STA302 Final Project

Linear Regression Application on Air Pollutants and Emissions on Crop Production

Minh Thi Nhat Dang

Introduction

My project aims to draw inferences about the relationships between the air pollutants: carbon monoxide (CO), nitrogen oxide (NOX), sulfur oxide (SOX), Non-methane volatile organic compound (NMVOC), air emissions: greenhouse gas (GHG), carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), hydrofluorocarbons (HFC), perfluorocarbons (PFC), sulfur hexafluoride (SF6), nitrogen trifluoride (NF3) and the crop productions (maize, rice, soybean, wheat). My research questions are:

1. Do these pollutants impact crop production?
2. Which air pollutants and greenhouse gases affect crop production?

It is important to understand the effects of pollutants on crop production is crucial for ensuring food safety, promoting sustainable agriculture, and maintaining crop yields.

After conducting literature reviews, it is widely acknowledged that there exists a relationship between specific air emissions and crop production. However, my project is distinctive in that it aims to apply linear regression to analyze the data collected from multiple countries and take a macro-level perspective. This contrasts with previous studies, which have primarily focused on data collected from specific regions or nations.

Methods

To obtain data, I merged three datasets from the Organisation for Economic Co-operation website. These datasets include information on across around 40 countries' crop production for four different crops, five different air pollutants, and greenhouse gas emissions measuring eight different types of greenhouse gases. By matching the countries and year of measurement, I combined these datasets to create a new dataset for 15 overlapping countries. I focused on data from 2012 to 2019 when production and emissions remained relatively stable, without external factors such as the 2008 economic recession or the 2020 pandemic. The units used in this dataset are in 'Thousand Tons'.

To mitigate the impact of influential points and improve the validity of the result, I split the data into 70% training (78 data points) and 30% testing datasets (33 data points). Using the training dataset, I perform MLR with most of the greenhouse gases as predictors, except for particulates 2.5, particulates 10, and unspecified mixed of HFCs and PFCs due to lack of recorded data, and the total crop production as the response variable. Residuals are examined for their behavior through QQ plots, residual plots, and residuals vs fitted values. Influential observations were identified, and model selection is performed using stepwise and shrinkage methods. AIC and

BIC values were calculated for each round of selection/elimination for stepwise method, while LASSO shrinks predictors to zero to find the model with the simplest expression for shrinkage method. Reduced model is then compared with the full model to verify the statistical significance using ANOVA. Lastly, cross validation is used to check overfitting and the model predictions accuracy.

Results

The MLR results on the train data indicate that the intercept is estimated to be 19.18 with significant at the 0.0001 level, while the other coefficients have very small values (from -0.002 to 0.004). This suggests that the intercept alone can explain the variation in the dependent variable, total crop production. Among the predictors only SOX has statistically significant coefficients ($p < 0.05$). The other predictors, CO, NMVOC, NOX, PFC, CH4, CO2, GHG, HFC, N2O, and SF6 indicate their lack of significance in predicting the dependent variable ($p \geq 0.05$).

The R^2_{adj} for this model is 0.1097, indicating that this model with the given predictors can only explain 10.97% of the variation in total crop production.

All predictors were checked for multicollinearity using VIF, and all are found to have very high VIF values compared to the level of cut off being 5. This shows a high degree of correlation among them. This is reasonable as all air emissions exist together, they all display a decreasing trend over the year. I will later compare a reduced model with all the predictors removed with the full model.

