# Energy Dataset - Applied Linear Model Project

Sveva Maria Martilotti

2024-03-07

## INTRODUCTION

This dataset consist of 46 observations of 11 different variables, dividing into vital aspects such as electricity access, renewable energy, carbon emissions, populations and economic growth, comparing european and not european nations in 2020.

More precisely,my dataset initially considered all countries in the world, but I decided to focus only on countries belonging to IEA(International Energy Agency) and OECD (Organization for Economic Cooperation and Development), this is because countries differ a lot not only in economic, demographic, social, but also in political terms, and I wanted to focus on countries that had at least some ideas in common: aligned energy policy, sustainability and energy efficiency.

Why IEA and OECD? The goal of understanding the factors influencing renewable energy production is shared by both the EU and the IEA and OECD. Involving countries from both the IEA and OECD provides a broader and more diverse view of the factors at play. This approach allows a wide range of contexts, policies, and economic conditions to be considered. In this way I can analyze whether or not I belong to the EU (and thus follow certain policy rather than others) and how it affects the response variable.

The dataset has been downloaded form Kaggle, but the data come from the World Bank, International Energy Agency, and ourworldindata.org (whose team of researchers is based at the University of Oxford), which are secure and institutional sources.

## GOAL

The goal is to understand what factors affect the amount of renewable energy (in Twh) produced in a country; for example, could be that richer countries have more resources to invest in renewable energy production. Or it is logical to assume that countries with a higher proportion of electricity from low-carbon sources also have higher renewable energy production.

## DATA

The variables I'm going to consider are the following:

1. **Access to electricity (wrt the population)**: percentage of the population that had access to electricity in a certain country, in 2020.
2. **Electricity from fossil fuels (TWh)**: The amount of electricity produced from fossil fuels (Coal, oil and natural gas) expressed in terawatt-hours in a certain country, in 2020.

3. **Electricity from nuclear (TWh)**: The amount of electricity produced from nuclear power stations expressed in terawatt-hours in a certain country, in 2020.

4. **Electricity from renewables (TWh)**: It's my response variable, and measure the electricity generated from renewable sources (hydro, solar, wind, geothermal and biomass) in terawatt-hours in a certain country, in 2020.

5. **Low-carbon electricity (%)**: The percentage of electricity produced using sources that generate a reduced amount of greenhouse gas emissions compared to traditional sources with respect to the total energy, in a certain country, in 2020. Includes both renewable and nonrenewable sources with lower carbon emissions (e.g., nuclear and natural gas with carbon capture technologies).

6. **Primary energy consumption per capita (Twh/person)**: Average amount of primary energy used per person, in a certain country, in 2020. The primary energy is the energy found in the nature that has not been subjected to any human engineered conversion process, espressed in Twh.

7. **Renewables (% equivalent primary energy)**: The percentage of renewable energy with respect to the primary energy. However, I decided to disregard this variable since it's calculated based on the response variable and this could lead to interpretation problems and insufficient study.

8. **Gdp_per_capita**: Gross domestic product (GDP) is a measure of the economic activity, defined as the value of all goods and services produced less the value of any goods or services used in their creation. It's calculated by dividing the GDP of a nation by its population and it's expressed in dollars.

9. **Density/n(P/Km2)**: Population density per square kilometer of a certain country.

10. **Land Area(Km2)**:Total land area in square kilometers of a certain country.

11. **Entities**: countries that we are taking into consideration for this analysis.

Among those variables one is categorical: **Entities**, but since each country is on its own, I replaced this variable with **Country members**, which groups countries according to its location (Europe, not Europe,where the latter means that the country does not belong to the European Union, but belongs to the IEA or to the OECD).

