Sigma J Eng Nat Sci, Vol. 44, No. 5, pp. 3357–3369, October, 2026

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2026.2162

Research Article

Robust regression analysis of carbon dioxide emissions dataset for Türkiye

Elgiz ASKEROĞLU[1], Fatma Zehra DOĞRU[2,*]

[1]*Department of Statistics, Institute of Science, Giresun University, Giresun, 28200, Türkiye*
[2]*Department of Data Science and Analytics, Giresun University, Giresun, 28200, Türkiye*

ARTICLE INFO

Article history
Received: 16 December 2024
Revised: 13 March 2025
Accepted: 15 June 2025

Keywords:
CO_2 Emissions; Generalized Least Squares; Least Squares; Robust Regression

ABSTRACT

This research examines Türkiye's carbon dioxide emissions from 1973 to 2021. As carbon dioxide emissions have become one of the key factors in climate change related policy discussion. This study aimed to determine which economic factors or industry sectors affect the level of carbon dioxide emissions. Therefore, robust regression, least squares and generalized least squares techniques are used on carbon dioxide emissions data set. Robust regression method outperforms traditional regression models in comparison to least squares and generalized least squares. Robust regression yields more reliable and efficient results by eliminating the effects of outliers and enhancing the stability of the model. Model performance of these regression estimation techniques are evaluated using mean squared error. Model diagnostics are performed to check the fundamental regression assumptions including heteroscedasticity, normality, autocorrelation, and multicollinearity. When multicollinearity is present, variables are eliminated using backward elimination method. After elimination of variables, all regression estimation techniques are used again to determine which model fits best. Autocorrelation and heteroscedasticity are corrected if there are present. Our findings show that carbon dioxide emissions are affected by per capita energy use, carbon footprint per capita, and secondary sector' share of gross domestic product. According to the mean squared error, the robust regression model outperforms all other models. We conclude that Türkiye's growth has been followed by increasing dependence on fossil fuels. Policies encouraging renewables and energy efficiency should become a priority.

Cite this article as: Askeroğlu E, Doğru FZ. Robust regression analysis of carbon dioxide emissions dataset for Türkiye. Sigma J Eng Nat Sci 2026;44(5):3357–3369.

INTRODUCTION

The industrial revolution witnessed major breakthroughs in terms of human welfare; technology; and material and commodities production. At the same time with these positive aspects of industrial revolution, some negative externalities started to draw more attention over time. Among the most concerning of these negative externalities is the effect of increasing carbon dioxide (CO_2) emissions

*Corresponding author.
*E-mail address: fatma.dogru@giresun.edu.tr

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey
© Author. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

on climate change. The purpose of this study is to review the Türkiye CO_2 emissions data and try to understand both negative and positive impacts of industrialization. This study aims to find out the major driving factors of Türkiye's CO_2 emissions, determine the relationship between energy consumption-economic growth-emission nexus, and check the effectiveness of environmental regulations on the reduction of emissions for sustainable policymaking.

Industrialisation is generally known to have three categories of effects on the environment: Atmosphere Pollution, Soil Pollution and Water Pollution. Atmosphere pollution affects the air humans breathe and significantly affects the greenhouse effect. Scientific prediction says that if we could reduce the average temperature of Earth by 1.5 degrees Celsius to 7.5 degrees Celsius the Earth would go fully glaciated. Also, Earth would have a tropical environment if we increase the temperature by 1.5 degrees Celsius to 5.5 degrees Celsius [1]. Prediction of increasing average temperature has been a large topic of recent publications due to large amounts of data we can get from glaciers. Ice caps and ice sheets preserve valuable information about the composition of the atmosphere on Earth, both chemically and physically. Information is generally collected from ice cores taken at certain locations [2].

One of the first papers to discuss how using fossil fuels could increase the temperature of our atmosphere was written by [3]. They believed that at the rate we were using fuel back then it would take thousands of years for the amount of CO_2 to be doubled in our atmosphere. [4] also talked about this topic. He discussed how if we continued to burn coal at our current rates it would affect the temperature of our planet. Showing that the earth could see an increase in its average temperature from us using pit coal.

After industrialization took effect and CO_2 started increasing many researchers started to take interest in this subject. [5] was one of these researchers who studied the temperature changes from 1880-1930's. He predicted that there would be a 0.3 degree increase in temperature around the world. By studying the industrial age he was able to discover these results. [6] researched the properties that CO_2 has on our atmosphere. Which allowed us to learn more about what causes global warming. His studies allowed for a better understanding on how climate works. [7] looked at how city's can affect the world climate. He stated that changes can have large impacts on your surrounding area. It is difficult to know if climate will continue to fluctuate or if we will cause changes. There are many problems that need to be looked into such as heat rejection and CO_2.

Temperature data for past 100 years has been used by [8] to show an increase in temperature throughout the world. This research has helped to understand how temperatures have changed over a longer period of time and its effect on climate. Research by [9] used data from the Berkeley Earth Surface Temperature (BEST) database to look at global temperature from 1850-2021. Their results showed that global temperature has increased over time similar to previous studies. Showing that we need to start working on solving this problem.

Climate change has become an issue that has taken over the world so much that in 2015 the UN held a conference in Paris to try and keep global temperature below 2°C [10]. They also wanted to aim for a maximum temperature increase of 1.5°C. Many people know about how harmful climate change can be to humans and the environment but [11] believes that these temperatures are unattainable due to how much carbon is being released into our atmosphere and how dependent the world is on fossil fuels. He also brings up that although many countries have promised to reach zero carbon emissions by 2050 there are no plans in place to do this. [12] predicted that all current power plants, cars and factories will release 496 gigatons (Gt) of CO_2 between the years of 2010-2060. Which they believe could cause global warming to rise by 1.3°C. If we continue to add more carbon based machines we could see our global emissions increase by 2986-7402 Gt CO_2 by the year 2100. Which will cause our temperature to increase by 2.4-4.6°C.

More needs to be done to help limit climate change. [13] discusses how we need to start lowering our emission of CO_2 asap in order to decrease climate change. There should be more laws put into effect for industry, travel and daily usage of CO_2 to lower our emission of Greenhouse gases and fight climate change. To allow us and the earth to have life for years to come.