Min	1Q	Median	3Q	Max
-14.8536	-1.7160	0.7513	2.5511	19.7503

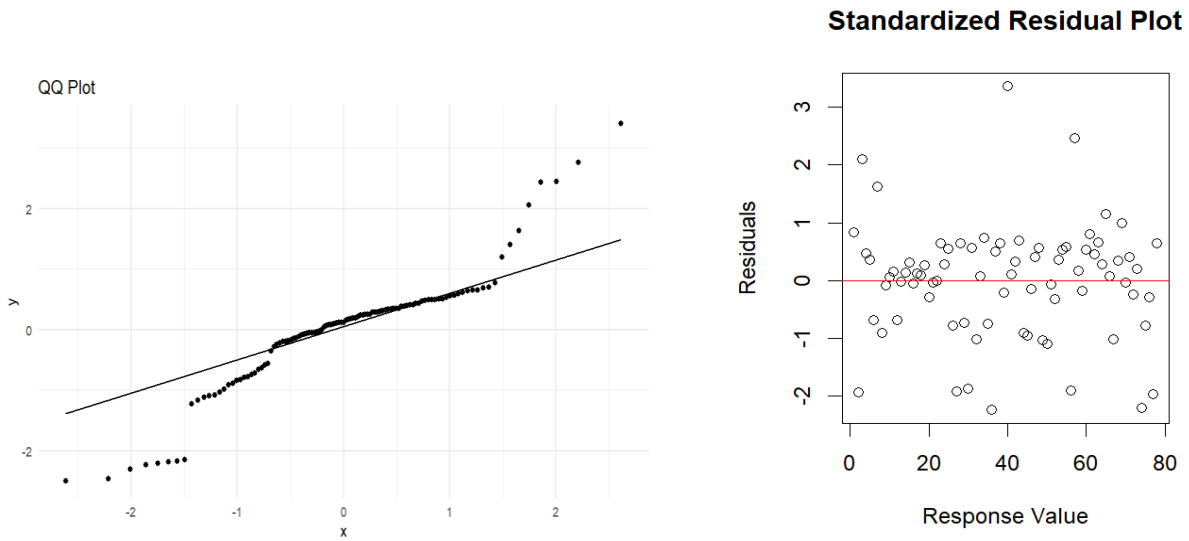


Figure 1. Residuals' distribution from full model

Figure 1 shows the residuals' range from -12.74 to 18.29, with a residual standard error of 6.20. Most of the residual's cluster around zero, with some exceptions due to extreme observations. The QQ plot confirms that most of the residuals follow a normal distribution, except for data points that deviate at the two ends. This deviation could be due to the left-skewed nature of the emissions data, with most measurements being very low, and only a few countries, particularly the USA, having very high emissions.

For the step wise model selection, both backward and forward, AIC and BIC values are calculated for each selection/elimination. Only four predictors are significant: SOX, CH4, CO2 and GHG. According to the ANOVA test, the null hypothesis is that there is no significant difference in the mean of the response values between the full model and the reduced model. The F-statistic is 0.6326, which is relatively low. The p-value is 0.727 > 0.05. This means that we fail to reject the null hypothesis and conclude that the reduced model is sufficient to explain the response. The R^2_{adj} for the reduced model also shows a better explanation for the response of being 0.14.

For the shrinkage method, LASSO and Elastic-net ($\alpha = 0.6$), no predictors are chosen. Thus, the ANOVA test the full model versus the LASSO model is used to verify if the reduced or full model are more significant. The p-value of 0.00609 suggests that the full model may not be significant. Furthermore, the residuals sum of squares of 2364.99 and mean square of 35.83 represent the variability of the response variable that is not explained by the predictors in the full model.

ANOVA test is then used to test the significance between AIC/BIC based and LASSO model. The p-value of 0.00431 suggests we fail to reject the null hypothesis and that the ACI/BIC based model is more suitable to explain the crop production response.

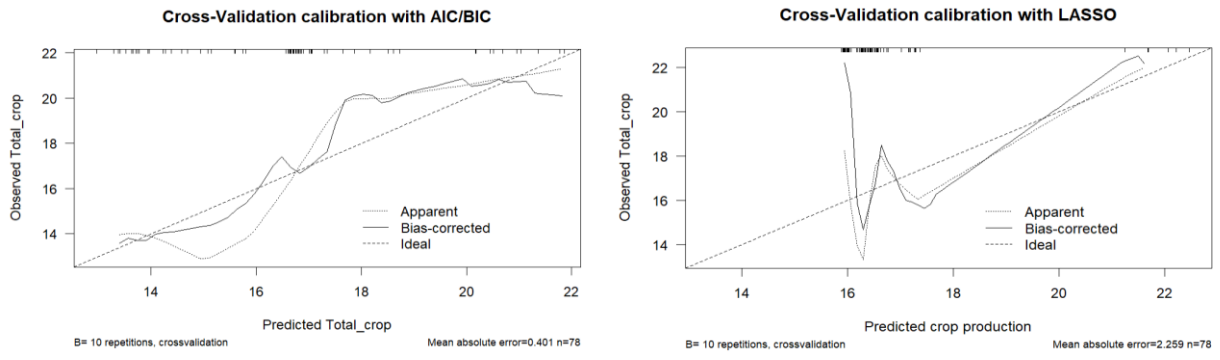


Figure 2. 10 folds cross validation with AIC/BIC and LASSO/Elastic-net ($\alpha = 0.6$) based selected model.

