Migration Letters

Volume: 20, No: 6, pp. 139-153

ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)
www.migrationletters.com

Optimizing Linear Regression Models with Lasso and Ridge Regression: A Study on UAE Financial Behavior during COVID-19

Samir K. Safi[1], Mouza Alsheryani[2], Maitha Alrashdi[3], Rawan Suleiman[4], Dania Awwad[5], Zainab Nasr Abdalla[6]

Abstract

This study examines the way individuals in the United Arab Emirates handled their finances during the COVID-19 pandemic, with a focus on optimizing linear regression models utilizing Lasso and Ridge Regression methods. Using the World Bank's enormous Global Findex dataset, we navigated the obstacles of multicollinearity and missing values. Our investigation showed that specific economic indicators and demographic characteristics, including gender and educational attainment, are crucial in affecting financial choices during crises. Our study carefully analyzed the performance of Lasso and Ridge Regression in the context of the Global Findex dataset. Ridge Regression was outperformed by Lasso, which is renowned for its feature selection and complexity reduction abilities. This research provides a comprehensive knowledge of the way advanced regression approaches might improve predictive modeling in such settings, shedding light on the complex dynamics of financial activity during a worldwide crisis.

Keywords: *Optimization, Ridge, Lasso, Predictive Modeling, COVID-19, UAE.*

1. INTRODUCTION

In multiple regression analysis, the presence of multicollinearity, which indicates the linear relationships among independent variables, can pose challenges in interpreting the results. However, knowledgeable researchers can overcome these challenges by employing various techniques to investigate the contributions of predictors to the regression model and their interrelationships. These techniques include correlation coefficients, beta weights, structure coefficients, all possible subsets regression, commonality coefficients, dominance weights, and relative importance weights. By utilizing multiple indices, researchers can gain a comprehensive understanding of the effects of predictors in the regression model, as well as their relationships with each other. Statistical software can provide support for implementing these analyses.

[1] Department of Statistics and Business Analytics, CBE, United Arab Emirates University, United Arab Emirates, ORCID Number: https://orcid.org/0000-0003-4385-6047
[2] Department of Accounting with a minor of Statistics and Data Analysis, CBE, United Arab Emirates University, United Arab Emirates, Email: 202007497@uaeu.ac.ae
[3] Department of Department of Statistics and Business Analytics, CBE, United Arab Emirates University, United Arab Emirates, Email: 202212623@uaeu.ac.ae
[4] Department of Department of Statistics and Business Analytics, CBE, United Arab Emirates University, United Arab Emirates, Email: 700037141@uaeu.ac.ae
[5] Department of Department of Statistics and Business Analytics, CBE, United Arab Emirates University, United Arab Emirates, Email: 700036873@uaeu.ac.ae
[6] Department of Department of Statistics and Business Analytics, CBE, United Arab Emirates University, United Arab Emirates, Email: 700041184@uaeu.ac.ae

Multiple regression analysis is widely used in various fields such as business administration, economics, engineering, and the social, health, and biological sciences. It aims to model the relationship between a dependent variable and multiple independent variables. However, multicollinearity can arise when the independent variables exhibit linear relationships among themselves. Collinearity occurs when two variables closely resemble perfect linear combinations of each other. Multicollinearity occurs when the regression model includes several variables that are significantly correlated not only with the dependent variable but also with each other.

Multicollinearity can lead to skewed or misleading results when attempting to determine the optimal utilization of each factor in predicting or understanding the response variable in a statistical model. It can also result in wider confidence intervals and less reliable likelihood values for the predictors, rendering the findings from a model with multicollinearity less trustworthy. Additionally, multicollinearity increases the standard errors of each coefficient in the model, altering the analysis results and potentially causing previously significant variables to become statistically insignificant. The increased variance of the regression coefficients due to multicollinearity further complicates the interpretation of the coefficients. Numerous studies have highlighted the problems associated with multicollinearity in regression models, including biased and uneven standard errors and impractical explanations of the results. To address these issues, researchers are encouraged to employ advanced regression procedures such as principal components regression, weighted regression, and ridge regression methods to detect multicollinearity.

The choice between Lasso and Ridge regression depends on the specific characteristics of the problem and the desired properties of the model. Lasso performs well when there are numerous features, but only a small fraction of them is truly important. In contrast, ridge regression is effective in the presence of multicollinearity and when all features impact the model's performance. Both approaches offer robust frameworks for enhancing model performance and interpretability, enabling researchers to design reliable and efficient regression models for a wide range of applications. By carefully considering the advantages and disadvantages of each technique, researchers can select the most appropriate regularization method for their specific regression analysis.

So, in summary, Multicollinearity in multiple regression analysis poses challenges in interpretation. Researchers can overcome this by using techniques like correlation coefficients, beta weights, and structure coefficients to understand predictor contributions and relationships. Multicollinearity can lead to misleading results, wider confidence intervals, and less reliable likelihood values. It increases standard errors, potentially rendering previously significant variables insignificant. Advanced regression procedures like principal components regression, weighted regression, and ridge regression help address multicollinearity.

Lasso and Ridge regression are effective strategies to handle multicollinearity and prevent overfitting. Lasso (L1 regularization) promotes simpler models with fewer parameters, useful for multicollinearity and automated model selection. Ridge regression (L2 regularization) reduces bias and variability of estimates, particularly suitable for multicollinearity. Choosing between Lasso and Ridge regression depends on problem characteristics and desired model properties. Lasso is suited for many features with few important ones, while ridge regression is effective when all features impact performance. These techniques enhance model performance and interpretability, enabling researchers to design reliable regression models for various applications.