The purpose of this research is to determine CO_2 emissions in Türkiye and analyze their effect on environment and economy. There has been an increase in the number of studies investigating the correlation of CO_2 emissions with energy use and economic growth. These studies allow us to understand what factors cause these emissions to increase. Studies by [14] and [15] showed that Türkiye's CO_2 emissions correlate to income and energy use. As a country's income grows so does its environmental pollution. In addition [16] looked at how population, fossil fuels and gross domestic product affect CO_2 emissions from energy consumption. They concluded that these were the primary causes of CO_2 emissions from energy. [17] studied different sectors of industry and showed that policy should be implemented to help lower emissions from each sector. They mentioned the agriculture, industrial and transportation sectors. [18] studied a certain region and calculated how much CO_2 emissions there are from city trash disposal in the Thrace region.

There are many studies that show that technology is able to help lower carbon emissions. [19] stated that with higher production technologies can lead to less environmental damage while allowing economies to grow. There are many studies that have found that policies have not been effective at lowering CO_2 emissions. [20] stated that although there are ecological tax's in place there have not been any decreases in CO_2 emissions. [21] showed that Türkiye is an importer of CO_2 emissions. With industries

such as education health and finance accounting for most of those emissions.

There are many research papers that support the idea of renewable energy reducing CO_2 emissions in Türkiye. In [22] they used quantile regression and found that economic growth and population increase CO_2 but renewable energy helped reduce emissions. Renewable energy has been found to reduce CO_2 emissions when paired with an increase in total factor productivity in [23]. While increases in income and imports have increased emissions. [24] researched Türkiyes energy policy and how they still rely on fossil fuels to help meet their emissions goal.

The research shown above concludes that Türkiye needs to continue with renewable energy development and use better policies that will allow it to progress towards a sustainable economy. Although growth is going to have a big impact on Türkiyes emissions. With new technology and cleaner energy methods and policies can decrease Türkiye's emission and help create a sustainable environment.

There are many different variables that can help explain CO_2 emissions. By looking at how they relate to each other we can learn about how much a country uses electricity and oil as well as its GDP and total CO_2 emission level. It is important to find these correlations so we can work towards solving our current problems. In this paper we will be taking these variables and learning more about what can cause CO_2 emissions. We hope to help further research on this topic and help create better policies for sustainable living.

For this purpose we will be using Least Squares(LS), Generalized Least Squares(GLS) and Robust Regression (RR) to test CO_2 emissions in Türkiye. Most regression models have some requirements that need to be satisfied before you can use them. These include normality, no outliers and homoscedasticity. Many times these conditions will not be met when working with data that pertain to our environment or economy. If there are extreme observations in your data set, heteroscedasticity or if your regressors are correlated. Your least squares estimator can become very sensitive to these problems. So in order to overcome these issues we will also be using robust regression which allows for more reliable results by decreasing the effect outliers have on the model.

This will be about the effect of different factors that can affect CO_2 emission levels. These factors include GDP, consumption of energy and laws regarding environment. We will also look at the different explanatory variables and test for any underlying assumption violations.

The rest of this study will consist of Section 2 which describes the LS, GLS and RR methods and eliminating variables. Section 3 will describe our CO_2 emissions data. In section 4 we will analyze our data using the LS, GLS and RR methods. We will test for any underlying assumption violations and eliminate variables. Section 5 will consist of our conclusions from the analysis.

MATERIALS AND METHODS

Regression analysis can be used to find relationships between different variables. There are many different types of regression that can be applied to almost any field of science. Engineering, Physics, Chemistry, Biology, Economics, Social Sciences, Life Science just to name a few. Below we will discuss what type of regressions were used in this paper.

Multiple Linear Regressions

Let $\{(x_1, y_1),\ldots,(x_n, y_n)\}$ represent a given sample, where y_i is the dependent variable, $\boldsymbol{x}_i$ is the vector of independent variables, $\boldsymbol{\beta}$ is the vector of unknown regression parameters, and ϵ_i denotes the random error term. Assuming a linear relationship between y_i and x_i, the multiple linear regression model is defined as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad (1)$$

or, equivalently,

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i,\ i = 1, \ldots, n\ , j = 1, \ldots, k. \quad (2)$$

Let $p = k + 1$, then the equation (2) can be written in a more compact form as:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i. \quad (3)$$

Alternatively, in matrix notation, let $\boldsymbol{X}$ be the $n \times p$ design matrix consisting of independent variables, $\boldsymbol{y}$ be the $n \times 1$ vector of dependent variables, $\boldsymbol{\beta}$ be the $p \times 1$ vector of unknown regression parameters, and ϵ be the $n \times 1$ vector of random errors. The regression model can then be expressed as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \quad (4)$$

To ensure that the linear regression model provides reliable estimates, the following core assumptions must hold:

-The dependent variable $\boldsymbol{y}$ should have a linear relationship with $\boldsymbol{X}$.

-Observations are independent of one another.

-Homoscedasticity meaning the variance of the errors is equal. $\mathrm{Var}(\epsilon) = \sigma^2 I$

- $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$

- $\boldsymbol{X}$ variables should not be correlated with each other.

-ϵ_i's should follow a normal distribution $\epsilon_i \sim N(0, \sigma^2)$.

If these are met we can get efficient estimates using LS.

Least Squares (LS) Method

LS method is a widely used technique for estimating the regression parameters in equation (4). This method aims to minimize the sum of squared errors, which represents the differences between the observed and predicted values of the dependent variable. The objective function for the least squares estimation is given by:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (y - \boldsymbol{X\beta})'(y - \boldsymbol{X\beta}). \quad (5)$$

Minimizing this function concerning $\boldsymbol{\beta}$ leads to the LS estimator:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

While LS is computationally efficient and provides the best linear unbiased estimates (BLUE) under normality, it is highly sensitive to outliers and violations of model assumptions. This sensitivity motivates the use of alternative estimation methods such as GLS and RR.

Generalized Least Squares (GLS) Method

The GLS method is a specialized version of the LS method, particularly useful when the dataset exhibits autocorrelation. The GLS estimator of $\boldsymbol{\beta}$ is obtained as follows:

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\Sigma^{-1}\boldsymbol{y}. \quad (6)$$

The matrix Σ in equation (6) is described as a positive and symmetric $n \times n$ matrix whose (i, j)th element is equal to $Cov(e_i, e_j)$. Here, $Cov(e_i, e_j)$ represents the covariance of the residuals of $\hat{\boldsymbol{\beta}}_{GLS}$. Furthermore, the GLS estimator can be obtained by transforming the LS method. Let S be the Cholesky decomposition of Σ. Then, the transformed model given in equation (3) becomes:

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \quad (7)$$

where $\boldsymbol{Y}^* = S^{-1}\boldsymbol{y}$, $\boldsymbol{X}^* = S^{-1}\boldsymbol{X}$ and $\epsilon^* = S^{-1}\,\epsilon$. Alternatively, the GLS estimator can be obtained as:

$$\widehat{\boldsymbol{\beta}}^* = (\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*\prime}\boldsymbol{Y}^*. \quad (8)$$

Robust Regression (RR) Method

In regression analysis, the presence of outliers or deviations from model assumptions can significantly affect parameter estimates, leading to biased or inefficient results. LS is known to be a best estimator when all assumptions are met. However it can become very inefficient when there are outliers present in the data. Therefore robust regression (RR) methods were created to allow us to get good estimates even with outliers and/or heteroscedasticity. These methods allow us to down weight extreme values while being efficient on normal data.

