Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 8, No. 6, 1670-1691
2024
Publisher: Learning Gate
DOI: 10.55214/25768484.v8i6.2307

Enhancing predictive accuracy of electoral outcomes: A comparative study of dimension reduction techniques in linear regression models

Zlatan Morić[1]*
[1]Algebra University Croatia; zlatan.moric@algebra.hr (Z.M.).

Abstract: The effectiveness of Principal Component Analysis (PCA) and Factor Analysis (FA) in enhancing the predictive accuracy and interpretability of linear regression models for electoral outcomes is examined in this study, with a focus on the Croatian presidential election. Challenges such as multicollinearity and model complexity were addressed through the application of PCA and FA to socio-demographic and political variables, resulting in their transformation into a simpler, more manageable form. Multicollinearity was significantly reduced by PCA, and model stability was improved. Deeper insights into voter behavior were provided by FA through the identification of latent variables. Improvements were noted; however, issues with homoscedasticity were identified, which affected the predictive reliability of the models. The optimized linear regression model, specifically designed to address homoscedasticity (LR2), was found to demonstrate superior performance in terms of predictive accuracy and stability when compared to models that utilized PCA and FA alone. The importance of maintaining homoscedasticity in regression models is underscored by the findings, and the potential of dimension reduction techniques to improve electoral predictions is highlighted. The integration of these techniques in forecasting models is advocated for, with the aim of assisting political analysts and policymakers. It is suggested that future research could explore the combination of these methods with other advanced modeling approaches to further enhance predictive capabilities.

Keywords: *Electoral prediction FA, Factor analysis, Homoscedasticity, Linear regression, PCA, Principal component analysis.*

1. Introduction

Accurate predictions are crucial for informed decision-making in various fields, including economics, finance, social sciences, and politics. Predictive modeling plays a vital role in data science, serving as a foundation for these applications. Linear regression is a commonly employed predictive technique because of its simplicity, ease of interpretation, and relatively low computational cost. Several factors can compromise the efficacy of linear regression models, especially when dealing with high-dimensional data. Factors like multicollinearity, which occurs when predictor variables have a high correlation, can result in inflated standard errors and unreliable coefficient estimates. This, in turn, can have a detrimental impact on the predictive power of the model.

Frequently employed in addressing the challenges of high-dimensional data in regression models are dimension reduction techniques such as Principal Component Analysis (PCA) and Factor Analysis (FA). PCA is a technique that can be used to transform a set of correlated variables into a smaller set of uncorrelated components. These components are able to retain a significant amount of the variance that is present in the original dataset. The effectiveness of this method lies in its ability to address multicollinearity, which occurs when predictor variables exhibit high correlation. It achieves this by combining correlated variables into independent principal components. In a recent study on environmental factors affecting mortality, researchers successfully utilized PCA to address

* Correspondence: zlatan.moric@algebra.hr

multicollinearity. The results of the study indicated that PCA was able to generate orthogonal components, thereby enhancing the interpretability of the regression results [1].

In the realm of statistical analysis, Factor Analysis (FA) serves as a valuable tool for uncovering hidden variables, known as factors, which shed light on the underlying patterns of correlations among observable variables. Uncovering the underlying structure of complex data is a particularly valuable application of FA, especially in the field of social sciences. A study conducted by García-Santillán et al. utilized factor analysis (FA) to effectively measure latent variables. The results demonstrated the usefulness of FA in reducing data dimensionality while preserving the meaningful underlying structures [2]. Using factor analysis can uncover hidden relationships that may not be immediately obvious, which enhances the effectiveness of regression models in terms of predictive accuracy and interpretability.

These techniques are frequently combined to improve the stability and interpretability of models. Through the utilization of PCA and FA, scientists have the ability to streamline intricate data sets, diminish interference, and enhance the resilience of their models. This ultimately results in more precise and dependable predictions [3].

In order to ensure the optimal performance and reliability of linear regression models, it is crucial to satisfy the conditions outlined in the Gauss-Markov theorem. These conditions play a fundamental role beyond just dimension reduction. A fundamental principle in linear regression theory is the Gauss-Markov theorem, which states that the Ordinary Least Squares (OLS) estimator is the Best Linear Unbiased Estimator (BLUE) when specific conditions are met. These conditions include linearity, independence of errors, homoscedasticity (constant variance of errors), and the absence of perfect multicollinearity. When these assumptions are satisfied, OLS estimators exhibit the smallest variance compared to other unbiased linear estimators. This property makes them highly efficient and reliable for inference [4].