As **Figure 2**, the mean squared error between the observed response and the predicted response from the testing data for the model selected based on AIC/BIC is only 0.41. However, the model tends to underestimate the data from 14 to 17 and overestimate the data from 17 to 22. For the LASSO-based model, the mean squared error is higher at 1.16. The model does not perform well for small values, but for the middle values, the observed and predicted values follow closely. This is because of the nature of the data having a wide range of crop production, while air emissions remain at a low level.

Discussion

The limitation of this study is that the data contains influential data points, and the violation of the linear regression assumption limiting the model's effectiveness. However, removing these points is not feasible, as they may hold crucial information that other data points do not capture. Removing them may oversimplify the model, leading to overfitting and inaccurate representation of the relationship between air emissions and crop production. Moreover, these chosen datasets may not be suitable for drawing inferences about the relationship, given the variations in production practices, climate, and soil conditions among different countries. Additionally, unmeasured confounding variables specific to each country can bias the results of regression analysis.

In conclusion, from the obtained models, all the predictors have very small value and most of them show no statistical significance in the full model. When comparing the full model with reduced models via ANOVA, there is evidence that the reduced model is better at explaining the crop production, especially for the stepwise based model with SOX, CH4, CO2 and GHG being selected as predictors. However, it should be noted that the coefficients for these emissions are very small. A more in-depth study is needed to further investigate and better understand the relationship between these emissions and crop production.

References

OECD (2023), "Air and climate: Air emissions by source", *OECD Environment Statistics* (database), https://doi.org/ 10.1787/data-00598-en

OECD (2023), Crop production (indicator). doi: 10.1787/49a4e677-en

Shrestha, R.K., Shi, D., Obaid, H. *et al.* (2022). Crops' response to the emergent air pollutants. *Planta* 256, 80 (2022). https://doi.org/10.1007/s00425-022-03993-1

Rai, R., Agrawal, M., & Agrawal, S. B. (2015). Gaseous air pollutants: A review on current and future trends of emissions and impact on agriculture. *Environmental Science and Pollution Research.* https://www.researchgate.net/publication/267726469_Gaseous_air_pollutants_a_review_on_current_and_future_trends_of_emissions_and_impact_on_agriculture

Rehman, A., Ma, H. & Ozturk, I. (2020). Decoupling the climatic and carbon dioxide emission influence to maize crop production in Pakistan. *Air Qual Atmos Health* 13, 695–707 (2020).https://doi.org/10.1007/s11869-oooooooi020-00825-7

Appendix

Full Model

Coefficients	Estimate	Std. Error	t-value	Pr (>|t|)
Intercept	16.91	1.407	12.014	< 2e-16
CO	0.0005765	0.0007589	0.760	0.45017
NMVOC	0.0007120	0.0023124	0.308	0.75911
NOX	-0.0009776	0.0016110	-0.607	0.54605
SOX	0.0043685	0.0015899	2.748	0.00773
CH4	-0.0028511	0.0017816	-1.600	0.11430
CO2	-0.0028360	0.0017748	1.598	0.11485
GHG	0.0028198	0.0017743	-1.589	0.11678
HFC	-0.0026379	0.0018229	-1.447	0.15260
N2O	-0.0028262	0.0018431	-1.533	0.12995
PFC	-0.0023219	0.0029317	-0.792	0.43119
SF6	-0.0021909	0.0017753	-1.234	0.22154

Residual standard error: 5.986 on 66 degrees of freedom
Multiple R-squared: 0.2368,
Adjusted R-squared: 0.1097
F-statistic: 1.862 on 11 and 66 DF,
p-value: 0.06091