Among various robust approaches [25-30], the M-estimator introduced by [25] is one of the most commonly used robust regression approaches. Although it is not fully resistant to extreme values, it is effective when outliers are primarily found in the dependent variable *y*.

The M-estimation method minimizes a function of residuals that grows at a slower than the LS function, thereby reducing the impact of large residuals. The M-estimator of $\boldsymbol{\beta}$ is obtained by minimizing the following objective function:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(\epsilon_i) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}), \quad (9)$$

where x_i' is the *i* th row of the design matrix $\boldsymbol{X}$, and $\rho(\cdot)$ is a function that determines how residuals are incorporated into the estimation. The LS method minimizes the sum of squared residuals, whereas RR methods use alternative loss functions that penalize large residuals less. This reduces the sensitivity to outliers. The M-estimator is obtained by the maximum likelihood approach and has the scale-invariant property as:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{\epsilon_i}{\sigma}\right) = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left[\rho\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right], \quad (10)$$

where σ is a robust estimate of scale, computed as:

$$\sigma = \frac{median|\epsilon_i - median(\epsilon_i)|}{0.6745}. \quad (11)$$

With the help of the following equation, $\boldsymbol{\beta}$ and σ parameters can be solved simultaneously:

$$Q(\boldsymbol{\beta}, \sigma) = \frac{1}{n}\sum_{i=1}^{n} \left[\rho\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) + \alpha\right]\sigma. \quad (12)$$

There are several ρ functions in the literature with different robustness properties. This study uses Tukey's bisquare function introduced by [31] as a particularly efficient choice

$$\rho_c = \left(1 - \left(\frac{x}{c}\right)^2\right)^3, \quad |x| \le c \qquad 1, \ |x| > c. \quad (13)$$

The Tukey's bisquare function is commonly used as it strikes a balance between efficiency and robustness. It downweights large residuals, thereby effectively reducing their impact on parameter estimation while retaining high efficiency for well-behaved data.

Multicollinearity and Variable Omission

All explanatory variables included in a regression model are not always statistically significant. Thus, selection of variables is an important part of regression analysis. One problem that can arise during the process of variable selection is multicollinearity, which is the phenomenon of explanatory variables being highly correlated with each other. If multicollinearity is detected then one of the common suggestions is to remove some of the correlated variables from the model. Often , variable elimination methods are used in such cases .

There are many variable elimination methods documented in the literature. For example, [32] proposed the leaps and bounds algorithm, which uses information calculated from regression to aid variable selection. However, techniques such as bootstrap resampling, predicted residual error sum of squares (PRESS), deviance or residual deviance,

adjusted R^2, or cross-validation can be computationally demanding, especially when applied under complex criteria. These more advanced methods typically need to be computed so they can not be applied to typical situations.

In addition to traditional methods, modern regularization techniques-including Ridge regression [33], Lasso regression [34], and Principal Component Regression (PCR) [35]-have been widely adopted to address multicollinearity and aid in variable selection. These methods shrink regression coefficients, and in some cases, set some coefficients exactly to zero, effectively identifying the most relevant predictors while enhancing model stability.

Many newer forms of regularization have also been implemented to deal with multi-collinearity. These include but are not limited to Ridge regression [33] and Lasso regression [34]. Both will reduce coefficient estimates and at times can make coefficients equal to 0 allowing for better predictor selection. Principal Component Regression (PCR) is another example [35]. Regularization can help with predictor selection when there are many variables.

But with small sample sizes these techniques can give worse results because of the high penalty values. Because our dataset had such a small sample size we decided to use variable elimination techniques. There are many ways to go about eliminating variables from your model. Step wise elimination will not always give you the best model but can allow you to reduce the complexity of your model. So we used a step wise variable selection technique called backward elimination.

Backward Elimination Method

The backward elimination method is a widely used variable elimination technique in the literature. This method is generally less negatively affected by correlations between regressors than other methods [36]. If a variable leads to a lower sum of squared errors (SSE), it should be retained. When a variable's effect on SSE reduction isn't substantial, you'd go ahead and remove it. After a variable is removed, you then assess which other variable is contributing least to the reduction of SSE. This iterative approach continues until only variables that significantly reduce your SSE remain, or until no variables are left. This approach ultimately leads to a statistically more precise model.

Data

CO_2 is one of the biggest problems for the enviroment. It's a menace to Earth and everything on it. The variables we will be looking at that can impact CO_2 emissions are Carbon Footprint, Fuel Consumption Per Capita, Electricity Consumption Per Capita, Primary Sector of GDP Percentage, Secondary Sector of GDP Percentage, Tertiary Sector of GDP Percentage and GDP Per Capita. By using these variables and CO_2 emissions, what affects CO_2 in Türkiye between 1973 and 2021 will be determined.

The per capita CO_2 emissions index is calculated by dividing the total tons of emissions by the population of the country, taken from Worldmeters, WDB (website), and Eurostat (website). Per capita fuel consumption is the amount of fuel consumed in kilograms by the country's population. Source: WDB, CEIC. Per capita electricity consumption means total electricity consumption in kWh divided by the population including industrial entities in the country (WDB, TURKSTAT (Turkish Statistical Institute)). The carbon footprint metric in this study is the land area needed to sequester all CO_2 emissions from human activities throughout their life cycle [37] and is from Footprintnetwork (website) on a per hectare basis.

A country calculates something called GDP to measure the value of the goods and services it produces during a specific period of time. Per capita GDP is found by dividing the calculated GDP value by the population of the country at the time. The variable of GDP per capita in this study is standardized by the exchange rate in 2017 dollars obtained from the WDB . The variable GDP per capita has different sectors such as the primary, secondary and tertiary. This measure includes variables that are derived from GDP per capita. It can be divided into three categories: primary, secondary and tertiary. Primary sector will be the percent of agri production to GDP. Our secondary sector will be the share of industry production in GDP. And finally, the percent of income in the service industry. It is tertiary sector. It has been obtained from WDB (Table 1).

