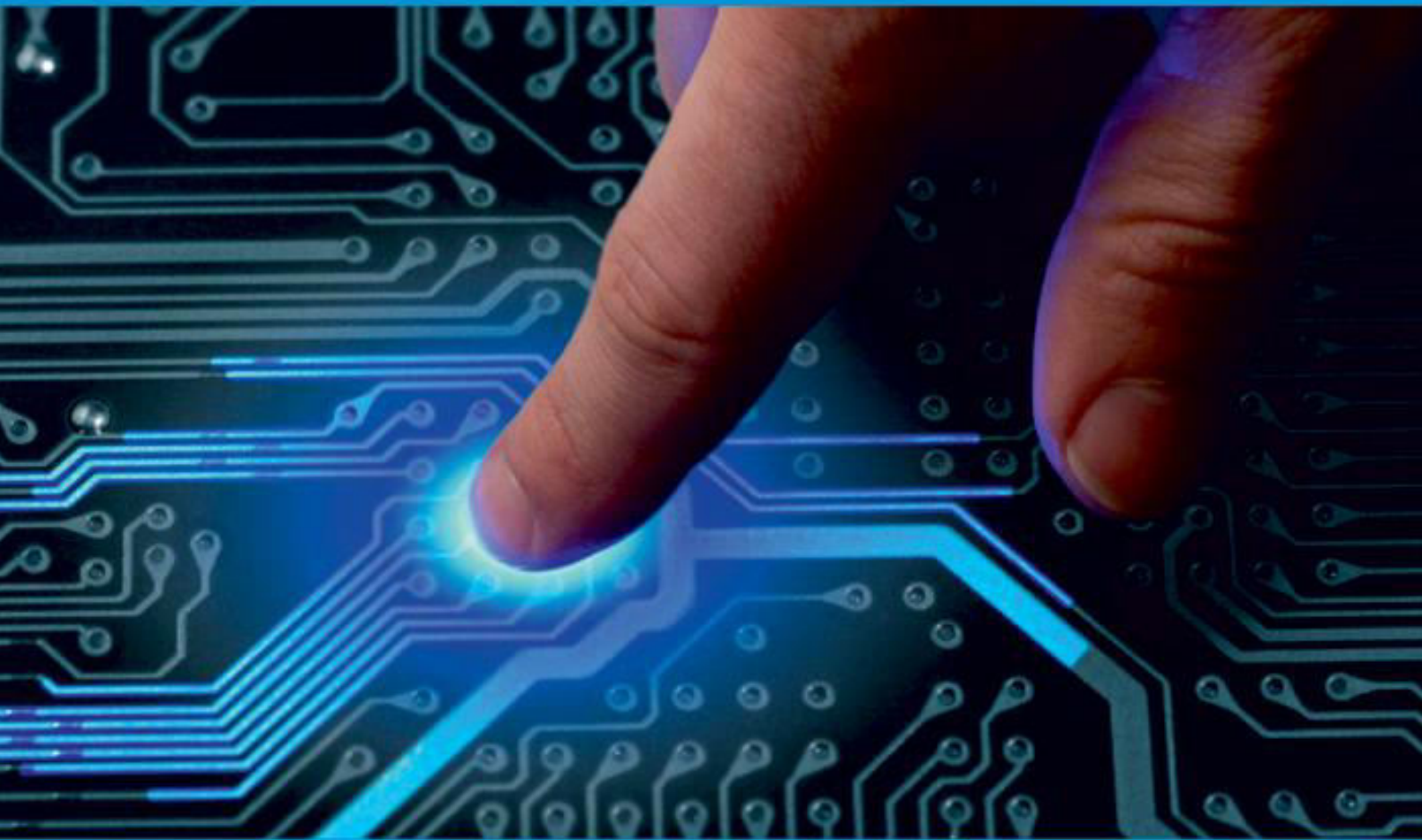




IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH


IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 7, July 2023

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.379

 9940 572 462

 6381 907 438

 ijircce@gmail.com

 www.ijircce.com

Comparative Study of Regression Models using Case Study of Limited Data

Joel Thomas Chacko

UG Student, Dept. of C.S., MVJ College of Engineering, Whitefield, Bengaluru, India

