Final Assignment

Asia Grillo, ID student: 5409650

1) Exploratory analysis

1.1) Dataset Description

For this project, I created a dataset by integrating data from multiple **sources**: *EDGAR* (Emission Database for Global Atmospheric Research), a European Commission project that provides consistent estimates of fossil CO_2 emissions worldwide; *WORLD BANK*, offering a range of socio-economic and environmental indicators; *ESA* (European Space Agency), which detects wildfire occurrences through the Sentinel-3 satellite.
The dataset was harmonized using country codes and all data refer to the same year, **2021**, ensuring temporal consistency across variables.

The **main goal** of this analysis is to investigate the associations between CO_2 emissions and a set of independent variables that capture the socioeconomic, demographic, and environmental characteristics of countries. By analyzing these relationships, the objective is to identify key factors driving CO_2 emissions and assess their potential impact.
To achieve this, the **response variable** in this study is **CO_2 emissions**, which represents the amount of carbon dioxide released into the atmosphere. It is expressed in *megatons (Mt)*, which corresponds to million metric tons. Summing the values in this column, we find that in 2021 the world emitted approximately 35,780 Mt of CO_2.

First of all, I removed four observations with **missing values** in GDP capita and excluded five predictors with more than 10% missing values. The remaining missing data, accounting for 0.7% of the total dataset, were imputed using MissForest, ensuring that the density function of the predictors remained nearly unchanged.
After these preprocessing steps, the final dataset consists of **174 countries**, which represent the statistical units of the analysis, and a total of **18 covariates**, which are:

- *GDP capita*: Gross Domestic Product divided by midyear population, expressed in U.S. dollars (USD).
- *Renewable energy*: Percentage of a country's total final energy consumption derived from renewable sources.
- *Total population*: Midyear estimates of all residents in a country.
- *Urban population*: Percentage of people living in urban areas.
- *Population density*: Number of people per square kilometer, calculated as the midyear population divided by the total land area of a country.
- *Population growth*: Annual percentage increase in a country's population, calculated using the exponential growth rate.
- *Agriculture GDP*: Economic contribution of primary sector activities, expressed as a percentage of a country's GDP.
- *Industry GDP*: Economic contribution of industrial activities, including construction and manufacturing, expressed as a percentage of a country's GDP.
- *Manufacturing GDP*: Economic contribution of the manufacturing sector, expressed as a percentage of a country's GDP.
- *Agricultural land*: Country's total land area used for agricultural purposes, expressed as a percentage.
- *Forest area*: Percentage of land covered by forests.
- *Energy intensity*: Amount of energy used per unit of economic output, measured in megajoules per U.S. dollar of GDP at 2017 purchasing power parity (PPP).
- *Electricity access*: Percentage of population with access to electricity.
- *Natural resource rents*: Revenues generated from natural resources, expressed as a percentage of a country's GDP.

- *Life expectancy*: Average number of years a newborn is expected to live under current mortality conditions.
- *Unemployment rate*: Percentage of the total labor force that is actively seeking employment in a country.
- *Fire spots*: Number of wildfire occurrences detected by the Sentinel-3 satellite in a country.
- *Continent*: A **categorical variable** indicating the geographical location of each country. Originally, this variable had six levels (Europe, Asia, Oceania, Africa, North America, South America). To simplify the analysis, North America and South America were merged into a single category, *America*, while Asia and Oceania were grouped together under *Asia-Oceania*. The final variable consists of **four levels**: *Europe* with 36 countries, *Africa* with 49 countries, *America* with 33 countries and *Asia-Oceania* with 56 countries. In the regression analysis, *Europe* is set as the **baseline** category.

Below is the table with the **summary statistics** of the quantitative variables that are not expressed as percentages, providing insights into their distribution and structure.

Table 1: Summary of same variables

CO2_emission	GDP_capita	Pop_tot	Energy_intensity	Life_expectancy	Fire_spots
Min. : 0.075	Min. : 214.1	Min. :1.778e+04	Min. : 1.090	Min. :52.68	Min. : 0.0
1st Qu.: 4.081	1st Qu.: 2185.5	1st Qu.:2.523e+06	1st Qu.: 2.960	1st Qu.:65.67	1st Qu.: 2.0
Median : 14.725	Median : 5341.3	Median :9.643e+06	Median : 4.020	Median :71.69	Median : 36.0
Mean : 205.633	Mean : 15706.1	Mean :4.414e+07	Mean : 4.769	Mean :71.06	Mean : 608.5
3rd Qu.: 76.670	3rd Qu.: 18566.9	3rd Qu.:3.232e+07	3rd Qu.: 5.765	3rd Qu.:76.30	3rd Qu.: 235.8
Max. :12466.316	Max. :133711.8	Max. :1.414e+09	Max. :18.720	Max. :84.45	Max. :23927.0

1.2) Graphical representation of data

To enhance the interpretability of the coefficients, all independent variables were **centered** by subtracting their respective means. Additionally, since the response takes only positive values, a **logarithmic transformation** was applied, allowing it to span the entire set of real numbers.
From the figure below, we observe that the asymmetry in the response variable has been corrected, making its distribution more symmetric.

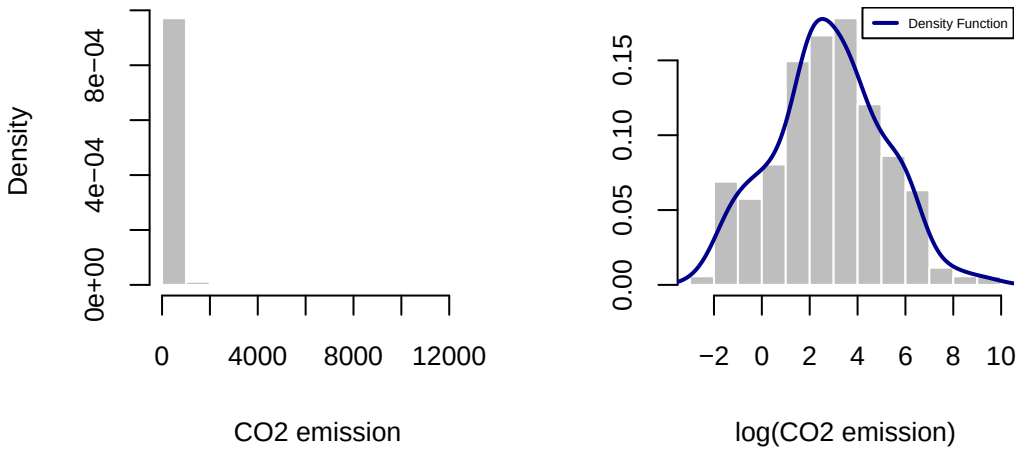

Figure 1: Histogram of CO2 emissions before and after the logarithmic transformation.