Table 1. Defining variables

Description	Variables	Source
CO_2	Carbon Dioxide Emission Per Capita	Worldometers (1973-2014) and World Data Bank (2015-2021)
EPC	Energy Consumption Per Capita	World Data Bank (1973-2014) and TURKSTAT (2015-2021)
CFP	Carbon Foot Print Per Capita	Footprintnetwork (1973-2021)
S1	Sector 1	World Data Bank (1973-2021)
EUO	Oil Consumption Per Capita	World Data Bank (1973-2015) and CEICDATA (2016-2021)
S2	Sector 2	World Data Bank (1973-2021)
S3	Sector 3	World Data Bank (1973-2021)
GDP	Gross Domestic Products Per capita	World Data Bank (1973-2021)

ANALYSES

The study examines variables that may affect CO_2 emissions, including energy consumption, carbon footprint, oil consumption, GDP and percentage of GDP in the primary, secondary and tertiary sectors. Table 2 presents the data used for Türkiye for the period 1972-2021. The mean for the CO_2 is 3.2 and standard deviation is ± 1.1. The EPC mean was found to be 1551.1 and SD ± 964.2. CFP had a mean of 1.3 with SD ± 0.4. EUO has a mean of 1151.8 and SD of ± 391.1 whereas GNP has a mean of 5084.2 and SD of ± 3924.3. S1, S2 and S3 which are percentage of primary, secondary and tertiary sectors have mean of 15.4, 27.5 and 50.2 with SD of ± 8.9, ± 2.9 and ± 4.7. Standardization will be employed before comparability between variables measured in different units is assured. The dataset has no missing observations. If there are missing observations then we could have used imputation method such as mean, median, mode, forward/backward fill. More information on missing values can be found here [38-42].

Figure 1 shows pairwise graphs to explore the relationship between the CO_2 emissions and the independent variables. These plots show roughly linear relationships between CO_2 emissions and several predictors, in particular: CO_2-EPC, CO_2-CFP, CO_2-S1 and CO_2-GDP. Taking into account the linear relationships observed, the linear regression method was considered suitable for the analysis of the data.

Figure 2 illustrates the correlation matrix. The correlation matrix delineates the relationship between CO_2 emissions and various economic and sectoral indicators.

Table 2. Descriptive statistics

Variables	Observation	Mean	Standard Deviation	Median	Mod	Minimum Value	Maximum Value	Standard Error
CO_2	49	3.2	1.1	3.2	1.5	1.6	5.3	0.1
EPC	49	1551.1	964.2	1418.7	1285.4	292.9	3386	137.7
CFP	49	1.3	0.4	1.3	0.6	0.6	2.014	0.1
S1	49	15.4	8.9	14.5	9.5	5.5	36.002	1.3
EUO	49	1151.8	391.1	1109.9	460.1	640.4	1979.4	55.9
S2	49	27.5	2.9	26.8	3.4	22.5	32.9	0.4
S3	49	50.2	4.7	51.1	3.9	38.5	57.2	0.7
GDP	49	5084.2	3924.3	3100.4	2569.7	676.4	12507.8	560.61

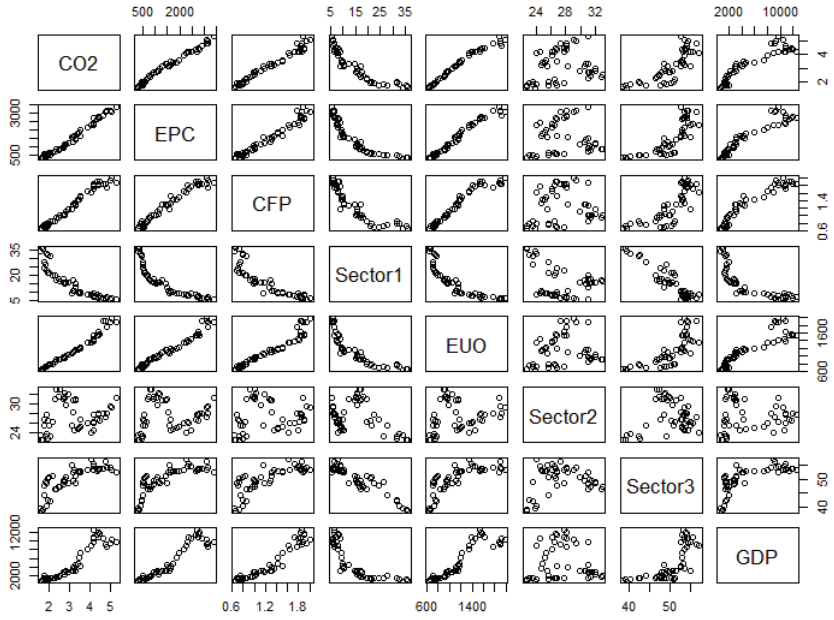

Figure 1. Pairwise graphs.

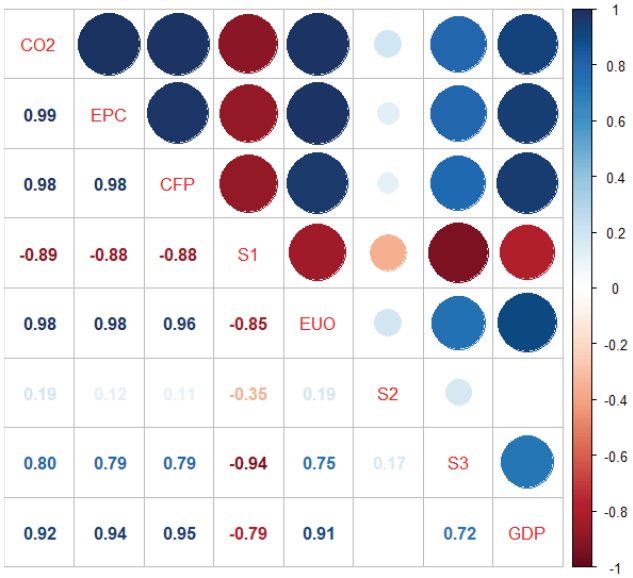

Figure 2. Correlation analysis.

Specifically, it is determined that GDP (0.92), CFP (0.98), and EPC (0.99) are significantly positively correlated with CO_2 emissions. The results suggest that energy consumption, economic growth, and carbon footprint are the most critical factors contributing to the increase in emissions in Türkiye. Additionally, the evidence indicates that the CO_2 emissions are strongly correlated with EUO (0.98), suggesting that the emissions are primarily sourced from fossil fuels.