ABSTRACT: In this comparative study, we evaluate and compare the performance of these regression models based on their accuracy, robustness and generalizability. In this study, we use a dataset of house prices to compare the performance of these models in predicting prices based on available independent variables. This study aims to identify the best regression models for predicting house prices and provides insight into the strengths and weaknesses of each model. This paper attempts to understand the behaviour of these regression models when trained on case studies containing limited data. The purpose of this research is to help researchers and practitioners choose the appropriate regression model for a particular application based on the properties of the data set and the specific research question. The findings may have implications for many fields, including economics, finance, social sciences, and engineering.

KEYWORDS: Regression Analysis, Linear Regression, Lasso Regression, Elastic Net Regression, Quantile Regression, Poisson Regression, Comparative Analysis.

I. INTRODUCTION

Regression analysis is a widely used statistical technique that helps identify and quantify relationships between one or more independent and dependent variables. Regression models are widely used in various fields such as economics, finance, social sciences, and engineering to predict the future values of dependent variables based on the values of independent variables. Regression models aim to capture the basic functional form that describes the relationships between variables, which can be used to make predictions and draw conclusions.

Choosing the right regression model for a particular dataset depends on several factors, including: B. The nature and distribution of data, the nature of relationships between variables, and specific goals of analysis. Various types of regression models have been developed over the years, each with their own strengths and weaknesses. Commonly used regression models include linear regression, logistic regression, polynomial regression, ridge regression, and random forest regression.

The purpose of this comparative study is to evaluate and compare the performance of five widely used regression models. They are-

1. Linear Regression:

- A linear regression model assumes a linear relationship between the dependent variable and one or more independent variables.
- It is a simple and widely used method that is easy to interpret and implement.
- It can be used for both continuous and categorical dependent variables.
- It may not capture complex relationships between variables and may not perform well when assumptions are violated.

2. Lasso Regression:

- A Lasso regression model is used to select important features and reduce the number of predictors in a model.
- It shrinks the regression coefficients towards zero, effectively reducing the size of the coefficients for some variables to zero, thereby eliminating those variables from the model.
- It is useful when dealing with high-dimensional datasets, where the number of predictors is large compared to the number of observations.
- It can be computationally expensive when the number of predictors is very large.

3. Quantile Regression:

- A quantile regression model estimates the conditional quantiles of the dependent variable based on the values of one or more independent variables.
- It is a useful method when the relationship between the variables is not linear, or when the distribution of the dependent variable is not normal.
- It can be used to estimate the median, quartiles, or any other quantile of the dependent variable.
- It is less sensitive to outliers than other regression models.

4. Elastic Net Regression:

- An Elastic Net regression model is a combination of Lasso and Ridge regression.
- It is used to select important features while reducing the size of the coefficients for some variables.
- It can handle highly correlated predictors better than Lasso regression.
- It is useful when dealing with high-dimensional datasets, where the number of predictors is large compared to the number of observations.

5. Poisson Regression:

- A Poisson regression model is used to model count data, where the dependent variable is the number of occurrences of an event.
- It assumes that the dependent variable follows a Poisson distribution.
- It is useful when dealing with non-negative integer values of the dependent variable.
- It can be extended to handle over-dispersed count data using a negative binomial regression model.

This comparative study evaluates the performance of these three regression models on the house price dataset and provides insight into the strengths and weaknesses of each model. This study is designed to help researchers and practitioners select a suitable regression model for a particular application.

II. RELATED WORK

1. Apostolos Nicholas Refenes, Achileas Zapranis, Gavin Francis (1994). Stock performance modeling using neural networks: A comparative study with regression models [1]

We examine the use of neural networks as an alternative to classical statistical techniques for forecasting within the framework of the APT (arbitrage pricing theory) model for stock ranking. We show that neural networks outperform these statistical techniques in forecasting accuracy terms, and give better model fitness in-sample by one order of magnitude. We identify intervals for the network parameter values for which these performance figures are statistically stable. Neural networks have been criticised for not being able to provide an explanation of how they interact with their environment and how they reach an outcome. We show that by using sensitivity analysis, neural networks can provide a reasonable explanation of their predictive behaviour and can model their environment more convincingly than regression models.