The importance of these assumptions has been extensively documented in econometric literature. In his 2010 publication, Wooldridge provides a comprehensive analysis of the Gauss-Markov theorem, emphasizing the crucial role that these assumptions play in maintaining the integrity of econometric models. When the assumption of homoscedasticity is violated, the standard errors of the OLS estimates become biased. This occurs when the variance of errors is not constant. Incorrect conclusions about the statistical significance of predictors can arise from this bias, which in turn undermines the validity of the model [5].

Another significant challenge arises from the issue of multicollinearity, which occurs when predictor variables are highly correlated. This study emphasizes the impact of multicollinearity on the variances of coefficient estimates, which can result in instability within the regression model. The instability of the model poses a challenge in determining the specific impact of each predictor variable, ultimately compromising the reliability of the model's conclusions [6].

In real-world scenarios, the assumptions that OLS relies on are frequently not met, which poses difficulties in ensuring the dependability and precision of regression models, despite their resilience under the Gauss-Markov conditions. When the assumption of homoscedasticity is violated, the standard errors may become biased, which can make the significance tests unreliable. In response to this, scientists have devised different solutions. A potential approach involves utilizing robust standard errors, which exhibit reduced sensitivity to heteroscedasticity. Consequently, this method yields more dependable inferences, even in cases where the Gauss-Markov conditions are not entirely satisfied. Groundbreaking study introduced a new approach that proved to be highly effective in addressing the issue of heteroscedasticity in a range of econometric models [7].

In OLS estimation, the stability of the estimates can be greatly affected by multicollinearity, which occurs when predictor variables are strongly correlated. The technique of ridge regression, first introduced in 1970, is widely recognized in the field for its ability to address the issue of multicollinearity. By introducing a controlled level of bias into the regression estimates, ridge regression effectively reduces variance and enhances the overall stability of the model [8].

In addition, researchers have acknowledged the effectiveness of incorporating dimension reduction techniques like Principal Component Analysis (PCA) and Factor Analysis (FA) into regression modeling as a proactive strategy for addressing multicollinearity. These techniques simplify the model by reducing the dimensionality of the predictor space, while still preserving the essential structure of the data. Hastie et al. explored the impact of these methods on enhancing the interpretability of models and improving their predictive performance, especially when dealing with intricate, multivariate data [9].

This study adds to the current discussion by analyzing how PCA and FA affect the predictive accuracy of linear regression models, specifically using Croatian presidential election data. It is hypothesized that superior predictive performance can be achieved through factor analysis, which effectively captures latent variables that reflect underlying voter behavior. This approach closely adheres to the Gauss-Markov conditions.

The methodology employed in the following sections is detailed, including data collection, preprocessing, and the application of PCA and FA. The results are presented and discussed in the context of the existing literature, providing valuable insights into the practical implications of dimension reduction techniques and the Gauss-Markov theorem in predictive modeling.

2. Research Goal and Methods

This research aims to evaluate the effect of dimension reduction techniques, specifically Principal Component Analysis (PCA) and Factor Analysis (FA), on the predictive accuracy of linear regression models. This study focuses on the Croatian presidential election and aims to predict voter behavior using a variety of socio-demographic and political variables. The study utilizes PCA and FA to decrease the dimensionality of the predictor variables, which helps to tackle problems like multicollinearity and enhances the stability and interpretability of the regression models. It is believed that factor analysis, by capturing latent variables, will result in better predictive performance when compared to models that utilize the original or PCA-transformed variables.

The study utilized a dataset comprising socio-demographic and political variables that were collected from various municipalities in Croatia. Population density, average income, education levels, unemployment rates, and historical voting patterns are among the variables considered. The data was obtained from official government statistics and election records, guaranteeing its accuracy and relevance for predictive modeling.

Ensuring the accuracy and reliability of any predictive modeling process relies heavily on the critical step of data preprocessing. Before conducting Principal Component Analysis (PCA) and Factor Analysis (FA), the dataset underwent various preprocessing techniques in this study. In this section, we will provide a detailed account of the preprocessing steps that were performed to prepare the data for analysis. We will also include references to relevant literature that support these steps. Prior to applying PCA and FA, the dataset underwent a series of preprocessing steps to ensure its suitability for analysis.

2.1. Missing Data Treatment

Large-scale socio-demographic studies often encounter the challenge of missing data in their datasets. Properly addressing missing data is crucial to prevent bias or compromising the integrity of the dataset. This study utilized multiple imputation methods to address missing values. The process of multiple imputation entails generating a range of plausible datasets by filling in missing values using the available data. These datasets are then combined to derive estimates and draw inferences that account for the uncertainty introduced by the imputations [10]. This method is considered superior to simpler techniques, like mean substitution, due to its ability to maintain the relationships between variables and take into account the variability in the missing data.