The rest of this article is organized as follows. Section 2 presents a brief overview of the existing literature. The data and methodology are outlined in Section 3. The data analysis of both Ridge and Lasso Regression models is discussed in Section 4, and the final section concludes by summarizing the fundamental results

2. LITERATURE

Regularization techniques, such as ridge, lasso, and their combination elastic net, are widely used in the data science and machine learning fields to improve upon the standard linear regression models. This review analyzes the use of traditional least squares methods, regularization techniques, and other hybrid methods to highlight regularization techniques' usage cases, capabilities, and limitations.

Ordinary least squares (OLS) is an old and commonly used technique in regression. However, according to (Kan et al., 2019) research, OLS performs worse when compared to other regularization techniques, and the research suggests implementing simple regularization models to create more effective risk adjustment models. Nevertheless, least square regression methods can aid regularization techniques by processing data and identifying relevant factors. For example, using PLS in a study on the quality of edible oils improved the marginal prediction errors and helped address collinearity in the model (Gilbraith et al., 2021).

(Schmidt, 2005) suggests that due to the need for precise prediction models with understandable or minimal representations, L1 regularization, when combined with the Least Squares objective function, has attracted substantial interest across a variety of areas. Techniques for parameter estimation include optimizing loss functions that are subject to the L1 penalty, choosing appropriate regularization parameter values, and dealing with orthogonal design matrix optimization. The study by (Schmidt et al., 2007) is supported by the paper. Due to its non-differentiability, L1 regularization presents optimization issues; however, they analyzed state-of-the-art optimization techniques to handle this problem. Numerous comparisons reveal that these methods regularly rank highly for efficiency and convergence speed as measured by the number of function evaluations needed. The comparison of optimization techniques to handle the issues brought on by the non-differentiability of L1 regularization is probably covered in both sources.

Regression and classification approaches are used in supervised learning, where the main objectives are to reduce empirical risk and, through regularization, address overfitting. New regularization techniques are proposed by (Shafieezadeh-Abadeh et al., 2019) based on notions for distributionally resilient optimization. On the other hand, despite issues with model divergence during training, Generative Adversarial Networks (GANs) have attracted substantial attention for applications like creating synthetic samples. Recent studies, such as (Shafieezadeh-Abadeh et al., 2019; Lee & Seok, 2020) concentrate on creating regularization methods to stabilize GAN training. The performance and stability of GAN models have improved as a result of these developments.

Regression models' stability and clarity can be negatively impacted by multicollinearity or the strong correlation across predictor variables. Ridge regression is particularly effective in addressing multicollinearity problems. Ridge regression mitigates the effects of multicollinearity and regulates the estimated coefficients by introducing a degree of bias to the regression estimates of the regression model (Schreiber-Gregory, 2018). He discussed the importance of Ridge regression in controlling multicollinearity and provided insightful explanations of its function in enhancing the accuracy of regression analysis. In addition, (Mohammadi, 2020) suggested a fresh approach to determining multicollinearity's negative effects. The significance levels of the coefficient estimations from the generalized ridge regression and traditional least squares are compared using this method. It classifies the level that results in false statistical inference as detrimental multicollinearity. Even with tiny sample sizes, the proposed test exhibits significant power and is effective in locating harmful multicollinearity. Considering the combined results, it is important to recognize and address multicollinearity in regression analysis, and Ridge Regression, and the suggested method may be helpful in reducing its effects.

(Melkumova & Shatskikh, 2017) performed different regression techniques on the wine quality data to compare the residual sum of squares of these techniques. The results showed that the smallest residual sum of squares in the training data was obtained by ordinary least square regression. However, in the testing data, the LASSO and ridge regression had a smaller residual sum of squares than the ordinary least square regression. Additionally, the statistical analysis performed on the wine quality data showed that as lambda increases, both ridge and LASSO shrink to zero, and thus reduce the variance. However, unlike ridge regression, the LASSO regression is easier to interpret as it carries out a variable selection, which produces zero coefficient estimates for some variables.

(Sirimongkolkasem & Drikvandi, 2019) investigated regularized methods, including Lasso, Elastic Net, and Ridge regression, for analyzing high-dimensional data. It explores the effects of data correlation, covariate location, and effect size on prediction, parameter estimation, and variable selection. The study finds that correlated data improves the performance of common regularized methods, with Elastic Net showing better variable selection in the presence of correlation. The de-biased Lasso performs well in low-dimensional data but faces issues like multicollinearity and multiple hypothesis testing. (Gao et al., 2020) focused on high-dimensional regression modeling and compared Ridge regression and Lasso regression on financial data. Lasso regression is found to be effective in big data modeling and coefficient compression, outperforming Ridge regression in terms of cross-validation mean square error and interpretability of the equations. The study uses fiscal revenue-related predictors and emphasizes the importance of fiscal revenue in evaluating China's economic development.

Regression models and regularization techniques are essential in statistical learning and have been studied in different contexts. (Salehi et al., 2019) introduced regularized logistic regression (RLR), which incorporates a convex regularizer to encourage structured parameter vectors. They provided a detailed analysis of RLR's performance, including explicit expressions for various metrics and optimization of the regularized parameter. On the other hand, the optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization is a paper that challenges the conventional wisdom of strong regularization for preventing overfitting in large models. It focuses on underdetermined linear regression situations, where predictors outnumber observations. Surprisingly, the paper shows that explicit positive ridge penalties may not improve the model and suggests that negative ridge penalties can be optimal in such cases. It highlights how low-variance directions implicitly contribute to regularization in high-dimensional data. By combining these perspectives, we gain a more comprehensive understanding of regularization in regression models. While (Salehi et al., 2019) emphasized the benefits of regularization in logistic regression, particularly in the high-dimensional regime, (Kobak et al., 2020) provided insights into cases where excessive regularization may hinder model performance, urging us to consider the variance and structure of predictors.

