T.R.
ONDOKUZ MAYIS UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF STATISTICS

ESTIMATION OF ELECTRICITY ENERGY PRODUCTION: KARBALA EXAMPLE

Master's Thesis

Ahmed Azeez ABDULHUSSEIN

Supervısor
Assoc. Dr. Hasan BULUT

SAMSUN
2023

T.R.
ONDOKUZ MAYIS UNIVERSITY
INSTITUTE OF GRADUATE STUDIES
DEPARTMENT OF STATISTICS

ESTIMATION OF ELECTRICITY ENERGY PRODUCTION: KARBALA EXAMPLE

Master's Thesis

Ahmed Azeez ABDULHUSSEIN

Supervisor

Assoc. Dr. Hasan BULUT

SAMSUN
2023

ACCEPTANCE AND APPROVAL OF THE THESIS

The study entitled **"ESTIMATION OF ELECTRICITY ENERGY PRODUCTION: KARBALA EXAMPLE"** was prepared by **Ahmed Azeez ABDULHUSSEIN**, and supervised by **Assoc. Dr. Hasan BULUT** was found successful and unanimously accepted by committee members as a Master thesis, following the examination on the date 12.1.2023.

	Title Name SURNAME **University** **Department/Art**	**Final Decision**
Chairman (Supervisor)	Prof. Dr. Necati Alp ERILLI Sivas Cumhuriyet University Department of Econometrics	☒ Accept ☐ Reject
Member	Prof. Dr. Kamil ALAKUŞ Ondokuz Mayıs University Department of Statistics	☒ Accept ☐ Reject
Member	Assoc. Prof. Dr. Hasan BULUT Ondokuz Mayıs University Department of Statistics	☒ Accept ☐ Reject

This thesis has been approved by the committee members that already stated above and determined by the Institute Executive Board.

.

Prof. Dr. Ali BOLAT
Head of Institute of Graduate Studies

DECLARATION OF COMPLIANCE WITH SCIENTIFIC ETHIC

I hereby declare and undertake that I complied with scientific ethics and academic rules in all stages of my Master's Thesis, that I have referred to each quotation that I use directly or indirectly in the study, and that the works I have used consist of those shown in the sources, that it was written in accordance with the institute writing guide and that the situations stated in the article 3, section 9 of the Regulation for TÜBİTAK Research and Publication Ethics Board were not violated.

Is Ethics Committee Necessary?

Yes ☐ (If it necessary, please add appendices.)

No ☒

… /… / 2023
Ahmed Azeez ABDULHUSSEIN

DECLARATION OF THE THESIS STUDY ORIGINALITY REPORT

Thesis Title : ESTIMATION OF ELECTRICITY ENERGY PRODUCTION: KARBALA EXAMPLE

As a result of the originality report taken by me from the plagiarism detection program on 20 /12. /2022 for the thesis titled above;

Similarity ratio : % 20

Single resource rate : % 2 has been released.

… /… / 2023
Assoc. Dr. Hasan BULUT

ÖZET

ELEKTRİK ENERJİSİ ÜRETİMİNİN TAHMİNİ: KERBALA ÖRNEĞİ

Ahmed Azeez ABDULHUSSEIN
Ondokuz Mayıs Üniversitesi
Lisansüstü Eğitim Enstitüsü
İstatistik Ana Bilim Dalı
Yüksek Lisans, Ocak/2023
Danışman: Accoc. Dr. Hasan BULUT

Regresyon analizinde temsil edilen istatistiksel tahmin süreçleri ve diğer istatistiksel süreçler, onun hakkında yazdıkları çok sayıda ve farklı türleri ve bölümlerine rağmen, yaşamın eklemlerinin gelişmesiyle daha önemlidir. Özellikle incelenen veriler Çoklu-Eş-doğrusallık probleminden muzdarip olduğunda. Ve günlük hayatımızda, özellikle ekonomik olanlar olmak üzere, bu sorundan muzdarip birçok olgunun varlığı, bu nedenle tez, Sıradan En Küçük Kareler yönteminin (OLS) regresyon modeli olan regresyon katsayılarını tahmin etme yöntemlerini karşılaştırmaya geldi ve sırt regresyon modeli. Çoklu doğrusal modellerin katsayıları, Sıradan En Küçük Kareler yöntemi (OLS) ve sırt regresyon yöntemi kullanılarak tahmin edildi.

Varyans şişirme faktörü VIF, problemin model verilerinde olup olmadığını belirlemek için kullanılmıştır. parametre.

Farklı yöntemler arasındaki karşılaştırma, ortalama kare hatası MSE, p değeri ve artıklar aracılığıyla yapılır.

Yöntemleri pratikte uygulamak için, elektrik enerjisi üretim istasyonlarından birinde elektrik enerjisi üretiminin 50 gözleminden oluşan rastgele bir örnek çizilmiş, elektrik enerjisi üretimini etkileyen faktörleri incelemek için y: Elektrik Enerjisi (tepki değişkeni) ve birkaç bağımsızdır. değişkenler (X_1:Kara Yağ, X_2: Gazyağı, X_3:Motor Yağ, X_4:Kimyasallar, X_5:Çalışma, X_6:İş için Ek Ücret, X_7:Ortalama Aylık Sıcaklık)ve çoklu -lineerlik probleminde, Çalışma kapsamındaki verilerimizin regresyon modelini tahmin etmede Ridge regresyon yönteminin sıradan en küçük kareler yöntemine göre daha iyi olduğu sonucuna ulaşılmış, en düşük MSE elde edilmiş ve verilerdeki çoklu doğrusallık sorunu çözülmüştür.

Anahtar Sözcükler: Regresyon analizi, çoklu bağlantı problemi, Varyans Enflasyon Faktörü VIF, Adi En Küçük Kareler yöntemi (OLS), Ridge Regresyon yöntemi.

ABSTRACT

ESTIMATION OF ELECTRICITY ENERGY PRODUCTION: KARBALA EXAMPLE

Ahmed Azeez ABDULHUSSEIN
Ondokuz Mayıs University
Institute of Graduate Studies
Department of Statıstıcs
Master, January/2023
Supervisor: Assoc. Prof. Dr. Hasan BULUT

This thesis presents a comparison and an evaluation of the performance of different regression analysis methods for estimating electricity energy production. In particular, it compares the performance of the OLS method and the Ridge regression method by analyzing the relationship among several independent variables and electricity energy production. The study employs metrics like Mean Squared Error, p-value, and Coefficient of Determination to compare the results. Additionally, the Variance Inflation Factor is used to identify any multicollinearity issues present in the data.

The study is based on a random sample of (50) observations of electricity energy production from a power station. Seven independent variables (black oil, additional wages, engine oil, chemicals, operating cost, maintenance fee, and kerosene) are considered in the analysis, and their relationship with the electricity energy production (response variable) is examined. The sample was chosen to represent a real-world scenario and to provide a representative view of the factors affecting electricity energy production.

The results of the study show that the Ridge regression method outperforms the OLS method in terms of MSE and the ability to handle multicollinearity issues. The use of Ridge regression allows for a more accurate estimation of electricity energy production and can be useful in making decisions related to energy production and consumption. The conclusion suggests that Ridge regression is a better option for estimating electricity energy production and can be used as a tool for managing energy resources.

Keywords: Regression Analysis, Multicollinearity problem, Variance Inflation Factor VIF, Ordinary Least Squares Method OLS, Ridge Regression Method.

ACKNOWLEDGEMENT

Thanks and appreciation

I would like to extend my deep thanks and appreciation to each professor who attributed to me the stages of a study that I have reached today, and who contributed to building my scientific personality and spreading in it the love of hard work and perseverance.

I also do not forget my family, who endured a lot of trouble during my travel period, and left them, to fulfill my dream of completing my master's studies. My wife and children have been my best support in completing this prestigious degree.

I also extend my thanks, gratitude, and love to my distinguished teacher, Dr. (Hassan Bulut), who was my brother, friend, and outstanding teacher, and whenever I need him, I find him with me.

Thank you, sir, and I wish you continued success.

I also thank my friends who have always offered their financial and moral support, and I will never forget their virtue and their positions with me.

Finally, I dedicate this humble effort to my dear father and dear mother, who are anticipating my arrival to this degree of study.

Ahmed Azeez ABDULHUSSEIN

CONTENTS

FIGURES LEGENDS

TABLES LEGENDS

1. INTRODUCTION

Statistics has become a widely popular and important science due to its ability to enter into various aspects of life, thanks to its connection with machine learning and artificial intelligence. It enables us to handle, read, analyze, and predict data in various fields. Regression analysis can be defined as a crucial statistical technique which helps in understanding the relationship among variables in an equation known as the regression model, by estimating its features, we can conclude the importance, strength, and direction of this relationship. It is one of the fundamental principles of data analysis, which aims to identify the relationship among the dependent variable as well as other independent variables. To achieve this, we must reduce error rates and prevent over-fitting, avoiding the use of too many coefficients. (Brownlee, 2018).

Regression analysis is a statistical approach that can be applied to check the relationship among a dependent variable and one or more independent variables. By building a regression model, we can understand the nature of the association between variables and quantify it. It is essential to identify the factors that affect this model and to establish a standard method for forecasting continuous values. Regression analysis is a powerful tool for data modeling and analysis. (Mosleh et al., 2022).

Regression analysis is not without its challenges, one of which is the problem of multicollinearity, which can occur when the independent variables in the model are highly correlated. This can lead to weak estimates and unreliable results, as the OLS method relies on the assumption that the independent variables are independent. Multicollinearity can also affect the interpretation of the significance and importance of the variables in the model. The statistical researcher may encounter this issue when using the method of ordinary least squares and must take steps to address it to produce accurate and reliable results.

The problem of multicollinearity often arises in the use of multiple linear regression analysis, which occurs when the explanatory variables are highly correlated. This correlation increases the variance of the estimates of the regression coefficients and makes the results of the OLS method unreliable. It is a basic assumption of the OLS method that the explanatory variables must be independent, with no correlation between them.

When a model is applied to new measurements, the presence of multicollinearity in the training data can greatly impair its generalization capabilities, causing it to perform unreliably. To address this problem, researchers have put forth numerous approaches such as the Ridge regression approach developed by Horel and Kennard (1970). This method works by reducing the impact and compressing the variables by adding a constant value k to the main diagonal matrix of the data. This helps to address the problem of multicollinearity and reduces the mean squares error (MSE) compared to the ordinary least squares (OLS) method. Other methods such as the variance inflation factor VIF can also be used to reveal the presence of multicollinearity. However, choosing the right constant value k can be a challenge in the Ridge regression method. Therefore, an estimated regression model is built and parameters are found based on the chosen value of k, thus making the results more valid, logical, and reliable. (Santos et al., 2009).

In addition to the Ridge regression approach, other methods such as Lasso regression and Principal Component Regression (PCR) might be applied to address the problem of multicollinearity. These methods work by shrinking or rotating the variables in the model and have been shown to be effective in reducing the variance of the estimates and improving the accuracy of the model. Furthermore, it is important to note that multicollinearity not only affects the estimation of the regression coefficients but also affects the interpretation of the coefficients, making it difficult to infer causality or interpret the impact of a single predictor on the response variable. Therefore, it is recommended to carefully check for the presence of multicollinearity first, then interpreting the results of a regression analysis.

1.1. Objective of Study

The main objectives of the proposed thesis are the following:

- To examine the presence of the multicollinearity problem and its impact on the regression model.
- To investigate the correlation among variables, analyzing the degree of correlation between the independent variables and determining their effect on the dependent variable.
- To determine the most effective method by comparing the estimators of Ridge regression and Ordinary Least Squares.
- To make a comparison of the performance of Ridge regression and ordinary least squares method using the MSE criterion and Coefficient of Determination R^2, for various correlation coefficients.

1.2. Literature Review

A thorough literature review has been conducted to investigate the various studies and investigations that have examined the parameters of the regression model, with a focus on those utilizing the ordinary least squares technique and the Ridge regression approach. The following studies will be presented as they relate to the research topic:

- Ragnar Frisch, a Norwegian statistician, is widely considered one of the early pioneers in the study of multicollinearity relationships in the field of econometrics. His work in the 1930s, specifically in the analysis of time series, brought attention to the potential risks of having a large number of explanatory variables in a regression model. Frisch recognized that this could lead to multicollinearity, a phenomenon where the independent variables in the model are highly correlated, which can greatly affect the estimates of parameters and their variances. His contributions laid the foundation for future research in this area and many researchers have built upon his work in the decades since to further understand and address the challenges presented by multicollinearity in regression analysis.
- In 1970, Horel proposed a novel estimator called Ridge, as an alternative to the ordinary least squares (OLS) estimator when there is a multicollinearity problem. The Ridge method works by adding a non-zero number k to the principal diagonal

of the studied matrix. The same year, Marquardt explained that the Ridge method is a biased estimation method and it is applied if the independent variables are non-orthogonal.

- In the year 2000, Al-Qasimi employed a simulation method to check and make a comparison among the performance of the Ridge regression analysis method against other techniques in handling multicollinearity problems. The study aimed to understand the effectiveness of the Ridge method in dealing with correlated independent variables and its impact on the accuracy of the estimates.
- Kibria (2003) conducted a study that aimed to make a comparison among the performance of different estimation methods for Ridge regression parameters. The study introduced new estimations based on the generalized Ridge regression method and compared it to the Ordinary Least Squares approach using a simulation method. The study used the mean square error (MSE) criterion to evaluate the performance of the different methods. The study also aimed to investigate the effect of different correlation coefficients on the performance of the Ridge regression method in comparison to the OLS method.
- (Al-Katib, 2004) conducted a study that involved the use of trend analysis in factorial experiments and developed a solution to address the problem of multicollinearity that was encountered when using the Ridge regression method.
- A study by Al-Saadoun (2005) evaluated the effectiveness of the Normal Ridge regression method and the Bayesian Ridge regression method for finding the equation of a multiple linear regression model. By utilizing simulation data and comparing the models based on the mean square error (MSE) criterion, the researcher found that the Bayesian Ridge regression method was superior to the Normal Ridge regression method.
- A study by AL-Hassan in 2010 explored the use of Ridge regression as a replacement for the traditional OLS method of estimation in cases where there is a multicollinearity issue among independent variables. The study suggested a novel approach for determining the Ridge parameter and evaluated its effectiveness using simulation data by comparing the mean square error (MSE) results.
- Wichern & Churc Kibriahill (2012) conducted a study that provided evidence that the Ridge regression method reduces the outliers of the ordinary least squares

method when data suffers from a multicollinearity problem. They used simulations and real datasets to show the validity of this method.

- In 2014, Mohammed conducted a study on the problem of multicollinearity and its impact on the performance of regression models. The study aimed to make a comparison among the effectiveness of the Ridge Regression method and the Principal Component Regression (PCR) approach in addressing the issue of multicollinearity. The study found that the intensity of multicollinearity had a direct impact on the Mean Squared Error (MSE) of the model, regardless of the sample size or variance level. The study also found that the PCR method performed better than the Ridge Regression method in addressing the issue of multicollinearity and recommended the use of the PCR method when dealing with complete or incomplete multicollinearity in the model.
- A study by Fitrianto and Yik (2014) evaluated the variance of the regression model in comparison to the ordinary least squares (OLS) approach. They found that the variance of the model was higher when utilizing OLS, indicating a high correlation among the independent variables in multiple linear regression models. The research further revealed that Ridge regression, a technique used to tackle multicollinearity, performed better in comparison to OLS. This study supports previous research that has shown the effectiveness of Ridge regression in addressing multicollinearity issues.
- Duzan and Shariff (2015) did research to check the effectiveness of the Ridge regression method in addressing the problem of multicollinearity. They used simulation data and applied the Ridge regression method as an alternative to the Ordinary Least Squares (OLS) method when dealing with highly correlated explanatory variables. The results of their study showed that the Ridge regression technique was able to reduce the outliers present in the OLS method when dealing with multicollinearity. This suggests that the Ridge regression approach is a useful tool for addressing the problem of multicollinearity and producing more reliable and accurate results.
- Owusu (2018) explored the use of multiple linear regression analysis to understand the relationship among various independent variables and a dependent variable. He highlighted the challenges that arise when the independent variables are highly correlated, and how this would give inconsistent estimates of the

regression coefficients. To simplify this, he presented the Ridge regression analysis as a regularization method that can mitigate the effect of high correlation among the independent variables and provide more stable and consistent estimates of the regression coefficients.