2. Kavitha S, Varuna S, Ramya R (2016). A comparative analysis on linear regression and support vector regression [2]

In business, consumers interest, behavior, product profits are the insights required to predict the future of business with the current data or historical data. These insights can be generated with the statistical techniques for the purpose of forecasting. The statistical techniques can be evaluated for the predictive model based on the requirements of the data. The prediction and forecasting are done widely with time series data. Most of the applications such as weather forecasting, finance and stock market combine historical data with the current streaming data for better accuracy. However, the time series data is analysed with regression models. In this paper, linear regression and support vector regression model is compared using the training data set in order to use the correct model for better prediction and accuracy.

3. Carlos Bastida, José Holguín-Veras (2009). Freight Generation Models: Comparative Analysis of Regression Models and Multiple Classification Analysis [3]

This paper conducts a comparative analysis of two alternative approaches to freight generation modeling: ordinary least square (OLS) and cross classification. OLS models were estimated to identify a functional relationship between the number of freight deliveries per day and a set of company attributes used as independent variables. Cross-classification techniques aim at identifying a classification structure that provides a good representation of the freight generation process. To that effect, multiple classification analysis was used to identify groups of independent variables explaining freight generation, which provided the basis for constructing cross-classification tables. In both cases freight generation

is explained as a function of company attributes. The model estimation process used data obtained from commercial establishments located in Manhattan and Brooklyn, New York. More than 190 different variables were tested as predictors for the number of deliveries received or carried per day. Six linear regression models found to be statistically significant and conceptually valid are discussed.

4. James W Wisnowski, Douglas C Montgomery, James R Simpson (2001). A Comparative analysis of multiple outlier detection procedures in the linear regression model [4]

We evaluate several published techniques to detect multiple outliers in linear regression using an extensive Monte Carlo simulation. These procedures include both direct methods from algorithms and indirect methods from robust regression estimators. We evaluate the impact of outlier density and geometry, regressor variable dimension, and outlying distance in both leverage and residual on detection capability and false alarm (swamping) probability. The simulation scenarios focus on outlier configurations likely to be encountered in practice and use a designed experiment approach. The results for each scenario provide insight and limitations to performance for each technique. Finally, we summarize each procedure's performance and make recommendations.

5. Soner Cankaya, G. Tamer Kayaalp, Levent Sangun, Yalcin Tahtali & Mustafa Akar(2006). A Comparative Study of Estimation Methods for Parameters in Multiple Linear Regression Model [5]

This paper investigated least squares method, non-parametric method and robust regression methods to predict the parameters of multiple regression models. To evaluate these methods, measurements of body weight, total length and fork length of fishes collected from *Serranus cabrilla* were used. In these regression models, body weight was dependent variable whereas total length and fork length were independent variables. The results show that non-parametric regression method, general additive model, has minimum R2 value and least median squares has maximum R2 value, 0.334 and 0.855, respectively.

III. METHODOLOGY

Choosing the right regression model for a particular dataset depends on several factors, including: B. The nature and distribution of data, the nature of relationships between variables, and specific goals of analysis. Various types of regression models have been developed over the years, each with their own strengths and weaknesses. Commonly used regression models include linear regression, logistic regression, polynomial regression, ridge regression, and random forest regression.

The methodology of analysing the following regressions: Linear Regression, Lasso Regression, Quantile Regression, Elastic Net Regression, and Poisson Regression are as shown below.

3.1 Different Regression Models

3.1.1 Linear Regression

The most extensively used modelling technique is linear regression, which assumes a linear connection between a dependent variable (Y) and an independent variable (X). It employs a regression line, also known as a best-fit line. The linear connection is defined as $Y = c + m \cdot X + e$, where 'c' denotes the intercept, 'm' denotes the slope of the line, and 'e' is the error term.