Expanding on this topic, (Sarma and Barma, 2019) proposed a method using LASSO regression to perform variable selection to analyze emotions ignoring the fact that emotions are extremely individualized. The extracted relevant features using the power spectral density from signals were subjected to LASSO to pick the best feature set for every subject. The reduced features acquired a high accuracy classification, which proves the efficiency of the proposed technique. Similarly, Yang et al., (2022) proposed a technique to estimate the parameters of the predistortion model by implementing LASSO regression. This study incorporated LASSO regression for the same reason as (Sarma and Barma, 2019), because of its ability to extract significant variables to reduce the complexity of the model by shrinking some variables' coefficients to zero. This method proved its success as it resulted in a 78% reduction of the model coefficients, which leaves only 28 out of 125.

Furthermore, a study conducted by (Neba et al., 2023) explored the utilization of regularized generalized linear models (GLMs), including Lasso regression, Ridge regression, and Elastic Net regression, for credit card fraud detection. The study compared the performance of these techniques on a dataset of credit card transactions and found that Ridge regression achieved the highest accuracy of 98%, followed by Lasso regression (93.2%) and Elastic Net (93.1%). This research demonstrated the effectiveness of regularized GLMs as valuable tools for fraud detection in high-dimensional data. (Tkachenko et al., 2021) simulated the average weight of Roman snails, focusing on two models, ridge and LASSO regression. It was proved that the ridge model showed the best results in modeling the average weight of Roman snails because it considered all quantitative and qualitative factors affecting the snails, which led to a significant error reduction, unlike in the LASSO model.

Additionally, a study on optimizing regularization in DRT illustrates that lasso helps determine relevant peaks in the graphical analysis (Saccoccio et al., 2014). Lasso can also assist variation regularization, a technique effective in image diagnosis and retaining critical structural information, with results showing that a L1 and total variation Quasi-Network Flow model was promising in creating desirable road profiles and reducing the waviness of vertical road alignment in case studies on existing roads (Iannantuono et al., 2023). Additionally, (Gilbraith et al., 2021) find that a lasso model performs slightly better than a ridge and is more favorable in the elastic net model than a ridge. (Kan et al., 2019) found similar results as lasso regression performed the best when predicting risk across high and low levels. Despite the Lasso winning favor, studies demonstrated that the Ridge is an advantageous graphical smoother and, in some instances, performs comparably to the Lasso. Ultimately, using both L1 and L2 regularization techniques significantly improves the model, and combining them in elastic net regularization enhances models by acquiring more accurate and insightful results (Saccoccio et al., 2014; Gilbraith et al., 2021).

Moreover, et al., (Kumar et al., 2019) investigated the relationship between adaptive Lasso and generalized Ridge estimators. Their research revealed that, in certain situations, the estimators derived from both procedures were identical, indicating a common underlying process. This study shed light on the practical applications of adaptive Lasso and generalized Ridge estimators in linear regression analysis while providing insights into the theoretical connections between them.

Machine learning (ML) is a technology that gives systems the ability to learn on their own through real-world interactions and generalizing from examples without explicit programming, as in the case of rule-based programming. It plays a key role in a wide range of critical applications. In machine learning, linear regression (LR) is a fundamental technique used to obtain a linear trend. However, linear regression is just one facet of comprehensive statistical and machine-learning algorithms. Review on Linear Regression Comprehensive in Machine Learning is a paper that highlights linear regression as one of the most common algorithms used to find linear relationships between predictors, including simple regression and multiple regression (MLR) (Maulud and Abdulazeez, 2020). Researchers have extensively studied linear and polynomial regression to optimize prediction and precision, comparing their performance. Notably, the emphasis lies on evaluating model efficiency by correlating it with the actual values obtained for the explanatory variables. Another study further adds to the discussion by noting that Support Vector Machines (SVMs) provide advanced features such as high accuracy and predictability in Stock Market Prediction using Linear Regression and Support Vector Machines, making them a valuable addition to the field (Gururaj et al., 2019). In summary, these combined papers provide a comprehensive overview of linear regression's importance in machine learning, its variations, and the ongoing research focused on optimizing its performance.

Li, Y., et al., (2018) introduced a regularized graph neural network (RGNN) approach that considers the biological topology among different brain regions in Electroencephalography (EEG) signals. It models inter-channel relations using an adjacency matrix inspired by neuroscience theories and proposes two regularizes to handle cross-subject EEG variations and noisy labels. The experiments demonstrate the superior performance of the RGNN model compared to state-of-the-art models. (Zhong et al., 2022) presented a novel regression model called graph regularized sparse linear regression (GRSLR) for EEG emotion recognition. GRSLR extends conventional linear regression by incorporating graph regularization and sparse regularization to learn a transform matrix that preserves the intrinsic manifold of the data samples. The proposed algorithm is evaluated on two databases, SEED and RCLS, showing superiority over classic baselines (Li et al., 2018). Additionally, the authors make the RCLS database publicly available for further research.