## Graphical rapresentation of the data

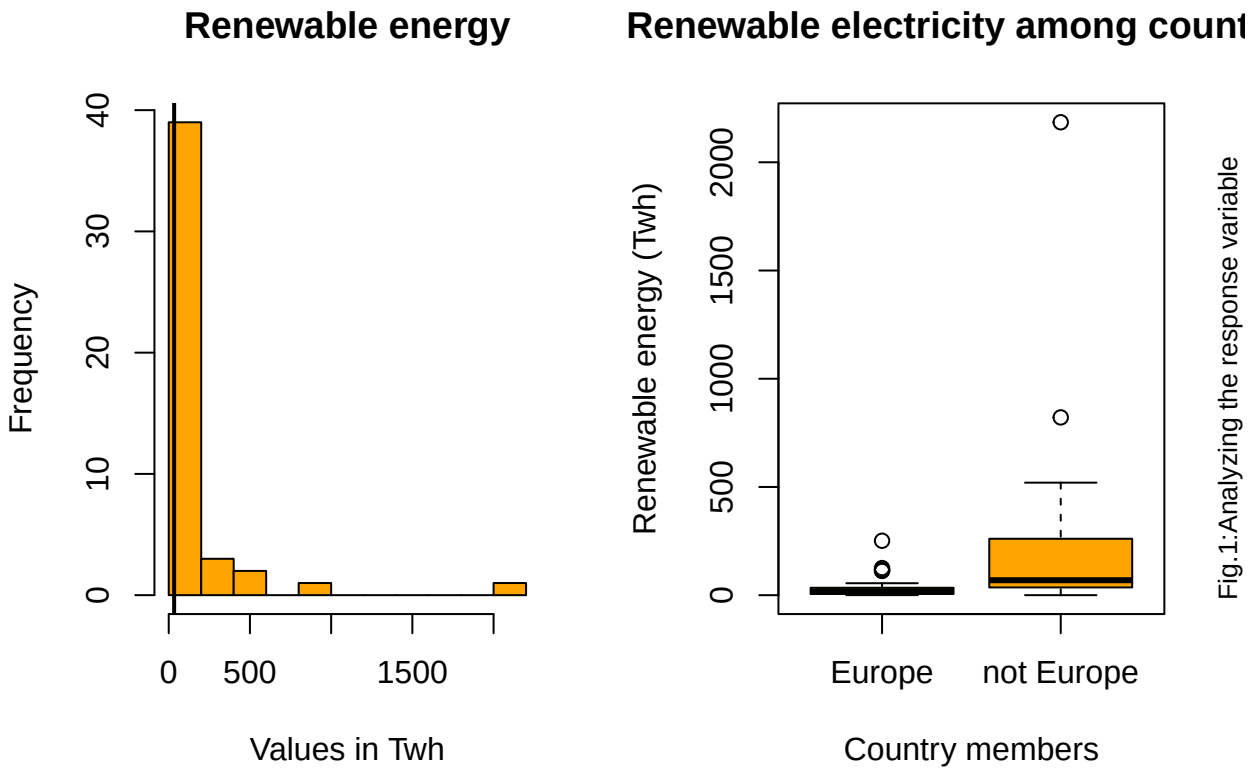


Fig.1:Analyzing the response variable

In *Fig.1* we can see two graphs:

In general, the *histogram* shows the distribution of the response variable and its frequency. The vertical line represents the median, that is, the central value of the response. We see that the distribution of renewable energy values is skewed to the right. In particular, most countries produce about 50 Thw renewable energy or less.

In the *boxplot*, we compare the distribution of renewable energy between the two categories. Note that there is greater variability among not European countries (compared to European Countries). But I already expected this since the countries I am considering are very different from each other, both in terms of wealth and size, but also in terms of legislation.

- Among non-Europeans there are also very large countries compared to the average size of European countries, such as China, United States, Indonesians and Australia, and I expect that these, being larger, will also produce more since they will need more energy.

The dots in the boxplot represent values that are very different from the range of values in the response:

For example, for Europe, one of the dots represents **Germany** that since 2000, thanks to a law (*Erneuerbare-Energien-Gesetz law*), established financial incentives to promote the development of renewable energy; now, despite being among the countries with the fewest hours of sunshine, Germany is one of the world's largest producers of solar energy, according to AIE (The International Energy Agency).

Other points:**France**, **Italy** and **Spain** and **Sweden** (fist for hydropower in Europe and for years has set a goal of achieving 100% renewable energy).

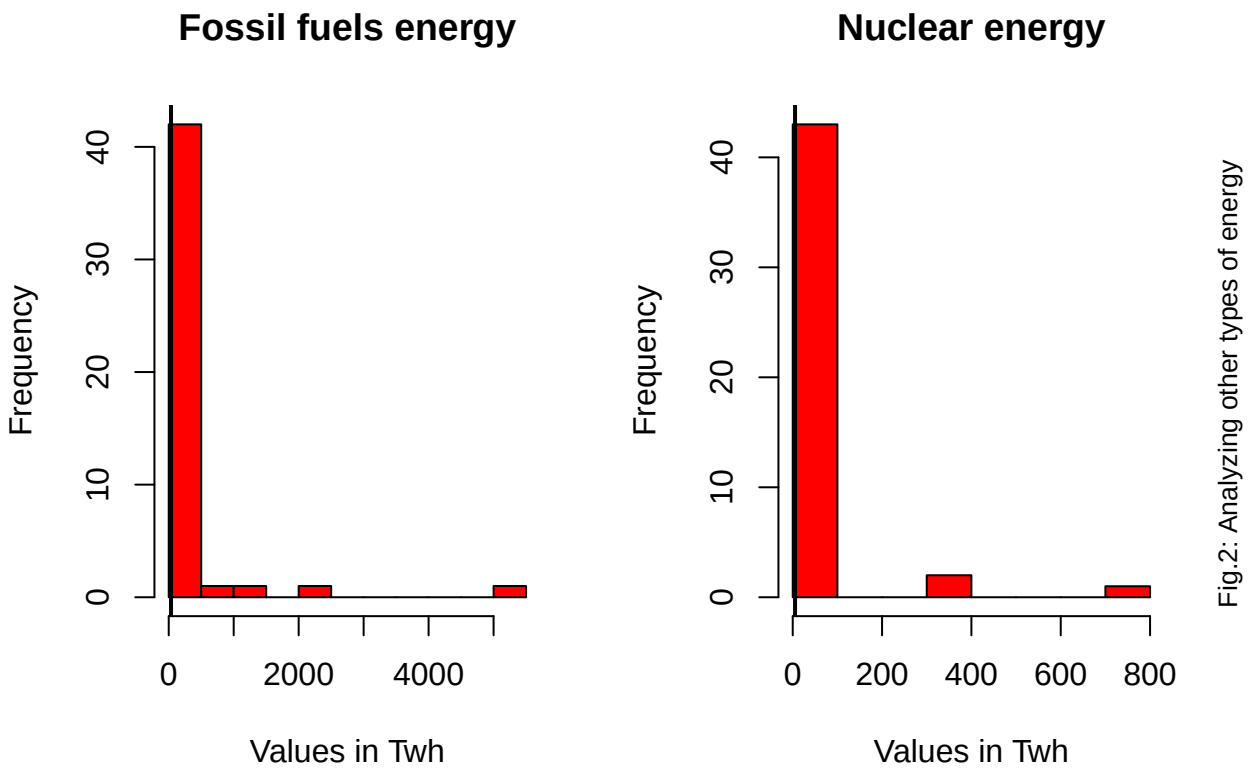


Fig.2: Analyzing other types of energy

In the *Fig.2* we learn something about the other types of energy: nuclear and fossil fuels.

Again, we can see from the histograms that the amount in Twh produced is very different from country to country, with more than most countries producing between 50 and 100 Twh of nonrenewable energy.

In particular, there are two countries that produce about 2000 and 5000 Twh of energy from fossil fuels, and they are the **U.S.** and **China**, respectively.