- Ali Sadig (2018) conducted a study to examine the factors impacting the production cost of the Muthanna Cement Plant by utilizing the Ridge regression method. The cost of production quantity was considered as the dependent variable, while the influencing factors such as price per ton, electrical energy, and consumed quantity were considered the independent variables that were affected by the problem of multicollinearity. The study found that the Ridge regression method was the most effective when analyzing financial and economic data. The analysis was performed using the R-Studio software.

2. METHOD AND MATERIAL

2.1. Basic Concepts

We will provide an overview of the key concepts and definitions that will be used throughout the study. Understanding these fundamental concepts is essential for interpreting and understanding the results and findings presented in the research. It will help the reader to have a clear understanding of the material and methods used in the study.

- Independence: The idea that giving information of the value of a variable does not provide information about the value of the other variable. In the presence of a correlation between variables, they are not independent.
- Variables: Quantities or attributes that can take on different values and are the focus of the study.
- Random variables: mathematical functions that take a set of possible outcomes (sample space) and map them to a set of numerical values (real numbers).
- Dependent variable: The variable of interest in the study, whose value is affected by one or more independent variables.
- Independent variable: this type would be defined by the variable whose effect on the dependent variable is being studied. Its value is not affected by other variables.
- Orthogonal variables: Variables that have no correlation with each other. In this case, multiple regression is not necessary to check the relationship among the dependent variable as well as independent variables.
- Limited vs infinite relationships: In limited relationships, changes in the dependent variable are caused by the influence of independent variables only. However, in infinite relationships, changes in the dependent variable may be caused by a combination of other variables of unknown origin.
- The concept of statistical relationship versus mathematical relationship highlights key differences between the two, such as the fact that mathematical relationships are complete and have known parameters, while statistical relationships are incomplete and have unknown parameters. Additionally, the values of variables in mathematical relationships are unknown, while in statistical relationships, the values of variables are known.

- Variance is a measure of the spread of data and is represented by the square of the standard deviation. It helps to understand how much variation exists within a set of data.
- Interaction refers to the mixed effect of a couple of independent variables on a dependent variable. It can be thought of as a third independent variable.
- Correlation is the study of the relationship among a couple of variables and helps to determine the results of this relationship. There are several types of correlation coefficients that one could apply to measure the relationship among variables, such as simple correlation (between two variables only), partial correlation (between two variables while excluding the influence of one or more variables), and multiple correlations.
- The Pearson correlation coefficient is a statistical measure used to determine the linear relationship between two variables. It is represented by the symbol "r" and can be calculated using various formulas. The coefficient ranges from -1 to 1, where a value of -1 indicates a strong negative correlation, 0 indicates no correlation, and 1 indicates a strong positive correlation. This coefficient is frequently used in simple linear regression analysis to examine the association between a dependent variable and an independent variable, it can be calculated by using different formulas:

$$\boldsymbol{r_{x1x2} = \frac{s_{xy}}{s_x s_y} = \frac{\sum(x_1-\bar{x}_1)\sum(x_2-\bar{x}_2)}{\sqrt{\sum(x_1-\bar{x}_1)^2(x_2-\bar{x}_2)^2}}} \quad (2.1)$$

$$\boldsymbol{= \frac{n\sum x_1x_2 - \sum x_1 \sum x_2}{\sqrt{n\sum x_1^2-(\sum x_1)^2}\sqrt{n\sum x_2^2-(\sum x_2)^2}}} \quad (2.2)$$

$$\boldsymbol{= \frac{\sum x_1x_2 - n\bar{x}_1\bar{x}_2}{\sqrt{n\sum x_1^2-(n\bar{x}_1)^2}\sqrt{n\sum x_2^2-(n\bar{x}_2)^2}}} \quad (2.3)$$

- Partial correlation coefficient: It checks the strength as well as the direction of the relationship among the variables after removing the effect of other variables included in the study. This coefficient is represented by the symbol "$\boldsymbol{r_{x1\,x2.x3}}$" and is calculated using a specific equation that takes into consideration the three variables involved.

$$r_{x1\ x2.x3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}. \qquad (2.4)$$

Since $r_{x_1x_2.x_3}$ is the partial correlation coefficient between the first and second variables, excluding the effect of the third variable, and r_{12}, r_{13}, r_{23} are the simple correlation coefficients between the different variables.

- Multiple correlation coefficient: It is used to measure the relationship among more than two variables and is symbolized by the symbol R, and the sign of the correlation coefficient here does not indicate the direction of the relationship, because this trend is not uniform for all variables and it takes only positive values.

$$R = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1-r_{23}^2}} \qquad (2.5)$$

$$R = \sqrt{1-(1-r_{12}^2)(1-r_{12.3}^2)} \qquad (2.6)$$

- Coefficient of Determination R^2: It measures and explains the percentage of total deviations or changes that occur in the dependent variable y_i, which accounts for the changes of the independent variable x_i, it is the percentage of the influence of the independent variables on the dependent variable. The following formula can calculate it:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{\sum \hat{e}^2}{\sum_{i=1}^{n}(yi-\bar{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}i-\bar{y})^2}{\sum_{i=1}^{n}(yi-\bar{y})^2} \qquad (2.7)$$

R^2 is one of the most important transactions that measure the correlation between two variables, the existence of such a relationship implies that one of these two variables depends on the change or occurrence of the other variable, the determination parameter is defined and belongs to the following domain.

$$0 \le R^2 \le 1$$

Difficulties in using R^2 as a measure of the quality because this coefficient depends on changes in y (explained and unexplained) and thus, the number of degrees of freedom does not explain the statistical problems.

- There is another parameter called the modified coefficient of determination (R_{Adj}^2), which can be used if there is a multicollinear problem in the data, which is defined by the following formula:

$$R^2_{Adj} = 1 - \frac{SSR/(n-k-1)}{SST/n-1} \qquad (2.8)$$

Where n the number of witnesses and $(k+1)$ is the number of estimated parameters and with a simple substitution, we find:

$$R^2_{Adj} = 1 - (1 - R^2)\left(\frac{n-k-1}{n-1}\right) \qquad (2.9)$$

From the above equation, the relationship among R^2 and R^2_{Adj} appears, where: $k > 1$,$R^2 > R^2_{Adj}$, $k = 1$,$R^2 = R^2_{Adj}$

If the sample size n is very large, then R^2 and R^2_{Adj} are close in value, but in small samples, if the number of independent variables is large compared to the sample size, then R^2_{Adj} is much less than R^2, and it can take negative values. In this case, it must be explained on the basis that its value is zero. Since R^2_{Adj} has a set of properties that make it a better means of measuring the quality of conciliation than R^2, it at least answers the questions of some researchers about the importance of increasing the number of variables for the model, without thinking about the reason for the emergence of these variables, in any case. Nevertheless, one should not think that R^2_{Adj} solves all the problems related to the R^2 scale for the quality of fit, since the decision about the possibility of some variables appearing in the model or not, remains dependent on other theoretical considerations. In economic measurement, the numerical value of R^2_{Adj} is very sensitive to the type of data or data used.

- R^2_{Adj}: This R-squared variant has been tweaked to account for the model and the number of predictors in it, which always falls short of the R^2, the modified R-squared might be applied to find if various regression models can fit together or not.
- The difference between R $and\, R^2$: The essential difference between the coefficient of determination and the coefficient of correlation lies in causation, where the correlation coefficient checks the relationship among a couple of variables regardless of the role that each variable plays (we do not need to specify the independent and dependent variables). As for the coefficient of determination, it also measures the correlation but takes into account the causation since the second variable x_i, is the one that explains the phenomenon y_i.

- Model: A representation and simulation of a real system used to analyze its behavior and improve its performance by finding the optimal formula for the system in the future.
- Modeling: The process of building a model that expresses the true dimensions of a particular system or an idea, provided that the representation process reflects a clear picture of the nature of the relationships that link the components and elements of the system.
- Graphical Representation: A chart that visually displays information, which helps in understanding numbers and comparisons between them.
- Predicted Values: The values that result from the prediction of the regression model after estimating parameters for each observation in the model.
- Residuals: represent the differences among the real values of the dependent variable as well as the other values that have been estimated by the regression model for each observation.

2.2. Regression models

A regression model is a mathematical tool which can be applied to understand the relationship among dependent variables as well as independent variables. This method can be used to identify patterns and trends in data and is widely applied in various fields such as business, economics, and finance to analyze the impact of different factors on a given outcome, like production, prices, and stock performance.

There are two main types of regression models: simple and multiple. Simple regression models involve a single independent variable and a dependent variable, while multiple regression models involve multiple independent variables. Additionally, regression models can be further divided into linear and non-linear models. The models of linear regression would suppose that the relation among the independent as well as dependent variables could be linear, while non-linear models allow for more complex relationships.

In data science and machine learning, there are different forms of regression models, each with its specific applications. However, at their core, all regression techniques aim to assess the effect of independent variables on a dependent variable. (Jeewantha, 2021).

2.2.1. Linear Regression Analysis (LRA)

Linear Regression Analysis (LRA) is a statistical technique used to investigate the connection between a dependent variable and one or more independent variables. The term "regression" was first used by statistician Francis Fulton in 1877 when he examined the patterns of height in human populations and noticed that the children of tall and short parents tend to move towards the average height (Krashniak et al, 2021).

Regression is a statistical technique that involves constructing a model that estimates the relationship among a dependent variable and one or more independent variables. The resulting equation can be used to make predictions about the dependent variable, based on the values of the independent variables (Gunst et al, 2018).

The main goal of regression analysis is to understand the nature of changes in variables by analyzing a set of data and estimating the parameters of the model. LRA can be used for a variety of tasks, such as describing the data, forecasting and estimating responses, and controlling the values of the dependent variable by determining the values of the independent variables that affect it. LRA can be further divided into two categories: simple linear regression and multiple regression. (Frost, 2017).

The three tasks that the regression model can be used for include:

1- Describing the data of a phenomenon by finding a regression equation that describes the phenomenon in general.

2- Forecasting and estimating the response to make appropriate decisions for the model.

3- Controlling the values of the dependent variable by determining the values of the independent variable affecting the response.

2.2.1.1. Simple Linear Regression (SLR)

Simple linear regression is the most basic type of regression model, which involves a relationship among a single independent variable and a single dependent variable, represented by the following equation:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 * X + \varepsilon_i \quad , \quad i = 1 , \dots , n} \qquad (2.10)$$

$$\boldsymbol{Y_i \sim N\ (\mu_i, \sigma^2) \quad ; \quad i = 1, 2, \dots, n} \qquad (2.11)$$

β_0 and β_1 are parameters of the regression model

Y: is the Target Variable, Label

X: is the Predictor Variable, Feature

β_0: is the bias, intercept of the regression line (can be obtained putting x=0),

β_1: is the weight in the regression equation, the slope of the regression line, or the Linear regression coefficient

And ε_i: is the random error (for a good model, it will be negligible);

ε_i Represents the error in interpreting y_i and can be written in the following form.

$$\boldsymbol{\varepsilon_i \sim N\ (0, \sigma^2)} \qquad (2.12)$$

$$\boldsymbol{\varepsilon_i = y_i - \beta_0 - \beta_1 x_i} \qquad (2.13)$$

The presence of an error term, also known as the residual term, in a regression model is a fundamental aspect of the model. This term represents the difference between the observed value of the dependent variable and the predicted value of the dependent variable based on the values of the independent variables. The error term is essential for understanding the model's accuracy and the presence of any systematic errors.

There are several reasons for the existence of the error term in a regression model. One of the main reasons is the presence of omitted variables. In other words, there may be independent variables that affect the dependent variable, but they were not included in the model. This can lead to a biased estimation of the regression coefficients, and a lack of fit in the model. Additionally, there may be measurement errors in the data collection process, which can also lead to errors in the estimation of the model's parameters.

Another reason for the error term is the incorrect mathematical formulation of the model. For example, if the model's assumptions are not met or the model is misspecified, the error term will be larger than expected. This can happen if the model contains unimportant independent variables, or if important independent variables were left out of the model.

Lastly, the error term is also considered as a substitute for a large number of variables that affect the dependent variable, but they were neglected. This is why the error term is considered a random error that follows a normal distribution with mean zero and constant variance (Sim et al, 2018).

There are several reasons why certain variables may not be included in the model, such as:

i. Lack of knowledge or awareness of these variables.
ii. Difficulty in obtaining data or information on these variables.
iii. Difficulty in measuring some variables statistically.
iv. Attempting to simplify the model for the purpose of the study.
v. Some variables may have a weak effect on the dependent variable and may not be easily measurable or quantifiable.
vi. Even if all variables affecting the phenomenon are known, there may still be challenges in obtaining accurate and complete data for all of them.
vii. Inaccuracies in describing the regression model.
viii. Presence of measurement errors or uncertainty in the values of the dependent and independent variables.

It is essential to understand that the discrepancy between the predicted values and the actual values is represented by the error term. This error can be positive or negative, and it is caused by factors not included in the model and other variables that may impact the model's precision.

2.2.1.2. Multiple Linear Regression (MLR)

Multiple linear regression is a statistical method used to study the relationship among a dependent variable and multiple independent variables. It is an extension of simple linear regression, which only considers one independent variable. In multiple linear regression, the model is represented by a mathematical equation that expresses the relationship among multiple independent variables $(x_1, x_2, x_3, \dots x_k)$ and a single dependent variable (y). The goal of multiple linear regression is to predict the value of the dependent variable based on the values of the independent variables. It is used when there are three or more variables and the dependent variable is one of the variables (y) and the other variables are independent variables $(x_1, x_2, x_3, \dots x_k)$. It is a prediction approach and is used in a wide range of fields such as business, economics, and finance to understand and evaluate the relationships between different influencing factors.

Understanding the functional relationships between the dependent and independent variables can also aid in identifying the causes of variation in the dependent variable. (Vandam et al, 2015).

The (MLR) fundamental model is:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon i} \qquad (\mathbf{2.14})$$

The equation (2.14) results in the determination of the matrix for the following formula:

$$\boldsymbol{\widehat{\beta}} = (\boldsymbol{X^T X})^{-1} \boldsymbol{X^T y} \qquad (\mathbf{2.15})$$

$$\text{where:} \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_0} \\ \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \vdots \\ \boldsymbol{\beta_m} \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{X_{11}} & \boldsymbol{X_{12}} & \cdots & \boldsymbol{X_{1m}} \\ \mathbf{1} & \boldsymbol{X_{21}} & \boldsymbol{X_{22}} & \cdots & \boldsymbol{X_{2m}} \\ \mathbf{1} & \boldsymbol{X_{31}} & \boldsymbol{X_{32}} & \cdots & \boldsymbol{X_{3m}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{1} & \boldsymbol{X_{n1}} & \boldsymbol{X_{n2}} & \cdots & \boldsymbol{X_{nm}} \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y_1} \\ \boldsymbol{Y_2} \\ \boldsymbol{Y_3} \\ \vdots \\ \boldsymbol{Y_n} \end{bmatrix} \qquad (\mathbf{2.16})$$

Johnson and Wichern (2004) note that while multiple linear regression assumes that the predictor variables are independent, they may actually be correlated. The extent of correlation between the predictor variables is known as multicollinearity. (Owusu et al, 2018).