The boxplot below shows the distribution of log-transformed CO_2 emissions across different continents. **Asia-Oceania** exhibits the highest variability, with a wider interquartile range and more extreme values, reflecting the fact that *China* alone contributes 51.07% of the continent's emissions and 34.84% of global emissions. **Europe** has a relatively high median CO_2 emission, with a more concentrated distribution, in

line with *Germany*'s 20.55% share of European emissions and 1.86% globally. **America** shows a moderate spread, with the *United States* contributing 69.30% of continental emissions and 13.28% globally. **Africa** has the lowest median emissions and the smallest variability, where *South Africa* accounts for 34.22% of African emissions but only 1.22% of global CO_2 emissions.

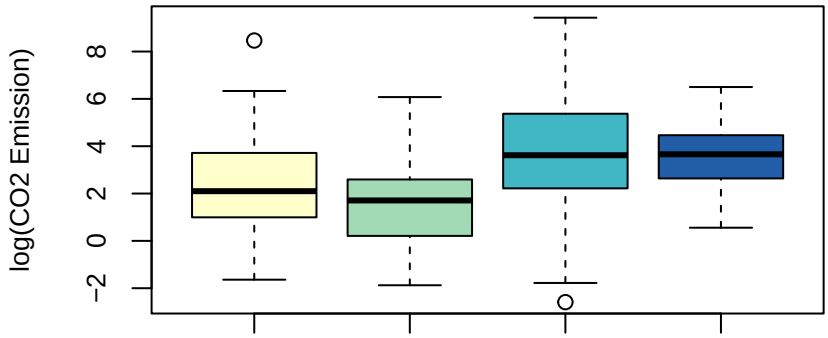

Figure 2: CO2 Emissions by Continent

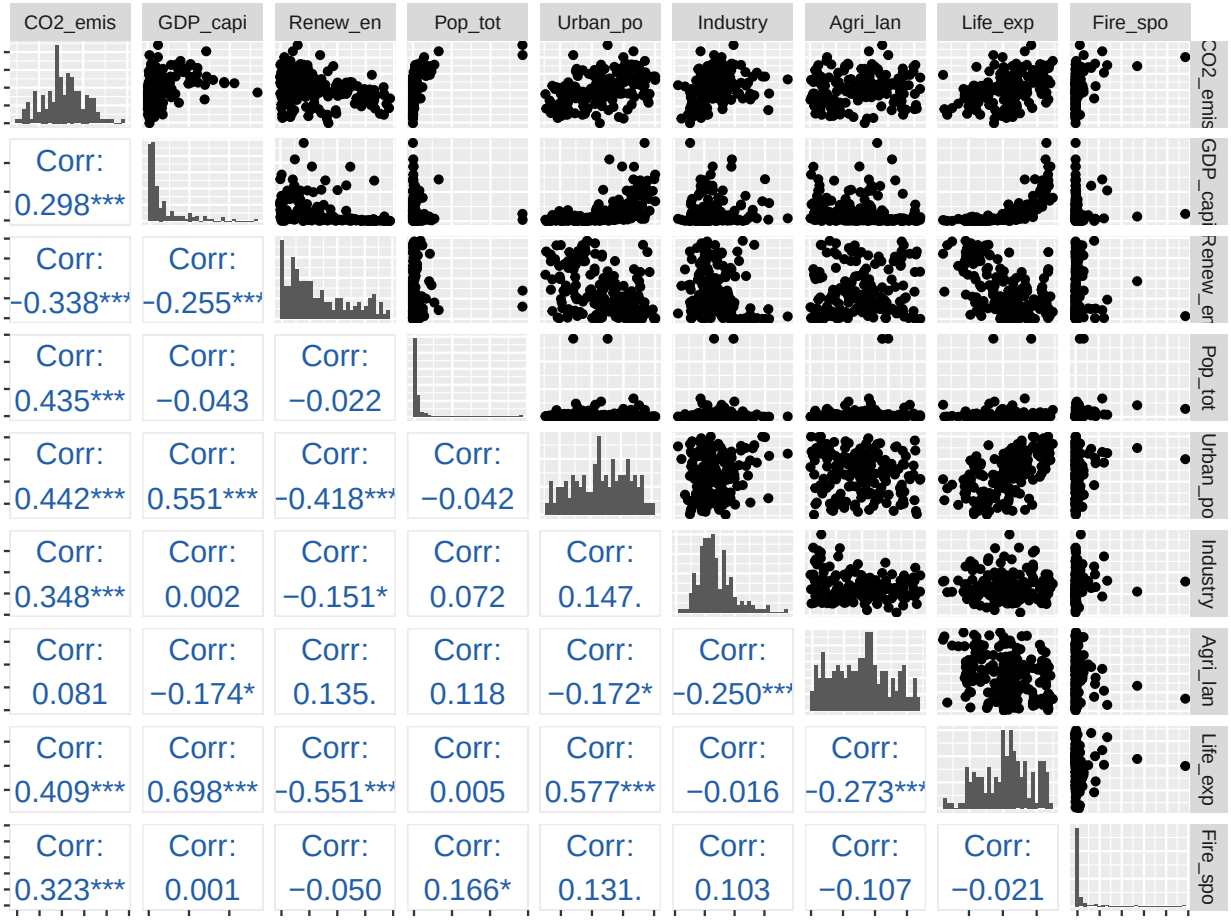

Figure 3: Scatter plots, Histograms and Correlation Matrix

The plot above shows, above the main diagonal, **scatter plots** between the response variable and some of the predictors. It is already apparent that a non-linear transformation of the {Fire spots} and *Pop tot* variables will likely be necessary to improve their linear relationship with the response. This is also noticeable from the skewed distribution visible in the **histograms** along the main diagonal for these two variables. A transformation could also improve the skewed distribution of *GDP capita*. However, since this variable was not selected in the initial variable selection, it will is not considered in the model. Below the main diagonal, we can find **correlation matrix**, with the highest correlations between 0.5 and 0.7 observed for *GDP capita* with *Urban pop* and *Life expectancy*, and between *Renewable energy* and *Life expectancy*.