Such huge values compared to others are not a mistake, but depend on the fact that both countries have significant resources of coal, oil and natural gas, have very large and developed economies that require a large amount of energy to fuel their industries and especially energy policies.

The same can be observed for nuclear energy.

# FIT THE MODEL

First I studied how my quantitative variables interacted with each other.

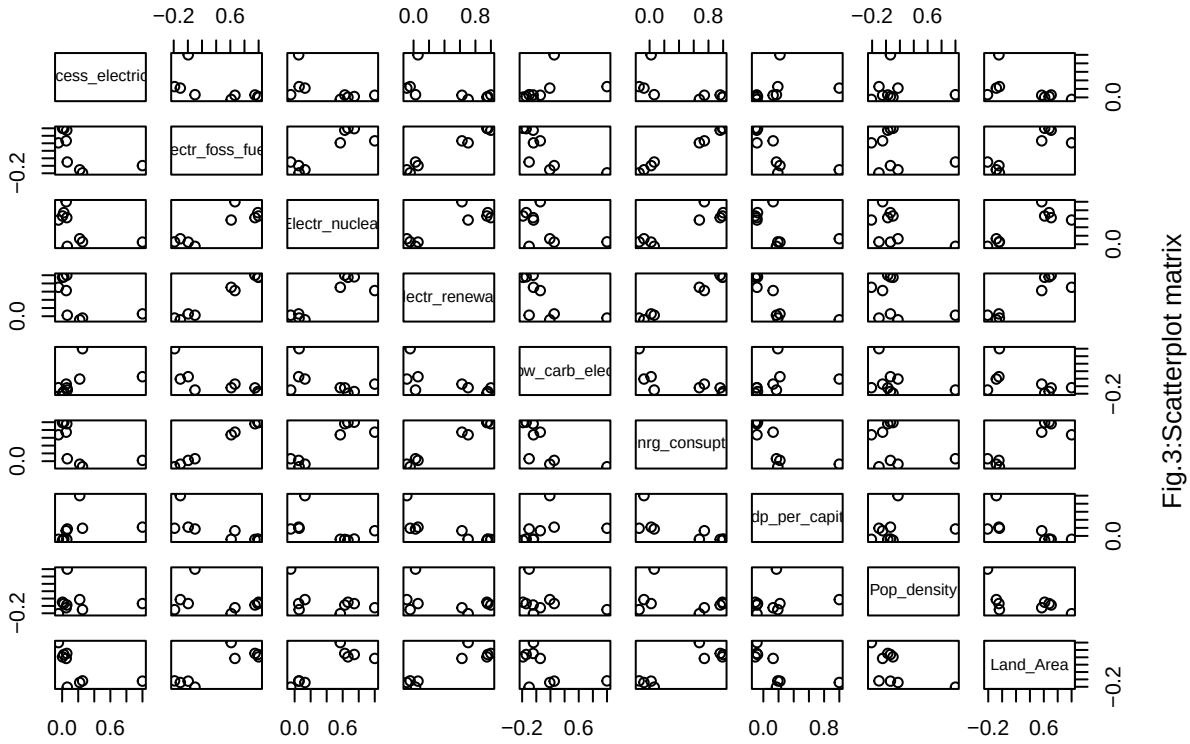


*Fig.3* shows both the bivariate relationship between renewable energy (response) and the indipendent quantitative variables, but also the relationship between the variables themselves. Moreover, it also shows unusual observations such as outliers, high leverage point and influential points.

Most of the variables are poorly correlated with each other: some show a slight positive or negative correlation with the exception of **Electric renewable**, **Fossil fuels** and **nuclear energy** that have a strong positive relationship.

Moreover,I decided to remove *Primary energy consumption per capita* because it has a strong multicollinearity wrt "Electric Fossil Fuels" and "Electric Nuclear Energy".

I decided not to conbine this variable with the others, because I wanted to consider their impact without the influence of the collinear variable, and anyway a lot of the information in the "Primary energy" variable is already given by "fossil fuels" and "nuclear".

Moreover, I decided to not combine "Electric fossil fuels" and "Electr_nuclear" into a single variable there are many countries that do not produce nuclear energy, but produce a lot of energy from fossil fuels, so I didn't want to equate the two. Anyway by removing "Primary energy" the multicollinearity problem is solved.

# The model

```
OLS=lm(Electr_renewabl~Access_electricity+dat_centered$Electr_nuclear+dat_centered$Electr_foss_fuels+Lo
```

*Note*: I centered the independent variables.

As we can see, my model contains 9 variables (including the intercept) and to avoid overfitting problems and to simplify interpretation I performed variable selection by looking for the best model.

I also considered doing an interaction between "fossil fuels" and "country members" and between "nuclear energy" and "country members", to understand how membership in a group of countries (Europe or non-Europe) affects the relationship between fossil energy/nuclear energy and renewable energy production. For example, European countries might have more renewable energy policies than non-European country and this could lead to higher production of energy.

Through an ANOVA test I saw that the interaction has a significant impact on the response (low p-value),this means that the effect of the production of fossil fuels electricity on the production of renewable electricity in Twh depends on whether or not the countries are members of the EU; however, by calculating the VIF I saw that the variables with the interaction were multicollinear with each other, so I decided not to include the interaction, to not affect the variance and the p-value of the estimated values.

# Best model

In order to determine which model is optinal to fit (and which regressor), I performed a subset selection taking in consideration different criteria: AIC, BIC, Adjusted R^2 and Cp Mallow.