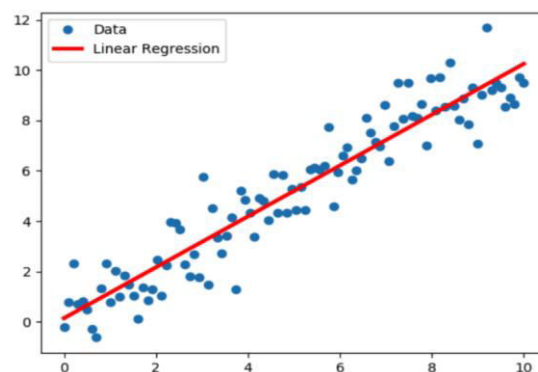


Fig. 3.1.1 Linear Regression

3.1.2 Lasso Regression

As with ridge regression, the lasso (Least Absolute Shrinkage and Selection Operator) technique penalizes the absolute magnitude of the regression coefficient. Additionally, the lasso regression technique employs variable selection, which leads to the shrinkage of coefficient values to absolute zero.

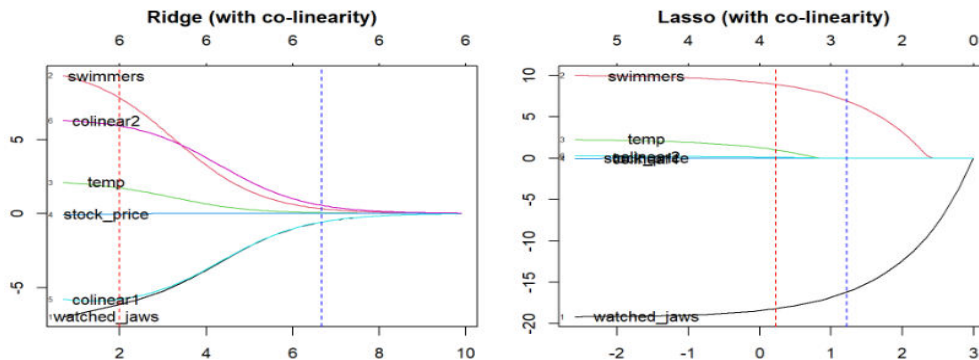


Fig 3.1.2. Lasso Regression

3.1.3 Quantile Regression

The quantile regression approach is a subset of the linear regression technique. It is employed when the linear regression requirements are not met or when the data contains outliers. In statistics and econometrics, quantile regression is used.

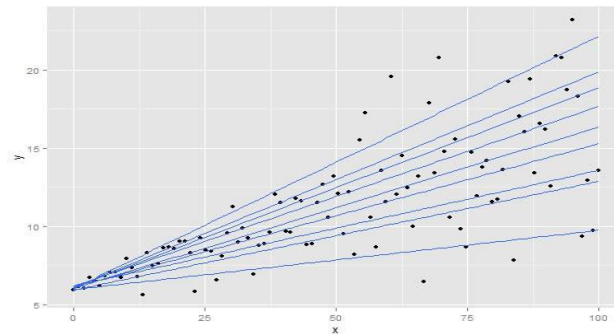


Fig 3.1.3 Quantile Regression

3.1.4 Elastic Net Regression

Elastic net regression combines ridge and lasso regression techniques that are particularly useful when dealing with strongly correlated data. It regularizes regression models by utilizing the penalties associated with the ridge and lasso regression methods.

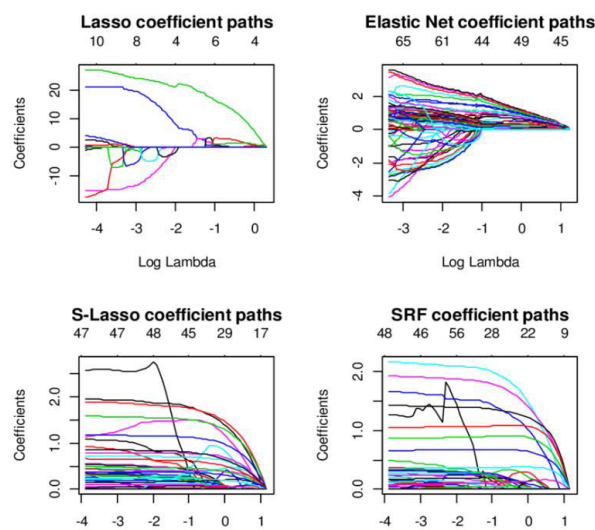


Fig 3.1.4 Elastic Net Regression

3.1.5 Poisson Regression

In statistics, Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the

logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model, especially when used to model contingency tables.