The regression model is based on a set of assumptions and conditions that can be summarized as follows:

i. The relationship among the dependent and independent variables is linear.
ii. The variance of the independent variable is greater than zero.
iii. The measurements of the dependent variable and the independent variables are accurate and without errors (there are no measurement errors).
iv. The number of observations is greater than the number of variables in the model.
v. The error term follows a normal distribution with a mean of zero and constant variance.
vi. The errors are independent of one another, meaning there is no autocorrelation problem.

vii. The error term is independent of the values of the independent variable, meaning there is no heteroscedasticity problem.

viii. The independent variables are independent of each other, meaning there is no multicollinearity problem.

2.2.2. Basic Assumptions of Multiple Linear Regression Model

The least squares method is a statistical modeling technique that is commonly used to fit linear models. The basic idea behind this approach is to identify a straight line that best fits the data points in a way that minimizes the sum of the squares of the distances between the points and the line or to determine a value that minimizes this sum. This method was first credited to the German mathematician Carl Friedrich Gauss (Stigler, 1981).

The technique is widely used due to its several advantageous statistical characteristics, including being the best unbiased linear estimator, which sets it apart from other techniques and makes it one of the most effective methods for regression (Owusu et al, 2018).

An example of the least squares method can be seen in the estimation of the following equations:

$$\boldsymbol{U = Y - X\beta} \qquad (\mathbf{2.17})$$

Equation (2.17) represents the real regression equation, that is, the real data before extracting the error limit and estimating the data.

$$\boldsymbol{\widehat{U} = \widehat{Y} - \widehat{X}\widehat{\beta}} \qquad (\mathbf{2.18})$$

$$\boldsymbol{U * \acute{U} = Y\acute{Y} - 2 * \beta\, \acute{X}\, \acute{Y} + \acute{\beta}\, \acute{X}X\beta} \qquad (\mathbf{2.19})$$

The equation (2.18) below represents the estimated regression equation. After we extract the error limit U_i, we estimate the parameters, so it is as in the above formula. (John, 2019).

We may use the least squares method to derive the equation and obtain the formula for the estimator of the regression model's coefficients, which can be represented as follows:

$$\boldsymbol{\frac{\partial U\, \acute{U}}{\partial \acute{\beta}} = 0 - 2\acute{X}\, Y + 2\, \acute{\beta}\acute{X}X} \qquad (\mathbf{2.20})$$

Building a linear regression model and estimating model coefficients using the least squares method must fulfill a number of hypotheses that can be summarized as follows:

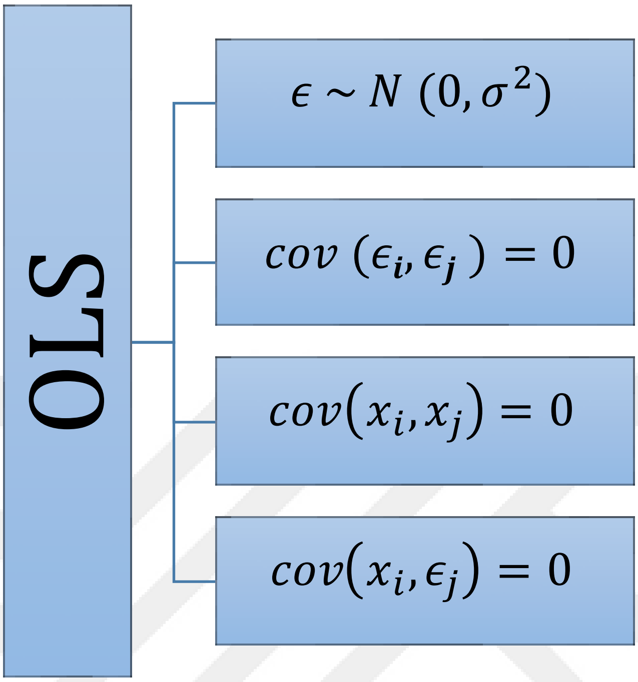

Figure 2.1 Assumptions of the ordinary least squares method 1

Where is the vector of unknown parameters?

$$\boldsymbol{\hat{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y} \tag{2.21}$$

The estimator is the best unbiased linear estimator if these assumptions are available, if any of the assumptions are defective, regression problems appear.

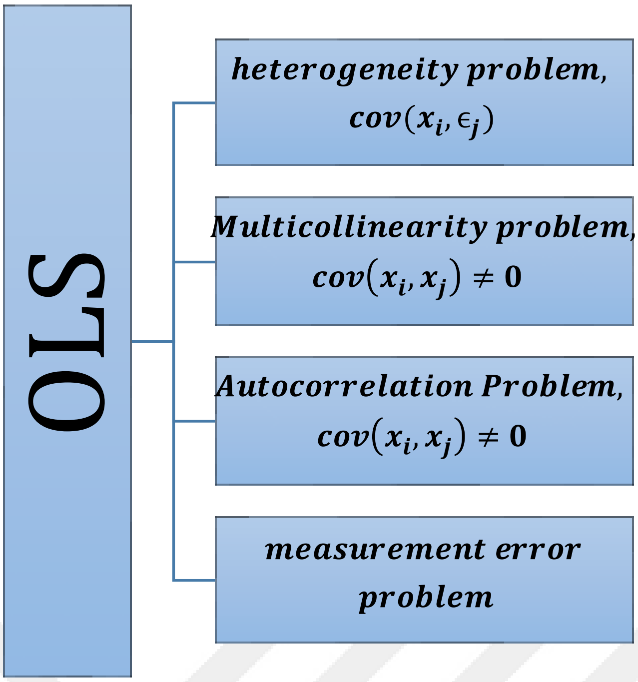

Figure 2.2 Assumptions of the ordinary least squares method 2

The least squares method is known for its desirable properties in terms of unbiasedness, consistency, and efficiency when the required assumptions are met. (Long, 2008).

Linear regression analysis is a statistical method that is used to study mathematical models by describing a linear relationship among a dependent variable and one or more independent variables. However, when there are measurement errors in the data, the traditional least-squares approach may not be suitable. In these situations, different methods may be used to account for measurement errors. These methods can have complex mathematical foundations and are not typically discussed in connection with straightforward regressions. When measurement errors impact the independent variables, but the dependent variable is observed without error, a simpler method can be applied.

One of the assumptions of linear regression is that the relationship among the dependent and independent variables is linear. To determine if this assumption is met, there are three methods that can be used. One is to fit a high-order polynomial function to the data and check for significant departures from zero in the regression coefficients for terms of second or higher order. Another method is to stratify the data based on the independent variable values and calculate a regression equation for each stratum, then determine the significance of the differences between each stratum.

Finally, the third method is to check the residuals of the linear regression model to see if it is linear or not. If the assumption of linearity is broken, it is necessary to choose other methods to estimate the model coefficients.

2.3. The Multicollinearity Problem

The multicollinearity problem is a common issue when constructing a regression model. It occurs when there is a linear correlation between the independent variables in the multiple linear regression model, which is a basic assumption that must be met when estimating the parameters of the model. When this assumption is not met, it can lead to estimators of the ordinary least squares (OLS) not having the efficiency property of having less variance. However, this assumption is often not met in practical applications, as variables are naturally intertwined and can affect each other. (Allison, 2010).

Multicollinearity is a compound term, with "multi" meaning multiple and "co" meaning combined or overlapping, while "linearity" refers to linear relationships. The problem of multicollinearity is one of degree rather than kind, as it is expected that there will be overlaps or multicollinearity between independent variables when studying the multiple linear regression model. If two or more variables are closely related to each other, a multi-linear relationship develops in the model, suggesting that one independent variable can be predicted based on the values of another independent variable. This makes it difficult to distinguish between the distinct effects of independent variables on the dependent variable. When two or more independent variables in a regression model have a strong relationship with one another, multicollinearity emerges. This means that by knowing the values of other independent variables, one independent variable can be predicted.

Multicollinearity can be problematic as it makes it challenging to discern between the various ways in which the independent factors affect the dependent variable. The model's accuracy may not be significantly impacted by the multicollinearity. However, there is a danger of losing faith in our ability to predict the impacts of specific model characteristics, which can make it difficult to understand the model's behavior. (Weaving et al., 2019).

We will use the linear equation shown below as an illustration:

$$\boldsymbol{y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2} \qquad (2.22)$$

The rise in y, to increase the unit in β_1 is called the coefficient of x_1 while keeping x_2 constant. However, since x_1 and x_2 are closely related, changes in one will affect changes in the other, and we will not be able to see their distinct effects on y, as a result, distinguishing between the impacts of $x_1$on y and the effects of x_2 on y, is difficult.

2.3.1. Types of Multicollinearity

There are two types of multicollinearity between predictive variables:

2.3.1.1. Exact Multicollinearity

When it becomes impossible to find the inverse of the information matrix, meaning the determinant of the matrix is zero, $|X'X| = 0$, and is smaller than the number of variables, it is known as multicollinearity. This occurs when there is a strong linear relationship among two or more independent variables and makes it impossible to estimate the parameters of the model, resulting in a situation called perfect multicollinearity. This level of collinearity indicates an exact linear relationship among multiple independent variables. (Doll, J. P, 1974)

To prove $|X'X| = 0$, we suppose that we have a model consisting of two explanatory variables and a dependent variable as follows:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i} \qquad (\mathbf{2.23})$$

$$X'X = \begin{bmatrix} \sum x_{i1}^2 \sum x_{i1}x_{i2} \\ \sum x_{i1}x_{i2} \sum x_{i2}^2 \end{bmatrix}$$

$$X_{ij} = (x_{ij} - \overline{x}_j)$$

If the perfect linear relation among these mentioned variables x_{i1}, x_{i2}, then the equation would be as follows:

$x_{i2} = kx_{i1}$

$k \neq 0$

The value of k is a constant

Substituting this relationship into the matrix $(X'X)$ we get:

$$X'X = \begin{bmatrix} \sum {x_{i1}}^2 & k\sum x_{i1}^2 \\ k\sum x_{i1}^2 & k^2\sum x_{i1}^2 \end{bmatrix}$$

Therefore :

$$|X'X| = K^2(\sum x_{i1}^2)^2 - K^2(\sum x_{i1}^2)^2 = 0$$

2.3.1.2. Non-Exact Multicollinearity

Non-exact multicollinearity is a situation that arises if a non-zero determinant of the information matrix appears but is relatively close to it. This condition is often referred to as high imperfect multicollinearity. When this occurs, the independent variables are highly correlated and tend to change together, either increasing or decreasing. This can lead to high variance in the estimated parameters and large standard errors, which can affect the results of $t-$test statistics.

Therefore, the approximation of the mentioned model parameters in this scenario may not accurately reflect the fact of the case being proposed, even if the coefficient of determination R^2 is high. This is because the model's parameters may be small and insignificant. (Kalnins, A, 2018). The equation for this is given by:

$$\boldsymbol{t_1 x_1 + t_2 x_2 + \ldots + t_m x_{m1} + W_i = 0} \qquad (\mathbf{2.24})$$

where $t_1, t_2, t_3, \ldots \ldots \ldots \ldots, t_m$ are constants, and at least one of them is not equal to zero and also $w_i \neq 0$.

In this case, the variance of the estimator is very high because the value is too small for the determinant of the matrix ($X'X$), and when performing a $t - test$, the significance of the parameters appears on the grounds that their variances are large, while these parameters are not significant, but the correlation of some variables with each other made the model unable to show the effect of each separately from the other.

2.3.2. Reasons For Multicollinearity

There are several reasons that can lead to multicollinearity in a regression model, some of which include:

i. Having fewer observations compared to the number of variables in the model.

ii. Incorporating lagged variables as explanatory variables, such as including the current price of a crop and its price in a previous period as variables in explaining crop production.

iii. Some variables tend to change together, for example, variables such as employee income, years of experience, age, and job salary often have a strong positive correlation with one another. (Chen, G. J, 2012).

2.3.3. Consequences Multicollinearity

The presence of multicollinearity can lead to a number of consequences that can negatively impact the regression analysis.

i. A decrease in the efficiency of estimators, as well as a decrease in the accuracy of the estimate, as it becomes difficult to estimate the relative effect of the independent variables.

ii. The standard errors were large, and the t-value was low in comparison with their tabularly, and this leads the researcher to a kind of error in deleting some variables in the model.

iii. Estimates become unstable and highly sensitive to any change in the sample. (Gorman, K, 2010).

Suppose we have the following model:

$$\mathbf{Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + . . . + \beta_k x_k + \mathcal{E}_i} \qquad (\mathbf{2.25})$$

While $\mathcal{E}_i \sim NID\ (0,\sigma^2)$ but, $cov\left(X_i, X_j\right) \neq 0$, in other words, there is a relationship among the two variables $\left(X_i, X_j\right)$, therefore, the matrix X will be incomplete, i.e. $r(x) < k+1$.

In this case, it is difficult to find the determinant, and then find the inverse of the matrix $(X'X\ )$, to clarify this problem, we address the multiple regression model that contains only two variables (x_1, x_2), defined in the formula below:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2} + \mathcal{E}_{\boldsymbol{i}} \qquad (\mathbf{2.26})$$

To estimate the previous model by regression method

$$\boldsymbol{y_i - \bar{y} = (\beta_0 - \beta_0) + \beta_1 (X_1 - \overline{X_1}) + \beta_2 (X_2 - \bar{X}_2)} + (\mathcal{E}_{\boldsymbol{i}} - \bar{\mathcal{E}}) \qquad (\mathbf{2.27})$$

Using the least squares method(OLS), the parameters of the model can be estimated, $\hat{\beta} = (X'X)^{-1}\,X'Y$.

2.3.4. The impact of collinearity on the OLS method

The implications of the existence of the problem of complete and incomplete multilinearity in estimating the parameters of the regression model using the method of least squares can be presented as follows:

- First case: If the determinant of the matrix =1, $r = 1$.

This indicates the existence of the relationship among the variables explained in the model means the presence of multicollinearity, which makes it difficult to isolate their individual effects on the dependent variable, and thus violates the terms of use OLS and are not statistically significant and may take the error sign even though R^2 may be in addition if the relationship among the variables is complete ($r = 1$), this means that the inverse of the matrix cannot be found, and therefore the OLS method cannot be used to estimate the parameters of the model. (Paul, R. K, 2006).

$$\boldsymbol{X'Y} = \begin{bmatrix} \boldsymbol{sp(x_1,y)} \\ \boldsymbol{p(x_2,y)} \end{bmatrix} = \begin{bmatrix} \sum (x_{1i}-\bar{x}_1)^2 (x_{2i}-\bar{x}_2) \\ \sum (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2) \sum (x_{2i}-\bar{x}_2)^2 \end{bmatrix} \qquad (\mathbf{2.28})$$

$$\boldsymbol{X'Y} = \begin{bmatrix} \boldsymbol{sp(x_1,y)} \\ \boldsymbol{p(x_2,y)} \end{bmatrix} = \begin{bmatrix} \sum (x_{1\iota}-\bar{x}_1)(y_{1\iota}-\bar{y}_1) \\ \sum (x_{2\iota}-\bar{x}_2)(y_{2\iota}-\bar{y}_2) \end{bmatrix} \qquad (\mathbf{2.29})$$

$$|\boldsymbol{X'X}\,| = \sum (x_{1\iota}-\bar{x}_1)^2 \sum (x_{2\iota}-\bar{x}_2)^2 [\sum (x_{1\iota}-\bar{x}_1)^2 (x_{2\iota}-\bar{x}_2)]^2 = \mathbf{0} \qquad (\mathbf{2.30})$$

Where the correlation coefficient is defined as:

$$\boldsymbol{r} = \frac{\sum (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\sqrt{\sum (x_{1i}-\bar{x}_1)^2 \sum (x_{2i}-\bar{x}_2)^2}} \qquad (\mathbf{2.31})$$

If the correlation is perfect, then $|X'X| = 0$, and thus the inverse of the matrix $X'X$ cannot be found, and in this case, we cannot estimate by the ordinary least squares method OLS.