In the following table, we can see for each model how many predictors, and which, the algorithm takes into account (looking to *).

```
ols<-regsubsets(Electr_renewabl~Access_electricity+Electr_foss_fuels+Electr_nuclear+Low_carb_electr+Pop
summary(ols)
```

```
## Subset selection object
## Call: regsubsets.formula(Electr_renewabl ~ Access_electricity + Electr_foss_fuels +
##     Electr_nuclear + Low_carb_electr + Pop_density + gdp_per_capita +
##     Land_Area + Country_members, data = dat_centered)
## 8 Variables  (and intercept)
##                            Forced in Forced out
## Access_electricity             FALSE      FALSE
## Electr_foss_fuels              FALSE      FALSE
## Electr_nuclear                 FALSE      FALSE
## Low_carb_electr                FALSE      FALSE
## Pop_density                    FALSE      FALSE
## gdp_per_capita                 FALSE      FALSE
## Land_Area                      FALSE      FALSE
## Country_membersnot Europe      FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          Access_electricity Electr_foss_fuels Electr_nuclear Low_carb_electr
## 1  ( 1 ) " "                "*"               " "            " "
## 2  ( 1 ) " "                "*"               " "            " "
## 3  ( 1 ) " "                "*"               " "            "*"
## 4  ( 1 ) " "                "*"               "*"            "*"
## 5  ( 1 ) "*"                "*"               "*"            "*"
## 6  ( 1 ) "*"                "*"               "*"            "*"
## 7  ( 1 ) "*"                "*"               "*"            "*"
## 8  ( 1 ) "*"                "*"               "*"            "*"
##          Pop_density gdp_per_capita Land_Area Country_membersnot Europe
## 1  ( 1 ) " "         " "            " "       " "
```

```
## 2  ( 1 ) " "         " "           "*"         " "
## 3  ( 1 ) " "         " "           "*"         " "
## 4  ( 1 ) " "         " "           "*"         " "
## 5  ( 1 ) " "         " "           "*"         " "
## 6  ( 1 ) "*"         " "           "*"         " "
## 7  ( 1 ) "*"         "*"           "*"         " "
## 8  ( 1 ) "*"         "*"           "*"         "*"
```

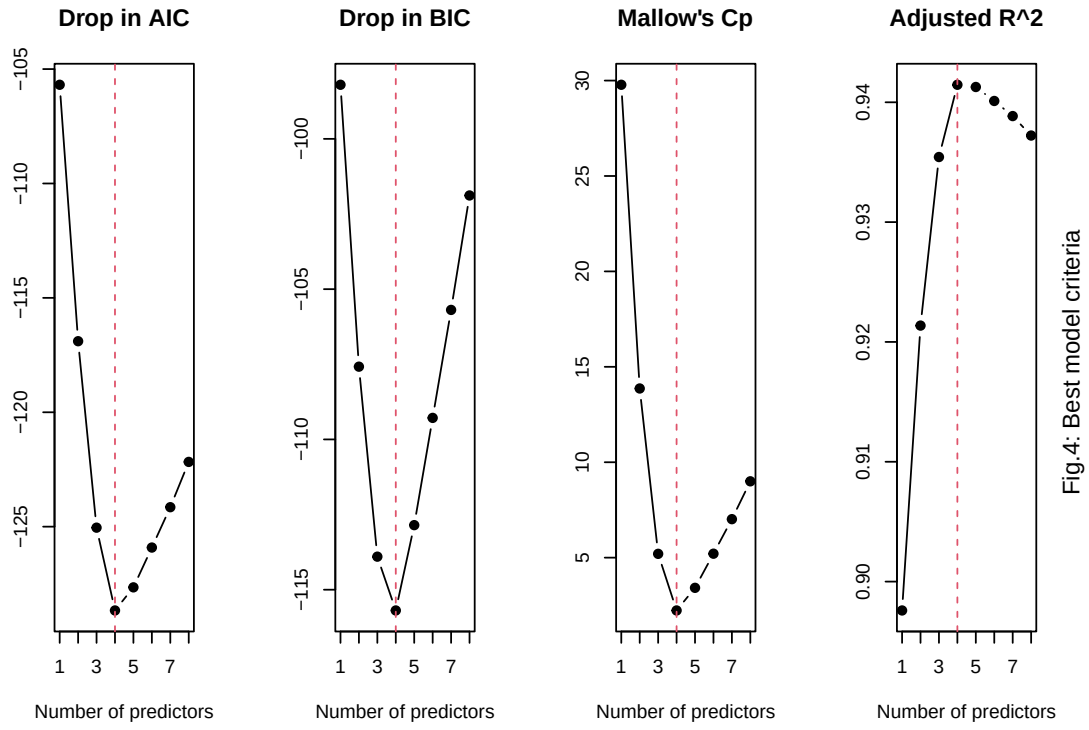


Fig.4: Best model criteria

*Fig.4* shows 4 different plots with the optimal value according to each indicator:AIC, BIC, Cp and Adjusted R^2.

Each indicator has a different goals e.g., BIC penalizes more complex models (with more variables), AIC better balancing complexity with data fit. Instead, Adjusted R^2 measure how much of the variability of the response variable is explained by the model, taking into consideration the numbers of indipendent variable (allowing comparison with different models ).

Trying also with the forward and backward method would see that the result are the same.

Since all 4 criteria confirmed a model with 4 variables my best model will be as follows:

renewable electricity ~ nuclear energy + fossil fuels energy +low carbon electricity +Land Area

## Cross Validation

So far we have looked at how the model fits the data, if we want to also consider its goodness of prediction then we need to look at CV errors.

The are 3 different methods to perform cross validation(validation set, LOOCV e k-fold method); I decided to perform the LOOCV cause I wanted to use all my data and thus increase precision.