2) Variable selection

2.1) Best Subset Selection

To select the most relevant predictors, **Best Subset Selection** was performed, evaluating all possible combinations to identify the best model based on the highest R^2. To reduce computational complexity, **Forward** and **Backward Stepwise Selection** were also applied, both leading to the same final model as Best Subset Selection.

Five selection criteria were considered to determine the optimal number of predictors. The **Mallow's C_p** statistic and **Akaike Information Criterion (AIC)** penalize model complexity, with lower values preferred. The **Bayesian Information Criterion (BIC)** applies a stronger penalty, favoring simpler models, so lower values are preferred. The **Adjusted R^2** corrects for the number of predictors and is preferred higher, as it indicates a better fit. Finally, the **Cross-Validation error** was computed, with lower values preferred to identify the best model.

After the initial variable selection, diagnostic analysis of the selected model revealed issues of **non-linearity**. To address this, quadratic terms were introduced for *Pop tot* and *Industry GDP*, along with a logarithmic transformation for *Fire spots*. Additionally, an interaction term was considered to allow for different slopes of *GDP capita* based on the continents.
Following these transformations, the variable selection process was repeated to reassess the optimal set of predictors.

```
n = nrow(dataset)
dataset$Fire_spots <- log(dataset$Fire_spots + (abs(min(dataset$Fire_spots)) + 1))
best_model <- regsubsets(CO2_emission ~ . + Continent*GDP_capita + I(Pop_tot^2)
                         + I(Ind_GDP^2), data = dataset[, -c(1,2)],  nvmax = 25)
summ <- summary(best_model)
aic_values = numeric(25)
for (k in 1:25) {
  aic_values[k] = summ$bic[k] - (k + 2) * log(n) + 2 * (k + 2)
}
```

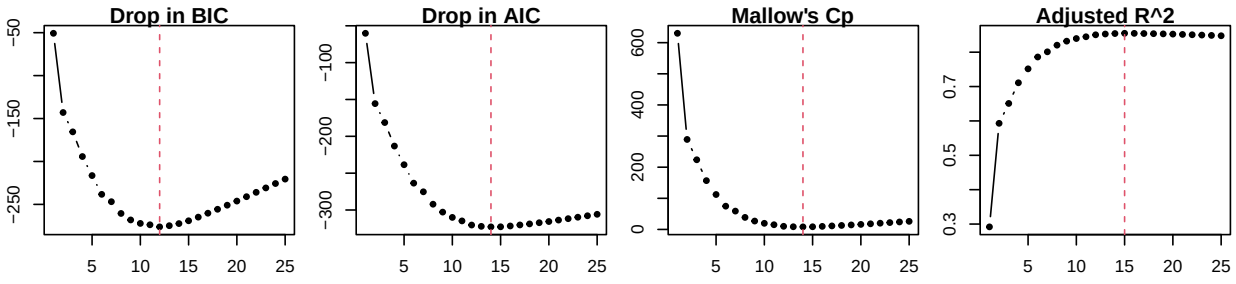

Figure 4: Model Comparison Criteria in the Best Subset Selection Approach

Based on the BIC graph, the model with **9 predictors** was selected:

(Intercept), *Ren_en*, *Pop_tot*, *Urb_pop*, *Ind_GDP*, *Agri_land*, *Life_exp*, *Fire_spots*, *I(Pop_tot^2)* and *I(Ind_GDP^2)*

Although the lowest BIC corresponds to a model with 12 predictors, the **principle of parsimony** suggests that, among models with similar BIC values, the one with fewer predictors should be preferred. This model also ensures compliance with the **principle of hierarchy**, as by selecting the previously added quadratic terms, it has also included their corresponding first-degree terms.
The model with 10 predictors would have added the categorical variable, but it was decided to exclude it in order to avoid adding three more predictors, thus keeping the model simpler.

2.2) Cross Validation error

We now apply **k-Fold Cross Validation**, which evaluates model performance by partitioning the data into k subsets. The model is iteratively trained on $k-1$ folds and validated on the remaining fold.
At each iteration, the *Mean Squared Error (MSE)* is calculated for each model with a given number of predictors and the MSE values for models with the same number of predictors are averaged. The model with the lowest average MSE is selected. In this case, each observation is assigned to one of $k = 10$ folds, and the results are used to assess model performance.

```
dataset.cv <- dataset[, -c(1,2)] #without Country name and Country code
p = 25
k = 10
set.seed(123)
folds = sample(1:k, n, replace = TRUE)
cv.errors = matrix(NA, k, p, dimnames = list(NULL, paste(1:p)))
for(j in 1:k) {
  best.fit = regsubsets(CO2_emission ~ . + Continent*GDP_capita + I(Pop_tot^2)
                        + I(Ind_GDP^2), data = dataset.cv[folds != j,],  nvmax = 25)
  for (i in 1:p) {
    mat = model.matrix(as.formula(best.fit$call[[2]]), dataset.cv[folds == j,])
    coefi = coef(best.fit, id = i)
    xvars = names(coefi)
    pred= mat[, xvars] %*% coefi
    cv.errors[j, i] = mean  ((dataset.cv$CO2_emission[folds == j] - pred) ^ 2)
  }
}
cv.mean = colMeans(cv.errors)
```