In contrast, S1 (-0.89) is highly negatively correlated with CO_2 emissions showing structural transformation of economy where emissions are increasingly coming from industrial and service sectors and not from agriculture. In addition, S3 (0.80) has a significant positive correlation with CO_2 which indicates the environmental implications of service based economic activities. Taken together, the trends suggest that CO_2 emissions are likely to increase with industrialization, the expansion of economies into services and the growth of economies driven by increasing energy use, economic growth and dependence on petroleum inputs.

Following the correlation analysis, the regression model is specified as follows:

$$CO2 = \beta_0 + \beta_1 EPC + \beta_2 CFP + \beta_3 S1 + \beta_4 EUO + \beta_5 S2 + \beta_6 S3 + \beta_7 GDP + \varepsilon. \quad (14)$$

To assess the overall significance of the model, the following hypotheses are tested:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1: \beta_j \neq 0 \; for \; at \; least \; one \; j.$$

The estimated values of the LS, GLS and RR methods are presented in Table 3. It gives estimates of coefficients, adjusted R2, standard error (SE), p-values, Durbin Watson test statistic (and p-values) and mean square error (MSE) to help compare models. Sign of relationships is same but the coefficient estimates differ by method. The analysis indicates that EPC and CFP are significant at α=0.001 and positively correlated with CO_2. S2 is significant (α=0.05) and positively correlated. Other variables are not statistically significant at 0.001, 0.01 and 0.05 level. However, variable EUO is important in the RR approach. However, S1 and S3 models are not significant statistically. The share of agricultural output in the GDP per capita has little impact on the CO_2 emissions. This is because it generates organic waste and is relatively small in comparison with other industries . For example, the services sector is thought to

Table 3. Estimation results for LS, GLS, and RR methods

	LS			GLS			RR (Bisquare)		
Variables	**Coefficients**	**SE**	**p**	**Coefficients**	**SE**	**p**	**Coefficients**	**SE**	**p**
Constant	-3.8056e-16	0.0160	1.000	0.0013	0.0158	0.9351	-0.0086	0.0102	0.4007
EPC	0.6139	0.1097	0.0001***	0.6363	0.1130	0.00001***	0.4297	0.1059	0.0001***
CFP	0.4981	0.0704	0.0001***	0.4577	0.0693	0.00001***	0.5228	0.0679	0.0001***
S1	0.0815	0.0743	0.2794	0.0761	0.0828	0.3635	0.0243	0.0717	0.7363
EUO	0.0709	0.0807	0.3851	0.0969	0.0850	0.2614	0.1579	0.0779	0.0493**
S2	0.0662	0.0208	0.002843**	0.0628	0.0253	0.0175*	0.0488	0.0201	0.0196**
S3	0.0575	0.0447	0.2058	0.0499	0.0487	0.3118	0.0363	0.0432	0.4055
GDP	-0.1725	0.0433	0.00028***	-0.1743	0.0523	0.0018***	-0.1386	0.0418	0.002**
Durbin-Watson p	0.0001***			0.0380			0.0190		
MSE	0.0742			0.0736			0.0716		
Adjusted R^2	0.9945			0.9884			0.9950		

Significance levels 0.05** 0.001***

make a negligible contribution to GDP per capita in terms of CO_2 emissions. This sector includes low carbon emission industries such as hospitality and dining and tourism and healthcare. All these factors point to a relatively low direct impact of both sectors on CO_2 emissions, when compared with industrial and energy-intensive sectors.

Moreover, the RR method exhibits lower standard errors compared to LS and GLS, and achieves the smallest MSE, indicating better model performance and robustness.

Although the RR method yields superior results, it is important to note that not all explanatory variables are statistically significant Additionally, the validity of a regression model depends not only on statistical significance but also on adherence to key model assumptions. As reported in Table 3, the Durbin-Watson test statistic across all estimation methods suggests the presence of autocorrelation. The null hypothesis, stating no autocorrelation, is rejected when using a 0.05 significance level. As there is a lack of independence in the errors we will continue to run more tests to see if our model is valid. The following tests will be applied one after another under the LS method. Allowing us to test for lack of normalcy, homoscedasticity, independent errors and if any influential points exist. P value will be used to determine our test statistic by using:

$$p = 2P(T > |t|) = 2\big(1 - F_t(|t|, df)\big), \qquad (15)$$

where $F_t(\cdot)$ is the cumulative distribution function of the Student t distribution and df is the degrees of freedom parameter.

Table 4. Tests of homogeneity and normalization

Kolmogorov-Smirnov		Breusch-Pagan	
p	6.014e-09	p	0.0859

Figure 3 shows residual vs fitted, Q-Q plot of Residuals, Standardized residual vs fitted value and cooks distance which are used to determine if there are outliers in the data. Outliers can be identified by using these figures.

The Breusch-Pagan and Kolmogorov-Smirnov tests are employed to test the assumptions of homoscedasticity and normality of the regression model. The results are reported in Table 4. The hypotheses for homogeneity of variance are:

H_0: *Homoscedasticity is present (the residuals are distributed with equal variance)*

H_1: *Heteroscedasticity is present (the residuals are not distributed with equal variance).*

The Breusch-Pagan test presented in Table 4 shows a p-value of 0.08595, indicating no significant heteroscedasticity. In contrast, the Kolmogorov-Smirnov test reports a p-value of 6.014e-09, indicating a significant departure from normality in the residuals.

Table 5 shows the VIF of all the independent variables. As we can see only S2 has a value less than 5 which means that all other variables have a very high value of VIF. EPC, CFP, S1 and EUO have very high multicollinearity.