This study involved a series of steps for the imputation process:

- Identifying Missing Data: The researchers began by scanning the dataset for any missing values in all variables. A calculation was performed to determine the percentage of missing data for each variable. Variables that had more than 20% missing data were identified as potential candidates for exclusion. This is because having a high amount of missing data can undermine the accuracy of the imputed values.
- Imputation: In order to address missing data, we utilized multiple imputation through the Markov Chain Monte Carlo (MCMC) method. This approach assumes that the missing data is random (MAR) and was applied to variables with less than 20% missing data. This method proves to be highly valuable when dealing with intricate datasets that exhibit non-monotone missing data patterns [11].
- Validation was conducted by comparing the datasets after imputation to assess consistency across imputations. We examined summary statistics, such as means, variances, and correlations, to verify the reasonableness of the imputed values and to ensure that they did not introduce any new biases.

The study demonstrates that utilizing multiple imputation through the Markov Chain Monte Carlo method effectively addresses missing data in large-scale socio-demographic studies, ensuring the preservation of variable relationships and the integrity of statistical inferences despite the presence of incomplete data.

2.2. Normalization

Normalization is an important preprocessing step that is necessary to ensure that variables are comparable. This is particularly crucial when utilizing techniques such as PCA and FA, which are sensitive to the relative scales of the input data. The study involved normalizing each variable by subtracting the mean and dividing by the standard deviation, which transformed the data into z-scores. The data is centered around zero and has a standard deviation of one in order to prevent variables with larger ranges from overpowering the analysis. [12].

The specific steps taken for normalization include:

- Calculation of Mean and Standard Deviation: For each variable, the mean and standard deviation were calculated. These statistics were then used to standardize the variable.
- Transformation to Z-scores: Each data point was transformed using the formula $Z = \frac{X-\mu}{\sigma}$, where X is the original value, μ is the mean, and σ is the standard deviation. This transformation ensures that all variables contribute equally to the analysis.

Normalization is particularly important in regression analysis, as it helps in achieving better model convergence and stability, especially when variables are measured on different scales [10].

2.3. Multicollinearity Check

Unreliable estimates of regression coefficients can arise when there is a high correlation among predictor variables in a regression model, which is known as multicollinearity. Prior to applying PCA and FA, the dataset underwent a thorough examination for multicollinearity through the utilization of the Variance Inflation Factor (VIF). The extent to which collinearity increases the variance of an estimated regression coefficient can be quantified by VIF.

- The VIF was calculated for each predictor variable. Variables that had VIF values exceeding 10 were identified as having high collinearity and were marked for additional investigation [13].
- In order to address the issue of multicollinearity, we took measures to either transform or combine variables with high VIF values. An example of the application of PCA was the creation of uncorrelated principal components, which effectively reduced the multicollinearity in the dataset.

- An assessment was conducted to evaluate the impact of multicollinearity on the regression models. The models were compared before and after the application of PCA. The effectiveness of the steps taken to mitigate multicollinearity was confirmed through this comparison.

By implementing these preprocessing techniques, the data was effectively prepared for further analysis, resulting in more dependable and understandable outcomes from the PCA and FA. Enhancing the robustness of statistical models and improving their predictive performance and generalizability is achieved through proper data preprocessing [14].

2.4. Principal Component Analysis (PCA)

A commonly employed technique in multivariate statistics is Principal Component Analysis (PCA), which involves transforming a group of correlated variables into a smaller set of uncorrelated components referred to as principal components. The maximum variance in the data is captured by these components, effectively reducing the dimensionality of the dataset. This makes PCA a valuable tool for addressing concerns like multicollinearity in regression models. In this section, we will outline the detailed steps involved in applying PCA in this study, with support from relevant literature.

PCA is commonly employed to simplify the complexity of high-dimensional data while preserving a significant amount of information. In this method, an orthogonal transformation is used to convert the original correlated variables into a set of linearly uncorrelated variables, known as the principal components. In the dataset, the first principal component captures the highest amount of variance, followed by each subsequent component capturing the next highest variance. It is important to note that each component must be orthogonal to the previous ones [15]. PCA offers several advantages:

- A reduction in the number of variables in the dataset can be achieved through PCA, while still retaining a substantial amount of information. Simplifying the models and enhancing computational efficiency is achieved through this reduction.
- The resolution of multicollinearity is addressed by PCA, which transforms correlated variables into uncorrelated components. This process enhances the stability and interpretability of regression models [16].
- The visualization of high-dimensional data in a lower-dimensional space is made possible by PCA. This technique allows for a clearer understanding and interpretation of the data structure, often in two or three dimensions.

PCA effectively reduces the complexity of high-dimensional data by transforming correlated variables into uncorrelated principal components, which simplifies models, resolves multicollinearity, and enhances both computational efficiency and data visualization.

