Statistical Regression and Modeling

Subhajit Pal,Sumit Banerjee,Amit Adhikari,Ankit Dalal

April 17, 2025

Contents

Abstract

This report presents a comprehensive statistical regression project in R, aimed at providing end-to-end insights from data preprocessing to the final predictive model.The primary goal is to demonstrate how data cleaning, exploratory data analysis (EDA), model building, evaluation and prediction can be implemented using R.

Key components of the workflow include:

- Data cleaning techniques such as handling missing values and outliers,
- Visualization techniques for understanding data distribution and relationships,
- Model development using linear, polynomial and interaction terms,
- Model evaluation using criteria like AIC, BIC, adjusted R^2 and cross-validation,
- Interpretation of results and prediction accuracy.

Through visualizations and statistical validation, this project illustrates a robust and reproducible approach to regression analysis using R programming.

1. Introduction

In this report, we outline the process of handling raw data, cleaning it, performing exploratory analysis, constructing models, validating the models, and making predictions using R.The project aims to explore the relationship between selected greenhouse gas (GHG) predictors and various outcomes by developing and evaluating different statistical models. The hypothesis is that GHG predictors can provide insights into specific patterns, potentially aiding in country-level classification or other analytical purposes. To test this, a linear regression model was built to analyze the most significant predictors, achieving a strong fit, though overfitting limited its predictive power. A Generalized Additive Model (GAM) was then employed to balance model flexibility and complexity, with constrained degrees of freedom to prevent overfitting; however, this approach was restricted by the limited number of predictors and interaction terms. Lastly, a logistic regression model was tested for its feasibility in using these predictors to classify countries but failed to achieve significance. Through this multi-model approach, the project seeks to understand the effectiveness of these statistical tools and the limitations of GHG predictors in predictive modeling.

1.1 Data Description

The dataset we used in this analysis is a Proprietary dataset from Dr. Shreya Banerjee. which has 80 countries, 75 features related to greenhouse gas emissions Data set: Denmark (1843-2023), Croatia-Combodia(1850-2023) Target variable: Adj.gdp (Billion USD, 2011 base) Predictors:
CoalCo2 (million tonnes), Gasco2 (million tonnes), Methane (million tonnes), Nitrous Oxide (million tonnes), OilCo2 (million tonnes) .

1.2 Linear Regression

- **Target**: For linear Regression, we take only Denmark country data, and The dependent variable or outcome that we wish to predict is Adj.gdp (per Billion USD for 2011 currency)
- **Predictor Variables**:These include numeric and categorical features that potentially influence the target variable are coal Co2(per million tonnes), gas Co2(million tonnes), methane(per million tonnes), nitrous oxide(per million tonnes), and oil Co2(per million tonnes).

1.3 Logistic Regression

- **Target**: For Logistic Regression, We take two country data, Denmark and Cambodia. The dependent variable here is the country.
- **Predictor Variables**: These include numeric and categorical features that potentially influence the target variable are coal Co2(per million tonnes), gas Co2(million tonnes), methane(per million tonnes), nitrous oxide(per million tonnes), and oil Co2(per million tonnes).