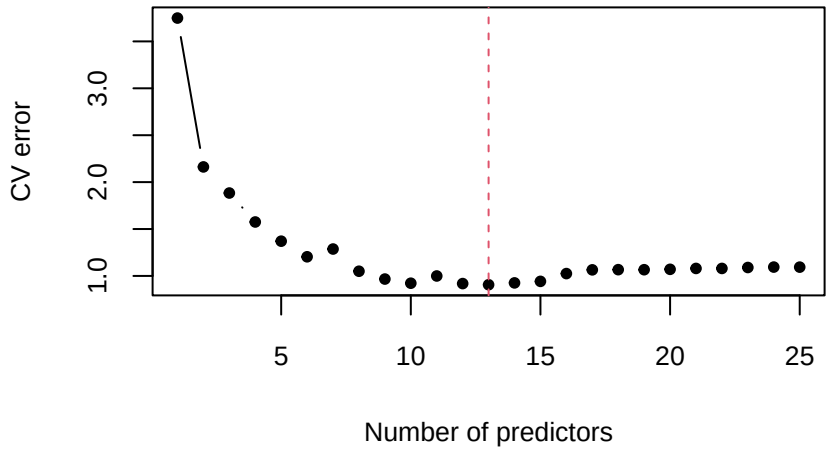

Figure 5: k-Fold Cross Validation error Criteria

The MSE starts to stabilize around the model with 10 predictors, but even in this case, the model with 9 predictors is preferred to ensure lower complexity and better alignment with the principle of parsimony.

The **final regression model fitted** for this analysis is:

$$\log(Y) = \beta_0 + \beta_1 x_{\text{Ren\_en}} + \beta_2 x_{\text{Pop\_tot}} + \beta_3 x^2_{\text{Pop\_tot}} + \beta_4 x_{\text{Urb\_pop}} + \beta_5 x_{\text{Ind\_GDP}} + \beta_6 x^2_{\text{Ind\_GDP}} + \beta_7 x_{\text{Agri\_land}} + \beta_8 x_{\text{Life\_exp}} + \beta_9 \log(x_{\text{Fire\_spots}}) + \epsilon$$

3) Collinearity

Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated, making it difficult to isolate the individual effect of each predictor. To identify potential collinearity issues, I computed the **Variance Inflation Factor (VIF)**, which quantifies how much the variance of an estimated regression coefficient is inflated due to collinearity in the model. The VIF is defined as:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ represents the coefficient of determination from the regression of X_j onto all other predictors. If R^2 is close to 1, it indicates that X_j is highly correlated with the other predictors, leading to a large VIF value. A VIF of 1 suggests no collinearity, while a VIF greater than 10 indicates a high level of multicollinearity.

Table 2: VIF values

Ren_en	Pop_tot	I(Pop_tot^2)	Urb_pop	Ind_GDP	I(Ind_GDP^2)	Agri_land	Life_exp	Fire_spots
1.626195	16.71272	15.44579	1.695499	1.781147	1.672997	1.24842	2.146622	1.800217

From the VIF values, we observe two high values. One of the main issues when adding a quadratic term is the potential correlation between the first-degree and second-degree predictors. While this does not occur for *Industry GDP*, it is present for *Pop tot*, which could lead to computational inaccuracies in the OLS estimation.

4) Diagnostics

A **diagnostic analysis** is performed on the model to assess whether the fundamental assumptions of linear regression hold and to identify any potential violations that could affect model performance.

4.1) Homoschedasticity

The variance of the residuals should remain constant across all levels of the predictors. The **residuals versus fitted values plot** is a *diagnostic tool* for assessing homoscedasticity. If the assumption holds, the residuals should have a constant spread around zero, forming a null plot. Any increasing or decreasing spread suggests **heteroscedasticity**, a violation of the constant variance assumption. This plot can also highlight potential non-linearity in the model's structure.

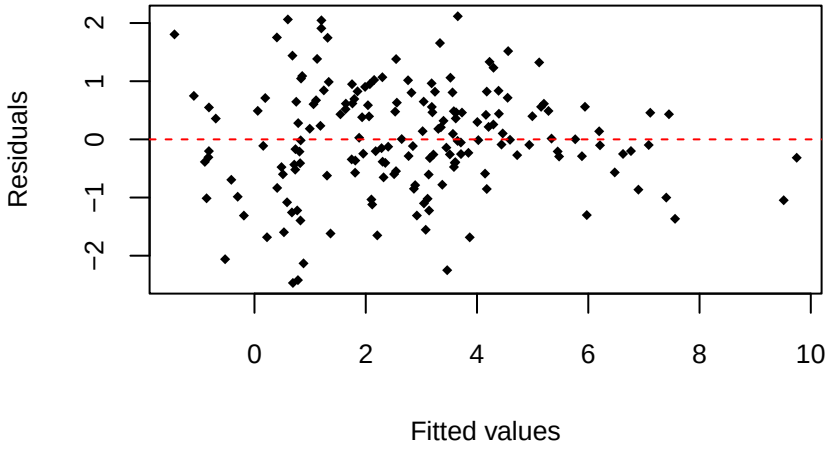

Figure 6: Residuals vs fitted value plot

The plot suggests that the assumption of linearity is approximately satisfied, thanks to the applied transformations. However, these modifications may have introduced *mild* heteroscedasticity, as indicated by the **left-opening megaphone** pattern.
A traditional solution for addressing heteroscedasticity would involve transforming the response variable using a *concave* function, which has already been applied in this case. Since the problem persists, it may be beneficial to try a **Weighted Least Squares (WLS)**, instead of the Ordinary Least Squares (OLS). However, it was preferred not to implement this solution, as the problem does not appear to be severe.

4.2) Linearity

The relationship between the predictors and the response variable should be linear. This assumption can be expressed as:

$$E(Y) = X\beta$$

where $E(Y)$ represents the expected value of the response variable, X is the matrix of predictors and β is the vector of coefficients. One of the most effective diagnostic tools for assessing the adequacy of this structural assumption is the **residual plots against individual predictors**. If the assumption of linearity holds, the residuals should be randomly scattered around zero, without systematic patterns.

