'Can Natural Factors Explain Any Cross-Country Differences in Carbon Dioxide Emissions?' Energy policy, 30 (1). pp. 7-12.

World Bank, World Development Indicators. (2016). 'CO2 emissions (metric tons per capita)' [Data file]. https://data.worldbank.org/indicator/EN.ATM.CO2E.PC

World Bank, World Development Indicators. (2019). 'GDP per capita (current US$)' [Data file]. https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

World Bank, World Development Indicators. (2016). 'Climate Change Knowledge Portal' [Data file]. https://climateknowledgeportal.worldbank.org/download-data

World Bank, World Development Indicators. (2013). 'Urban land area (sq. km)' [Data file]. https://data.worldbank.org/indicator/AG.LND.TOTL.UR.K2

World Bank, World Development Indicators. (2018). 'Land area (sq. km)' [Data file]. https://data.worldbank.org/indicator/AG.LND.TOTL.K2

World Bank, World Development Indicators. (2015). 'Renewable energy consumption (% of total final energy consumption)' [Data file]. https://data.worldbank.org/indicator/EG.FEC.RNEW.ZS

World Bank, World Development Indicators. (2015). 'Fossil fuel energy consumption (% of total final consumption)' [Data file]. https://data.worldbank.org/indicator/EG.USE.COMM.FO.ZS