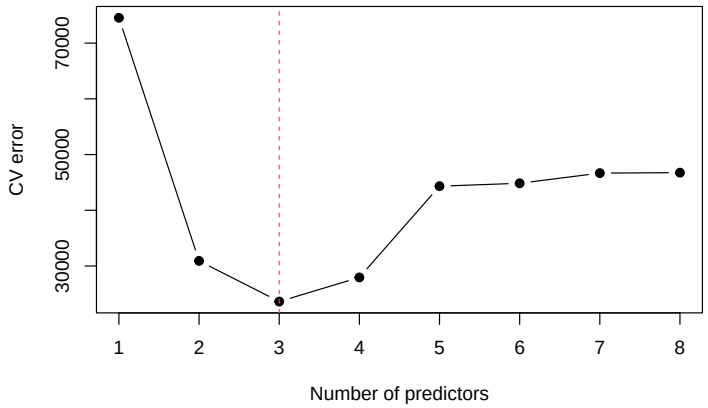

Fig.5: LOOCV method

*Fig.5* shows the value of CV error for each model with k variables. The best is the one to which the lowest value corresponds, in this case the model with 3 predictors. Since prediction is not the main goal of this analysis and the CV error value for the model 4 is quite close to the excellent one, I keep model 4.

## Collinearity issue

To study the presence of collinearity between the independent variables, we can use the following tools:

1. **correlation matrix**(heatmap): shows correlation only between two variables.
2. **VIF**: more accurate, since it takes into account the whole regression model.

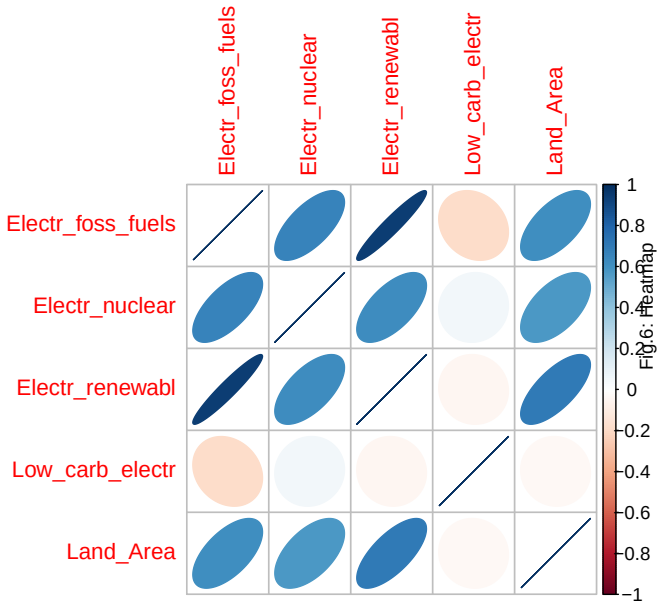

Fig.6: Heatmap

```
OLS_best<-lm(Electr_renewabl~Electr_nuclear+Electr_foss_fuels+Low_carb_electr+Land_Area,data=dat_center
vif(OLS_best)
```

```
##    Electr_nuclear Electr_foss_fuels   Low_carb_electr         Land_Area
##          2.051619          2.293626          1.105733          1.732478
```

Since the VIF values are all around 1 and 2, there is no multicollinearity issue in my model.

# DIAGNOSTIC

## Linearity assumption

Each graph represents a variable in my model against the Pearson residuals(a standardized version of the model residuals): this tool is useful to see how the residuals are distribuited wrt the fitted values.

```
residualPlots(OLS_best,test=F)
mtext("Fig.6:Linearity for each regressor", side=3, line=-22.3, outer=TRUE, cex=0.8, col="black")
```

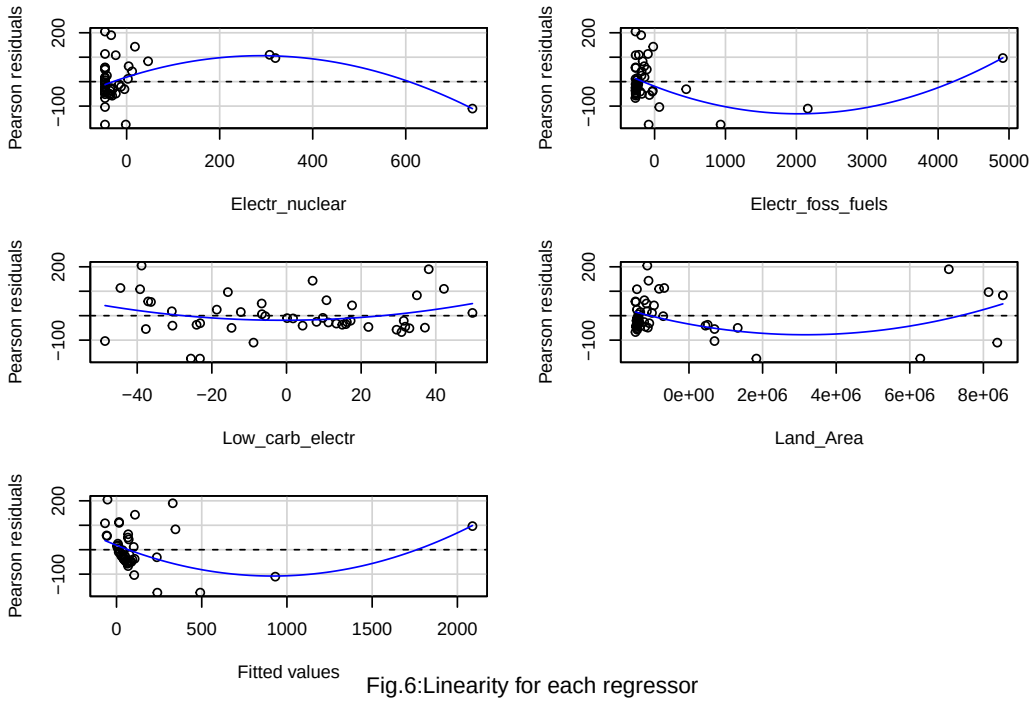


Fig.6:Linearity for each regressor

From *Fig.6* we can see the plot of the residuals for each regressor and their trend: variables do not seem to have a linear relationship, except for "Low Carbon Electricity".

In order to improve linearity, we can try transforming the indipendent variables.

## Homoscedasticity and normality

To check for homoscedasticity, which is the constancy of error variance, we can use again a residuals plot. This tool is also useful to check the linearity assumption.