- The second case: If the matrix determinant is non-zero, but it approaches zero, i.e.

$$|\boldsymbol{X'X}| = \sum (x_{1\iota}-\bar{x}_1)^2 \sum (x_{2\iota}-\bar{x}_2)^2 - \left[\sum (x_{1\iota}-\bar{x}_1)^2 (x_{2\iota}-\bar{x}_2)\right]^2 \approxeq \mathbf{0} \qquad (\mathbf{2.32})$$

This means that the correlation between the two variables is incomplete In this case, the inverse of the matrix $X'X$, can be found, but the elements of this matrix will be large, Since the least squares estimator $\hat{\beta}$ where $\hat{\beta} = (X'X)^{-1}X'Y$ the estimator vector elements $\hat{\beta}$ will also be large, and this means that the estimation is biased and on it $E(\hat{\beta}) \neq \beta$, since the variance of the estimators is given by the following formula $V(\hat{\beta}) = \acute{x}x^{-1}\sigma^2$ and the matrix elements contain large values, this means that this method gives Incompetent estimations, which does not give the optimal model (Mela et al, 2002).

2.4. Investigation of the Multicollinearity Problem

If we have the following model:

$$\boldsymbol{Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon_i} \qquad (\mathbf{2.33})$$

There are several ways to investigate the existence of a multicollinearity problem, and each of them has its own advantages and details. In this study, we use the methods of variance inflation factor VIF, and Correlation Matrix so we will mention them in detail. (Duzan et al, 2015).

2.4.1. Variance Inflation Factor VIF

It is one of the widely used methods for investigating the problem of multilinearity, as this measure was proposed by the scientist Marquaradt in the year (1970), and it is one of the measures used in investigating the problem of multilinearity and its severity. If the problem is little or no, the value of VIF is an approximation of one (VIF$\sim$1), if the problem is medium, the value of VIF ranges between ($1 < VIF < 5$), but if it is large, it may reach ten or even greater than that, a variance inflation factor is calculated for each of the independent variables, as it is used to measure the correlation of each independent variable with the other independent variables. If the value of ($VIF \geq 5$), indicates the presence of multicollinearity between the explanatory variables. (Akinwande et al, 2015).

The formula for calculating VIF is:

$$\boldsymbol{VIF = \frac{1}{(1 - R^2{}_j)}} \qquad (\mathbf{2.34})$$

$$, j = 1,2,...,p \quad , \quad 0 \leq VIF \leq \infty \ , \ 0 \leq R_j^2 \leq 1$$

Where : p: The number of independent variables.

R_j^2: It is the coefficient of determination of the independent variable x_j

Accordingly, if the model contains p, of independent variables, it means that there is p of VIF, we note through the mathematical formula to it:

1. In the event that there is a perfect linear correlation between the independent variable x_j number j and the rest of the independent variables, the value of VIF infinity, $VIF = \infty$.
2. While in the absence of a perfect linear correlation between the independent variable x_j number j and the rest of the independent variables, the value of VIF

equals 1, and this measure is a sufficient reason to neglect the variable x_j. (Emmanuel et al, 2021).

2.4.2. Correlation Matrix

One way to identify multicollinearity is by analyzing the correlation matrix of the explanatory variables. If there is a strong correlation between two variables, x_i, and x_j, and the absolute value of their correlation coefficient, $|r_{ij}|$, is close to one, then the regression model is likely experiencing multicollinearity. However, it is important to note that a weak correlation between variables does not necessarily mean there is no multicollinearity present. The determinant of the correlation matrix can be used as a measure of the degree of multicollinearity in the model. (Rockwell, 1975).

2.5. Remedy of Multicollinearity

There are several suggested solutions to address the problem of multicollinearity. These solutions are based on the possibility of finding other sources of data, the importance of the factors that caused its emergence, and the goal for which the function under study is being estimated. If the multiplicity does not actually affect the capabilities of the model, some researchers suggest neglecting its presence in the model. One way to avoid multicollinearity is to expand the sample size. For example, annual data can be converted into seasonal or monthly data if possible. Another solution is to eliminate linearity by dropping or deleting the variable that causes the problem, but this method can create other problems. Some researchers suggest adding additional information to the model as well. (Lindner et al, 2020).

The presence of multicollinearity can pose challenges in distinguishing the effect of individual independent variables on the dependent variable. To overcome this issue, additional information, such as prior knowledge from economic theory, can be used to impose restrictions on certain parameters. In summary, the problem of multicollinearity can be summarized as follows:

i. Eliminating correlated variables that are causing the issue and then using the OLS method to estimate.
ii. Increasing the sample size.
iii. Adding new variables to the model.
iv. Transforming variables by taking the ratio of one variable to another.
v. Utilizing external information about certain variables.

vi. Implementing estimation methods that are based on a smaller number of principal variables such as:

a. The restricted least squares technique.
b. The approach of merging time series with cross-sectional data.
c. The use of biased methods of estimation, the most important of which are:
    1. Principle Component Analysis
    2. Ridge regression method
    3. Adjusted regression method: where the new regression coefficient is known to find the value of $\hat{\beta}$ we use different values for the matrix.

Note * One of the disadvantages of this method is to find the value of the constant c, which makes us obtain the best estimate in terms of unbiased and least variance. (Bagheri et al, 2009).

2.6. Ridge Regression

Ridge regression is a method for estimating multiple linear regression coefficients that were developed in the 1970s and 1980s. Despite being introduced earlier in the literature by researchers Crone, Hoerl, and Draper, it gained widespread popularity after the publication of an article by Hoerl and Kennard in Technometrics. (Hoerl et al, 1970). This method is a variation of ordinary least squares (OLS) that can be used when there is multicollinearity or high correlation among the independent variables. It works by adding a penalty term to the OLS objective function to shrink the regression coefficients towards zero, which helps to reduce the variance of the estimates and improve the model's stability.

The issue of multicollinearity is commonly encountered in multiple linear regression, to address this problem, various researchers have proposed alternative estimators to the OLS estimator, with the Ridge regression estimator being one of the most notable solutions. (Marquardt et al, 1975).

Ridge regression is a technique for adjusting a multiple linear regression model that can be applied to data with multicollinearity. The method, which was first proposed by Hoerl and Kennard, utilizes Tikhonov regularization for all the regression parameters in order to control the inflation and instability typically associated with least-squares estimates, resulting in the Ridge estimator. (Zhang, 2019).

Tikhonov regularization is a method for resolving ill-posed problems by introducing a small value, called the regularization parameter, which is added to the original problem to make it well-posed. This method is commonly used in various fields such as mathematics, physics, and engineering to stabilize the solution and prevent overfitting.

$$\boldsymbol{\hat{\beta}} = (\boldsymbol{X'X})^{-\mathbf{1}}\boldsymbol{X'Y} \qquad (\mathbf{2.35})$$

Ridge regression, which is a regression model tuning approach, is commonly used to address the problem of multicollinearity in data. The method, which was first introduced by Hoerl and Kennard, utilizes Tikhonov regularization for all the regression parameters to control inflation and general instability associated with the least-squares estimates. The Ridge penalty parameter, denoted by k, controls the amount of regularization applied to the model. Large values of k tend to decrease the size of the predicted regression coefficients, resulting in fewer useful model parameters. (Hemmerle, W. J, 1975).

This approach is used to produce regularization and is particularly useful when multicollinearity is a concern. The ordinary least squares method is unbiased, but variances are large, leading to predicted values that are distant from the actual values. (Akdeniz, 2019).

Regression estimators generally advise contraction methods to lessen the impact of collinearity for both linear and nonlinear regression models. The Ridge regression estimator is one of the most well-known methods for solving the multicollinearity problem in a linear regression model, as it reduces the value of the objective function.

It is possible to infer through the OLS estimator that there are strong multicollinear correlations between the predictor variables when this method is not effective. To address this problem, many writers have developed a number of estimators as alternatives to the OLS estimator, most notably the Ridge regression estimator. (William et al, 1975).

According to (Bager, 2018), Ridge regression is one of the most widely used and common methods for dealing with the problem of multicollinearity between predictive variables. Hoerl and Kennard explained the purpose of adding the positive constant k to the elements of the diameter of the matrix, and then clarified that the costs are calculated according to the following formulas:

$$\boldsymbol{\beta_1} = \frac{\boldsymbol{r_{1y} - r_{12}\, r_{2y}}}{\boldsymbol{1 - r^2{}_{12}}}$$

$$\boldsymbol{\beta_2} = \frac{\boldsymbol{r_{2y} - r_{12} r_{2y}}}{\boldsymbol{1 - r^2{}_{12}}}$$

To calculate the values of $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$, and other parameters using the Ridge regression method, we use the following formulas:

$$\boldsymbol{\beta_1} = \frac{\boldsymbol{r_{1y} - \left[\frac{r_{12}}{1+k}\right] r_{2y}}}{\boldsymbol{1\text{-}[r_{12} \,/\, (1+k)]^2}} \left(\frac{\boldsymbol{1}}{\boldsymbol{1+k}}\right) \qquad (\mathbf{2.36})$$

$$\boldsymbol{\beta_2} = \frac{\boldsymbol{r_{2y} - \left[\frac{r_{12}}{1+k}\right] r_{1y}}}{\boldsymbol{1\text{-}[r_{12} \,/\, (1+k)]^2}} \left(\frac{\boldsymbol{1}}{\boldsymbol{1+k}}\right) \qquad (\mathbf{2.37})$$

We note that the main difference between the letter regression method and the method of OLS is $r_{12}/k + 1$ instead of r_{12}, and after adding the constant k to the elements of the matrix diameter, we get the equation:

$$(\boldsymbol{X'X + k})\widehat{\boldsymbol{\beta}} = \boldsymbol{X'Y} \qquad (\mathbf{2.38})$$

In the case of $k = k_{IP}$, the equation (2.38) becomes of the form:

$$(\boldsymbol{X'X + k_{IP}})\widehat{\boldsymbol{\beta}} = \boldsymbol{X'Y} \qquad (\mathbf{2.39})$$

And by multiplying both sides of the equation (2.39) by the matrix $(\acute{X}X + K_{IP})^{-1}$, we get:

$$\boldsymbol{c}\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\boldsymbol{X'X + k_{IP}})^{-1}\boldsymbol{X'Y} \qquad (\mathbf{2.40})$$

2.6.1. Properties of the Ridge Regression

i. Bias

In the case of $k = k_{IP}$, we substitute $X'Y$ from equation (2.39) with equation (2.40) and we get:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\boldsymbol{X'X + k_{IP}})^{-1}(\boldsymbol{X'X})\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}}$$

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}} = \boldsymbol{w}\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} \qquad (\mathbf{2.41})$$

Whereas:

$$\boldsymbol{cw} = (\boldsymbol{X'X + k_{IP}})^{-1}(\boldsymbol{X'X}) \qquad (\mathbf{2.42})$$

And by taking the expectation from both sides of equation (2.41), we get:

$$\boldsymbol{E}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{R}}) = \boldsymbol{wE}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{OLS}})$$

$$\boldsymbol{E}(\boldsymbol{\beta_R}) = \boldsymbol{w\beta} \qquad (\mathbf{2.43})$$

We notice from this equation that the Ridge estimator is biased for the original parameter and to find the amount of bias we substitute equation (2.41) with this equation as follows:

$$\boldsymbol{E(\widehat{\beta}_R) = (X'X + k_{IP})^{-1}(X'X)\beta}$$

By adding and subtracting the K_{IP} matrix to the $\boldsymbol{x'x}$ matrix we get:

$$\boldsymbol{E(\widehat{\beta}_R) = (X'X + k_{IP})^{-1}(X'X + k_{IP} - k_{IP})\beta}$$

$$\boldsymbol{E(\widehat{\beta}_R) = \beta - (X'X + k_{IP})^{-1}k_{IP}\beta}$$

In the case of $k = \text{diag } k_i$, the bias amount is:

$$\boldsymbol{E(\widehat{\beta}_R) = \beta - (X'X + kI_P)^{-1}k\beta}$$

(Dai et al, 2018).

ii. Variance

$$\boldsymbol{Var(\widehat{\beta}_k) = Var(X'X + k_{IP})^{-1}X'Y} \qquad \textbf{(2.44)}$$

$$\boldsymbol{Var(\widehat{\beta}_k) = Var[(X'X + k_{IP})^{-1}x'(x\beta + \mathcal{E})]}$$

$$\boldsymbol{= Var[\,(X'X + k_{IP})^{-1}X'X\beta + (X'X + k_{IP})^{-1}x'\mathcal{E}]}$$

$$\boldsymbol{= \sigma^2(X'X + k_{IP})^{-1}(X'X)(X'X + k_{IP})^{-1}}$$

$$\boldsymbol{= \sigma^2(X'X + k_{IP})^{-1}(X'X)(X'X + k_{IP})^{-1}}$$

Assuming $\boldsymbol{X'X} = \wedge$

$$\boldsymbol{= \sigma^2 V(\wedge + k_{IP})^{-2} \wedge \acute{V}}$$

$$\boldsymbol{= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} V_i \acute{V}_j} \qquad \textbf{(2.45)}$$

The variance of the parameter estimator β_{ik} is as follows:

$$\boldsymbol{Var(\widehat{\beta}_{ik}) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} V_{ij}^2} \qquad \textbf{(2.46)}$$

2.6.2. Estimation Via Ridge Regression Method

When the predictive variables are not orthogonal, the multicollinearity problem arises. This causes difficulty in estimating model parameters using the ordinary least squares method, due to inflation and instability in the estimated coefficients discrepancies. (Kibria, 2020).

$$\boldsymbol{\widehat{\alpha}_k = \left(X^{'}X\right)^{-1}X^{'}Y} \qquad \textbf{(2.47)}$$

To address the issue of multicollinearity and obtain more stable estimates, researchers have proposed various methods, including the Ridge regression method. This method operates by adding a constant k to the main diagonal elements of the matrix $X'X$ before taking its inverse. This approach is considered a modification of the least squares method, aimed at achieving the best model with estimators that possess a specific feature and the smallest mean squared error, while also being biased. The Ridge regression method can be estimated using the following equation:

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} = \left(\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{k}\right)^{-\mathbf{1}} \boldsymbol{X}^{'}\boldsymbol{Y} \qquad (\mathbf{2.48})$$

$Where \quad 0 \leq K \leq 1$

Applying this formula $\mathbf{k} = \boldsymbol{k}_{\boldsymbol{IP}}$ to the above equation:

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{K}} = \left(\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{k}_{\boldsymbol{IP}}\right)^{-\mathbf{1}} \boldsymbol{X}^{'}\boldsymbol{Y} \qquad (\mathbf{2.49})$$

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{k}} = \sum_{\boldsymbol{j}=\mathbf{1}}^{\boldsymbol{p}} (\boldsymbol{\lambda} + \boldsymbol{k})^{-\mathbf{1}} \boldsymbol{V}_{\boldsymbol{i}}\boldsymbol{V}_{\boldsymbol{j}}\boldsymbol{X}^{'}\boldsymbol{Y} \qquad (\mathbf{2.50})$$

$$Where \qquad X^{'}X = Diag\left(\lambda_1 + \lambda_2 + \cdots\cdots\cdots\cdots + \lambda_p\right)$$

We have $p = m + 1$, m being the number of explanatory variables for the regression model obtained through a set of derivations back to the model estimates by the OLS method, if we insert the following matrix $\boldsymbol{k}_{\boldsymbol{IP}}$ its dimensions $p \times p$ into the matrix before taking its inverse, the result will be the following equation:

$$\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{K}_{\boldsymbol{IP}})\hat{\boldsymbol{\alpha}}_{(\boldsymbol{k})} \qquad (\mathbf{2.51})$$

2.6.3. Methods of Choosing Ridge Parameter

In this part, we will go over the main methods for choosing the Ridge regression parameter. The following two ways below will be explained.