Stepwise Selection Model with AIC/BIC

Coefficients	Estimate	Std. Error	t-value	Pr (>|t|)
Intercept	16.79	0.89	18.765	< 2e-16
SOX	0.0024	0.00104	2.311	0.0236
CH4	-0.000146	0.000042	-3.100	0.0275
CO2	-0.000167	0.000016	-3.641	0.0005
GHG	0.0001163	0.000032	3.625	0.0005

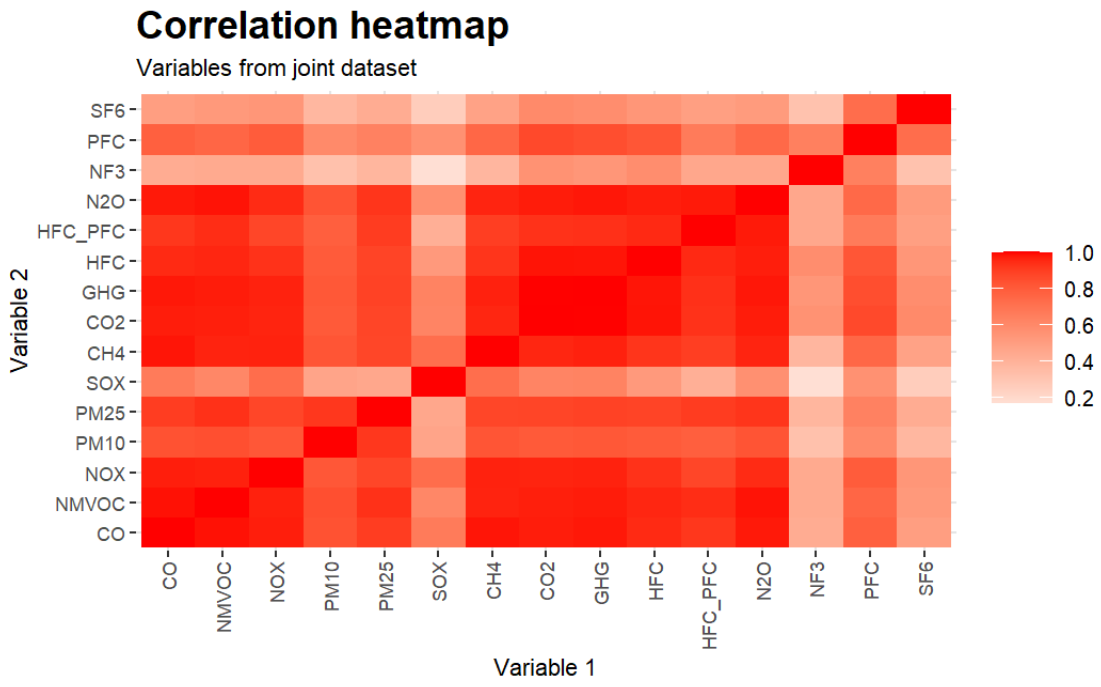
Residual standard error: 5.88 on 73 degrees of freedom
Multiple R-squared: 0.1856,
Adjusted R-squared: 0.141
F-statistic: 4.16 on 4 and 73 DF,
p-value: 0.004319

ANOVA

Model 1: Total_crop ~ CO + NMVOC + NOX + SOX + CH4 + CO2 + GHG + HFC + N2O + PFC + SF6
Model 2: Total_crop ~ SOX + CH4 + CO2 + GHG

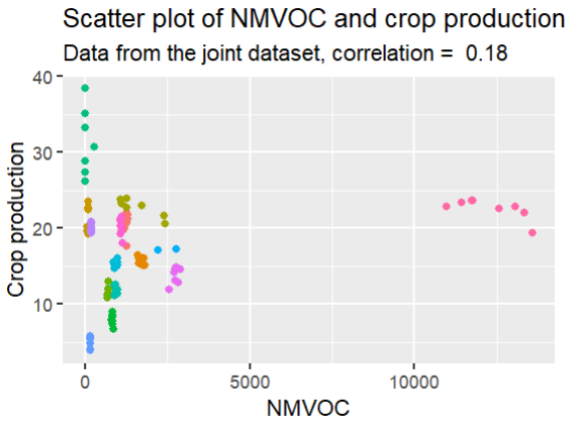
Res. Df	RSS	Df	Sum of Sq	F	Pr (>F)
66	2365.0				
73	2523.7	-7	-158.67	0.6326	0.7273