However, in the plot we see residues, not errors: because the homoscehdasticity assumption is satisfied, the residues must be randomly located around 0.

From *Fig. 7* we can see that the assumption of linearity is not satisfied, while regarding heteroscedasticity, we are unable to discern a clear path due to the number of observations.

Normality, however, does not seem to be entirely satisfied: Looking at the Normal Q-Q Plot, most of the points appear to align along the straight line, suggesting that the residuals might be normally distributed; however, there are some points that deviate from the line, suggesting the presence of some residuals that do not follow the normal distribution.

## Improve the model

```
OLS_sqrt <- lm(sqrt(Electr_renewabl) ~ Electr_nuclear+Electr_foss_fuels+ Low_carb_electr + Land_Area, da
```

First of all I tried to fix the linearity with a square transformation of nonlinear variables and I didn't solve the problem. Same for the polynomial, that solved the linearity but penalized too much the normality.

However, as we can see in *Fig.7* and *Fig.8*, I found a better balance between normality and linearity using a sqrt transformation of the response variable, as the normality was not fully satisfied, and this also fixed the linearity.

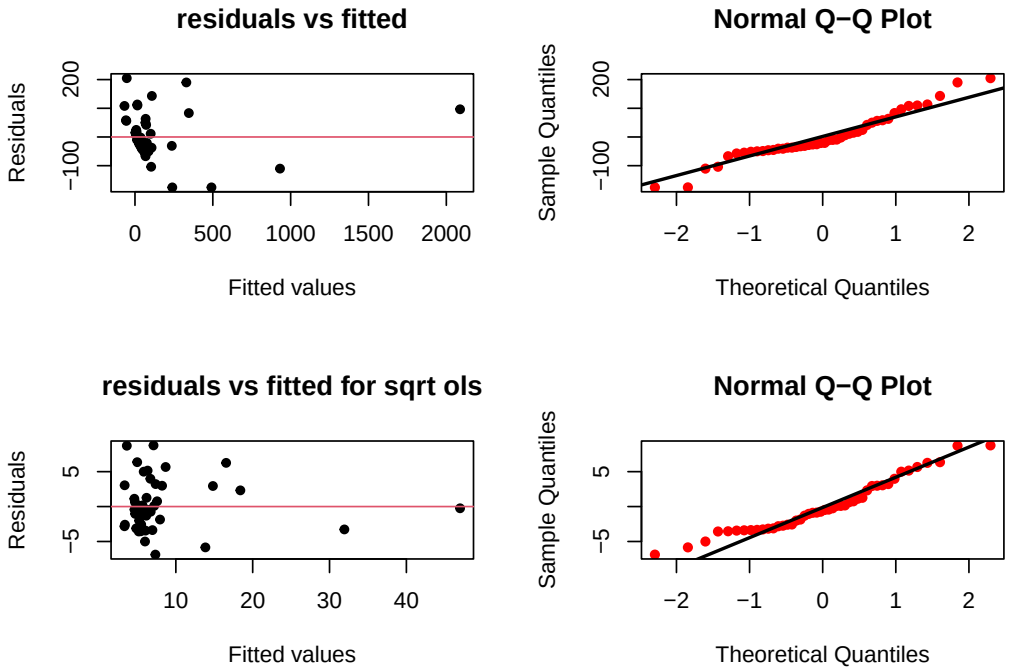

Fig.7:Homoscedasticity and nornality of the model

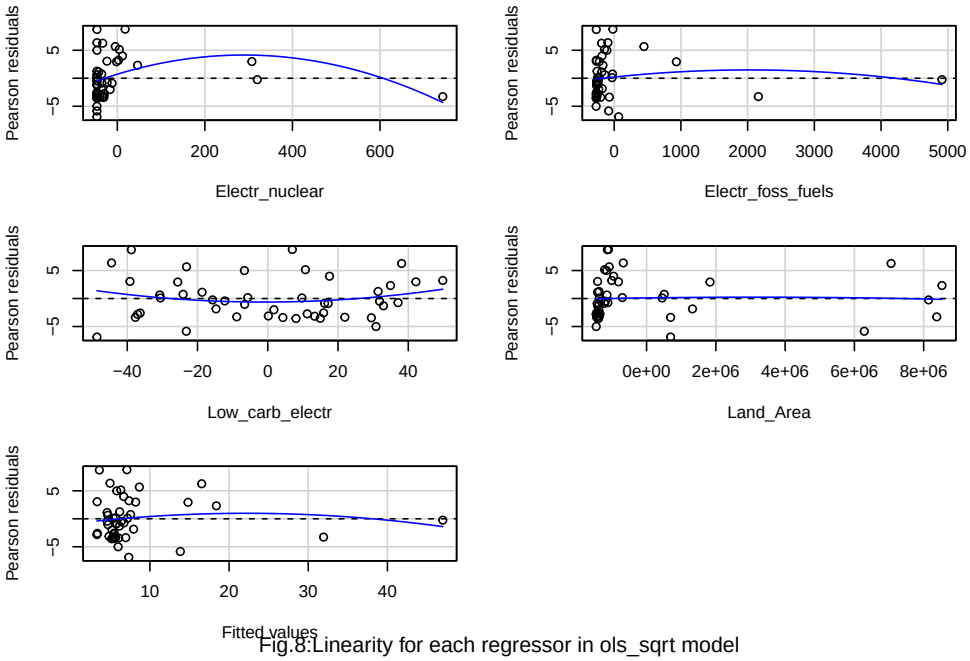


Fig.8:Linearity for each regressor in ols_sqrt model

## Unusual observations: outliers, high leverage point and influential points

**Outliers** are defined as points that don't fit the data well.

Specifically, a point is an outlier If its studentized residual is bigger then 3 or lower then -3.

A **leverage point** is a point whose its hii>2(p+1)/n = 2(4+1)/46=0.21. For hii we refers to the values on the diagonal of the hat matrix.