By combining these papers, we introduce a comprehensive framework for EEG-based emotion recognition that leverages the topology of EEG channels and incorporates regularization techniques. The proposed methods, RGNN and GRSLR, highlight the importance of capturing inter-channel relations and preserving intrinsic data structures. The experiments conducted on multiple datasets validate the effectiveness of both approaches, demonstrating improved performance compared to existing methods. The availability of the RCLS database further encourages future research in the field of emotion recognition.

Recently, new methods have been developed to capitalize on penalization capabilities and improve outcomes by combining them with other machine learning models. For instance, researchers developed a ranking-based regularization method when dealing with critical rare classes. The penalization techniques work in conjunction with neural networks to reduce the imbalance and bias of the data in machine learning associated with critical rare classes by increasing caution within the network in labeling a false positive. The approach consistently improved the models and outperformed other techniques implemented (Kiarash et al., 2023).

The above-mentioned studies collectively emphasize the importance and effectiveness of regularized regression methods such as Ridge regression and Lasso regression in addressing multicollinearity, enhancing prediction precision, performing variable selection, and improving risk assessment and fraud detection models. They also provide insights into the practical applications and theoretical connections between these methods in various fields of research.

Although the researched literature emphasizes the advantages and efficacy of regularized regression techniques like Ridge regression and Lasso regression, it is crucial to recognize their limits and consider possible areas for additional study. Regularization techniques are significant and effective in varying fields. Current literature investigates utilizing regularization in machine learning, specifically when dealing with high-dimensional data. Regardless, fields such as clinical medicine rarely apply penalization in medical analyses and the body of work regarding common business concepts, such as the critical rare positive, is still limited (Friedrich et al., 2023; Kiarash et al., 2023). Hence, our research explores regularization techniques to provide a comprehensive grasp of their potential in linear regression models.

The assumption of linearity between the predictors and the response variable is a drawback of regularized regression techniques. despite the fact that these techniques have frequently been successful, they could not work as well when the connection between the predictors and the response is remarkably nonlinear. To overcome this constraint, future studies may examine the use of nonlinear transformations or the use of alternative machine-learning approaches.

The effect of the regularization parameter choice on the effectiveness of the models should also be considered. Results can be strongly impacted by the regularization parameter selection, such as the alpha value in Ridge regression or the lambda value in Lasso regression. It would be helpful to conduct more research on techniques for automatically or data-driven choosing the regularization value.

Additionally, the majority of the research that was examined used regularized regression techniques in certain fields, such as banking, fraud detection, and risk assessment. Even though these applications offer insightful information, more studies should examine how well these techniques work with other domains and datasets. Understanding how regularized regression techniques may be applied across a variety of areas will help them become more widely used and practicable. We may better comprehend regularized regression methods and their wider application in other disciplines by resolving these constraints and examining the possible topics for future study.

3. METHODOLOGY

The World Bank provides the Global Findex dataset, which includes precise information on the UAE's degree of financial availability and accessibility through surveying one thousand adults' monetary behavior and financial decisions. The dataset contains, on an individualistic level, eighty-four econometric and demographic variables. Additionally, data weighing is employed to ensure a more accurate representation of the UAE's economy by accounting for household size, data collection errors, and several demographic factors (United Arab Emirates, 2023). Ultimately, the dataset provides an extensive account of the financial services used in the COVID-19 pandemic.

The original dataset contained 84 variables and 1000 observations. However, the dataset was very messy and had a lot of missing values, 36.6% of the data was missing. In order to clean the dataset, some variables were eliminated to facilitate the process of cleaning. The dataset ended up with 18 variables out of 84. R software was used to clean the dataset.

We fit the data to a traditional linear regression model using the “lm” function in R. The model analyzes the relationship between the dependent variables (e.g., financial behavior or decisions) and the independent variables (e.g., demographic factors). Even though the model might suffer from overfitting and other difficulties when dealing with the possible presence of non-linear relationships in the data, the least-squares model functions as a starting point to identify the relationships in the dataset.

Lasso and Ridge Regression are usually applied when multicollinearity is present. The function glmnet()from the glmnet package is used to fit the lasso and ridge regression. The value of alpha determines whether we are fitting ridge or lasso regression. Alpha equals to zero corresponds to using ridge regression, while setting alpha to one corresponds to implementing lasso regression.

For the selection of the appropriate regularization parameters, we employ cross-validation, where the model is trained using the training dataset. Then, the model’s performance is assessed, and the best regularization parameter is selected using the validation data subset and established performance metrics.

The regression model’s prediction was compared with the actual data to evaluate its performance. The comparison showed that the model predictions are very close to the actual response values. However, some outliers were spotted. The model overestimated and underestimated some values. Similarly, when comparing the lasso model with the actual data, only a few data points were underestimated or overestimated.

4. DATA ANALYSIS

In this section, we explore the interpretative side of our study and explain how Ridge and Lasso Regression performed on our massive dataset. Here, we analyze the outcomes of our results, giving a detailed analysis of the influence of predictor variables, the effect of regularization, and the performance differences between the two regression techniques.

We aim to methodically analyzing the coefficient estimates, prioritizing factors according to their importance, and illuminating the different coefficient trajectories as regularization increases. Our goal is to identify key information that helps with making informed choices when using predictive modeling. Furthermore, the prediction accuracy, residual pattern behavior, and a comparison of the performance of the Ridge and Lasso Regression models will be examined.

We started the data analysis by firstly, the data cleaning process was performed, secondly, the data was then divided into two sets: a training set with 80% and a testing set with 20%.