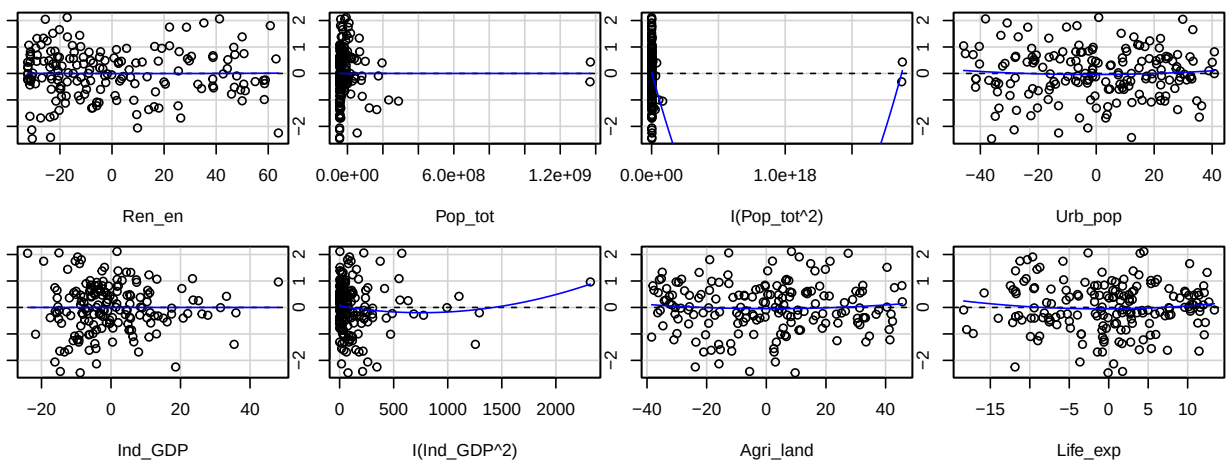

Figure 7: Residuals vs Individual Predictors plot

The residual plots indicate that the inclusion of quadratic terms has helped linearize the relationship between *Pop tot*, *Ind GDP* and the response variable. The residuals for the linear terms of these predictors exhibit a linear pattern, suggesting that the quadratic transformations have successfully captured part of the non-linearity. However, the residual plots for the quadratic terms themselves still show non-linearity, as evidenced by the curved trends. For the other predictors, no strong patterns emerge. The variable *Fire spots* was not plotted due to space limitations, but it satisfies the linearity assumption due to the previous logarithmic transformation.

4.3) Normality Assumption

The residuals should follow a normal distribution. The **Q-Q plot** is commonly used to assess this assumption by comparing the empirical quantiles of the residuals to those of a normal distribution. If the residuals are normally distributed, the points will align along a straight diagonal line, with deviations indicating potential issues like skewness or heavy tails.

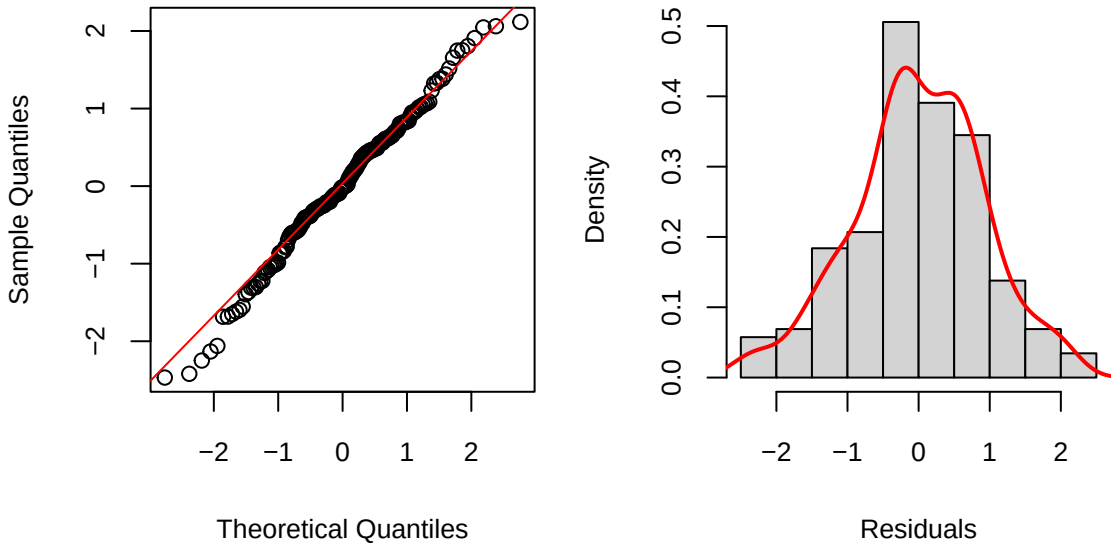

Figure 8: Normal Q-Q Plot on the right and the Histogram of Residuals on the left

As shown in the figure, the Q-Q plot indicates that the residuals approximately follow a normal distribution, as the points align closely with the red diagonal line. The histogram of the residuals further supports this observation, showing a bell-shaped distribution that reasonably fits the normal curve. Therefore, the assumption of normality for the residuals can be considered valid.

In addition to graphical diagnostics, the **Shapiro-Wilk test** provides a formal statistical evaluation of the normality assumption. This test assesses the null hypothesis that the residuals follow a normal distribution:

$$\begin{cases} H_0 : \epsilon \sim N(0, \sigma^2 I_n) \\ H_1 : \text{Otherwise} \end{cases}$$

Table 3: Shapiro-Wilk Normality Test

Test statistic	P value
0.9913	0.3736

The test result did not reject the null hypothesis, reinforcing the assumption of normality in the residuals.

This assumption allows us to make inferences about the model, supporting hypothesis testing and the construction of confidence intervals.

4.4) Correlation

Residuals should be independent and not exhibit autocorrelation. The spatial distribution map helps detect potential **spatial correlation** in the residuals. Ideally, residuals should be randomly distributed across regions without any specific pattern.