Finally, a point is **influential** if its Cook's Distance it's higher than 1.

In order to check for unusual observations, we set the tresholds mentioned above and plot them.

```
r_standard <- rstandard(OLS_sqrt)
hat <- hatvalues(OLS_sqrt)
cook <- cooks.distance(OLS_sqrt)
```

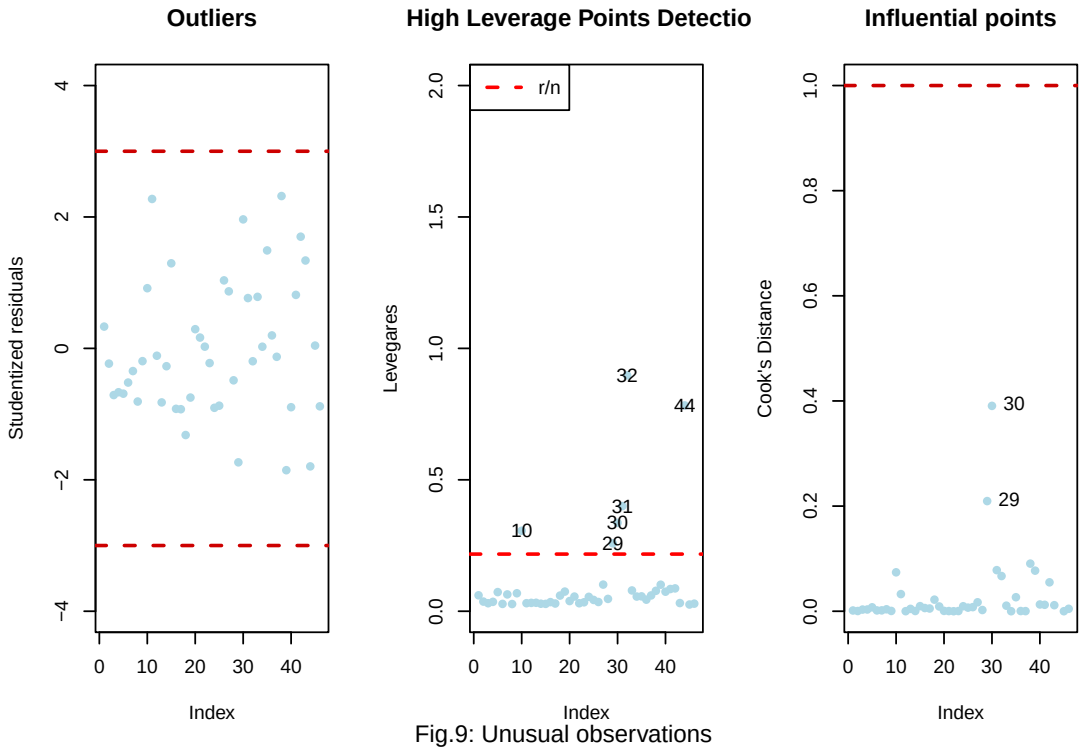

Fig.9: Unusual observations

*Fig. 9* shows that there are no outliers, but there are six **high leverage pioints**: 10th (*France*), 29th (*Australia*), 30th (*Brazil*), 31th (*Canada*), 32th (*China*) and 44th (*United States*).

We also see that there are no influential points, so the high leverage points don't have any significant impact on the model, hence fitting without each one of them is the same as fitting with them. For this reason all the points can be kept in the model.

# INTERPRETATION OF THE COEFFICIENTS

Table 1: Table 1: Model Coefficients

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 8.1580165 | 0.5770131 | 14.1383560 | 0.0000000 |
| Electr_nuclear | 0.0021155 | 0.0062155 | 0.3403553 | 0.7353267 |
| Electr_foss_fuels | 0.0059930 | 0.0010496 | 5.7099760 | 0.0000011 |
| Low_carb_electr | 0.0394675 | 0.0228141 | 1.7299607 | 0.0911597 |
| Land_Area | 0.0000011 | 0.0000003 | 4.2308527 | 0.0001275 |

**Note**: we can use square to remove the root.

The interpretation for each regressor is the following:

$\hat{\beta}_0$=8.15: The expected value of the *renewable energy production* is about 65 Twh(=$(8.1580165)^2$) when the other variables are equal to their average.

$\hat{\beta}_3$=0.0394675: An increase of one unit of nuclear energy production, results in an increase of about 1 Twh (=$(0.0394675)^2$) of renewable energy, holding all other variables constant.

All the other coefficients relative to quantitative variables can be interpreted in the same way: They represent the expected change in sqrt(renewable energy) for a one-unit increase in the predictor, holding all other predictors constant.

**sigma**:

```
## [1] 3.913493
```

It's the residual standard error, in other words it measures the mean error committed by the model. It's also a measure of the accuracy of the estimates: smaller is the value, more precise are the estimates.

In this case the value seems small enough, but it would be more useful to compare it with other models.

**confidence inerval**:

```
confint(OLS_sqrt)
```

```
##                           2.5 %       97.5 %
## (Intercept)        6.992715e+00 9.323318e+00
## Electr_nuclear    -1.043703e-02 1.466801e-02
## Electr_foss_fuels  3.873374e-03 8.112678e-03
## Low_carb_electr   -6.606508e-03 8.554145e-02
## Land_Area          5.989228e-07 1.692888e-06
```