4.1. Ridge Trace Plot

To investigate the influence of altering lambda values on model performance, ridge regression was used. The ideal lambda (λ) that reduced the mean squared error was determined via cross-validation. We wanted to see how different predictor factors affected the response variable 'wgt' in the Ridge Regression study. We started by looking at the Ridge Trace Plot, which shows how the coefficients vary when the penalty parameter lambda changes. The graph is shown in Figure 1. The ideal lambda for minimizing cross-validation error is represented as the lowest point on the curve, specifically at log lambda value 0.09132721. This point represents a significant crossroads in our study, denoting the time at which the regularization's favorable influence on the model's bias-variance trade-off is greatest. The optimum lambda effectively fine-tunes the influence of the coefficients, preventing them from exploding uncontrolled or being too limited.

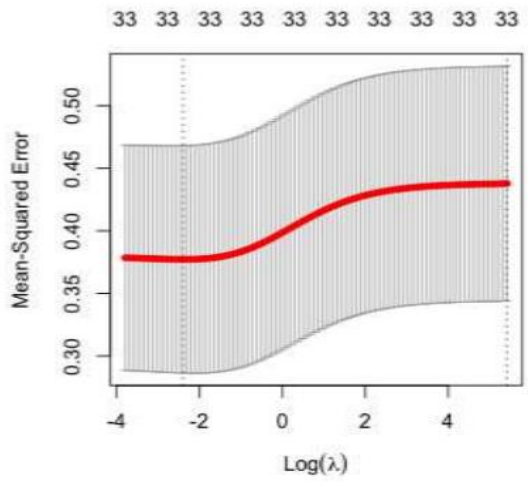

Figure 1: Ridge Trace Plot

4.2. Estimated Coefficients and Absolute Coefficient Values

The magnitudes of the coefficients were calculated and approximated. The absolute coefficient values of the variables were used to rank them. The computed coefficients denote the correlations between predictor variables and responder variables (wgt). A positive association is shown by positive coefficients, whereas a negative relationship is indicated by negative coefficients. The size of these coefficients indicates the extent to which they impact the response variable.

The absolute coefficient values were critical in the process of prioritizing predictors. The purpose of this ranking process is to reveal the most influential predictions in the ridge regression model. Specifically, the interpretation demonstrates the hierarchy of impact among predictors: “completed primary school or less” appears as the most significant, having a coefficient of 0.2829623736, closely followed by “completed secondary school” at 0.2196697912 in the subsequent position, “saved using an account at a financial institution” at 0.1946915575 in the third, “respondent is female” at 0.1883179959 in the fourth, and “has a mobile money account” at 0.1571741562 in the fifth. This order of significance underlines that predictors with bigger absolute coefficients have a stronger effect on the response variable, emphasizing their importance in determining model outputs.

4.3. Ridge Coefficients Visualization

The Ridge Regression Coefficients graphic shows how coefficients vary when lambda's logarithm changes. The vertical line at the logarithm of the optimal lambda value indicates that the coefficients have attained stability and that future lambda reductions will have no effect on them.

Several essential conclusions may be obtained from the offered graphic depiction which is shown in Figure 2. To begin, the vertical demarcation along the x-axis at -2.3 corresponds to the logarithm of the optimal lambda value for coefficient regularization. This line marks the precise location of the most advantageous lambda value on the logarithmic scale. This ideal lambda value is critical in determining the degree of regularization applied on the coefficients. A higher lambda value, in particular, produces a stronger regularization effect, causing a significant contraction of coefficients towards the zero point. Furthermore, the vertical direction of this line shows that the coefficients have reached a condition of equilibrium, indicating that future decreases in lambda are unimportant.

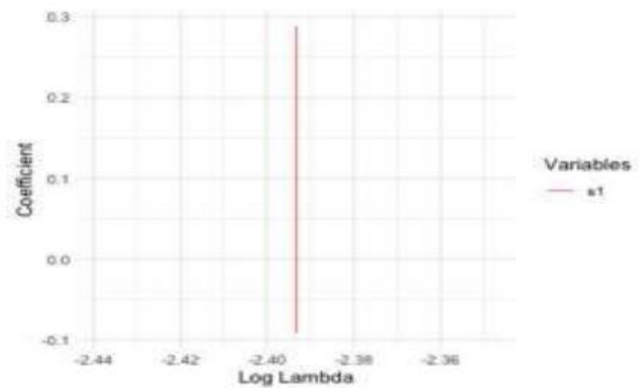

Figure 2: Ridge Regression Coefficients Plot

4.4. Lasso Regression Analysis

Lasso regression, like ridge regression, was done with various lambda values, and cross-validation assisted in determining the ideal lambda. The performance of the model was analyzed, and a residual plot was created to assess prediction precision.

The Lasso cross-validation graphic, which is shown in Figure 3, illustrates the link between the logarithm of the regularization parameter (log λ) and the performance metric mean squared error (MSE). Moving down the x-axis, the MSE decreases gradually until it reaches a turning point as the log λ grows. The MSE reaches its smallest value on the curve at log = 0.01834977, indicating the best point for balancing the trade-off between model complexity and simplicity. Beyond this point, the MSE tends to increase, demonstrating the difficult balance between model complexity and simplicity. The slope of the chart indicates the model's sensitivity to changes in log, demonstrating the

possibility for major swings in the MSE with even little modifications. The lowest point on the curve acts as a guidepost for finding the ideal log, enabling a successful fit to the training data while minimizing overfitting risk.