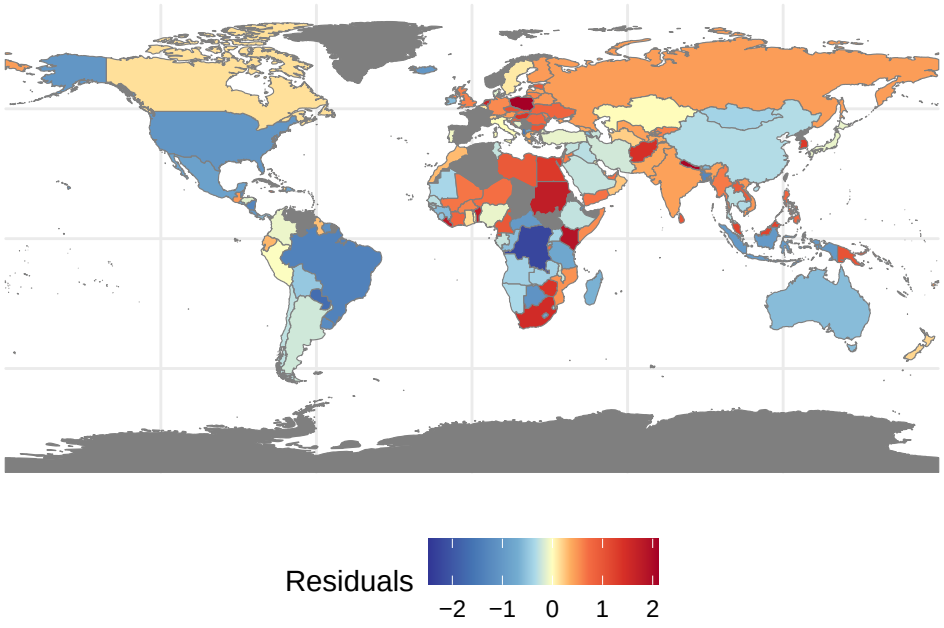

Figure 9: Spatial Distribution of Residuals

From the map, groups of similar residual values emerge, meaning that countries in the same geographical region tend to have residuals of the same sign and magnitude. Notably, patterns can be observed in regions like Europe, Africa and parts of South America. The presence of these spatial patterns indicates that the assumption of uncorrelated errors may be violated.
As a consequence, the estimated standard errors will tend to be *underestimated*, leading to narrower confidence and prediction intervals. Additionally, the p-values associated with the model will be lower than they should be, potentially leading to the *erroneous conclusion* that some parameters are statistically significant. A common remedy for dealing with correlated errors is to use **Generalized Least Squares (GLS)**, which adjusts for the correlation between residuals.

4.5) Unusual observation

Outliers, high-leverage and influential points should be identified to assess their impact on the model.

4.5.1) Outliers

An outlier is an observation that does not fit well within the assumed regression model. Outliers are typically detected using **standardized residuals**, which assess how far an observation deviates from the model's expected behavior. They are computed as:

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{jj}}}$$

A point is considered an outlier if its absolute studentized residual exceeds a certain threshold. Typically, a cutoff of 3 is used, meaning that if $|r_i| > 3$ the observation is likely an outlier.

In the model, the minimum and maximum values of the standardized residuals are:

-2.691 and *2.319*

Since they do not exceed the rule of thumb threshold, no outliers were detected.

4.5.2) High Leverage Points

A fundamental aspect of regression diagnostics is identifying **high leverage points**, which are data points with extreme values for the predictor variables and located far from the mean of the other data points in the predictor space. Leverage points are determined by the *hat values*, the diagonal elements of the hat matrix H, where $h_{ii} = x_i^T (X^T X)^{-1} x_i$. High leverage points are identified when the leverage value exceeds the threshold:

$$\text{Leverage Threshold} = \frac{2(p+1)}{n}$$

where p is the number of predictors and n is the number of observations.

In this model, the high leverage points are:

Brunei Darussalam, *China*, *Gabon*, *Indonesia*, *India*, *Lebanon*, *Libya*, *Nigeria*, *Pakistan*, *Timor-Leste* and *United States*

4.5.3) Influential Points

An **influential point** is a data point whose removal would significantly alter the fitted model. Unlike outliers, influential points impact the estimated regression coefficients and can distort the overall model interpretation. One of the most commonly used measures for detecting influential observations is **Cook's Distance**, which quantifies the influence of each data point on the estimated regression coefficients.

It is defined as:

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})^T(\hat{y} - \hat{y}_{(i)})}{p\hat{\sigma}^2} = \frac{r_i^2}{p}\frac{h_i}{1-h_i}$$

where r_i represents the standardized residual, h_i is the leverage of the observation and p is the number of predictors. The rule of thumb to detect influential points is: $0.5 \leq D_i < 1$ indicates that the observation is moderately influential; while $D_i \geq 1$ suggests that the observation is highly influential and could significantly affect the regression estimates.

Table 4: Max Cook Distance

Cook_Distance	Country
0.1277	Libya

Since the above value is less than 0.5, it can be said that no observation can be considered an influential point.

5) Best Model obtained

The summary of the best model obtained is the following:

```
##
## Call:
## lm(formula = CO2_emission ~ Ren_en + Pop_tot + I(Pop_tot^2) +
##     Urb_pop + Ind_GDP + I(Ind_GDP^2) + Agri_land + Life_exp +
##     Fire_spots, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.46833 -0.53990 -0.01517  0.61059  2.11620
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.277e+00  1.725e-01  13.197  < 2e-16 ***
## Ren_en       -1.161e-02  3.293e-03  -3.527 0.000546 ***
## Pop_tot       1.733e-08  1.870e-09   9.270  < 2e-16 ***
## I(Pop_tot^2) -1.075e-17  1.403e-18  -7.662 1.51e-12 ***
## Urb_pop       1.678e-02  4.143e-03   4.049 7.91e-05 ***
## Ind_GDP       6.486e-02  8.047e-03   8.060 1.51e-13 ***
## I(Ind_GDP^2) -1.675e-03  3.418e-04  -4.900 2.29e-06 ***
## Agri_land     1.678e-02  3.558e-03   4.717 5.10e-06 ***
## Life_exp      1.176e-01  1.402e-02   8.389 2.15e-14 ***
## Fire_spots    2.883e-01  3.648e-02   7.902 3.77e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9407 on 164 degrees of freedom
## Multiple R-squared:  0.8408, Adjusted R-squared:  0.832
## F-statistic: 96.22 on 9 and 164 DF,  p-value: < 2.2e-16
```

5.1) Intepretation of the Parameters

To facilitate interpretation, a table is provided showing the corresponding means of the predictors, which can be used for better understanding after centering.