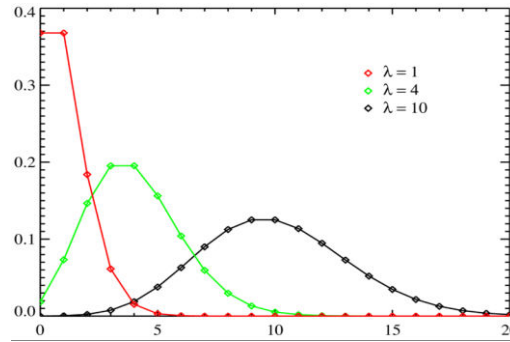


Fig 3.1.5 Poisson Regression

3.2 Mathematical Representations

$$Y=a+bX$$

$$a = \frac{(\sum Y) (\sum X^2) - (\sum X) (\sum XY)}{n (\sum x^2) - (\sum x)^2}$$

$$b = \frac{n (\sum XY) - (\sum X) (\sum Y)}{n (\sum x^2) - (\sum x)^2}$$

3.3 Regression Steps

The linear regression model can be simple (with only one dependent and one independent variable) or complex (with numerous dependent and independent variables) (with one dependent variable and more than one independent variable).

3.3.1 Data Collection:

Collect data on a relevant topic of your choice. Ensure that the data includes a dependent variable and one or more independent variables. It is also important to ensure that the data is representative and contains enough observations.

3.3.2 Data Pre-processing:

Clean and pre-process the data to remove missing values, outliers, and other errors that could affect the accuracy of the regression models. This may include data transformations such as scaling, centering, or normalization.

3.3.3 Model Selection:

Select the five regression models to be compared: Linear Regression, Lasso Regression, Quantile Regression, Elastic Net Regression, and Poisson Regression.

3.3.4 Model Fitting:

Fit each of the selected regression models to the pre-processed data using a suitable algorithm. For Linear Regression, the model is fitted using the ordinary least squares (OLS) method. For Lasso Regression, the model is fitted using the L1 regularization method. For Quantile Regression, the model is fitted using the quantile regression algorithm. For Elastic Net Regression, the model is fitted using a combination of L1 and L2 regularization methods. For Poisson Regression, the model is fitted using the maximum likelihood estimation method.

3.3.5 Model Evaluation:

Evaluate the performance of each model based on the chosen evaluation metric(s). Common metrics include mean squared error, R-squared, root mean squared error, and mean absolute error. You could also consider other metrics based on your research question.

3.3.6 Model Comparison:

Compare the performance of each model using statistical tests such as the F-test or cross-validation. This will help you determine which model performs best in predicting the dependent variable.

3.3.7 Results and Interpretation:

Present the results of the analysis and interpret the findings in the context of your research question. Highlight the strengths and limitations of each model, and provide recommendations for future research.

Overall, it is important to ensure that the methodology is transparent, and the analysis is reproducible. This can be achieved by providing detailed documentation and making the code and data publicly available.

IV. RESULTS

The dataset used in the study contains information on R&D, administrative, and marketing costs for 50 start-ups. This dataset is likely a panel dataset, where each observation represents a company over time.

The variable R&D spending measures the amount each firm invests in research and development activities. This may include costs related to product development, scientific research and innovation. Control variables measure the amount each company spends on overhead costs such as rent, salaries, and utilities. The marketing spending variable measures the amount each company spends on marketing and promotional activities such as advertising campaigns, sponsorships and media placements.