2. Methodoloy

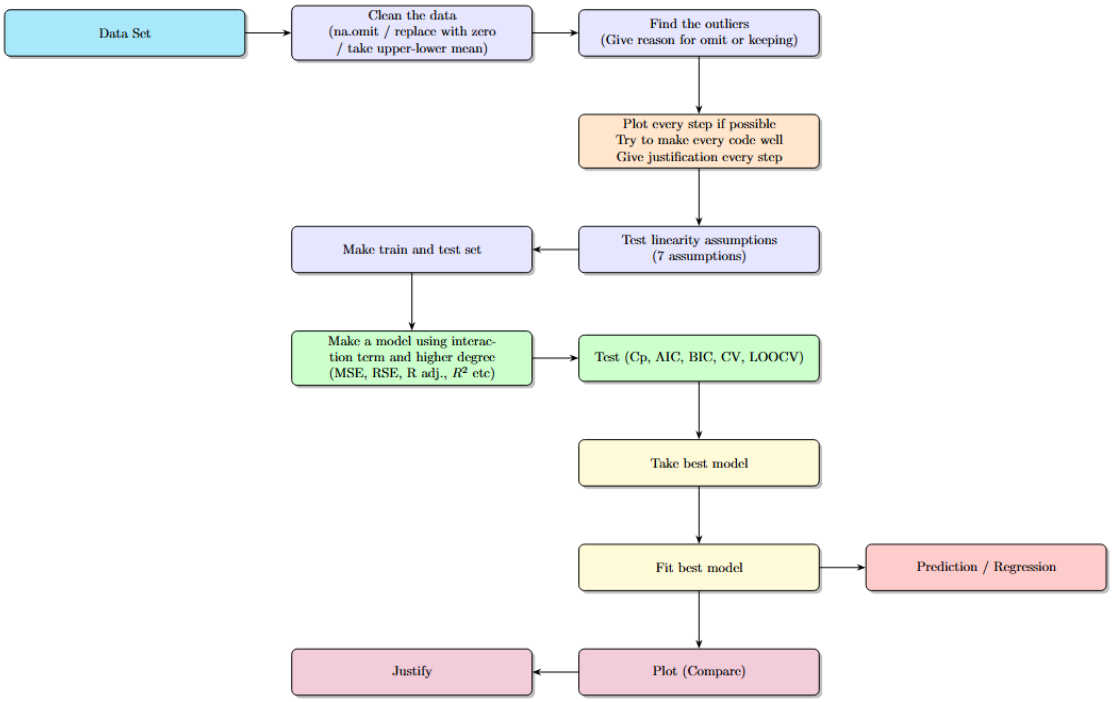


Figure 1: The Methodology Flowchart

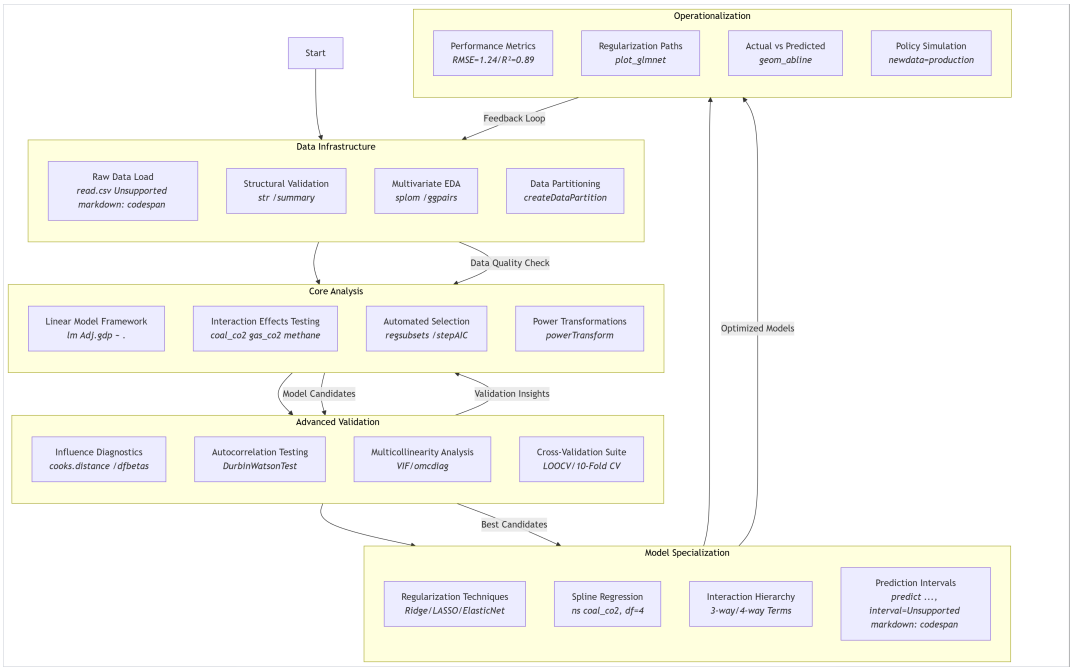


Figure 2: Brief Linkage Diagram

3. Analytical Workflow

Phase	Components
Data Preparation	Missing values, Outliers
EDA	Correlation, Distributions
Model Building	Linear, GAM, Polynomial
Model Selection	AIC, BIC, Adj. R^2
Validation	LOOCV, 10-fold CV
Regularization	Ridge, Lasso, Elastic Net
Evaluation	RMSE, R^2

4. Data Cleaning and Preprocessing

Data cleaning involves several key steps to ensure the dataset is ready for analysis. These include handling missing values, detecting and dealing with outliers, and transforming variables.

4.1 Handling Missing Values

As it is an ISO-certified data set, the missing values here are taken as zero, which is a standard operating procedure for GHG accounting.

4.2 Outlier Detection

Both Outliers Influential points and Leverage points are identified using Hat values, Cook distance, Dfbetas, Dffits, and Covratio. Leverage points are detected by hat values, and Influential points are detected by Cook distance, Dfbetas, and Dffits.

4.3 Most inclusive model with highest significant predictor

The model with coefficient value is given

```
Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  3.661262   2.556235   1.432  0.15388
coal_co2                                     1.837003   0.281771   6.519 7.54e-10 ***
gas_co2:methane                             10.283159   0.619352  16.603  < 2e-16 ***
coal_co2:oil_co2                             0.458853   0.154096   2.978  0.00332 **
gas_co2:oil_co2                             -3.550570   0.291061 -12.199  < 2e-16 ***
methane:oil_co2                              0.115876   0.022266   5.204 5.50e-07 ***
coal_co2:methane:oil_co2                    -0.034598   0.013303  -2.601  0.01011 *
coal_co2:gas_co2:methane:nitrous_oxide      -0.050277   0.007323  -6.865 1.16e-10 ***
coal_co2:gas_co2:oil_co2:nitrous_oxide       0.017618   0.002962   5.948 1.48e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.22 on 172 degrees of freedom
Multiple R-squared:  0.9545,    Adjusted R-squared:  0.9524
F-statistic: 451.3 on 8 and 172 DF,  p-value: < 2.2e-16
```

Figure 3: The coefficient summary of the model

5. Exploratory Data Analysis (EDA)

EDA is the process of analyzing data sets to summarize their main characteristics, often visualizing them.