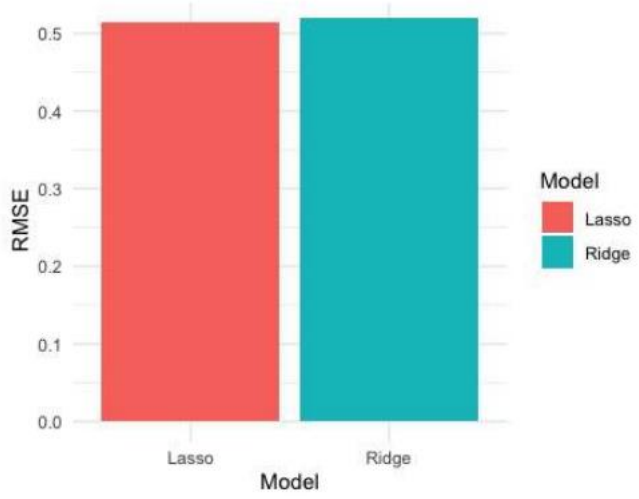

Figure 3: Model Performance Comparison Plot

4.5. Model Performance Analysis

Using a test dataset, we employed Ridge and Lasso regression models to predict conclusions. After making predictions using both models, we used the Root Mean Squared Error (RMSE) measure to assess how effectively they performed. The RMSE values generated indicate how effectively the Ridge and Lasso models predict the desired parameter on the test data. Lower RMSE values indicate that the model's predictions are more accurate since they are closer to the actual values. A graphical depiction is shown in Figure 4. Ridge had an RMSE of 0.5207947 while Lasso had an RMSE of 0.5150169. Because of the minor difference in these values, the Lasso model surpasses Ridge in producing predictions on this dataset. This suggests that Lasso's method of dealing with data features, such as selecting relevant components and lowering complexity, fits better here, which explains why its RMSE is smaller. However, it's important to note that the difference in RMSE between the two models is relatively small. This discovery demonstrates how regularization approaches such as Lasso may prevent overfitting and make models more realistic.

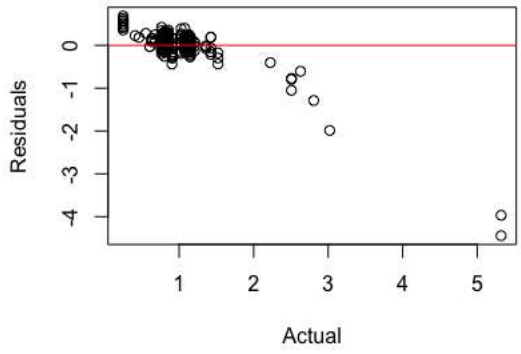

Figure 4: Ridge Regression Residual Plot

4.6. Residual Analysis

The residual analysis of both the Ridge and Lasso regression models helped us to investigate further each model's prediction precision. The variations between the actual response variable values and the expected values produced by the models are referred to as residuals.

The residual graph for the Ridge regression model, shown in Figure 5, demonstrated that the bulk of predictions were close to the actual values. This meant that the model's predictions were mostly correct. A deeper look revealed an aptitude for minor overestimating before specific groupings of data clusters. Furthermore, certain places had higher variances, especially where strong negative residuals were recorded. These discrepancies showed that the Ridge model might be unable to capture particular trends in the data at times, resulting in larger errors in prediction.

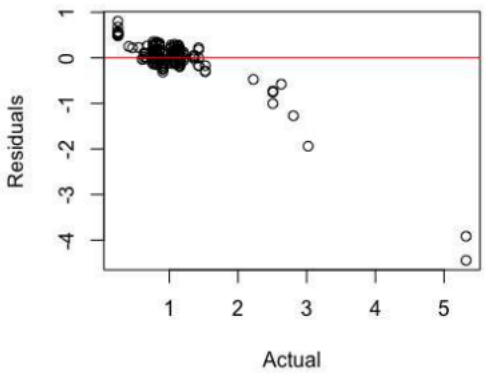

Figure 5: Lasso Regression Residual Plot

The residual plot of the Lasso regression model, shown in Figure 6, revealed similar findings. The majority of forecasts were close to the actual values, demonstrating the model's general accuracy. Similarly, there was some inaccurate estimation before certain clusters of data points, and more substantial variations occurred in locations with significant negative residuals. This behavior revealed that the Lasso model and the Ridge model had similar prediction capabilities and issues.

The occurrence of greater negative residuals in both situations may indicate that the models fail to accurately predict circumstances where the actual response values are significantly lower than anticipated. This might be due to the complexity of the data that the models are not successfully capturing, or it could be impacted by outliers impacting the projections. Despite both models performing well in terms of precision, the appearance of these patterns in the residual plots reveals areas where further model improvement or data pretreatment might possibly increase the precision of predictions.

The Lasso Regression Prediction vs. Actual plot, illustrated in Figure 6, shows a highly comparable pattern to the Ridge model. The data points closely follow the diagonal line, demonstrating the Lasso regression model's ability to reliably project response numbers. Once again, a few points deviate slightly off the diagonal line, indicating slight differences between the model's predictions and the actual values.