Using this dataset, we can explore the relationship between these three independent and dependent variables. A dependent variable can be an indicator of company performance, such as profit or sales. By analysing this relationship, you can identify the variables that have the greatest impact on performance and determine how your company can optimize spending to improve financial results.

	R&D Spend	Administration	Marketing Spend	Profit
count	50.000000	50.000000	50.000000	50.000000
mean	73721.615600	121344.639600	211025.097800	112012.639200
std	45902.256482	28017.802755	122290.310726	40306.180338
min	0.000000	51283.140000	0.000000	14681.400000
25%	39936.370000	103730.875000	129300.132500	90138.902500
50%	73051.080000	122699.795000	212716.240000	107978.190000
75%	101602.800000	144842.180000	299469.085000	139765.977500
max	165349.200000	182645.560000	471784.100000	192261.830000

Fig 4.1. Dataset Description

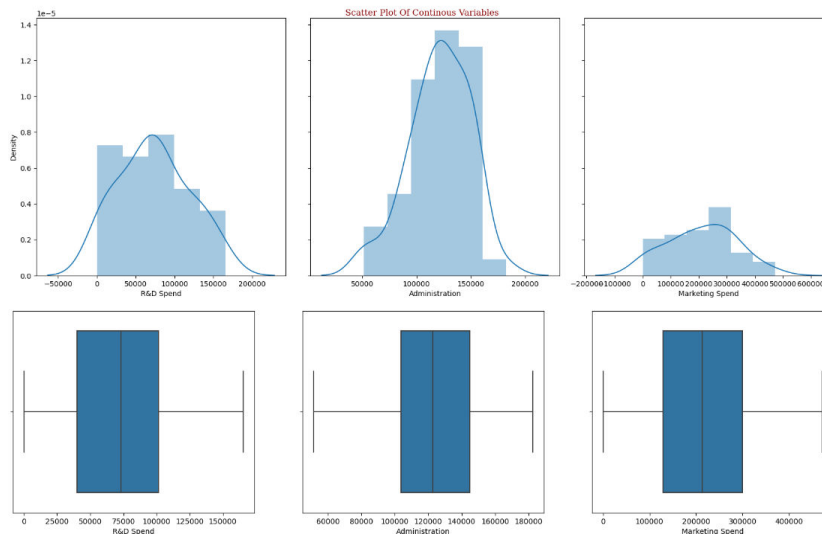


Fig 4.2 Scatter Plot of Continuous Variables

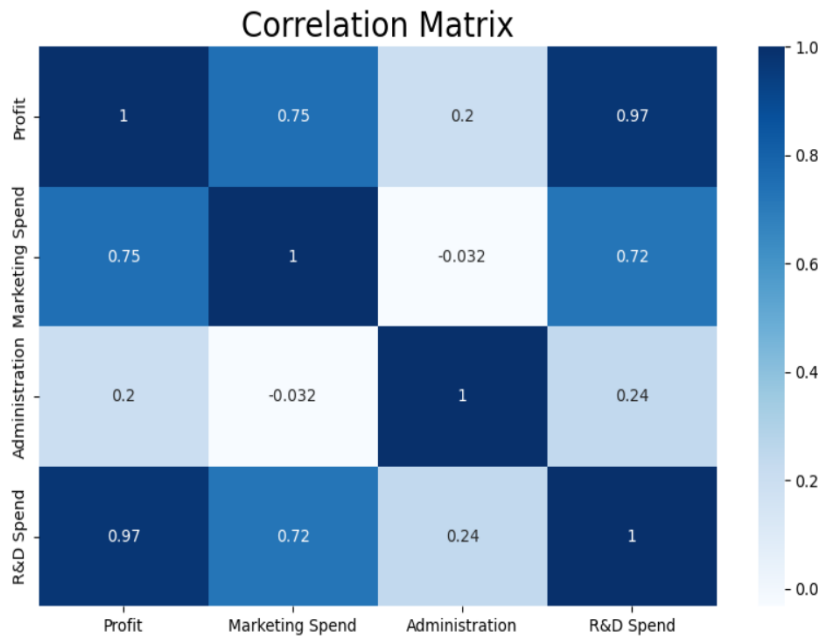


Fig 4.3 Correlation Matrix

The following were the results obtained from Regression Performed on the Dataset.