5.1 Autocorrelation Analysis:

We are doing the Durbin-Watson test to find autocorrelation. The Durbin Watson value came as $DW = 1.8057$. Since it is close to 2 this means there is a very slight autocorrelation.

5.2 Normality Check:

For normality, we can plot and then check in the graph. We can check by Q-Q plot and P-P plot. Here, we give a Q-Q plot, and we see that both tails are not in the line; this implies the error is not normally distributed. To solve this problem, we will transform it.

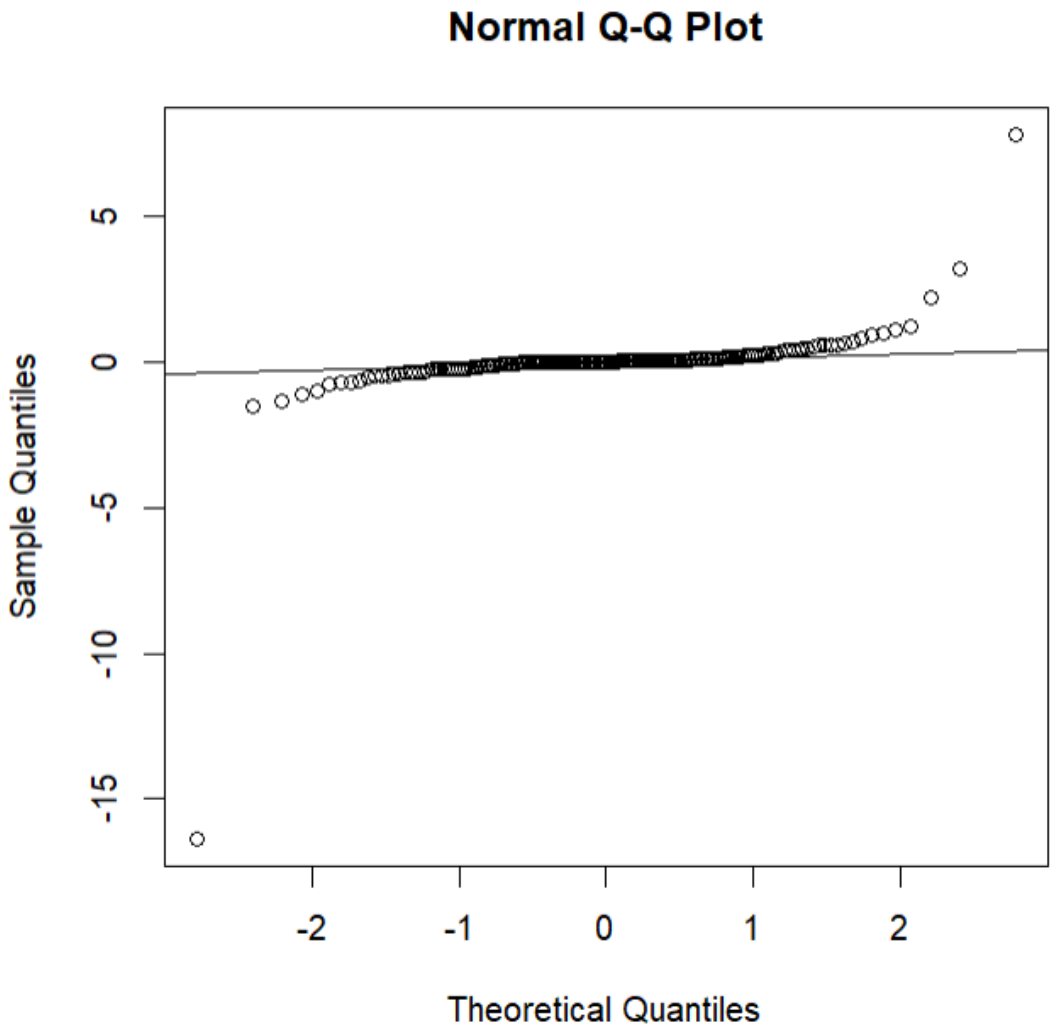


Figure 4: Q-Q Plot

5.3 Multicolllinearity:

To check Multicollinearity, we use a Correlation matrix and Variance Inflation Factor. (VIF). We see that there is some collinearity of some interaction terms. It can affect our future predictions.

5.4 Check for non-linearity:

We check non-linearity by plotting Partial regression and adding variable plots. We see that in the AV plot, it is non-linear data.