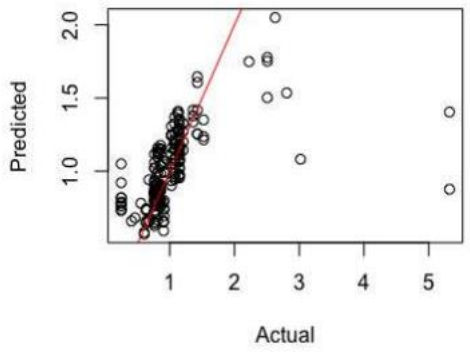

Figure 6: Lasso Regression Prediction vs. Actual Plot

4.7. Prediction vs. Actual Plot

In the Prediction vs. Actual graphs, we directly compare the regression models' predictions with the actual response values. These charts provide a clear visual assessment of how well the models' predictions match the data.

A significant number of the points in the Ridge Regression Prediction vs. Actual plot, depicted in Figure 7, observe a diagonal line with a slope of 1, reflecting the optimum case in which predictions precisely match the data in reality. This alignment shows that the Ridge regression model's predictions are quite close to the actual response values. It's worth noting, though, that a few points deviate slightly above and below the diagonal line. These deviations represent cases in which the model's predictions significantly exceed or underestimate the actual values. Despite these outliers, the general trend shows that the Ridge regression model gives reliable predictions, effectively capturing the underlying patterns in the data.

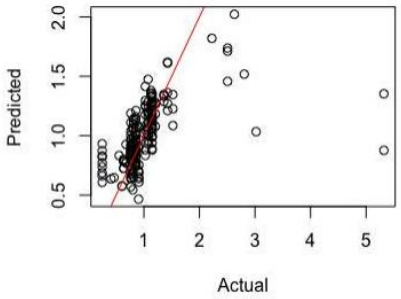

Figure 7: Ridge Regression Prediction vs. Actual plot

4.8. Coefficient Path Plot

The Ridge and Lasso regression coefficient pathways show how a regression model's coefficients vary when the regularization parameter (λ) is raised. The regularization parameter governs how much shrinkage is performed to the coefficients. The coefficients shrink towards zero as λ rises.

The coefficients start off big, but as λ grows, they shrink towards zero. Because some of the coefficients are finally set to zero, the Lasso regression is referred to as a sparse model. This indicates that the Lasso regression only considers a subset of the characteristics to be significant predictors. This implies that these characteristics are the most important weight predictors in the dataset. The Lasso regression model, presented in Figure 8, indicates that completed secondary school, completed primary school or less, the respondent is female, saved using an account at a financial institution, and bought something online using the Internet are the most influential predictors of weight, in that order are the most influential predictors of weight, in that order.

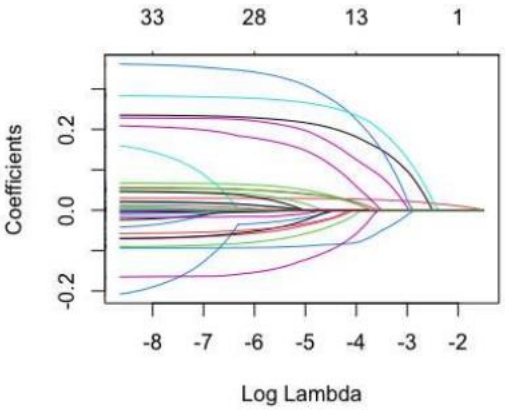

Figure 8: Lasso Regression Coefficient Path Plot

On the other hand, in the Ridge coefficient path plot, presented in Figure 9. The coefficients start off large, but as λ grows, they shrink towards zero. The coefficients, however, never approach zero, which is why the Ridge regression is referred to as a non-sparse model. This indicates that the Ridge regression incorporates all of the model's characteristics, but the coefficients are reduced to zero to decrease the variance of the estimations. The most influential predictors in the Ridge regression model are completed primary school or less, completed secondary school, saved using an account at a financial institution, the respondent is female, and has a mobile money account, in that order.

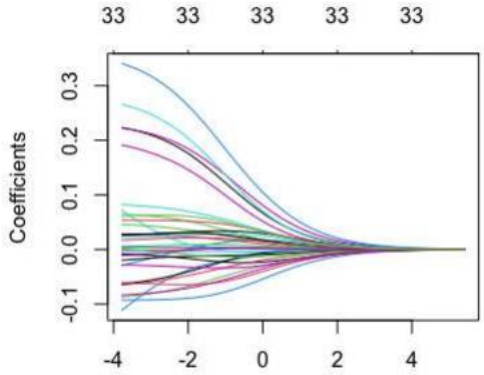

Figure 9: Ridge Regression Coefficient Path Plot

5. CONCLUSION

As a whole, our research comprises a thorough investigation into how people in the United Arab Emirates handled their finances during the difficult time of the COVID-19 outbreak. We sought to maximize our grasp of the complex dynamics influencing financial decisions in this precise setting by utilizing the effectiveness of linear regression models, particularly Lasso and Ridge Regression.

We used a methodical approach to data preparation, variable selection, and the use of advanced regression techniques. We were successful in identifying substantial demographic influences on financial behavior during the epidemic, including education levels, gender, and particular financial variables.

We discovered through model performance analysis that the Lasso and Ridge Regression models both produced insightful predictions of financial behavior. However, the slight advantage Lasso had in terms of RMSE revealed that this dataset was a good fit for its feature selection skills. This highlights the significance of using the appropriate modeling tools when analyzing financial behavior in the context of major economic shocks such as the COVID-19 crisis.

Our examination of the residuals showed subtleties in the models' precision. Although both models had good overall performance, the existence of outliers and variances in particular locations revealed areas in which additional study or data pretreatment could improve predicted accuracy. These results shed light on the limitations of the model and suggest possible directions for development.