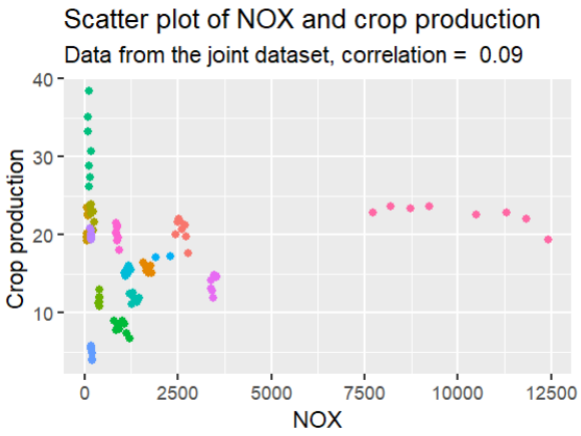
Figure 3. Statistics on the full model, reduced model based on stepwise method and ANOVA test for the two models.

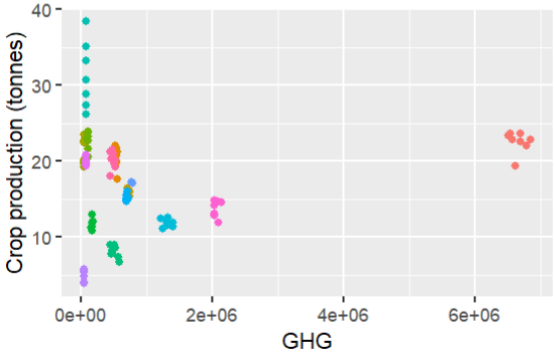


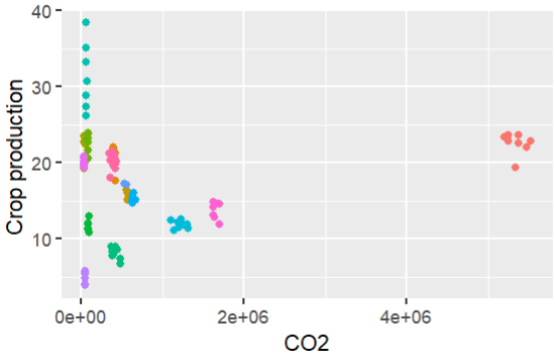
Coefficients	CO	NMVOC	NOX	CH4	CO2	GHG	HFC	N2O	PFC	SF6
VIF	191	113	38	216,688	13,282,270	20,362,590	11,423	101,311	51	43

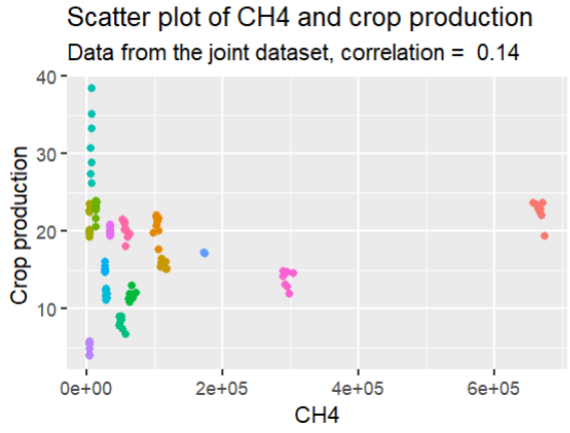
Figure 4. Correlation and VIF of the coefficients.


Scatter plot of NMVOC and crop production
Data from the joint dataset, correlation = 0.18
Crop production
NMVOC
0
5000
10000
10
20
30
40


Scatter plot of NOX and crop production
Data from the joint dataset, correlation = 0.09
Crop production
NOX
0
2500
5000
7500
10000
12500
10
20
30
40


Scatter plot of GHG and crop production
Data from the joint dataset, correlation = 0.12
Crop production (tonnes)
GHG
0e+00
2e+06
4e+06
6e+06
10
20
30
40


Scatter plot of CO2 and crop production
Data from the joint dataset, correlation = 0.12
Crop production
CO2
0e+00
2e+06
4e+06
10
20
30
40