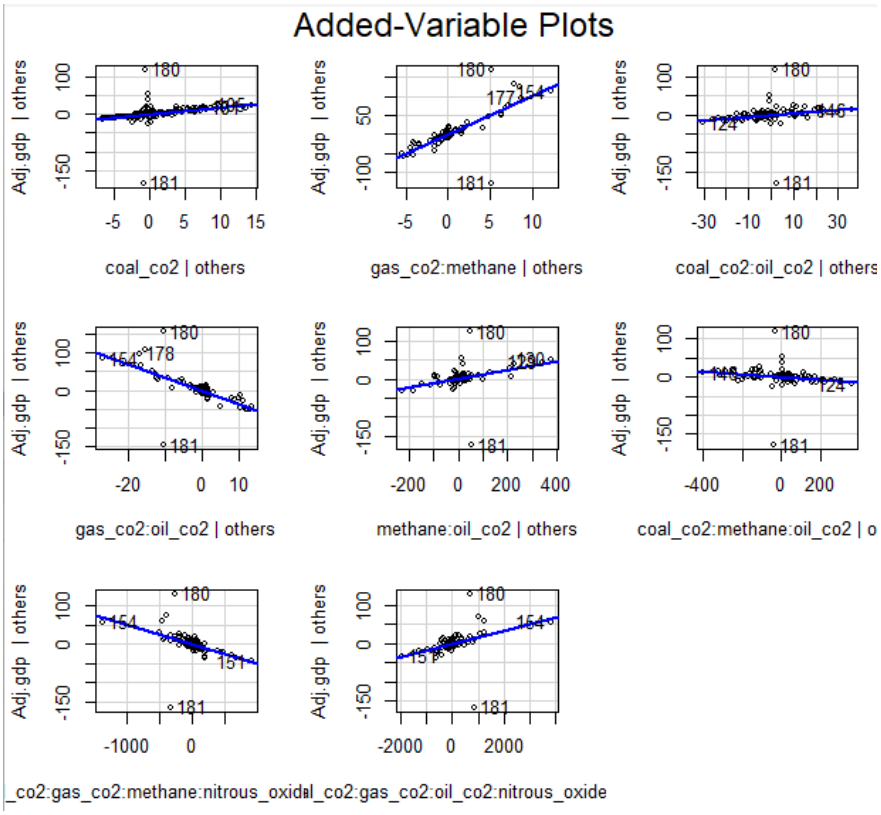


Figure 5: AV Plot

6. Transfomation of the data

We Transforma of the response variable to fix non-normality and non-constant variance. By Box-Cox method and Yeo-Johnson method.The round of lamda value came for our data is $\lambda = 0.33$. But we can see that it giving almost same compare to the previous model.

7. Cross validation

By Cross validation we find the optimal model.we use Loocv and K-fold approach for cross validation.For this we need to split the data.

7.1 Data Splitting

The data is split into a training set and a test set to ensure that the model is evaluated on unseen data. The typical split ratio is 80/20, with 80% of the data used for training and 20% for testing.

7.2 Interaction Model

An interaction term is added to see if the relationship between one predictor and the target depends on the value of another predictor:
Example:

```
model4=lm(yjPower(Adj.gdp,p3$roundlam)~coal_co2+gas_co2:methane+gas_co2:oil_co2+

methane:oil_co2+coal_co2:gas_co2:methane:nitrous_oxide+

coal_co2:gas_co2:nitrous_oxide:oil_co2,data=den)
```

8. Shrinkage Method:

We apply there Shrinkage Method for our data Ridge Regression,Lasso,Elastic net. To apply these methods first we need to find the the tuning parameter λ value. When $\lambda = 0$, the penalty term has no effect, and ridge regression will produce the least squares estimates. However, as $\lambda \to \infty$, the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero. After finding λ value we apply Loocv in both train and test set. Then we plot the coefficient and lambda value test and train for each Regression.