2.6.3.1. Ridge Trace Method

The Ridge Trace Method is a significant technique as it deals with the issue of multicollinearity in multiple linear regression and its practicality is demonstrated by analyzing a plot of the estimated regression coefficients in relation to the Ridge parameter. (Crouse et al, 1995).

Determining the optimal value for the Ridge parameter or selecting the most appropriate estimator within this class of estimators for a specific problem is the primary challenge in Ridge regression. (Niemelä-Nyrhinen et al, 1975).

The problem of multicollinearity is often ignored in the structural equation regression model, this is unfortunate as multicollinearity may lead to erroneous estimates of the path coefficient and lead non-statistical estimates. (Niemelä et al, 2014).

The Ridge Trace Method is one of the criteria used for selecting explanatory variables by deleting certain variables from the regression model. This method includes:

- Omitting the explanatory variable that contains coefficients that cannot accurately reflect its true amplitude and leads to zero.
- Omitting part of the variables whose coefficients are unstable.
- Omitting some variables if they have small predictive energy, even if they are stable and inferred from their small standard coefficient. (Kibria & Banik, 2016).

2.6.3.2. Methods of Estimation k

The selection of the optimal value of the Ridge coefficient in this method is based on the values of the response variable. It is considered a more objective method compared to the Ridge trace method. The aim of this method is to find the optimal value of k, which can greatly reduce the variance and increase the bias. (Dorugade & Kashid, 2010).

2.6.4. Exploratory Variable Selection using Ridge Regression

Choosing the best variables for the regression equation when there is a lack of orthogonality among them and a multicollinearity issue is present can be challenging. Traditional methods such as forward selection, backward deletion, and sequential steps regression may not be effective in this scenario and lead to suboptimal results in parameter estimation using ordinary least squares.

Selecting the most effective explanatory variables can be improved by taking into account the orthogonality of the variables and applying a bias factor. Using

Ridge regression in combination with other methods, such as the forward selection technique, backward deletion method, or sequential steps regression method, as suggested by Hoerl & Kennard, can enhance the selection process. (Arashi et al, 2021).

3. APPLICATION

In this chapter, we will apply the methods discussed in the previous chapters, namely the Ordinary Least Squares (OLS) technique, and the Ridge Regression technique, to a real-world data set related to electrical energy production. The goal of this application is to demonstrate how to address the problem of multicollinearity by identifying the most significant influencing factors and achieving accurate results.

3.1. Description of Data

In this study, a sample of data was collected from an electrical energy production station located in Karbala, Iraq, as a representation of similar stations in other cities. The sample includes data from 72 motors distributed across three sectors, with each sector containing 4 blocks and each block containing 6 engines. The data were collected monthly over a five-year period from April 1, 2017, to April 1, 2022, and analyzed to make predictions about electrical energy production and identify the factors that influence it. To facilitate analysis and improve precision when presenting results, the original data was divided into 1000 units.

3.2. Study Variables

In this study, a sample of 50 observations was selected using a simple random sampling method. These observations were chosen based on their potential impact on electrical energy production, as outlined in Appendix 1. The variables used in this research are defined in the following table. A total of 8 variables were considered, with one being the response variable (y_i) and the remaining 7 being independent variables (x_i):

Table 3.1 Explanation of Model Variables

Code	Variables
y	Electrical Energy Output / MW
x_1	Heavy Fuel (Black Oil) / liters
x_2	Additional Wages for Work / Dinar
x_3	Engine Oils / liters
x_4	Chemicals / liters
x_5	Average Operating Hours / hours/month
x_6	Maintenance fee / Dinar
x_7	Light Fuel (Kerosene) / liters
u_i	Random Errors

3.1. Software

In this study, various software programs were utilized to analyze and interpret data. These programs include R-Studio, Statistical Packages for Social Sciences (SPSS), Python, and Excel. These tools were used to achieve the objectives of the study, answer the research questions, and generate the desired results.

3.2. General significance test of the model and parameters

Here, we will perform some descriptive statistics on the data in order to understand the distribution of the sample and the characteristics of the variables under consideration. We will also divide the sample based on the explanatory variables to examine their individual impact on the production of electrical energy. Furthermore, we will analyze the statistical significance of the explanatory variables by evaluating important criteria such as mean, standard deviation, and correlation.

Table 3.2 Results significance test of the Full model:

Variables	**t**	$\boldsymbol{P\_value}$
y: Electrical Energy	6.646	0.000
x_1: Black Oil	3.721	0.001
x_2: Additional Wages for Work	-1.615	0.114
x_3: Engine Oil	-1.784	0.082
x_4: Chemicals	-0.625	0.536
x_5: Operating	1.435	0.159
x_6: Maintenance fee	-5.169	0.000
x_7: Kerosene	-0.08	0.000
R= 0.85, R^2= 0.73, $R^2_{Adj} = 0.69$, Std.Error =0.259, F=16.54, MSE=0.067, $P\_value =$ 0.00		

In the table above, we evaluated the overall significance and relevance of the statistical model by calculating the Std Error, adjusted Coefficient of Determination, Coefficient of Determination, and F-values. The table shows that R = 0.85, R^2= 0.73, and $R^2_{Adj} = 0.69$.

Independent variables seem to explain 73% of the electrical energy production for the overall model, which is a large proportion.

It seems like the model is logical, and also the value P_value = 0.00 the value F = 16.54, which is less than 5%, noting that the model is significant, thus there is a correlation between the energy production variable and other variables, and here the impact of each variable evident by testing each parameter individually, extracting the value of importance and the value of $t - test$ for each variable and knowing the relevance of each and which is more important.

When the value P_value is less than 5%, it is a variable that has statistical effect and significance, so the study model is important and statistically significant. (Garba et al, 2021).

3.3. The Interrelationship among The Variables

In order to understand the relationships between variables, it is important to examine the correlation matrix. This allows us to determine the degree of correlation between each pair of variables. (Taiyun et al, 2021).

The correlation matrix for our data set can be found in Table No. (3-3). This matrix was generated using the Python program, as outlined in Appendix No. (2) and (3).

Table 3.3 Correlation Matrix Results

	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
y	1	-0.069	0.135	0.722	0.075	0.655	-0.728	-0.029
p	-	0.63	0.34	0.00	0.60	0.00	0.00	0.84
x_1	-	1	-0.558	-0.278	-0.314	-0.188	0.539	0.237
p	-	-	0.00	0.05	0.02	0.19	0.00	0.09
x_2	-	-	1	0.432	0.148	0.188	-0.546	-0.259
p	-	-	-	0.002	0.30	0.19	0.00	0.69
x_3	-	-	-	1	0.094	0.838	-0.904	-0.111
p	-	-	-	-	0.51	0.00	0.00	0.44
x_4	-	-	-	-	1	0.034	-0.244	-0.201
p	-	-	-	-	-	0.81	0.08	0.16
x_5	-	-	-	-	-	1	-0.705	-0.074
p	-	-	-	-	-	-	0.00	0.60
x_6	-	-	-	-	-	-	1	0.165
p	-	-	-	-	-	-	-	0.25
x_7	-	-	-	-	-	-	-	1

```
plt.figure(figsize=(15,10))
df_plot=sns.heatmap(file.corr(),cmap='YlGnBu',annot=True)
plt.show()
```

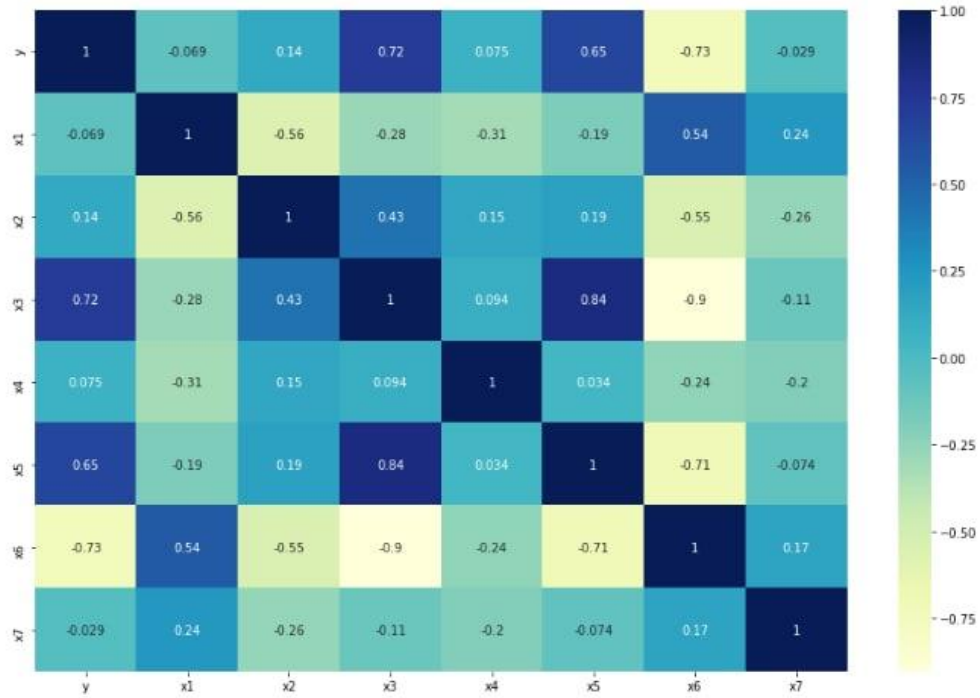

Figure 3.1 Correlation Matrix

The correlation matrix in the above Figure 3.1 and Appendix No: (3) reveal that there is a statistically significant correlation at a 5% level between the electric energy production variable (y) and the independent variables. The correlation coefficient between y and x_1 (black oil) is -0.06, indicating a weak inverse correlation. On the other hand, there is a weak direct correlation between y and x_2 (additional wages for work) with a coefficient of 0.13. The correlation between y and x_3 (engine oils) is strong and direct with a coefficient of 0.72. A weak direct correlation is also observed between y and x_4 (chemicals) with a coefficient of 0.07. The correlation between y and x_5 (average operating hours) is strong and direct with a coefficient of 0.65, and a strong inverse correlation is observed between y and x_6 (maintenance fee) with a coefficient of -0.72. Lastly, there is a weak inverse correlation between y and x_7 (kerosene) with a coefficient of -0.2.

As seen in the presented results above, the relationship among the variables is as follows:

The variable x_1 black oil has a weak reverse correlation with y electric energy production (-0.06) and a medium reverse correlation with x_2 additional wages for work (-0.55). It also has a weak reverse correlation with x_3 engine oil (-0.27), x_4 chemicals (-0.31), x_5 average operating hours (-0.18), and a medium inverse correlation with x_6 maintenance fee (-0.53) and a weak reverse correlation with x_7 kerosene (-0.23).

The variable x_2 additional wages for work has a weak direct correlation with y electric energy production (0.13) and a weak direct correlation with x_3 engine oil (0.43), a weak direct correlation with x_4 chemicals (0.14), a weak direct correlation with x_5 average operating hours (0.18), a medium inverse correlation with x_6 maintenance fee (-0.54) and a weak reverse correlation with x_7 kerosene (-0.25).

The variable x_3 engine oil has a strong direct correlation with y electric energy production (0.72) and a weak direct correlation with x_2 additional wages for work (0.43), a weak direct correlation with x_4 chemicals (0.09), a strong direct correlation with x_5 average operating hours (0.83), a strong inverse correlation with x_6 maintenance fee (-0.90) and a weak reverse correlation with x_7 kerosene (-0.11).

The variable x_4 chemicals have a weak direct correlation with y electric energy production (0.07), a weak direct correlation with x_3 engine oil (0.09), a weak direct correlation with x_5 average operating hours (0.03), a weak reverse correlation with x_6 maintenance fee (-0.24) and a weak reverse correlation with x_7 kerosene (-0.20).

The variable x_5 average operating hours has a strong direct correlation with y electric energy production (0.65) and a weak direct correlation with x_4 chemicals (0.03), a strong inverse correlation with x_6 maintenance fee (-0.70) and a weak reverse correlation with x_7 kerosene (-0.7).

The variable x_6 maintenance fee has a strong inverse correlation with y electric energy production (-0.72) and a weak direct correlation with x_7 kerosene (0.16). The variable x_7 kerosene has a weak reverse

3.4. Estimation Methods

Before applying estimation methods, it is important to first test the data to determine if there is a problem with multicollinearity.

3.4.1. Investigation of Multicollinearity Problem

In our study, we used the variance inflation factor VIF to investigate the problem of multicollinearity and evaluate the degree of correlation between variables, and this method is the most used, and the value of VIF starts from 1 and has no upper limit.

The general rule of thumb for interpreting VIF can be summarized as follows:

- A value of 1 indicates that there is no linear correlation between the explanatory variables in the model.
- If the value is between 1 and 5, it indicates that there is a medium correlation between the explanatory variables in the model.
- A value greater than 5 indicates that the correlation is severe between the explanatory variables in this model. In this case, the values of this model's estimates are not reliable.

In calculating the VIF, we used the Car function in the R package titled VIF, and we get the results as follows:

Table 3.4 Variance Inflation Factor test results VIF

Variables	VIF
x_1: Black Oil	2.557476
x_2: Additional Wages for work	1.957510
x_3: Engine Oil	14.544471
x_4: Chemicals	1.210497
x_5: Operating	4.073871
x_6: Maintenance fee	11.926665
x_7: Kerosene	1.114292

We find in the above table that the value of the variance inflation factor scale exceeded 5 for each of the following variables:

x_3: Engine Oil, x_6 Maintenance fee, as for the rest of the variables, they are close to the extent of VIF as well as their correlation was evident in the correlation matrix, the regression model of the OLS method when compared to the Ridge regression model would be indicative of this.

The visualization of the VIF values for each predictor variable is created by a simple horizontal bar chart and a vertical line at 5 is added so we can see how many VIF values exceed 5, see Figure 3.2 below.