LINEAR REGRESSION

TRAIN:
 MSE : 77960288.462268
 RMSE : 8829.512356991636
 R2 Score : 0.9546762374572824

TEST:
 MSE : 100604263.51952028
 RMSE : 10030.16767155566
 R2 Score : 0.9028338938059199

ELASTIC NET REGRESSION

TRAIN:
 MSE : 77960288.46226797
 RMSE : 8829.512356991634
 R2 Score : 0.9546762374572824

TEST:
 MSE : 100604263.45499416
 RMSE : 10030.167668339058
 R2 Score : 0.9028338938682409

LASSO REGRESSION

TRAIN:
 MSE : 77960288.46226795
 RMSE : 8829.512356991634
 R2 Score : 0.9546762374572824

TEST:
 MSE : 100604263.66351916
 RMSE : 10030.167678733947
 R2 Score : 0.9028338936668422

POISSON REGRESSION

TRAIN:
 MSE : 77960288.46226797
 RMSE : 8829.512356991634
 R2 Score : 0.9546762374572824

TEST:
 MSE : 100604263.45499416
 RMSE : 10030.167668339058
 R2 Score : 0.9028338938682409

QUANTILE REGRESSION

TRAIN:
 MSE : 83481264.39562127
 RMSE : 9136.808217075659
 R2 Score : 0.9514665084126229

TEST:
 MSE : 100581998.95983914
 RMSE : 10029.057730407136
 R2 Score : 0.902855397472809

Poisson Regression and Lasso Regression performed well in predicting the dependent variable based on the independent variables, while Linear Regression performed moderately. Quantile Regression was the least effective at predicting the dependent variable of the dataset.

V. CONCLUSION AND FUTURE WORK

In summary, our investigation aimed to conduct a comparative analysis of five different regression algorithms - Linear Regression, Lasso Regression, Quantile Regression, Elastic Net Regression, Poisson Regression. Using a dataset containing information on R&D, administrative, and marketing expenses for 50 startups, we examined the relationship between these variables and company performance.

After running our analysis, we found that each regression model has its own strengths and weaknesses. Poisson Regression and Lasso Regression performed well in predicting the dependent variable based on the independent

variables, while Linear Regression performed moderately. Quantile Regression was the least effective at predicting the dependent variable of the dataset. We also found that R&D spending and marketing spending were the most important independent variables in predicting company performance and were weakly influenced by management. These results provide insight into how companies can optimize spending to improve financial results and help researchers and practitioners make informed decisions about the regression models they use in their research and practice. helps you get down.

In summary, our work has contributed to the field of regression analysis by comparing the performance of different regression models on real datasets. We hope that this study will stimulate further research in this area and help practitioners make more accurate predictions about their work.

REFERENCES

- [1] Apostolos Nicholas Refenes, Achileas Zapranis, Gavin Francis (1994). Stock performance modeling using neural networks: A comparative study with regression models
- [2] Kavitha S, Varuna S and Ramya R, "A comparative analysis on linear regression and support vector regression," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), Coimbatore, 2016, pp. 1-5, doi: 10.1109/GET.2016.7916627.
- [3] Carlos Bastida, José Holguín-Veras (2009). Freight Generation Models: Comparative Analysis of Regression Models and Multiple Classification Analysis
- [4] James W Wisnowski, Douglas C Montgomery, James R Simpson (2001). A Comparative analysis of multiple outlier detection procedures in the linear regression model
- [5] Soner Cankaya,G. Tamer Kayaalp,Levent Sangun,Yalcin Tahtali &Mustafa Akar(2006). A Comparative Study of Estimation Methods for Parameters in Multiple Linear Regression Model



INNO  **SPACE**
SJIF Scientific Journal Impact Factor
Impact Factor: 8.379



ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 **9940 572 462**  **6381 907 438**  **ijircce@gmail.com**



www.ijircce.com

Scan to save the contact details