Scatter plot of CH4 and crop production
Data from the joint dataset, correlation = 0.14
Crop production
CH4
0e+00
2e+05
4e+05
6e+05
10
20
30
40


Scatter plot of SF6 and crop production
Data from the joint dataset, correlation = -0.02
Crop production
SF6
0
3000
6000
9000
10
20
30
40

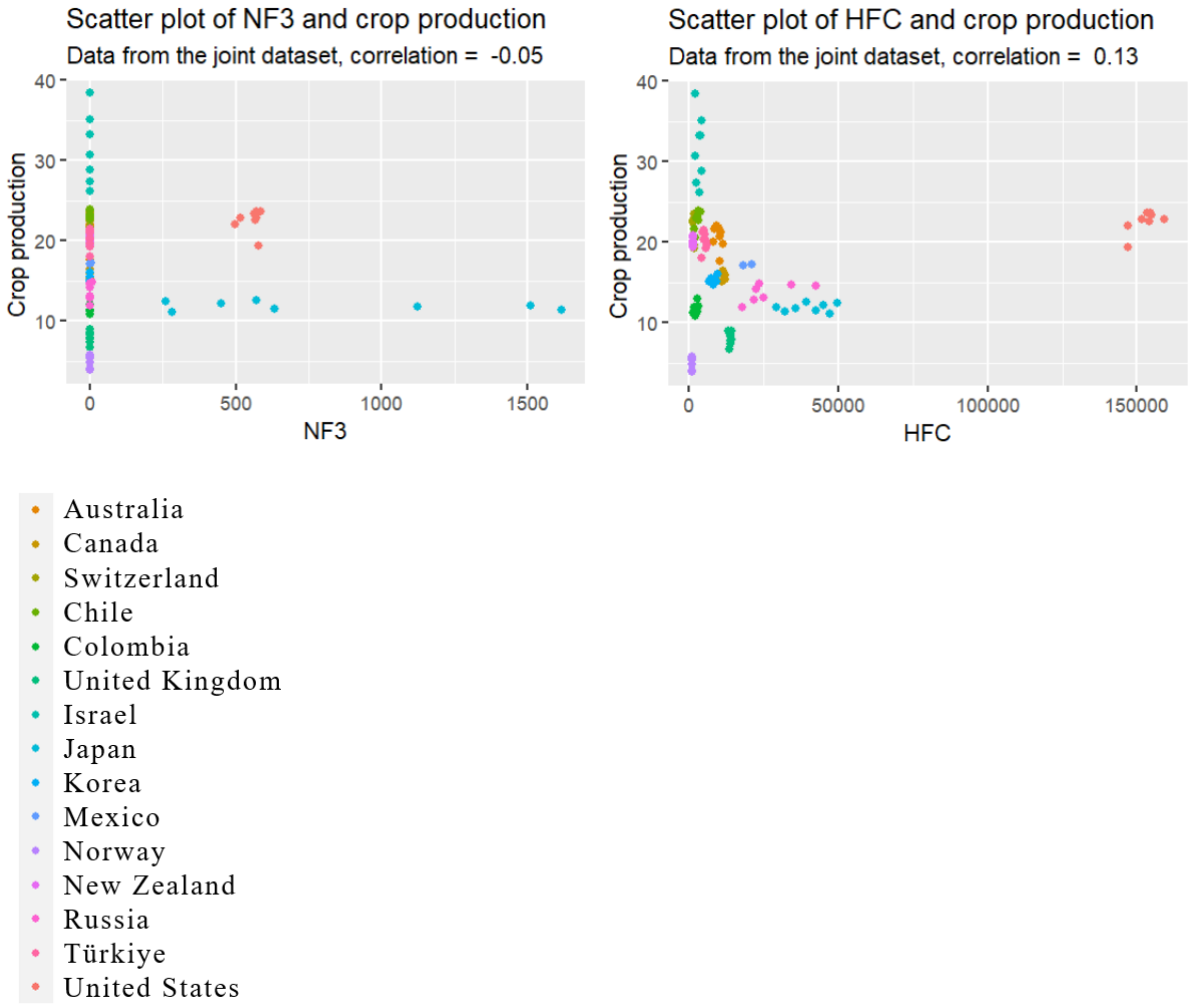


Figure 5. Relationship of the emissions and croproduction