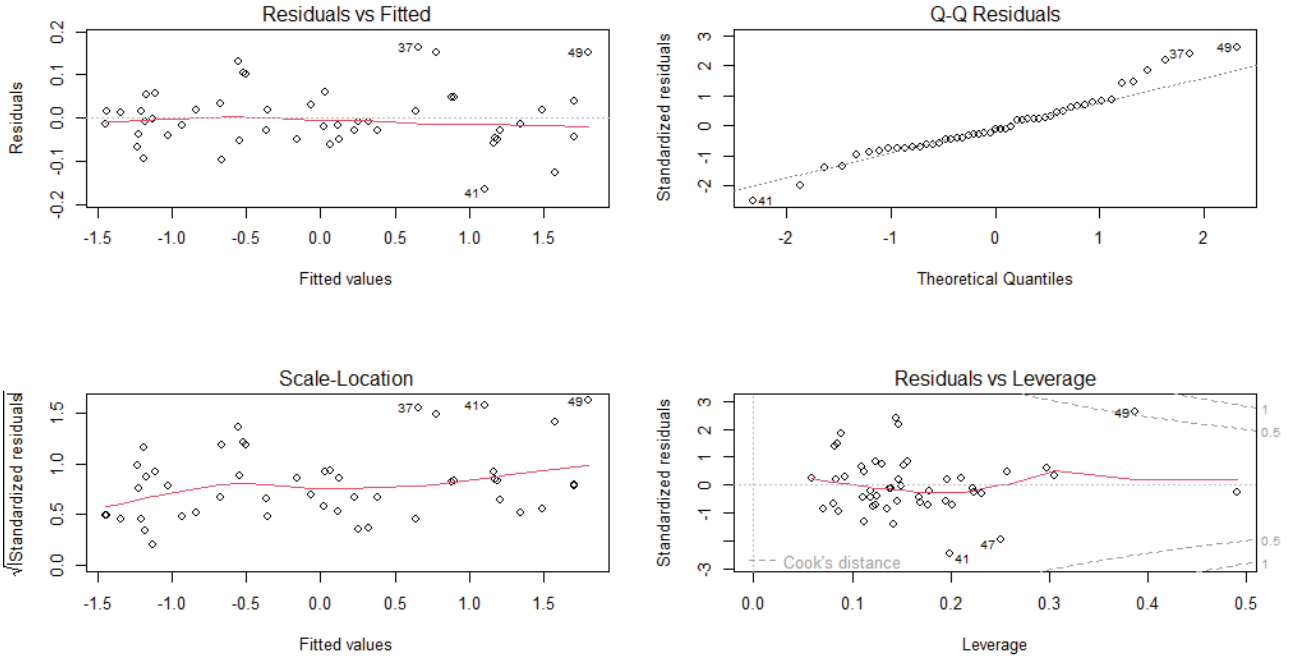

Figure 3. Residual analysis.

Table 5. Investigation of factors affecting variance

	EPC	CFP	S1	EUO	S2	S3	GDP
VIF	104.96	43.23	48.23	56.88	3.79	17.47	16.41

Table 6. Variance decomposition ratios

	Variance Ratios							
CI	β_0	**EPC**	**CFP**	**S1**	**EUO**	**S2**	**S3**	**GDP**
1.000	0.000	0.000	0.001	0.001	0.001	0.000	0.001	0.002
2.262	0.000	0.000	0.000	0.001	0.000	0.220	0.000	0.003
2.326	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3.688	0.000	0.001	0.002	0.007	0.005	0.042	0.080	0.011
8.297	0.000	0.008	0.001	0.001	0.078	0.058	0.001	0.438
12.437	0.000	0.001	0.217	0.163	0.045	0.108	0.283	0.200
19.551	0.000	0.127	0.689	0.383	0.006	0.157	0.425	0.043
30.807	0.000	0.863	0.090	0.445	0.866	0.416	0.210	0.303

As can be seen from Table 6 there is evidence of multi-collinearity as well. With the 8th CI value being greater than 30 (CI= 30.807). There are also high proportions of variance for the coefficients of EPC, CFP and EUO (0.863, 0.090 and 0.866 respectively). Consequently, these coefficients appear to be most susceptible to the issue of multicollinearity.

To address multicollinearity, backward elimination was applied using the Bayesian Information Criterion (BIC) [43]. The results of this elimination method are given in Table 7 we note that similar results were obtained using the forward elimination method.

The results, detailed in Table 7, identify the optimal subset of predictors: GDP, S2, CFP, and EPC. This model is expressed as:

$$\widehat{CO2} = \hat{\beta}_0 + \hat{\beta}_1 GDP + \hat{\beta}_2 S2 + \hat{\beta}_3 CFP + \hat{\beta}_4 EPC. \quad (16)$$

The estimated results of the model identified by backward elimination using LS, GLS and RR methods are given

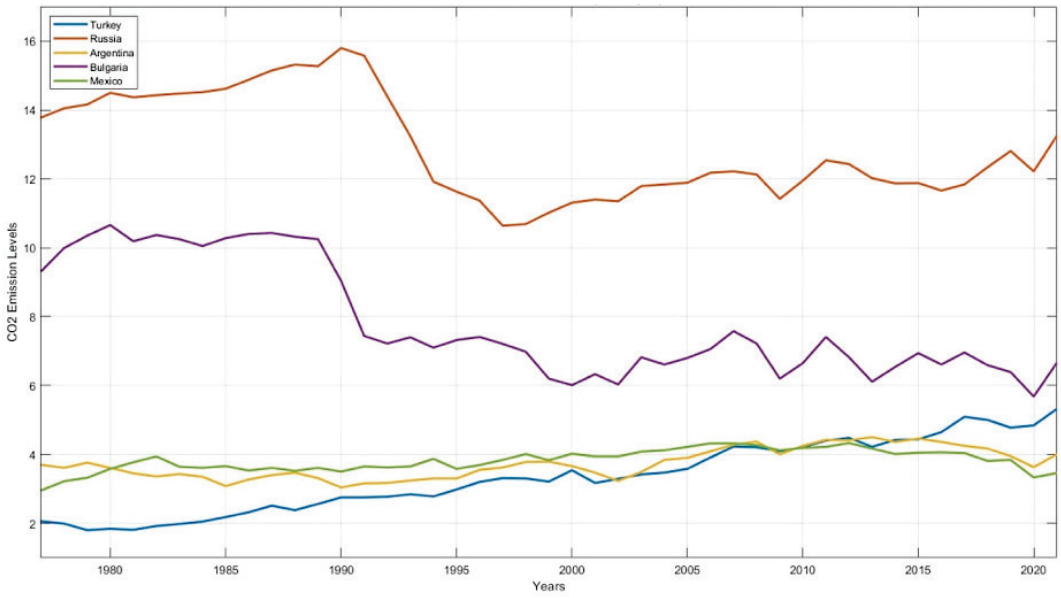

Figure 4. CO_2 emissions trends for selected countries.