In light of the aforementioned, this study advances our understanding of financial behavior under unusual conditions and emphasizes the significance of advanced modeling techniques, such as Lasso and Ridge Regression, in enhancing our comprehension of these intricate dynamics.

6. ACKNOWLEDGEMENTS

Funding for this paper was provided by the Summer Undergraduate Research Experience (SURE) PLUS Grant 2023 (SURE 2023) at United Arab Emirates University.

Declarations

The authors declare that all works are original and this manuscript has not been published in any other journal.

Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Data availability statement: The data supporting the findings of this study are available upon request. Interested researchers may contact the corresponding author to obtain access to the data for further analysis and validation.

References

Friedrich, S., Groll, A., Ickstadt, K., Kneib, T., Pauly, M., Rahnenführer, J., & Friede, T. (2023). Regularization approaches in clinical biostatistics: A review of methods and their applications. Statistical Methods in Medical Research, 32(2), 425-440.

Gao, L., Ding, Y., & Zhang, L. (2020). High dimensional regression coefficient compression model and its application. In Journal of Physics: Conference Series (Vol. 1437, No. 1, p. 012119). IOP Publishing.

Gilbraith, W. E., Carter, J. C., Adams, K. L., Booksh, K. S., & Ottaway, J. M. (2021). Improving prediction of peroxide value of edible oils using regularized regression models. Molecules, 26(23), 7281.

Gururaj, V., Shriya, V. R., & Ashwini, K. (2019). Stock market prediction using linear regression and support vector machines. Int J Appl Eng Res, 14(8), 1931-1934.

Iannantuono, A., Hare, W., & Lucet, Y. (2023). Optimization with regularization to create sensible vertical alignments in road design. Decision Analytics Journal, 6, 100183.

Kan, H. J., Kharrazi, H., Chang, H. Y., Bodycombe, D., Lemke, K., & Weiner, J. P. (2019). Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. PloS one, 14(3), e0213258.

Kobak, D., Lomond, J., & Sanchez, B. (2020). The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. The Journal of Machine Learning Research, 21(1), 6863-6878.

Kumar, D. (2019). Ridge Regression and Lasso Estimators for Data Analysis.

Lee, M., & Seok, J. (2020). Regularization methods for generative adversarial networks: An overview of recent studies. arXiv preprint arXiv:2005.09165.

Li, Y., Zheng, W., Cui, Z., Zong, Y., & Ge, S. (2019). EEG emotion recognition based on graph regularized sparse linear regression. Neural Processing Letters, 49, 555-571.

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), 140-147.

Melkumova, L. E., & Shatskikh, S. Y. (2017). Comparing Ridge and LASSO estimators for data analysis. Procedia engineering, 201, 746-755.

Kiarash, M., He, Z., Zhai, M., & Tung, F. (2023). Ranking Regularization for Critical Rare Classes: Minimizing False Positives at a High True Positive Rate. arXiv preprint arXiv:2304.00049.

Mohammadi, S. (2022). A test of harmful multicollinearity: A generalized ridge regression approach. Communications in Statistics-Theory and Methods, 51(3), 724-743.

Saccoccio, M., Wan, T. H., Chen, C., & Ciucci, F. (2014). Optimal regularization in distribution of relaxation times applied to electrochemical impedance spectroscopy: ridge and lasso regression methods-a theoretical and experimental study. Electrochimica Acta, 147, 470-482.

Salehi, F., Abbasi, E., & Hassibi, B. (2019). The impact of regularization on high-dimensional logistic regression. Advances in Neural Information Processing Systems, 32.

Sarma, P., & Barma, S. (2019, June). Emotion Analysis Based on LASSO. In 2019 IEEE Region 10 Symposium (TENSYMP) (pp. 72-77). IEEE.

Schmidt, M. (2005). Least squares optimization with L1-norm regularization. CS542B Project Report, 504, 195-221.

Schmidt, M., Fung, G., & Rosales, R. (2007). Fast optimization methods for l1 regularization: A comparative study and two new approaches. In Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18 (pp. 286-297). Springer Berlin Heidelberg.

Schreiber-Gregory, D. N. (2018). Ridge Regression and multicollinearity: An in-depth review. Model Assisted Statistics and Applications, 13(4), 359-365.

Shafieezadeh-Abadeh, S., Kuhn, D., & Esfahani, P. M. (2019). Regularization via mass transportation. Journal of Machine Learning Research, 20(103), 1-68.

Sirimongkolkasem, T., & Drikvandi, R. (2019). On regularisation methods for analysis of high dimensional data. Annals of Data Science, 6, 737-763.

Tkachenko, S. N., Tkachenko, I. A., Shpilevaya, S. G., & Dedkov, V. P. (2021, May). Modelling the production process of Roman snail using RIDGE and LASSO regression. In Journal of Physics: Conference Series (Vol. 1902, No. 1, p. 012134). IOP Publishing.

United Arab Emirates - Global Financial Inclusion (Global Findex) Database 2021. (2023, March 15). Microdata.worldbank.org. https://microdata.worldbank.org/index.php/catalog/4722/data-dictionary

Yang, X., Ren, J., Wang, X., & Song, Q. (2022, January). Reduce the dimension of the predistortion model coefficients by lasso regression. In 2022 IEEE International Conference on Consumer Electronics (ICCE) (pp. 1-3). IEEE.

Zhong, P., Wang, D., & Miao, C. (2020). EEG-based emotion recognition using regularized graph neural networks. IEEE Transactions on Affective Computing, 13(3), 1290-1301.