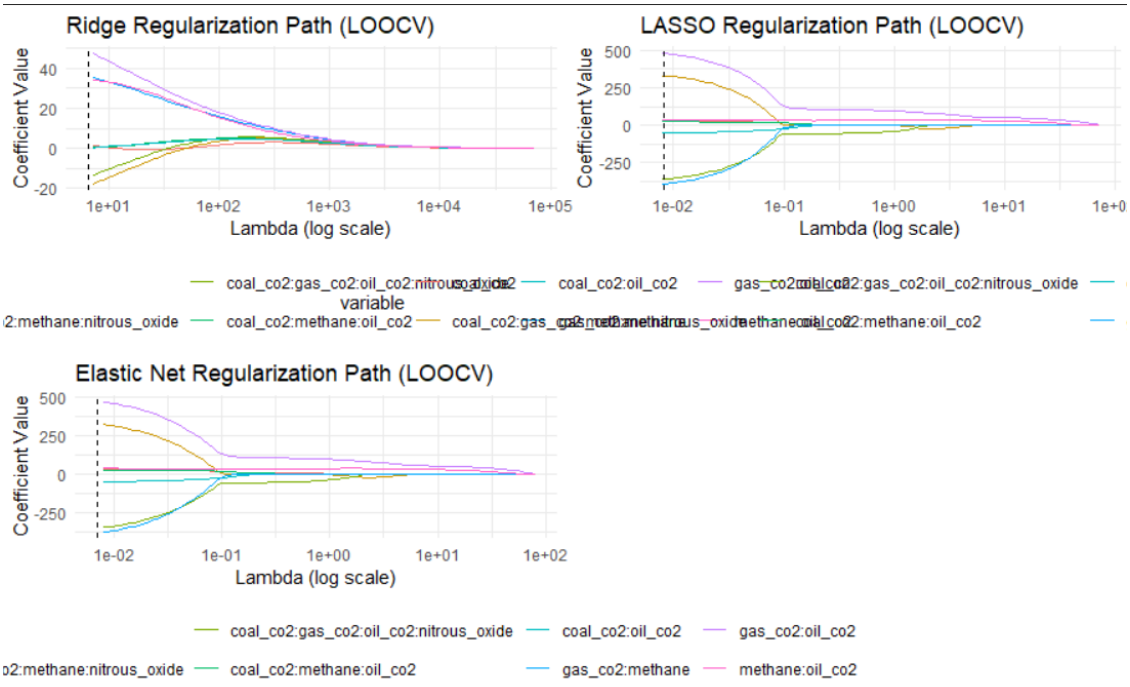


Figure 6: Test plot

8.1 Actual vs predicted model

Now we are ploting actual vs predicted model for the best model. We choose the best model in term of Adjusted R square.

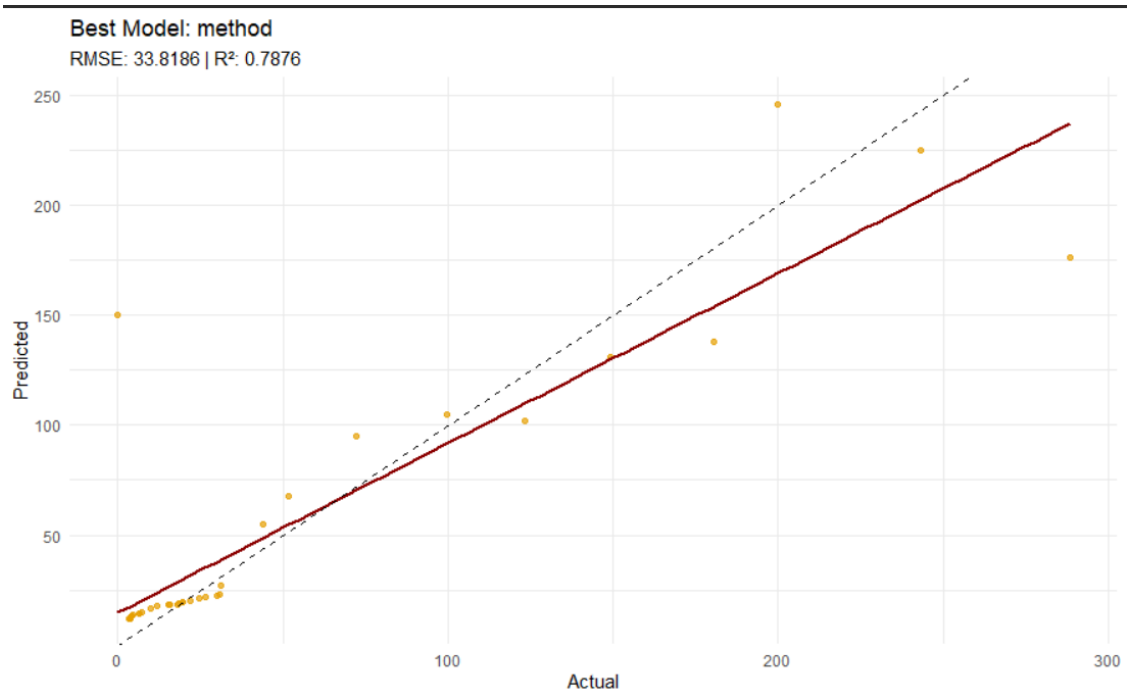


Figure 7

8.2 Polynomial Model

A polynomial model is created by adding quadratic terms to capture nonlinearity. In this project, first, we fit the data by spline, one by one predictor variable, then we use the GAM(Generalized Additive model) to predict the response variable by all its predictor variables.