Table 7. Backward elimination method results

Equation	**1**		
	Sum of Squares	**Residual Sum of Squares**	**Bayes Information Criteria**
EUO	0.004243	0.23000	-235.47
S1	0.006614	0.23237	-234.97
S3	0.009099	0.23486	-234.45
S2	0.055506	0.28126	-225.61
GDP	0.087097	0.31285	-220.40
EPC	0.172381	0.39814	-208.58
CFP	0.275530	0.50129	-197.29
Equation	**2**		
Bayes Information Criteria	**-235.47**		
	Sum of Squares	**Residual Sum of Squares**	**Bayes Information Criteria**
S3	0.01390	0.24390	-236.49
S1	0.01625	0.24626	-236.02
S2	0.10687	0.33687	-220.66
GDP	0.12554	0.35554	-218.02
CFP	0.33189	0.5618	-195.59
EPC	0.83695	1.06695	-164.17
Equation	**3**		
Bayes Information Criteria	**-236.49**		
	Sum of Squares	**Residual Sum of Squares**	**Bayes Information Criteria**
S1	0.00237	0.24628	-239.90
S2	0.10846	0.35237	-222.35
GDP	0.11892	0.36283	-220.92
CFP	0.32737	0.57127	-198.68
EPC	0.85398	1.09788	-166.66
Equation	**4**		
Bayes Information Criteria	**-239.9**		
	Sum of Squares	**Residual Sum of Squares**	**Bayes Information Criteria**
GDP	0.11680	0.36308	-224.78
S2	0.11980	0.36608	-224.37
CFP	0.34396	0.59024	-200.97
EPC	0.86944	1.11572	-169.77

Table 8. Estimation results after the elimination

LS				**GLS**			**RR (Bisquare)**		
Variables	**Coefficients**	**SE**	**p**	**Coefficients**	**SE**	**p**	**Coefficients**	**SE**	**p**
Constant	-7.52e-16	0.0106	1.000	3.2053	0.0176	0.0001***	3.2092	0.0179	0.0001***
EPC	0.6950	0.0557	0.0001***	0.7813	0.0690	0.0001***	0.7478	0.0699	0.0001***
CFP	0.4668	0.0595	0.0001***	0.5057	0.0675	0.0001***	0.4939	0.0684	0.0001***
S2	0.0544	0.0117	0.0001***	0.0517	0.0164	0.0029***	0.0565	0.0167	0.0014***
GDP	-0.1778	0.0389	0.0002***	-0.214	0.0538	0.0002***	-0.1641	0.0545	0.0043***
Durbin-Watson p	0.0001			0.0990			0.0490		
MSE	0.0748			0.0807			0.0587		
Adjusted R^2	0.9944			0.9987			0.9971		

Significance levels 0.05** 0.001***

in Table 8. This table summarizes coefficients, adjusted R^2, SE, p-values of significance, Durbin-Watson p-values and MSE values of all the methods. All coefficient estimates for the three methods are statistically significant at =0.001. Moreover, the autocorrelation problem previously reported by the Durbin-Watson test seems to be solved when GLS and RR are applied, with values of p of 0.0990 and 0.0490, respectively. The RR model is the best of the three with the lowest MSE. The better performance of RR as well as the resolution of multicollinearity and autocorrelation show that this model gives the most reliable and robust representation of the relationships between CO_2 emissions and the identified economic and sectoral indicators.

Comparative Study of CO_2 Emission Trends Between Economically Similar Countries

It is important to see the trends of CO_2 emission in the countries which have a similar economic structure with Türkiye. Unlike other countries with similar GDP level such as Argentina, Bulgaria and Mexico where trends have been fluctuating or stable, Türkiye's per capita CO_2 emissions have been on a steady upward trend (Fig. 4). Historically, Russia has had higher emissions, but there was a large drop in the 1990s, probably because of economic changes following the collapse of the Soviet Union.

This divergence indicates that Türkiye's economic growth has been tightly coupled with rising fossil fuel consumption, while emissions in similar economies have flattened or declined. With Türkiye's CO_2 emissions still on the rise, the country should continue to implement policies that will enable the use of renewable resources and improve the environment. This comparison indicates that there are many ways to grow our economy and improve our environment.

CONCLUSION

There are many ways in which we can harm our environment. We need to figure out what caused this so we can do a better job of keeping it from happening again. CO_2 emissions are one thing that can have a negative effect on our environment. In this paper we will study the relation between CO_2 and different sectors of life using model equations.

The results above show that RR is better than LS and GLS in terms of MSE. Also interesting is the observation that the RR model shows per capita oil consumption to be an important variable, which is not the case for the LS or GLS models. As noted before, the non-significant variables were predicted by the LS method. This shows that classical regression assumptions such as multicollinearity may have been violated. The model was improved by removing non-significant variables using a backward elimination approach.

The RR solved the problem of autocorrelation, more efficient than the estimates. The results show that it is possible to draw a conclusion on Türkiye. There is a negative correlation between electricity consumption, fuel consumption and industrial CO_2 and GDP per capita. One unit of EPC will increase CO_2 by 0.7478 units. for CO_2, an increase of one unit will increase CO_2 by 0.4939% with a significance level of 0.001 and S2 will increase by 0.0565 units for each unit increase. CO_2 has a negative effect on GDP with a value of -0.1641 which may lead to an increase in emissions. We can grow our economy and grow our emissions. Clearly, the continuous development of the industrial sector, built on these energies, will result in the degradation of our environment and the slowing down of our economy. So, EPC is the main driver of CO_2 emission growth. We need to have policies to reduce energy consumption. "This will help lower emissions and grow our economy."

This is consistent with other studies that have found that fossil fuels are a major source of our energy supply and hence a major source of CO_2 emissions. Variable S3 was also eliminated in the elimination of variables, which can be explained by the service industry consuming a large amount of plastic products. This results in an increased import of goods with plastics or use of petroleum based methods.

Policymakers need to help Türkiye to grow economically with renewable energies. Türkiye's increasing investment in renewable energy should lower its CO_2 emissions – and today's evidence supports that. Such renewable energies include solar, wind and hydropower. All of which have been shown to cut emissions. This will not only cut emissions but will help Türkiye become more self-sufficient.

Another limitation is that this research might not fully capture the dynamic or lag effect. As this study uses cross section with aggregated variables we can not say that results are dynamic. In addition, the complex interaction between energy, economy and environment has not been considered. An important goal is to apply the RR model to time-series and panel data. This enables us to track these dynamics over time and across countries. And by applying variable elimination with robust techniques we are also able to guaranty that the model is not affected by any extreme points. To this model you could add more indicators. Future research could take into account metrics like the share of renewable energy, energy prices and technological advances.

ACKNOWLEDGEMENTS

The insightful comments and helpful suggestions provided by the associate editor and reviewers have significantly enhanced this manuscript. We are grateful. As part of the "2224-B-Grant Program for Participation in Scientific Meetings within the Country," the Scientific and Technological Research Council of Türkiye (TUBITAK) provided financial support to the primary author (grant number: 1919B022403795).