Table 5: Mean of predictors before centering

Renew_energy	Pop_tot	Urban_pop	Industry_GDP	Agri_land	Life_expectacy	Fire_spots
32.41	44135651	59.22	26.75	38.94	71.06	608.5

The interpretation of parameters is as follows:

- $\beta_0 = 2.277 \iff e^{2.277} \approx 9.75$: This represents the expected CO_2 emissions (in megatons) for a typical country, where all predictors are at their mean value.
- $\beta_1 = -0.01161$: A **1 percentage point** increase in renewable energy consumption is associated with approximately a **1.16% decrease** in CO_2 emissions. This aligns with the expectation that a higher share of renewable energy reduces reliance on fossil fuels, leading to lower emissions.
- $\beta_2 = 1.733 \times 10^{-8}$. To better interpret the result, we consider an increase in population of **1 million**, which corresponds to a **1.73% increase** in CO_2 emissions. This implies that countries with larger populations tend to emit more CO_2, likely due to greater energy demand, transportation, and industrial production.
- $\beta_3 = -1.075 \times 10^{-17}$: The quadratic term indicates that the effect of population on CO_2 emissions is not linear, but the impact is extremely small, as β_3 is very close to zero. Therefore, the quadratic effect of population is negligible compared to the much more significant linear term β_2.
- $\beta_4 = 0.01678$: An increase in the urban population by **1 percentage point** leads to a **1.68% increase** in CO_2 emissions. Cities tend to generate higher emissions due to industrial activity, energy consumption and higher vehicle density. For example, if a country's urban population rises from 60% to 70%, CO_2 emissions are expected to increase by approximately $10 \times 1.68\% = 16.8\%$.
- $\beta_5 = 0.06486$: A **1 percentage point** increase in the share of GDP from industry leads to a **6.5% increase** in CO_2 emissions. This confirms that industrialization is a major driver of CO_2 emissions. So, if a country increases the industrial share of GDP from 30% to 40%, its CO_2 emissions would rise by: $10 \times 6.5\% = 65\%$.
- $\beta_6 = -0.001675$: The quadratic term suggests that the relationship between Ind_GDP and CO_2 emissions is **non-linear**. Initially, as Ind_GDP increases, CO_2 emissions also rise due to growing industrial activity. However, after a certain point, the effect of industrial GDP on CO_2 emissions becomes negative, indicating that further industrial growth leads to reduced emissions. This shift is likely due to the adoption of more efficient technologies and environmental policies in highly industrialized countries. To determine the point at which the effect changes sign, we computed the *derivative* of the regression function with respect to Ind_GDP. The derivative of $\beta_5 \cdot x_{\text{Ind\_GDP}} + \beta_6 \cdot x^2_{\text{Ind\_GDP}}$ was set equal to zero to find the turning point:

  $$\frac{\partial}{\partial x_{\text{Ind\_GDP}}} \left(\beta_5 \cdot x_{\text{Ind\_GDP}} + \beta_6 \cdot x^2_{\text{Ind\_GDP}}\right) = 0$$

  This leads to the equation:

  $$\beta_5 + 2 \cdot \beta_6 \cdot x_{\text{Ind\_GDP}} = 0$$

  Solving for Ind_GDP, we find that the *turning point* occurs at approximately 19.36%. This means that when the industrial share of GDP exceeds 19.36%, the effect of further industrialization on CO_2 emissions becomes negative. Before this point, industrialization increases emissions, while beyond this

threshold, the effect is reversed, and emissions decrease. Thus, countries with an industrial GDP share greater than 19.36% may experience a reduction in emissions as industrialization progresses, typically reflecting more economically developed nations. This decrease is likely due to improvements in efficiency, the adoption of sustainable practices, and the implementation of environmental policies. In contrast, emerging economies experience significant increases in emissions with industrialization, while in more advanced economies, the marginal impact of industry on emissions is lower, often due to better infrastructure and emission reduction policies.

- $\beta_7 = 0.01678$: A **1 percentage point** increase in agricultural land leads to a **1.68% increase** in CO_2 emissions. This is likely due to deforestation and the intensive use of fertilizers, both of which contribute to the release of greenhouse gases such as methane.
- $\beta_8 = 0.1176$: An increase of **1 year** in life expectancy is associated with a **11.76% increase** in CO_2 emissions. This suggests that countries with longer life expectancy are often more economically developed, probably leading to higher energy consumption, industrial activity and transportation needs.
- $\beta_9 = 0.2883$: A **1 unit** increase in **Fire spots** is associated with a **28.83% increase** in CO_2 emissions. This highlights the significant impact of forest fires on atmospheric CO_2 levels, as large-scale burning releases substantial amounts of carbon stored in vegetation.

5.2) Uncertainties

The **standard errors** of the coefficients are relatively small, suggesting precise estimates. The **Residual Standard Error (RSE)** of 0.9407 is also small in relation to the response variable's scale, indicating a good fit of the model.

Additionally, the **confidence intervals**, obtained using the confint() function, show that all variables are significant since their intervals do not include zero. This indicates that their effect on the dependent variable is statistically different from zero at the 95% confidence level. Furthermore, the confidence intervals are consistent with the sign of the relationship between each predictor and the response variable, aligning with the results of the corresponding t-tests, which will be performed subsequently.

5.3) t-Test

To assess whether a predictor significantly affects the dependent variable, we perform a **t-Test**, testing if a coefficient $\hat{\beta}_j$ is different from zero. The hypotheses are:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

To reject H_0, we examine the **p-value** in the `summary()` output. If $p < 0.05$, we reject H_0, indicating that $\hat{\beta}_j$ is **statistically significant**. If $p \geq 0.05$, we fail to reject H_0, suggesting that $\hat{\beta}_j$ is **not significant**. Since all p-values in our model are below 0.05, all predictors are highly statistically significant, supporting the conclusion that they meaningfully affect the response.