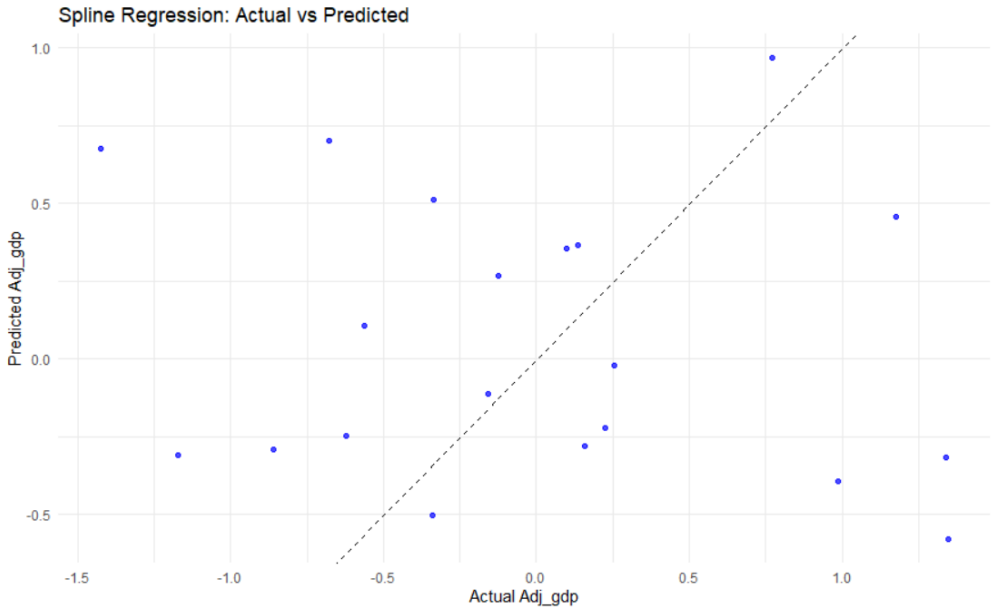


Figure 8

Now, the plot below is for nitrous oxide to fit the data point in the polynomial model. we did the same thing for each of the predictor one by one

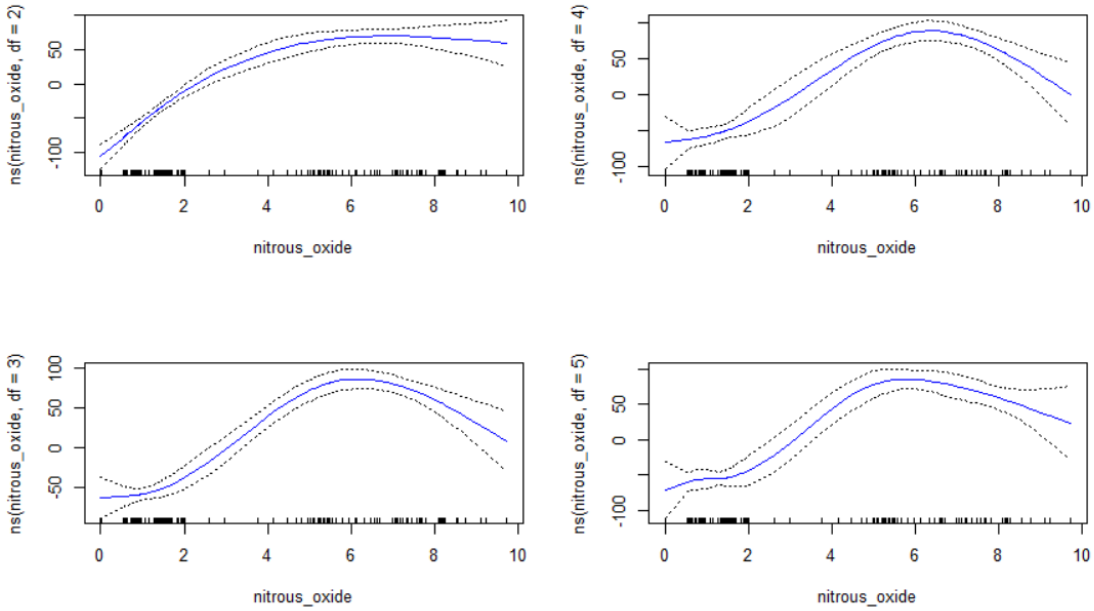
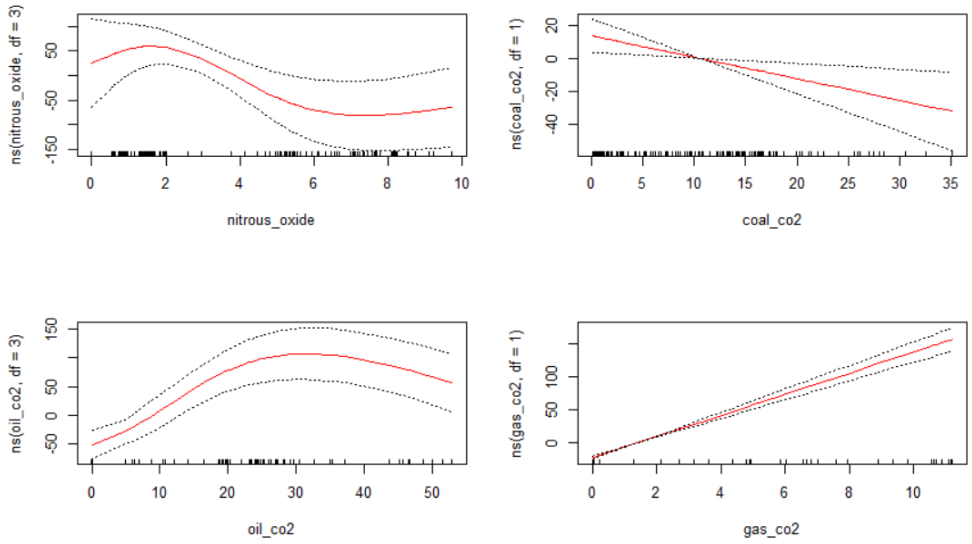


Figure 9

8.3 GAM

We are taking all significant predictors for the Generalized Additive Model with all combinations of different degrees of free of freedom.



