STAT 499 Undergraduate Research (Directed Reading Program Spring 2024)

Understanding Greenhouse Gas Emissions through Rashomon Effect

Wenxin(Ariel) Xia

Introduction

This research project delves into the complex and multifaceted nature of greenhouse gas emissions, leveraging the Rashomon Effect to understand the variability and contradictions in data interpretation. Named after Akira Kurosawa's 1950 film "Rashomon," the Rashomon Effect highlights how different perspectives can lead to varying accounts of the same event. By employing linear regression and Lasso models, this study compares and contrasts the outcomes generated from these statistical methods to illustrate the diverse narratives presented by the data. Through meticulous analysis and model comparison using the merged dataset, the research aims to underscore the significance of model selection and its implications on our understanding of environmental data, ultimately advocating for a more nuanced approach to interpreting factors that influence greenhouse gas emissions the most.

Datasets Selection & Dataset Merging Process

The dataset selection and merging process for this research project was a critical step in ensuring the reliability and comprehensiveness of the analysis. We utilized datasets from reputable sources, including the World Bank and Our World in Data, focusing on key variables such as agricultural land percentage, electric and hybrid car sales, forest area, GDP per capita, meat production, municipal waste recycling rates, energy consumption per capita, urban population percentage, rural population percentage, CO2 and greenhouse gas emissions per capita as the response variable.

To integrate these diverse datasets, a meticulous merging process was undertaken. This involved aligning the datasets based on common identifiers such as country and year to ensure consistency. In order to process the "NA" values, we firstly filtered the data from 1990 to 2023, which are the years that mostly contain data, and replaced all the NA values with 0 values to lower the data processing complexity. By carefully curating and merging these datasets, we aimed to create a robust foundation for the subsequent application of linear regression and Lasso models, allowing for a comprehensive analysis of the factors influencing greenhouse gas emissions and their various interpretations.

Linear Regression & Lasso Model

Fitting the datasets into linear regression and Lasso models yielded insightful results, highlighting the importance of different predictors in understanding greenhouse gas emissions. The linear regression model identified recycled waste, meat production, energy consumption per capita, forest area, GDP per capita, and urban population as significant predictors. This model aimed to minimize the mean squared error (MSE) and generated subsets based on various combinations of these predictors, ultimately selecting those with the lowest MSE values for a robust analysis.

On the other hand, the Lasso model, which incorporates regularization to enhance prediction accuracy and interpretability, identified a slightly different set of significant predictors. These included recycled waste, forest area, farmland, energy consumption per capita, GDP per capita, and meat production. The Lasso model's inclusion of a penalty term to shrink less important variables to zero proved useful in focusing on the most impactful predictors.

Rashomon Effect

The comparison between the linear regression and Lasso models highlighted the Rashomon Effect in data analysis, where different methodologies provide varying perspectives on the same issue. While both models emphasized the importance of recycled waste and GDP per capita, their differences in other predictors underscore the complexity and multi-faceted nature of interpreting greenhouse gas emissions data. This comparison

underscores the need for careful model selection and consideration of multiple analytical approaches to fully understand and address environmental challenges.