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Plass GN. Carbon dioxide and the climate. Am Sci 1956;44:302–316.

[2] Thompson LG. Ice core evidence for climate change in the Tropics: implications for our future. Quat Sci Rev 2000;19:19–35. [CrossRef]

[3] Arrhenius S. On the influence of carbonic acid in the air upon the temperature of the ground. Philos Mag 1896;41:237–276. [CrossRef]

[4] Ekholm N. On the variations of the climate of the geological and historical past and their causes. Q J R Meteorol Soc 1901;27:1–62. [CrossRef]

[5] Callendar GS. The artificial production of carbon dioxide and its influence on temperature. Q J R Meteorol Soc 1938;64:223–240. [CrossRef]

[6] Plass GN. The carbon dioxide theory of climatic change. Tellus 1955;8:140–154. [CrossRef]

[7] Landsberg HE. Man-made climatic changes: man's activities have altered the climate of urbanized areas and may affect global climate in the future. Science 1970;170:1265–1274. [CrossRef]

[8] Hansen J, Lebedeff S. Global surface air temperatures: update through 1987. Geophys Res Lett 1988;15:323–326. [CrossRef]

[9] Wang L, Wang L, Li Y, Wang J. A century-long analysis of global warming and earth temperature using a random walk with drift approach. Decis Anal J 2023;7:100237. [CrossRef]

[10] United Nations. Paris Agreement. Available at: https://www.un.org/en/climatechange/paris-agreement. Accessed November 11, 2026.

[11] Rothenberg G. A realistic look at CO2 emissions, climate change and the role of sustainable chemistry. Sustain Chem Clim Action 2023;2:100012. [CrossRef]

[12] Davis SJ, Caldeira K, Matthews HD. Future CO2 emissions and climate change from existing energy infrastructure. Science 2010;329:1330–1333. [CrossRef]

[13] Craig MT, Jaramillo P, Hodge BM. Carbon dioxide emissions effects of grid-scale electricity storage in a decarbonizing power system. Environ Res Lett 2018;13:014004. [CrossRef]

[14] Halicioglu F. An econometric study of CO2 emissions, energy consumption, income and foreign trade in Türkiye. Energy Policy 2009;37:1156–1164. [CrossRef]

[15] Ozturk I, Acaravci A. CO2 emissions, energy consumption and economic growth in Türkiye. Renew Sustain Energy Rev 2010;14:3220–3225. [CrossRef]

[16] Aydin G. The development and validation of regression models to predict energy-related CO2 emissions in Türkiye. Energy Sources Part B Econ Plan Policy 2015;10:176–182. [CrossRef]

[17] Akbostancı E, Tunç Gİ, Türüt-Aşık S. Drivers of fuel based carbon dioxide emissions: the case of Türkiye. Renew Sustain Energy Rev 2018;81:2599–2608. [CrossRef]

[18] Dinçer AR, Aral İF. Carbon dioxide emission in Thrace region and neighboring countries as a result of aerobic stabilization of urban treatment sludges. Sigma J Eng Nat Sci 2020;38:1729–1739.

[19] Bodur S, Küpeli M, Alp İ. Decoupling analysis of environmental pressures from economic growth in the EU-27 and Türkiye. Sigma J Eng Nat Sci 2021;39:29–38.

[20] Telatar OM, Birinci N. The effects of environmental tax on ecological footprint and carbon dioxide emissions: a nonlinear cointegration analysis on Türkiye. Environ Sci Pollut Res Int 2022;29:44335–44347. [CrossRef]

[21] Mangır N, Şahin ÜA. An environmentally extended global multi-regional input-output analysis of consumption-based and embodied import-based carbon emissions of Türkiye. Environ Sci Pollut Res Int 2022;29:54813–54826. [CrossRef]

[22] Türkyılmaz S, Öztürk KN. Analysis of factors affecting CO2 emissions in Türkiye using quantile regression. Sustainability 2024;16:9634. [CrossRef]

[23] Mukhtarov S. Do renewable energy and total factor productivity eliminate CO2 emissions in Türkiye? Environ Econ Policy Stud 2024;26:307–324. [CrossRef]

[24] Bolat E, Yildiz YA. Energy demand and CO2 emission forecast model for Türkiye with deep learning and machine learning algorithms. Elektron Elektrotech 2025;31:12–21. [CrossRef]

[25] Huber PJ. Robust regression: asymptotics, conjectures and Monte Carlo. Ann Stat 1973;1:799–821. [CrossRef]
[26] Andrews DF. A robust method for multiple linear regression. Technometrics 1974;16:523–531. [CrossRef]
[27] Rousseeuw PJ, Yohai V. Robust regression by means of S-estimators. In: Franke J, Härdle W, Martin RD, editors. Robust and nonlinear time series analysis. New York: Springer; 1984. p. 256–274. [CrossRef]
[28] Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. Robust statistics: the approach based on influence functions. New York: Wiley; 1986. 502 p.
[29] Yohai VJ. High breakdown-point and high efficiency robust estimates for regression. Ann Stat 1987;15:642–656. [CrossRef]
[30] Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. John Wiley & Sons; 2003.
[31] Beaton AE, Tukey JW. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics 1974;16:147–185. [CrossRef]
[32] Furnival G, Wilson R. Regression by leaps and bounds. Technometrics 1974;16:499–511. [CrossRef]
[33] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 1970;12:55–67. [CrossRef]
[34] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58:267–288. [CrossRef]
[35] Jolliffe IT. A note on the use of principal components in regression. J R Stat Soc Ser C Appl Stat 1982;31:300–303. [CrossRef]
[36] Myers RH, Montgomery DC. Response surface methodology: process and product optimization using designed experiments. New York: Wiley; 2002.
[37] Pandey D, Agrawal M, Pandey JS. Carbon footprint: current methods of estimation. Environ Monit Assess 2011;178:135–160. [CrossRef]
[38] Little RJ, Rubin DB. Statistical analysis with missing data. John Wiley & Sons; 2019. [CrossRef]
[39] Little RJ. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227–1237. [CrossRef]
[40] Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255–1264. [CrossRef]
[41] Schafer JL. Analysis of incomplete multivariate data. CRC Press; 1997. [CrossRef]
[42] Vach W. Logistic regression with missing values in the covariates. Vol. 86. Springer Science & Business Media; 2012.
[43] Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6:461–464. [CrossRef]