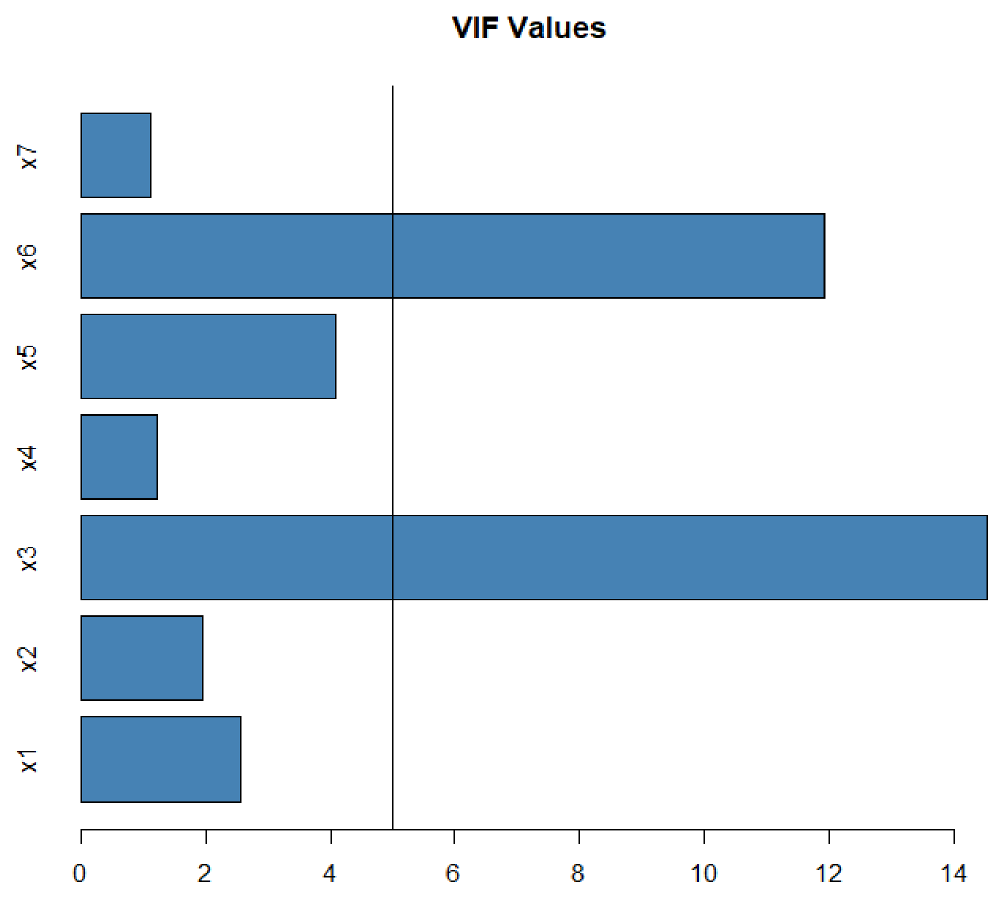

Figure 3.2 $VIF\ Values$

This indicates that the model suffers from a multicollinearity problem

3.4.2. Ordinary Least Squares Method

In this section, we apply the Ordinary Least Squares (OLS) method to our data set. We use Python programming language, importing necessary libraries such as NumPy, pandas, math, CSV, matplotlib, seaborn, and sklearn to perform the regression analysis and calculate the key parameters of the data. The implementation of the codes can be found in Appendix No. (4) and (5). We also provide general information about the data in Appendix No. (6).

To begin, we define the dependent and independent variables in our Python program and split them into training and test sets using the train_test_split function from the sklearn library, as shown in Appendix No. (7). We then perform the regression analysis using the OLS method and calculate the regression coefficients, correlation value, coefficient of determination, and sum of squared error. The results of this analysis can be found in Table No. (3.5) and the codes used can be found in Appendix No. (8).

The results of the OLS analysis show that the coefficient of determination, R^2, is 0.61, indicating that the model explains 61% of the variance in the dependent variable (electrical energy production). This is considered a medium percentage, as it suggests that 61% of the variance in the dependent variable can be explained by the independent variables (influencing factors). Additionally, the correlation coefficient is 0.78, indicating a strong direct correlation between the production of electrical energy and the independent variables. The mean squared error (MSE) is also calculated, with a value of 0.26, further indicating the model's accuracy

According to the variables included in the study, the linear regression equation is as follows:

$$\boldsymbol{\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7} \qquad (\mathbf{3.1})$$

Through Linear Regression, we got the parameters of the multiple linear regression by the OLS method and the parameters were as follows:

Table 3.5 Ordinary Least Squares Estimates OLS Results

Variables	**Parameter**	**Estimated**
Constant	β_0	4.898
X_1	β_1	1.063
X_2	β_2	-5.565
X_3	β_3	-0.109
X_4	β_4	-0.0009
X_5	β_5	12.7
X_6	β_6	-0.056
X_7	β_7	-17.3
$MSE = 0.263$, $R^2 = 0.619$, $R = 0.787$		

The estimated regression equation by Ordinary Least Squares method can be written as:

$$\hat{y} = (4.898) + (1.063x_1) - (5.565x_2) - (0.109x_3) - (0.0009x_4) + (12.7x_5) - (0.056x_6) - (17.3x_7) \qquad (3.2)$$

Figure 3.3 below was implemented in the Python program, within the codes for that, see appendix No. (9).

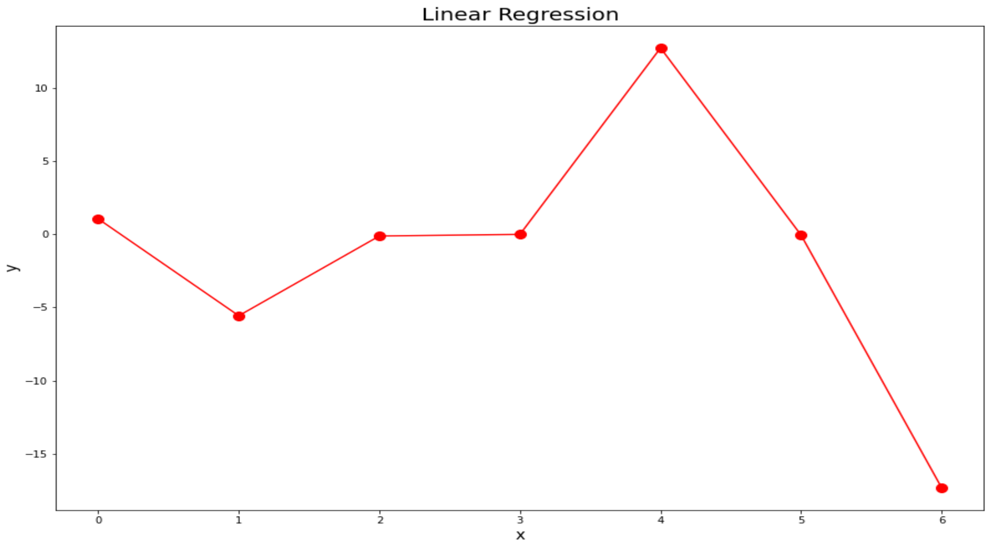

Figure 3.3 Representation of parameter values by OLS method

Based on the results of the OLS estimation method and Figure 3.3, we can see that the coefficient for the variable x_4 chemicals is -0.0009, which has a negative value indicating an inverse relationship. This suggests that as the amount of chemicals decreases, the electrical energy output increases, which may not align with reality. Additionally, we can observe that the estimates from the least squares method produced inaccurate results for the regression coefficients. This may be due to the presence of multicollinearity among the variables, which makes it difficult for Regression to accurately reflect the individual effects of each variable.

3.4.3. Ridge Regression Method

When there are multiple linear relationships in the data, we can fit the regression model using the Ridge regression technique.

This technique seeks to find coefficient estimates that minimize both the residual sum of squared values (RSS) and the sum of the squared coefficients.

$$\boldsymbol{RSS} = \sum(\boldsymbol{yi} - \widehat{\boldsymbol{yi}})^{\mathbf{2}} \qquad (\mathbf{3.3})$$

Note that:

Σ: denotes a Greek symbol meaning the sum.

yi: The value of the true response variable of the observation i.

$\widehat{yi}$: the expected value of the response variable.

Conversely, Ridge regression seeks to reduce the values of the following:

$$\boldsymbol{Ridge} = \boldsymbol{RSS} + \mathbf{k\Sigma} \sum \boldsymbol{\beta}_{\boldsymbol{j}}^{\mathbf{2}} \qquad (\mathbf{3.4})$$

where j is a predictor variable with a value between 1 and p and $\lambda \geq 0$.

The second term in the equation, the contraction penalty, is introduced to control the magnitude of the coefficients. The value that gives the lowest possible mean squared error (MSE) is chosen in the Ridge regression model. The Ridge regression method can solve the problem of multicollinearity in multiple linear regression by reducing the variance by adding a bias constant k, resulting in more stable estimations of the regression parameters despite being biased. (Yanuar et al, 2018).

To perform a Ridge regression analysis, we will use functions from the Ridge package and implement them in Python. (Jerome et al, 2010).

The Ridge regression parameters, R^2, MSE, and R were obtained when executing the codes in Python, as can be seen in Appendix No. (10). (Cule et al, 2022).

Table 3.6 Parameters of Ridge regression analysis Results

Variables	**Parameter**	**Estimated**
Constant	β_0	3.891
X_1	β_1	1.045
X_2	β_2	-0.435
X_3	β_3	-0.034
X_4	β_4	-0.0004
X_5	β_5	0.318
X_6	β_6	-0.045
X_7	β_7	0.012
MSE = 0.185 , R^2 = 0.811 , R = 0.9		

The estimated regression equation can be written using the Ridge regression method as follows:

$$\hat{y} = (3.891) + (1.045x_1) - (0.435x_2) - (0.0347x_3) - (0.0004x_4) + (0.318x_5) - (0.0454x_6) + (0.0123x_7) \quad (3.5)$$

```
plt.figure(figsize=(15,10))
plt.title('Ridge Regression',fontsize='20')
plt.plot(Rr.coef_,color='blue',alpha=3,marker='d',markersize='10',label=r'Ridge;$\alpha=0.1$')
plt.xlabel('x',fontsize='15')
plt.ylabel('y',fontsize='15')
plt.show()
```

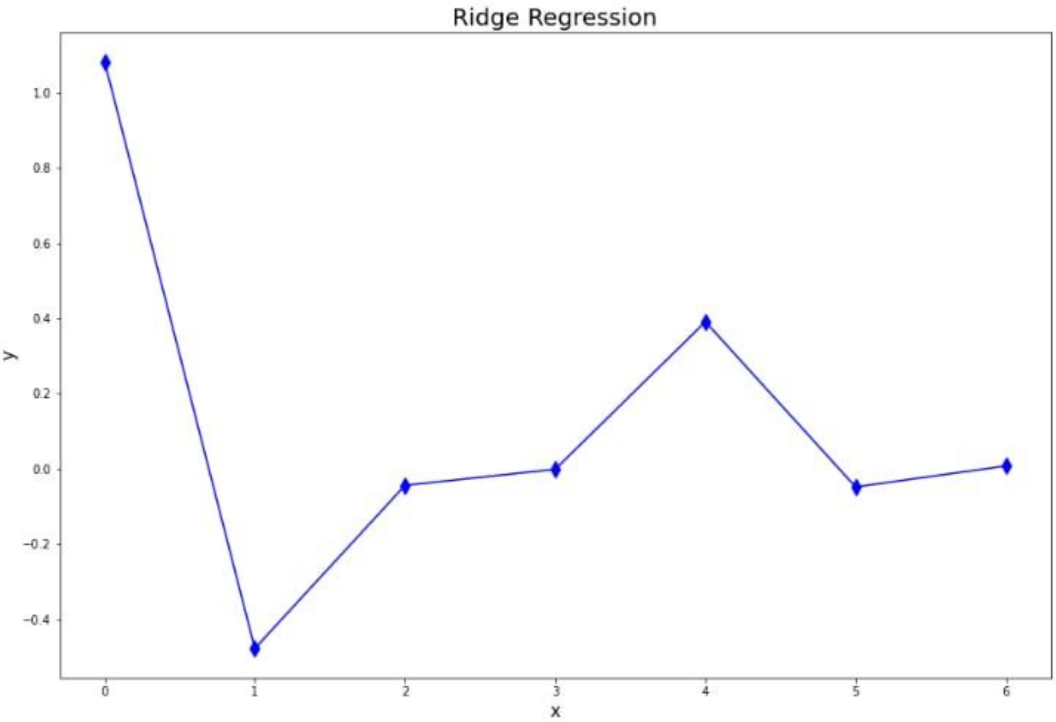

Figure 3.4 Representation of parameter values by Ridge regression

From the above Figure 3.4, we note that the results we obtained in estimating the model using the Ridge regression method were more accurate and better than the results we obtained from the OLS method. Therefore, the results according to the Ridge regression were as follows:

The value of the coefficient of determination R^2 = 0.81, is a measure of the accuracy of the model Ridge, which means that 81% of the variance in the dependent variable (electric energy production) can be predicted from the independent variables, which are larger than the value of the method determination parameter OLS, which was 0.61

The value of R = 0.9, and means that the production of electrical energy has a strong direct relationship with the independent variables affecting it. It is greater than the value of the correlation coefficient of the OLS method, which is equal to 0.7.

MSE of Ridge regression is 0.18, which was calculated, which is less than the MSE of OLS method, which is 0.26. This indicates that the Ridge regression method has addressed the problem of multicollinearity that existed in the data and gave a better regression model than the OLS method. Therefore, Ridge regression is considered the best for this study. The R-squared turns out to be 0.81 that is, the best model was able to explain 81% of the variation in the response values of the training data.

```
plt.figure(figsize=(15,10))
plt.title('Linear and Ridge Regression',fontsize='20')
plt.plot(lr.coef_,color='red',alpha=3,marker='o',markersize='10',label='Linear Regression')
plt.plot(Rr.coef_,color='blue',alpha=3,marker='d',markersize='10',label=r'Ridge;$\alpha=0.1$')
plt.xlabel('x',fontsize='15')
plt.ylabel('y',fontsize='15')
plt.legend(fontsize=12,loc=1)
plt.show()
```

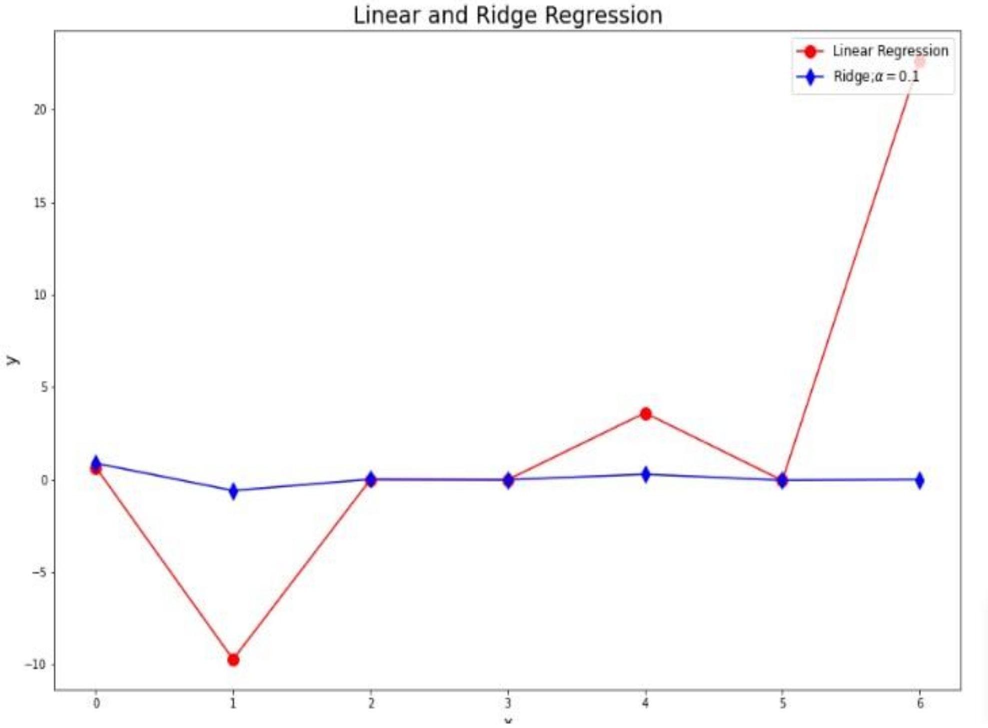

Figure 3.5 The comparison of the two models

We can see from Figure 3.5 that the parameter values for the two methods are summed. This indicates that the Ridge regression method has addressed the problem of multicollinearity that existed in the data, and gave a better regression model than the ordinary least squares method, that is MSE_{RIDGE} is less than MSE_{OLS}, so Ridge regression is considered the best for this study.