GAM ANOVA Summary ANOVA for Parametric Effects

Term	Df	Sum Sq	Mean Sq	F-value	p-value
ns(coal_co2, df = 2)	2	247451	123725	194.6068	$< 2 \times 10^{-16}$
ns(gas_co2, df = 2)	2	831581	415831	654.0297	$< 2 \times 10^{-16}$
ns(methane, df = 3)	3	128553	42851	67.3999	$< 2 \times 10^{-16}$
ns(nitrous_oxide, df = 3)	3	115393	38465	60.1387	$< 2 \times 10^{-16}$
ns(oil_co2, df = 3)	3	20624	6875	10.8130	2.14×10^{-6}
Residuals	131	83286	636		

- AIC = 1352.526

9. Logistic Model

Logistic regression is a statistical method used for modeling binary outcomes, where the response variable takes only two possible values (e.g., yes/no, 0/1). It estimates the probability that a given input point belongs to a particular category. The model uses the logistic (sigmoid) function to map predicted values to probabilities. Logistic regression is widely used in classification problems . Here in our project, we use it to classify the countries Denmark and Cambodia with respect to the predictor variable. The summary of the model with all interaction terms is given

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.611e+01	5.019e+04	0.001	0.999
coal_co2	8.455e+00	2.377e+05	0.000	1.000
gas_co2	3.269e+01	8.777e+05	0.000	1.000
methane	-1.924e+01	1.836e+04	-0.001	0.999
nitrous_oxide	6.362e+01	1.860e+05	0.000	1.000
oil_co2	-1.780e+01	2.062e+05	0.000	1.000
coal_co2:gas_co2	-6.587e+00	2.413e+05	0.000	1.000
coal_co2:methane	6.102e+00	2.476e+04	0.000	1.000
gas_co2:methane	3.159e+00	3.532e+04	0.000	1.000
coal_co2:nitrous_oxide	-3.233e+01	1.230e+05	0.000	1.000
gas_co2:nitrous_oxide	-2.093e+01	2.837e+05	0.000	1.000
methane:nitrous_oxide	1.261e+00	1.088e+04	0.000	1.000
coal_co2:oil_co2	3.951e+00	6.763e+04	0.000	1.000
gas_co2:oil_co2	1.773e+00	5.598e+04	0.000	1.000
methane:oil_co2	-1.220e+00	2.224e+04	0.000	1.000
nitrous_oxide:oil_co2	3.576e+00	7.798e+04	0.000	1.000
coal_co2:gas_co2:methane	-4.923e-01	9.603e+03	0.000	1.000
coal_co2:gas_co2:nitrous_oxide	3.398e+00	5.289e+04	0.000	1.000
coal_co2:methane:nitrous_oxide	4.139e-03	1.667e+03	0.000	1.000
gas_co2:methane:nitrous_oxide	-1.419e-02	2.505e+03	0.000	1.000
coal_co2:gas_co2:oil_co2	-4.206e-02	1.132e+04	0.000	1.000
coal_co2:methane:oil_co2	-7.234e-02	2.450e+03	0.000	1.000
gas_co2:methane:oil_co2	-6.949e-03	2.395e+03	0.000	1.000
coal_co2:nitrous_oxide:oil_co2	3.896e-01	1.048e+04	0.000	1.000
gas_co2:nitrous_oxide:oil_co2	8.042e-02	1.829e+04	0.000	1.000
methane:nitrous_oxide:oil_co2	9.690e-03	1.024e+03	0.000	1.000
coal_co2:gas_co2:methane:nitrous_oxide	-5.443e-03	3.116e+02	0.000	1.000
coal_co2:gas_co2:methane:oil_co2	7.948e-03	1.771e+02	0.000	1.000
coal_co2:gas_co2:nitrous_oxide:oil_co2	-5.788e-02	1.047e+03	0.000	1.000
coal_co2:methane:nitrous_oxide:oil_co2	-2.417e-03	7.157e+01	0.000	1.000
gas_co2:methane:nitrous_oxide:oil_co2	-7.489e-06	7.830e+01	0.000	1.000
coal_co2:gas_co2:methane:nitrous_oxide:oil_co2	1.992e-04	7.794e+00	0.000	1.000

Here, we see that all coefficients ofthe predictor are insignificant since the p-vlaue is high. So we can't do logistic Regression for our data.

10. Model Evaluation

The models are evaluated using multiple criteria:

- AIC (Akaike Information Criterion)
- BIC (Bayesian Information Criterion)
- Adjusted R^2
- Cross-validation

10.1 AIC and BIC

Lower values of AIC and BIC indicate better-fitting models:

```
AIC(model1)
BIC(model2)
```

10.2 Adjusted R^2

Higher values of Adjusted R^2 indicate better-fitting models.

11. Prediction and Results

The final model is used to make predictions on the test set: .