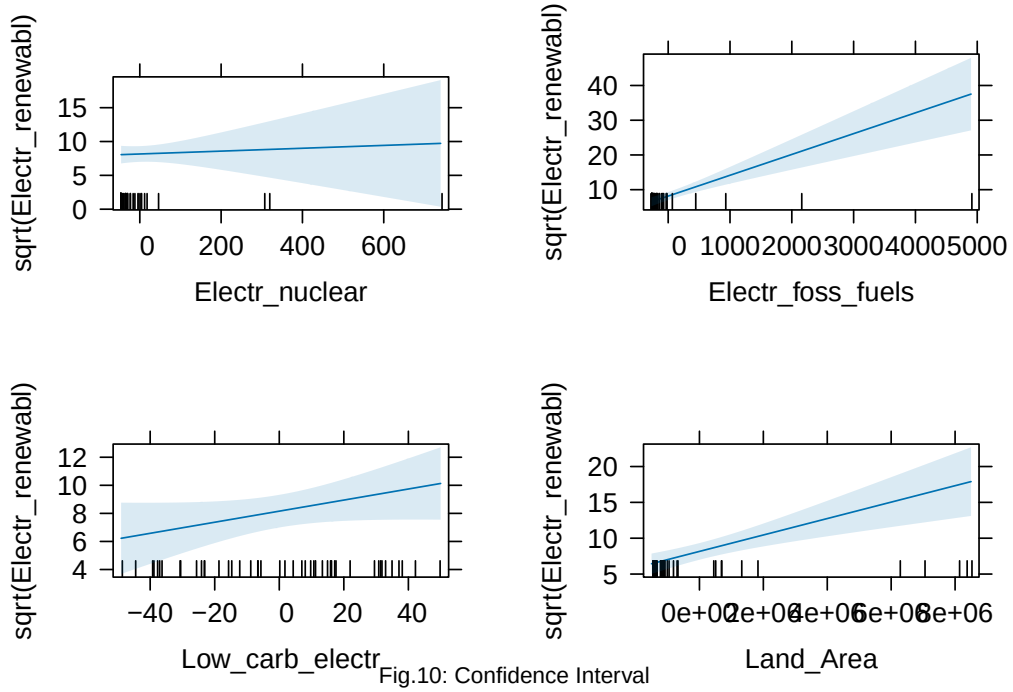


Fig.10: Confidence Interval

From *Fig.10*, we can see 4 graphs representing the relationship between the *sqrt(renewable electricity)* and each indipendent variable.

As we expected from the interpretation of the coefficients, all the plots display a positive relationship wrt the production of renewable electricity. The line in blue represents the centerline of the relationship, while the shaded area is the confidence interval,that is, how confident that our estimates values are between the inverval.

**CI for Electr_nuclear**: The plot represents the positive relationship between the sqrt(renewable electricity) and the amount of nuclear energy produced. In this case the slope is very flat, but this make sense since, as we have seen above, its impact on the response it's not significant. So, as nuclear energy increases, so does renewable energy, and the uncertainty increases.

Since it's 95% confidence interval, we can say with a confidence level of 95% that the "true" value of nuclear electricity is between -0.01 and 0.015 Twh.

The interpretation is the same for the other continuous variables; but the overall interpretation is that each variable has a positive relationship with the response variable, and as the values increase, the uncertainty also increases.

## Test each coefficient

Since testing each coefficient means looking at their significance, the hypotheses are as follows:

H0: $\hat{\beta}_j$=0 where j=1,...,4. H1: $\hat{\beta}_j$!=0

From *Table 1*, we can already see the result of this test looking for "t value" and "Pr(>|t|)". A coefficient is significant when its associated p-value is less than 0.05, i.e. $\hat{\beta}_2$ and $\hat{\beta}_4$.

Since the p-value associated with $\hat{\beta}_0$,$\hat{\beta}_1$ and $\hat{\beta}_3$ is bigger, we accept the null hypothesis, in other words they are equal to zero.

## Test a group of regressors

H0: the two models are equivalent H1: the two models are not equivalent

```
OLS_sqrt2 <- lm(sqrt(Electr_renewabl) ~ Electr_foss_fuels + Land_Area, data = dat_centered)
anova(OLS_sqrt,OLS_sqrt2)
```

```
## Analysis of Variance Table
##
## Model 1: sqrt(Electr_renewabl) ~ Electr_nuclear + Electr_foss_fuels +
##     Low_carb_electr + Land_Area
## Model 2: sqrt(Electr_renewabl) ~ Electr_foss_fuels + Land_Area
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     41 627.93
## 2     43 682.62 -2   -54.692 1.7855 0.1805
```

The p-value of the ANOVA test is bigger then 0.05: this suggest that there is not enough evidence against the null hypothesis, hence the full model and the nested model are equivalent.

The conclusion is that we can simplify our model with the nested model, hence the model without **Electr_nuclear**,**Low_carb_electr**.

## Goodness of fit of the model

R^2 is a measure of how well a regression model fits the observed data, computed by ratio between the variablity explained by the regression model (SSreg) and the total variability of the response (SSy).

**R^2**:

```
## [1] 0.7891359
```

In this case, a value of 0.78 indicates that the model is able to explain about the 80% of variability of the response *(sqrt of Renewable energy)*.

# Prediction

Suppose we observe a new country, e.g. Thailand, and we want to predict the amount of electricity produced from renewables resources and its associated uncertainty.

In particualar, the amount of energy produced from fossil fuels produced by Thailand in 2020 was 154.52 Twh and its land area is 513.120 Kmˆ2.

Table 2: Predicted value and Confidence Interval

| fit | lwr | upr |
|---|---|---|
| 5.707916 | -2.441759 | 13.85759 |

According to the regression model, the predicted amount of renewable energy produced by Thailand is about 32 Twh (=5.707916ˆ2).

Note that the prediction interval is wider than the confidence interval to account for the additional uncertainty due to predicting an individual response, and not the mean, for a given value of $X$.

# Conclusion

In this project, I explored the factors influencing renewable energy production in various IEA and OECD member countries, with a specific focus on the differences between European and non-European nations. After carefully selecting the most relevant variables through a model selection process, the final model highlighted that energy production from nuclear and fossil sources, along with the percentage of low-carbon electricity and territorial area, are the main determinants of renewable energy production.

This study contributes to a deeper understanding of the factors that can be influenced to promote a more sustainable energy future.