5.4) Comparing Models

I compared two models to evaluate the impact of adding quadratic terms for specific predictors, corresponding to the coefficients β_3 and β_6, using an **ANOVA test**. So the *narrower* model will be:

$$\log(Y) = \beta_0 + \beta_1 x_{\text{Ren\_en}} + \beta_2 x_{\text{Pop\_tot}} + \beta_3 x_{\text{Urb\_pop}} + \beta_4 x_{\text{Ind\_GDP}} + \beta_5 x_{\text{Agri\_land}} + \beta_6 x_{\text{Life\_exp}} + \beta_7 \log(x_{\text{Fire\_spots}}) + \epsilon$$

The hypothesis test is defined as follows:

$$\begin{cases} H_0 : \beta_3 = 0, \quad \beta_6 = 0 \\ H_1 : \beta_3 \neq 0, \quad \beta_6 \neq 0 \end{cases}$$

A **small p-value** from the ANOVA test leads to rejecting H_0, indicating that the quadratic terms significantly improve the model. Conversely, a **large p-value** suggests that the simpler model is preferable.

```
## Analysis of Variance Table
##
## Model 1: CO2_emission ~ Pop_tot + Urb_pop + Ind_GDP + Agri_land + Life_exp
## Model 2: CO2_emission ~ Ren_en + Pop_tot + I(Pop_tot^2) + Urb_pop + Ind_GDP +
##     I(Ind_GDP^2) + Agri_land + Life_exp + Fire_spots
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    168 397.83
## 2    164 145.12  4    252.71 71.395 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test results show that adding the quadratic terms for *Pop tot* and *Industry GDP* significantly improves the model. With an exceptionally high F-statistic and a small p-value, well below the 0.05 threshold, we reject the null hypothesis. This strong evidence confirms that the quadratic terms significantly contribute to explaining CO_2 emissions. Consequently, the model with quadratic terms is preferred over the simpler linear model.

5.5) Goodness of fit

To assess how well our model explains the variability in the dependent variable, we use the **coefficient of determination** R^2, which measures the proportion of variance in Y explained by the predictors:

$$R^2 = \frac{\mathrm{SS}_{\mathrm{Reg}}}{\mathrm{SS}_{YY}} = 1 - \frac{\mathrm{RSS}}{\mathrm{SS}_{YY}}$$

where RSS is the residual sum of squares and SS_{YY} is the total variance in Y. Since R^2 increases with more predictors, we also consider the **adjusted** R^2, which accounts for model complexity by penalizing unnecessary variables:

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

In our model, $R^2 = 0.8408$ and $R^2_{\mathrm{adj}} = 0.832$, meaning that **84.04%** of the variability in CO_2 emissions is explained, with the adjusted value confirming that the selected predictors are meaningful without overfitting.

6) Prediction

Suppose now that we have a new observation, representing a new country, with a population of **35,765,987**, an urban population percentage of **65%**, an industrial GDP share of **48%**, an agricultural land percentage of **35%**, a life expectancy of **76 years** and fire spots equal to **650**.

Since our model was trained using centered predictors, we must apply the same transformation to ensure consistency. This means subtracting the mean of each variable from the original dataset before making predictions.

```r
prediction <- predict(model, newdata = newdata, interval = "prediction", level = 0.95)
```

Table 6: Prediction

fit	lwr	upr
4.358	2.463	6.253

The predicted value and its confidence interval are based on input values that have been adjusted by subtracting their mean, as our model was trained with centered predictors. The means were taken from the original dataset before the variables were centered. This ensures that the predictions are consistent with the transformation applied during model training. The estimated CO_2 emissions for the selected values is **11.2**, with a **95% prediction interval** ranging from **8.781 to 13.63**.

7) Simulate n data points

The scatterplot compares the observed CO_2 emissions (x-axis) with the **simulated response** values (y-axis), which are generated using the estimated model parameters. Each point represents an observation, while the red dashed line represents the ideal relationship 1:1, where the simulated values perfectly match the observed ones.

```r
set.seed(123)
X <- model.matrix(model)
beta_hat <- coef(model)
sigma_hat <- summary(model)$sigma
Y_sim <- X %*% beta_hat + rnorm(n, mean = 0, sd = sigma_hat)
```

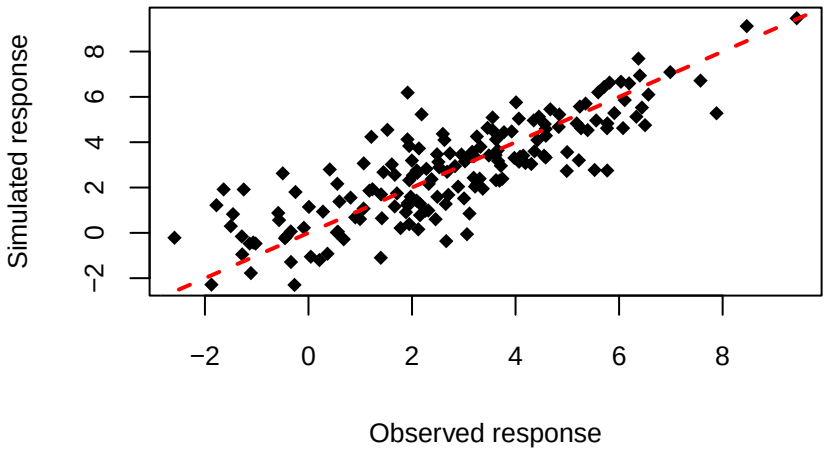

Figure 10: Scatterplot of Y obs and Y sim

The results suggest that the simulated responses align reasonably well with the observed values, indicating that the model captures the overall trend in the data. However, deviations from the red line suggest that some variability in CO_2 emissions is not fully explained. This may be due to omitted relevant variables or violations of model assumptions, such as correlated errors.