- **Strong interactions:**
  - Gas CO_2 + Methane: Positive ($\beta = 10.28$)
  - Coal CO_2 + Oil CO_2: Positive ($\beta = 0.46$)
  - Gas CO_2 + Oil CO_2: Negative ($\beta = -3.55$)
  - Methane + Oil CO_2: Positive ($\beta = 0.12$)
  - Coal CO_2 + Methane + Oil CO_2: Negative ($\beta = -0.035$)
  - Coal CO_2 + Gas CO_2 + Methane + Nitrous Oxide: Negative ($\beta = -0.050$)
  - Coal CO_2 + Gas CO_2 + Oil CO_2 + Nitrous Oxide: Positive ($\beta = 0.018$)

12. Model Justification

The model selection process considers several aspects, such as statistical metrics (AIC, BIC), predictive power (R^2, cross-validation), and the residual plots. The chosen model explains a significant portion of the variance in the target variable and satisfies most assumptions of linear Regression.

```
> print(test_metrics)
                   Model     RMSE        R2
Ridge_LOOCV   Ridge_LOOCV 33.81864 0.7813337
Ridge_10Fold Ridge_10Fold 33.81864 0.7813337
Lasso_LOOCV   Lasso_LOOCV 34.19846 0.7873607
Lasso_10Fold Lasso_10Fold 34.19898 0.7873574
Enet_LOOCV     Enet_LOOCV 34.12748 0.7875975
Enet_10Fold   Enet_10Fold 34.12748 0.7875975
> # Enhanced coefficient plot function
> plot_coef_path <- function(model, title) {
```

```
+ )
> print(test_metrics)
                   Model     RMSE        R2
Ridge_LOOCV   Ridge_LOOCV 33.10907 0.7943175
Ridge_10Fold Ridge_10Fold 33.10907 0.7943175
Lasso_LOOCV   Lasso_LOOCV 33.31385 0.7969370
Lasso_10Fold Lasso_10Fold 33.31385 0.7969370
Enet_LOOCV     Enet_LOOCV 33.16133 0.7979443
Enet_10Fold   Enet_10Fold 32.76823 0.8004933
> plot_glmnet <- function(model, title) {
+   plot(model$finalModel, xvar = "lambda", mai
```

13. Discussion

In this section, we discuss the limitations and potential improvements to the model:

- Model Complexity: As the model was complex, it seems that if we added more significant variables, our output could be better.
- Future scope: we could add more predictors can be more precise for model fitting and prediction.

14. Conclusion

The statistical regression analysis project in Denmark demonstrates a comprehensive approach to building predictive models using R, offering valuable insights into the relationship between greenhouse gas emissions and adjusted GDP. The project employed a methodical workflow, beginning with data cleaning and preprocessing, where missing values were handled according to ISO-certified standards for GHG accounting. The exploratory data analysis revealed slight autocorrelation, non-normally distributed errors, and presence of

multicollinearity in interaction terms that could potentially affect predictions. Data transformation using the Box-Cox and Yeo-Johnson methods was implemented to address non-normality and heteroscedasticity, yielding results comparable to the original model, suggesting robust underlying relationships in the data.

Linear Model: The selected model, based on the most significant predictors, achieves an adjusted R2 value of 0.9524. While this indicates a strong fit to the training data, it also suggests a potential case of overfitting. Despite its high R2, the model does not perform well in prediction.

GAM (Generalized Additive Model): The model was built to avoid overfitting by limiting the degrees of freedom (no more than 3). However, due to a limited number of predictors (only 5), not all interaction combinations could be explored. Therefore, while the model fits reasonably well, it cannot be claimed as the best possible fit.

Logistic Model: All combinations of individual and interaction predictors resulted in p-values higher than the significance threshold. Hence, it is concluded that predicting countries using only the selected GHG predictors is not feasible with logistic regression.

The high explanatory power of the interaction terms in the linear models suggests complex interrelationships between different greenhouse gas emissions in their effect on adjusted GDP. Particularly noteworthy is the strong positive interaction between gas CO2 and methane emissions, contrasted with the negative interaction between gas CO2 and oil CO2. These findings could inform environmental policy decisions by highlighting which combinations of emission reductions might minimize economic impact.

Several limitations merit consideration, such as the proprietary nature of the dataset limits reproducibility, and the focus on only five predictors from a larger set of 80 variables raises questions about potential omitted variable bias. Future research directions could explore more complex modeling approaches such as random forests or neural networks that might capture additional non-linear relationships not identified in the current models. Expanding the analysis to include more countries would test the generalizability of the findings and potentially identify country-specific emission-GDP relationships.

15. Reference

References

[1] Our World in Data. *Regions*. 2023. Available at: https://ourworldindata.org.

[2] International Organization for Standardization. *Regions*. 2023.

[3] Bolt, J., and van Zanden, J. L. *Maddison Project Database, version 2023*. Groningen Growth and Development Centre, University of Groningen. Available at: https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2023

[4] Global Carbon Project. *Global Carbon Budget 2024*. Available at: https://globalcarbonbudget.org/

[5] Jones, C. et al. *National contributions to climate change (2024)*. Available at: https://zenodo.org/records/7636699/latest

[6] W3Schools. *R Tutorial*. Available at: https://www.w3schools.com/r/default.asp

[7] R Core Team. *An Introduction to R*. Available at: https://cran.r-project.org/doc/manuals/R-intro.pdf

[8] James, G., Witten, D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning*. Available at: https://www.statlearning.com/