Table 3.7 Estimated regression results OLS and Ridge regression

Methods And Test and Parameters	OLS Regression	Ridge Regression
R^2	0.619	0.811
R	0.787	0.900
MSE	0.263	0.185
P_value	0.000	0.000
β_0	4.898	3.891
β_1	1.063	1.045
β_2	-5.565	-0.435
β_3	-0.109	-0.034
β_4	-0.0009	-0.0004
β_5	12.75	0.318
β_6	-0.056	-0.045
β_7	-17.34	0.012

By comparing the results in Table (3.7), we find that the Ridge regression method gave the best results as it had the highest coefficient of determination R^2 is 0.81, and the lowest of MSE = 0.18.

As for the OLS method, it gave a lower R^2 0.61, and a greater than MSE = 0.26.

Therefore, the Ridge regression method is the best, through which an optimal model can be built according to the data of this study.

4. CONCLUSIONS AND FUTURE RECOMMENDATIONS

In conclusion, the use of conventional parameter estimation methods, such as Ordinary Least Squares (OLS), can be unreliable when dealing with a high number of predictor variables or when those variables are highly correlated with each other. This can lead to large variances in the model parameters and inaccurate results. The presence of multicollinearity is a significant issue that must be addressed in order to obtain reliable results. In this study, we found that the Ridge Regression method was a more effective solution for addressing the problem of multicollinearity and producing accurate results. The Ridge Regression method resulted in a lower mean square error (MSE) than the OLS method, and therefore, it is considered the optimal model for predicting electrical energy output.

Based on the research conducted and the results obtained, the following recommendations are proposed:

- Further studies should be conducted to compare different methods for addressing the problem of multicollinearity in order to determine when each method is most appropriate to use.
- Emphasis should be placed on using statistical tests to investigate the problems of estimating the parameters of a regression model and to obtain accurate predictions.
- The use of the ordinary least squares method alone should be avoided as it does not provide an optimal model and does not fully address the issue of multicollinearity, particularly in economic studies.
- Further research should be conducted to investigate the features and capabilities of the Ridge regression method, and to compare it with other methods that address multicollinearity.
- The Ridge regression method should be applied to non-economic data to demonstrate its effectiveness in addressing multicollinearity issues.

REFERENCES

Adeboye, N. O., Fagoyinbo, I. S., & Olatayo, T. O. (2014). Estimation of the effect of multicollinearity on the standard error for regression coefficients. Journal of Mathematics, 10(4), 16-20.

Akdeniz, F., & Roozbeh, M. (2019). Generalized difference-based weighted mixed almost unbiased ridge estimator in partially linear models. Statistical Papers, 60(5), 1717-1739.

Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis. Open Journal of Statistics, 5(07), 754.

Al-Hassan, Y. M. (2010). Performance of a new Ridge Regression Estimator. Journal of the Association of Arab Universities for Basic and Applied Sciences, 9(2), pp. 4350.

Allison, P. (2022) When can you safely ignore multicollinearity?, Statistical Horizons. Available at: https://statisticalhorizons.com/multicollinearity/.

Arashi, M., Roozbeh, M., Hamzah, N. A., & Gasparini, M. (2021). Ridge regression and its applications in genetic studies. Plos one, 16(4), e0245376.

Authors (2020) P values, Analysis INN. Analysis INN. Available at: https://www.analysisinn.com/post/p-values/.

Bager, A. S. M., Mohammed, B. K., & Odah, M. H. (2018). Ridge Regression for Addressing of the Multicollinearity Problem with Application in Cost of Production. LUMEN Proceedings, 3, 56-63.

Bagheri, A., & Midi, H. (2009). Robust estimations as a remedy for multicollinearity caused by multiple high leverage points. Journal of Mathematics and Statistics, 5(4), 311.

Chen, G. J. (2012). A simple way to deal with multicollinearity. Journal of Applied Statistics, 39(9), 1893-1909.

Crouse, R. H., Jin, C., & Hanumara, R. C. (1995). Unbiased Ridge estimation with prior information and ridge trace. Communications in Statistics-Theory and Methods, 24(9), 2341-2354.

Cule E, Moritz S, Frankowski D (2022). _ridge: Ridge Regression with Automatic Selection of the Penalty Parameter_. R package version 3.3, https://CRAN.R-project.org/package=ridge

Cule, E., Vineis, P., & De Iorio, M. (2011). Significance testing in ridge regression for genetic data. BMC bioinformatics, 12(1), 1-15.

Dai, L., Chen, K., Sun, Z., Liu, Z., & Li, G. (2018). Broken adaptive ridge regression and its asymptotic properties. Journal of Multivariate Analysis, 168, 334-351.

Doll, J. P. (1974). On Exact Multicollinearity and the Estimation of the Cobb-Douglas Production Function. American Journal of Agricultural Economics, 56(3), 556-563.

Dorugade, A. V., & Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. Applied Mathematical Sciences, 4(9), 447-456.

Duzan, H., & Shariff, N. S. B. M. (2015). Ridge regression for solving the multicollinearity problem: review of methods and models. Journal of Applied Science.

Emmanuel, B. O., & Maureen, N. T. (2021). Investigating heteroscedastic disturbances in some transformed economic models while eliminating multicollinearity.

Fitrianto, A. and Yik, L. C. (2014). Performance of Ridge Regression Estimator Method on Small Sample size By Varying correlation coefficients: A simulation study. Journal of Mathematics and Statistics 10 (1), pp. 25 – 29.

Frost, J. (2017). Choosing the Correct Type of Regression Analysis-Statistics By Jim. Statistics By Jim.

Garba, N., Danchadi, N. S., and Abdulmumin, M. K. (2021). Evaluating the Performance of Ordinary Least Square and Polynomial Regression with Respect to Sample Size. International Journal Of Science for Global Sustainability, 7(4).

Gorman, K. (2010). The consequences of multicollinearity among socioeconomic predictors of negative concord in Philadelphia. University of Pennsylvania Working Papers in Linguistics, 16(2), 9.

Gunst, R. F., & Mason, R. L. (2018). Regression analysis and its application: a data-oriented approach. CRC Press.

Hemmerle, W. J. (1975). An explicit solution for generalized ridge regression. Technometrics, 17(3), 309-314.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Jeewantha, J. (2021). ASSUMPTIONS OF CLASSICAL LINEAR REGRESSION MODEL (CLRM). Available at SSRN 3972911.

Krashniak, A., & Lamm, E. (2021). Francis Galton's regression towards mediocrity and the stability of types. Studies in History and Philosophy of Science Part A, 86, 6-19.

Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22.

Kalnins, A. (2018). Multicollinearity: How common factors cause Type 1 errors in multivariate regression. Strategic Management Journal, 39(8), 2362-2385.

Kibria, B. M., & Banik, S. (2016). Some ridge regression estimators and their performances. Journal of Modern Applied Statistical Methods, 15(1), 12.

Kibria, B. M., and Lukman, A. F. (2020). A new ridge-type estimator for the linear regression model: Simulations and applications. Scientifica, 2020.

Lindner, T., Puck, J., & Verbeke, A. (2020). Misconceptions about multicollinearity in international business research: Identification, consequences, and remedies. Journal of International Business Studies, 51(3), 283-298.

Long, R. G. (2008). The crux of the method: assumptions in ordinary least squares and logistic regression. Psychological reports, 103(2), 431-434.

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. The American Statistician, 29(1), 3-20.

Mela, C. F., & Kopalle, P. K. (2002). The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. Applied Economics, 34(6), 667-677.

Mosleh, M., Roshani, S., & Coccia, M. (2022). Scientific laws of research funding to support citations and diffusion of knowledge in life science. Scientometrics, 127(4), 1931–1951. https://doi.org/10.1007/s11192-022-04300-1 .

Niemelä-Nyrhinen, J., & Leskinen, E. (2014). Multicollinearity in marketing models: notes on the application of ridge trace estimation in structural equation modeling. Electronic Journal of Business Research Methods, 12(1), pp3-15.

O'Brien, T. E., Tarasinski, B., & Terhal, B. M. (2019). Quantum phase estimation of multiple eigenvalues for small-scale (noisy) experiments. New Journal of Physics, 21(2), 023022.

Owusu-Ansah, Boateng. (2018). L1-L2 Regularization of Collinear Data (Doctoral dissertation, University of Ghana).

Paul, R. K. (2006). Multicollinearity: Causes, effects, and remedies. IASRI, New Delhi, 1(1), 58-65.

Rockwell, R. C. (1975). Assessment of multicollinearity: The Haitovsky test of the determinant. Sociological Methods & Research, 3(3), 308-320.

Sim, S. F., Chai, M. X. L., & Jeffrey Kimura, A. L. (2018). Prediction of lard in palm olein oil using simple linear regression (SLR), multiple linear regression (MLR), and partial least squares regression (PLSR) based on Fourier-transform infrared (FTIR). Journal of Chemistry, 2018.

Stigler, S. M. (1981). Gauss and the invention of least squares. the Annals of Statistics, 465-474. https://www.jstor.org/stable/2240811#metadata_info_tab_contents

Taiyun Wei and Viliam Simko (2021). R package 'complot': Visualization of a Correlation Matrix (Version 0.92). Available from https://github.com/taiyun/corrplot

Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Ullah, M. I., Aslam, M., Altaf, S., & Ahmed, M. (2019). Some new diagnostics of multicollinearity in linear regression model. Sains Malaysiana, 48(9), 2051-2060.

VanDam, M., Oller, D. K., Ambrose, S. E., Gray, S., Richards, J. A., Xu, D., and Moeller, M. P. (2015). Automated vocal analysis of children with hearing loss and their typical and atypical peers. Ear and hearing, 36(4), e146.

W. Wichern, Dean; A. Churchill, Gilbert, (2012)."A Comparison of Ridge Estimators'', T Econometrics Journal, 301-311.

Weaving, D., Jones, B., Ireton, M., Whitehead, S., Till, K., & Beggs, C. B. (2019). Overcoming the problem of multicollinearity in sports performance data: A novel application of partial least squares correlation analysis. PLoS One, 14(2), e0211776.

Weissfeld, L. A., & Sereika, S. M. (1991). A multicollinearity diagnostic for generalized linear models. Communications in Statistics-Theory and Methods, 20(4), 1183-1198.

Welling, M. (2013). Kernel ridge regression. Max Welling's classnotes in machine learning,

William J. Hemmerle (1975) An Explicit Solution for Generalized Ridge Regression, Technometrics, 17:3, 309-314, DOI: 10.1080/00401706.1975.10489333.

Yanuar, F., Tillah, M., & Devianto, D. (2018). Modeling of Human Development Index Using Ridge Regression Method. Eksakta: Berkala Ilmiah Bidang MIPA (E-ISSN: 2549-7464), 19(2), 1-11.

Zach (2020) Ridge regression in R (step-by-step), Statology. Available at: https://www.statology.org/ridge-regression-in-r/.

Zach (2021) How to calculate variance inflation factor (VIF) in R, Statology. Available at: https://www.statology.org/variance-inflation-factor-r/.

Zhang, F., Zhang, R., Wang, W., Yang, W., Li, L., Xiong, Y., ... & Du, Y. (2019). Ridge regression combined with model complexity analysis for near - infrared (NIR) spectroscopic model updating. Chemometrics and Intelligent Laboratory Systems, 195, 103896.

APPENDIX

APPENDIX 1: The Study Data, Which Represents The Production Of Electrical Energy And Eight Factors Affecting It.

y	X1	X2	X3	X4	X5	X6	X7
2.24	0.448	0.0223	15.3	92.2	0.0752	32.5	0.00222
2.71	0.435	0.0552	12.8	77.3	0.0629	32.5	0.00246
2.78	0.392	0.0669	13.2	79.1	0.0646	32.5	0.00316
2.77	0.384	0.0362	13.1	79	0.0642	32.5	0.00398
2.68	0.37	0.0336	12.7	76.8	0.0622	32.5	0.00444
2.2	0.305	0.0156	10.4	63.1	0.0512	32.5	0.00515
1.56	0.216	0.0412	7.39	44.8	0.0363	44.2	0.00538
1.56	0.216	0.0762	7.39	44.8	0.0363	44.2	0.0055
1.57	0.219	0.0446	7.45	44.8	0.0365	44.2	0.00503
1.46	0.2	0.0429	6.93	42	0.0339	44.2	0.00433
1.31	0.185	0.0602	6.24	37.4	0.0305	55.9	0.00316
1.01	0.143	0.0573	4.82	28.8	0.0236	55.9	0.00246
1.12	0.129	0.0549	4.36	26.6	0.0214	55.9	0.00222
1.08	0.826	0.0346	2.79	16.9	0.0137	67.4	0.00246
1.24	0.315	0.0316	1.08	6.58	0.00526	67.6	0.00292
1.22	0.312	0.0211	1.07	65.8	0.00526	67.6	0.00363
1.57	0.787	0.0354	2.68	16.3	0.0131	67.5	0.00433
1.46	0.636	0.0253	2.18	12.9	0.0106	67.5	0.00468
1.5	0.687	0.0208	2.35	140	0.0115	67.5	0.00491
1.47	0.656	0.0192	2.25	13.4	0.011	67.5	0.00514
1.42	0.59	0.0178	1.99	12.3	0.00971	67.6	0.00491
1.27	0.426	0.0226	1.28	77.3	0.00632	67.6	0.00433
1.27	0.455	0.0218	1.25	77.3	0.00608	67.5	0.00304
1.26	0.338	0.0175	1.24	76.1	0.0364	67.4	0.00222
1.56	0.212	0.0739	7.4	44.7	0.0113	43.9	0.0021
1.57	0.196	0.0412	2.25	440	0.00526	43.6	0.00245
1.46	0.197	0.0406	7.4	41.9	0.0062	44.2	0.00281
1.31	0.178	0.0591	5.45	374	0.0106	55.5	0.00363
1	0.131	0.0557	2.35	28.5	0.0115	55.4	0.00433
1.92	0.121	0.0526	1.26	26.3	0.011	55.8	0.00503
1.59	0.821	0.0253	2.35	16.6	0.0106	66.9	0.00175
1.23	0.26	0.0289	1.28	65.6	0.0304	66.9	0.00292
1.46	0.636	0.0254	2.18	12.9	0.00947	67.4	0.00386
1.49	0.688	0.0208	2.35	14	0.0116	67.3	0.00409
1.47	0.657	0.0193	2.25	13.5	0.0113	67.6	0.00467
1.42	0.59	0.0178	1.99	12.3	0.00994	67.5	0.00491
1.34	0.836	0.0349	2.77	17	0.0142	67.4	0.0028
1.03	0.347	0.0324	1.09	67	0.00561	67.6	0.00316
1.11	0.323	0.0209	1.08	67	0.00561	67.6	0.00385
1.58	0.802	0.0361	2.67	16.4	0.0127	67.5	0.00467
1.24	0.641	0.026	2.34	14.1	0.011	67.6	0.00503

1.27	0.675	0.0207	2.47	14.1	0.0106	67.5	0.00468
1.56	0.668	0.0199	2.36	13.6	0.0113	67.6	0.0055
1.46	0.61	0.0181	2.08	14.4	0.01	67.5	0.00573
1.46	0.415	0.0229	1.41	79.1	0.00678	67.6	0.00479
1.43	0.442	0.0223	1.33	78.6	0.00584	67.7	0.00327
1.45	0.347	0.0178	1.34	76.3	0.037	66.3	0.00211
1.31	0.224	0.0779	7.27	45.9	0.0115	44.6	0.00257
2.58	0.206	0.0408	2.36	45.2	0.00526	44.8	0.00233
2.16	0.203	0.041	7.09	41.1	0.00619	42.9	0.00257

APPENDIX 2: Define Variables (X, Y).

```
x=file.drop('y',axis=1).values
y=file['y'].values
print(x,y)
```

```
[[4.48e-01 2.23e-02 1.53e+01 9.22e+01 7.52e-02 3.25e+01 2.22e-03]
 [4.35e-01 5.52e-02 1.28e+01 7.73e+01 6.29e-02 3.25e+01 2.46e-03]
 [3.92e-01 6.69e-02 1.32e+01 7.91e+01 6.46e-02 3.25e+01 3.16e-03]
 [3.84e-01 3.62e-02 1.31e+01 7.90e+01 6.42e-02 3.25e+01 3.98e-03]
 [3.70e-01 3.36e-02 1.27e+01 7.68e+01 6.22e-02 3.25e+01 4.44e-03]
 [3.05e-01 1.56e-02 1.04e+01 6.31e+01 5.12e-02 3.25e+01 5.15e-03]
 [2.16e-01 4.12e-02 7.39e+00 4.48e+01 3.63e-02 4.42e+01 5.38e-03]
 [2.16e-01 7.62e-02 7.39e+00 4.48e+01 3.63e-02 4.42e+01 5.50e-03]
 [2.19e-01 4.46e-02 7.45e+00 4.48e+01 3.65e-02 4.42e+01 5.03e-03]
 [2.00e-01 4.29e-02 6.93e+00 4.20e+01 3.39e-02 4.42e+01 4.33e-03]
 [1.85e-01 6.02e-02 6.24e+00 3.74e+01 3.05e-02 5.59e+01 3.16e-03]
 [1.43e-01 5.73e-02 4.82e+00 2.88e+01 2.36e-02 5.59e+01 2.46e-03]
 [1.29e-01 5.49e-02 4.36e+00 2.66e+01 2.14e-02 5.59e+01 2.22e-03]
 [8.26e-01 3.46e-02 2.79e+00 1.69e+01 1.37e-02 6.74e+01 2.46e-03]
 [3.15e-01 3.16e-02 1.08e+00 6.58e+00 5.26e-03 6.76e+01 2.92e-03]
 [3.12e-01 2.11e-02 1.07e+00 6.58e+01 5.26e-03 6.76e+01 3.63e-03]
 [7.87e-01 3.54e-02 2.68e+00 1.63e+01 1.31e-02 6.75e+01 4.33e-03]
 [6.36e-01 2.53e-02 2.18e+00 1.29e+01 1.06e-02 6.75e+01 4.68e-03]
 [6.87e-01 2.08e-02 2.35e+00 1.40e+02 1.15e-02 6.75e+01 4.91e-03]
 [6.56e-01 1.92e-02 2.25e+00 1.34e+01 1.10e-02 6.75e+01 5.14e-03]
 [5.90e-01 1.78e-02 1.99e+00 1.23e+01 9.71e-03 6.76e+01 4.91e-03]
 [4.26e-01 2.26e-02 1.28e+00 7.73e+01 6.32e-03 6.76e+01 4.33e-03]
 [4.55e-01 2.18e-02 1.25e+00 7.73e+01 6.08e-03 6.75e+01 3.04e-03]
 [3.38e-01 1.75e-02 1.24e+00 7.61e+01 3.64e-02 6.74e+01 2.22e-03]
 [2.12e-01 7.39e-02 7.40e+00 4.47e+01 1.13e-02 4.39e+01 2.10e-03]
 [1.96e-01 4.12e-02 2.25e+00 4.40e+02 5.26e-03 4.36e+01 2.45e-03]
 [1.97e-01 4.06e-02 7.40e+00 4.19e+01 6.20e-03 4.42e+01 2.81e-03]
 [1.78e-01 5.91e-02 5.45e+00 3.74e+02 1.06e-02 5.55e+01 3.63e-03]
 [1.31e-01 5.57e-02 2.35e+00 2.85e+01 1.15e-02 5.54e+01 4.33e-03]
 [1.21e-01 5.26e-02 1.26e+00 2.63e+01 1.10e-02 5.58e+01 5.03e-03]
 [8.21e-01 2.53e-02 2.35e+00 1.66e+01 1.06e-02 6.69e+01 1.75e-03]
 [2.60e-01 2.89e-02 1.28e+00 6.56e+01 3.04e-02 6.69e+01 2.92e-03]
 [6.36e-01 2.54e-02 2.18e+00 1.29e+01 9.47e-03 6.74e+01 3.86e-03]
```

APPENDIX 3: Correlation Matrix

```
file.corr()
```

	y	x1	x2	x3	x4	x5	x6	x7
y	1.000000	-0.069136	0.135277	0.721969	0.074976	0.654707	-0.728352	-0.028872
x1	-0.069136	1.000000	-0.557630	-0.277882	-0.314273	-0.187514	0.539021	0.236507
x2	0.135277	-0.557630	1.000000	0.431709	0.147802	0.188003	-0.546121	-0.259044
x3	0.721969	-0.277882	0.431709	1.000000	0.093648	0.837697	-0.903646	-0.110698
x4	0.074976	-0.314273	0.147802	0.093648	1.000000	0.034022	-0.243946	-0.200872
x5	0.654707	-0.187514	0.188003	0.837697	0.034022	1.000000	-0.705035	-0.074458
x6	-0.728352	0.539021	-0.546121	-0.903646	-0.243946	-0.705035	1.000000	0.165024
x7	-0.028872	0.236507	-0.259044	-0.110698	-0.200872	-0.074458	0.165024	1.000000

APPENDIX 4: Codes Importing Packages, Functions, Classes

```
import numpy as np
import pandas as pd
import math
import csv
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression,Ridge
```

APPENDIX 5: Data Import Supplement

```
file=pd.read_csv('Downloads/Book1.csv')
file
```

	y	x1	x2	x3	x4	x5	x6	x7
0	2.24	0.448	0.0223	15.30	92.20	0.07520	32.5	0.00222
1	2.71	0.435	0.0552	12.80	77.30	0.06290	32.5	0.00246
2	2.78	0.392	0.0669	13.20	79.10	0.06460	32.5	0.00316
3	2.77	0.384	0.0362	13.10	79.00	0.06420	32.5	0.00398
4	2.68	0.370	0.0336	12.70	76.80	0.06220	32.5	0.00444
5	2.20	0.305	0.0156	10.40	63.10	0.05120	32.5	0.00515
6	1.56	0.216	0.0412	7.39	44.80	0.03630	44.2	0.00538
7	1.56	0.216	0.0762	7.39	44.80	0.03630	44.2	0.00550
8	1.57	0.219	0.0446	7.45	44.80	0.03650	44.2	0.00503
9	1.46	0.200	0.0429	6.93	42.00	0.03390	44.2	0.00433
10	1.31	0.185	0.0602	6.24	37.40	0.03050	55.9	0.00316
11	1.01	0.143	0.0573	4.82	28.80	0.02360	55.9	0.00246
12	1.12	0.129	0.0549	4.36	26.60	0.02140	55.9	0.00222
13	1.08	0.826	0.0346	2.79	16.90	0.01370	67.4	0.00246
14	1.24	0.315	0.0316	1.08	6.58	0.00526	67.6	0.00292
15	1.22	0.312	0.0211	1.07	65.80	0.00526	67.6	0.00363
16	1.57	0.787	0.0354	2.68	16.30	0.01310	67.5	0.00433

17	1.46	0.636	0.0253	2.18	12.90	0.01060	67.5	0.00468
18	1.50	0.687	0.0208	2.35	140.00	0.01150	67.5	0.00491
19	1.47	0.656	0.0192	2.25	13.40	0.01100	67.5	0.00514
20	1.42	0.590	0.0178	1.99	12.30	0.00971	67.6	0.00491
21	1.27	0.426	0.0226	1.28	77.30	0.00632	67.6	0.00433
22	1.27	0.455	0.0218	1.25	77.30	0.00608	67.5	0.00304
23	1.26	0.338	0.0175	1.24	76.10	0.03640	67.4	0.00222
24	1.56	0.212	0.0739	7.40	44.70	0.01130	43.9	0.00210
25	1.57	0.196	0.0412	2.25	440.00	0.00526	43.6	0.00245
26	1.46	0.197	0.0406	7.40	41.90	0.00620	44.2	0.00281
27	1.31	0.178	0.0591	5.45	374.00	0.01060	55.5	0.00363
28	1.00	0.131	0.0557	2.35	28.50	0.01150	55.4	0.00433
29	1.92	0.121	0.0526	1.26	26.30	0.01100	55.8	0.00503
30	1.59	0.821	0.0253	2.35	16.60	0.01060	66.9	0.00175
31	1.23	0.260	0.0289	1.28	65.60	0.03040	66.9	0.00292
32	1.46	0.636	0.0254	2.18	12.90	0.00947	67.4	0.00386
33	1.49	0.688	0.0208	2.35	14.00	0.01160	67.3	0.00409
34	1.47	0.657	0.0193	2.25	13.50	0.01130	67.6	0.00467
35	1.42	0.590	0.0178	1.99	12.30	0.00994	67.5	0.00491
36	1.34	0.836	0.0349	2.77	17.00	0.01420	67.4	0.00280
37	1.03	0.347	0.0324	1.09	67.00	0.00561	67.6	0.00316
38	1.11	0.323	0.0209	1.08	67.00	0.00561	67.6	0.00385
39	1.58	0.802	0.0361	2.67	16.40	0.01270	67.5	0.00467
40	1.24	0.641	0.0260	2.34	14.10	0.01100	67.6	0.00503
41	1.27	0.675	0.0207	2.47	14.10	0.01060	67.5	0.00468
42	1.56	0.668	0.0199	2.36	13.60	0.01130	67.6	0.00550
43	1.46	0.610	0.0181	2.08	14.40	0.01000	67.5	0.00573
44	1.46	0.415	0.0229	1.41	79.10	0.00678	67.6	0.00479
45	1.43	0.442	0.0223	1.33	78.60	0.00584	67.7	0.00327
46	1.45	0.347	0.0178	1.34	76.30	0.03700	66.3	0.00211
47	1.31	0.224	0.0779	7.27	45.90	0.01150	44.6	0.00257
48	2.58	0.206	0.0408	2.36	45.20	0.00526	44.8	0.00233
49	2.16	0.203	0.0410	7.09	41.10	0.00619	42.9	0.00257

APPENDIX 6: General Information About The Data

38	1.11	0.323	0.0209	1.08	67.00	0.00561	67.6	0.00385
39	1.58	0.802	0.0361	2.67	16.40	0.01270	67.5	0.00467
40	1.24	0.641	0.0260	2.34	14.10	0.01100	67.6	0.00503
41	1.27	0.675	0.0207	2.47	14.10	0.01060	67.5	0.00468
42	1.56	0.668	0.0199	2.36	13.60	0.01130	67.6	0.00550
43	1.46	0.610	0.0181	2.08	14.40	0.01000	67.5	0.00573
44	1.46	0.415	0.0229	1.41	79.10	0.00678	67.6	0.00479
45	1.43	0.442	0.0223	1.33	78.60	0.00584	67.7	0.00327
46	1.45	0.347	0.0178	1.34	76.30	0.03700	66.3	0.00211
47	1.31	0.224	0.0779	7.27	45.90	0.01150	44.6	0.00257
48	2.58	0.206	0.0408	2.36	45.20	0.00526	44.8	0.00233
49	2.16	0.203	0.0410	7.09	41.10	0.00619	42.9	0.00257

```
file.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 8 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   y       50 non-null     float64
 1   x1      50 non-null     float64
 2   x2      50 non-null     float64
 3   x3      50 non-null     float64
 4   x4      50 non-null     float64
 5   x5      50 non-null     float64
 6   x6      50 non-null     float64
 7   x7      50 non-null     float64
dtypes: float64(8)
memory usage: 3.2 KB
```

APPENDIX 7: Split The Data

```
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3)
len(x_train),len(x_test),len(y_train),len(y_test)
```

```
(35, 15, 35, 15)
```

APPENDIX 8: Linear Regression Model OLS, MSE, R^2, And Correlation.

```
Ir=LinearRegression().fit(x_train,y_train)
print('Intercept:',Ir.intercept_)
print('Slop:',Ir.coef_)
print('Determination Coefficient:',Ir.score(x_test,y_test))
print('Correlation Coefficient:',math.sqrt(Ir.score(x_test,y_test)))
y_pred=Ir.predict(x_test)
MSE_1=np.sqrt(mean_squared_error(y_test,y_pred))
print('MES_1:',MSE_1)
```

```
Intercept: 4.898897978344561
Slop: [ 1.06325118e+00 -5.56534915e+00 -1.09835215e-01 -9.56863948e-04
  1.27532303e+01 -5.68720265e-02 -1.73433290e+01]
Determination Coefficient: 0.6194852902489836
Correlation Coefficient: 0.7870738785202972
MES_1: 0.26362002088695363
```

APPENDIX 9: Linear Regression Chart

```
plt.figure(figsize=(15,10))
plt.title('Linear Regression',fontsize='20')
plt.plot(Ir.coef_,color='r',marker='o',markersize='10',label='Linear Regression')
plt.xlabel('x',fontsize='15')
plt.ylabel('y',fontsize='15')
plt.show()
```

APPENDIX 10: Ridge Regression Analysis Codes

```
Rr=Ridge(alpha=0.1).fit(x_train,y_train)
print('Intercept:',Rr.intercept_)
print('Slop:',Rr.coef_)
print('Determination Coefficient:',Rr.score(x_test,y_test))
print('Correlation Coefficient:',math.sqrt(Rr.score(x_test,y_test)))
y_pred=Rr.predict(x_test)
MSE_2=np.sqrt(mean_squared_error(y_test,y_pred))
print('MES_2:',MSE_2)
```

```
Intercept: 3.891364947350688
Slop: [ 1.04507537e+00 -4.35313823e-01 -3.47304512e-02 -4.97398518e-04
  3.18816759e-01 -4.54572920e-02  1.23324697e-02]
Determination Coefficient: 0.811652717312363
Correlation Coefficient: 0.9009177084020288
MES_2: 0.18546943309083538
```

CURRICULUM VITAE

Ahmed Aziz Abdul-Hussein Graduated from Al-Thabat Preparatory School / Scientific Branch (2008-2009) and Graduated from Karbala University / College of Administration and Economics / Department of Statistics (2013-2014). He joined Ondokuz Mayis University / Graduate Institute, Statistics Department, Master's Program (English Language) in (2020). In 2014, he worked in the Iraqi Ministry of Health in the Karbala City Health Department and held the position of director of the information bank in the main department in Iraq and is continuing his work.

Contact Information :

ORCID ID : *https://orcid.org/0000-0002-0964-5599*

Publications :

1. Participated in the scientific conference (The 14th International Scientific Research Congress) in August for the period from (20,21-2022) with the article (A STUDY ON PREDICTING THE PRODUCTION OF ELECTRICAL ENERGY: KARBALA EXAMPLE).