2.4.1. Covariance Matrix

Calculating the covariance matrix of the normalized dataset is the initial step in PCA. Insights can be gained from the covariance matrix regarding the variation of variables in the dataset. A strong relationship between two variables is indicated by a high covariance, which PCA aims to capture and summarize into fewer components.

The formula for calculating the covariance matrix involves the use of the data matrix, mean vector, and transpose operation. It is given by formula:

$$\mathrm{Cov}(X) = \frac{1}{n-1}(X-\mu)^{T}(X-\mu)$$

The matrix provided forms the foundation for identifying the principal components [17].

We examined the covariance matrix to identify pairs of variables that exhibited high covariance, suggesting redundancy in the data. By transforming these variables into a smaller set of uncorrelated components, the redundancy is effectively reduced through the use of PCA.

2.4.2. Eigenvalue Decomposition

In order to determine the principal components, it is crucial to perform eigenvalue decomposition of the covariance matrix. The researchers conducted an eigenvalue decomposition to acquire the eigenvalues and their corresponding eigenvectors from the covariance matrix. The amount of variance explained by each principal component is represented by the eigenvalues, while the direction of the new feature space is defined by the eigenvectors.

In order to form the principal components, the eigen-vectors were carefully selected after sorting the eigenvalues in descending order. In general, researchers tend to retain only the components with eigenvalues greater than 1, following the Kaiser criterion. This is because these components explain a higher amount of variance compared to the original variables [18]. A scree plot was created to visualize the eigenvalues, aiding in the determination of the optimal number of principal components to keep. An additional criterion for determining the number of components was used, based on the point in the scree plot where the eigenvalues start to level off [19].

2.4.3. Building Principal Components

After identifying the principal components, they were constructed by combining the original variables linearly, using the eigenvectors as coefficients. The construction of each principal component involved a linear combination of the original variables, with the coefficients being derived from the corresponding eigenvector. The transformation yields a fresh set of variables, known as the principal components, which exhibit no correlation among themselves.

The proportion of total variance explained was calculated by each principal component. Understanding the contribution of each component to the overall data structure is crucial in this context. The calculation was also performed to determine the cumulative variance explained by the retained components, ensuring that a satisfactory amount of the original variance was preserved.

An examination was conducted on the loadings of each variable on the principal components to interpret the components. The loadings demonstrate the correlation between the original variables and the principal components, with higher loadings indicating a more robust relationship. The principal components were named to reflect the underlying constructs they represent, based on the loadings. As an example, a component that has strong associations with variables linked to socio-economic status could be referred to as the "Socioeconomic Component." An assessment was conducted to determine the impact of dimensionality reduction on the interpretability and predictive power of the regression models. In subsequent regression models, the performance of models using the reduced set of principal components was compared to those using the original variables.

2.4.4. Application in Regression Modeling

Regression modeling has found practical application in various fields. Its usefulness lies in its ability to analyze and predict relationships between variables. By examining the relationship between a dependent variable and one or more independent variables, regression modeling can provide valuable insights and make accurate predictions. This has made it a popular tool in scientific research and data analysis.

Ultimately, the predictor variables in linear regression models were evaluated to determine their impact on enhancing predictive accuracy and model stability. The construction of regression models involved the utilization of principal components as predictors. The dimensionality of the model is reduced, and multicollinearity is mitigated, resulting in regression coefficients that are more stable and interpretable. An evaluation was conducted to compare the performance of the PCA-based regression model with models that utilized the original variables. The models' predictive accuracy and goodness of fit were assessed using metrics such as R-squared, Adjusted R-squared, and Root Mean Squared Error (RMSE) [14].

This study seeks to improve the predictive power of regression models and simplify them by systematically applying PCA. PCA is a highly effective tool that not only tackles multicollinearity but also enhances our comprehension of the underlying structure of the data, making it a crucial technique in contemporary data analysis.

2.5. Factor Analysis (FA).

A multivariate statistical technique known as Factor Analysis (FA) is commonly employed to uncover latent variables, referred to as factors, and identify underlying relationships between observed variables. The correlations among the observed variables can be explained by these factors, which represent underlying dimensions or constructs. In this study, the researchers utilized factor analysis (FA) to decrease the complexity of the dataset and uncover underlying variables that could potentially offer better insights into predicting voter behavior during the Croatian presidential election. This section outlines the steps used to apply FA in this study, with support from relevant literature.

The aim of Factor Analysis is to elucidate the correlations between observed variables by positing that these variables are impacted by a limited number of unobservable factors. The data suggests that these factors are responsible for the covariation of the observed variables. The main goals of FA are to decrease the number of variables while preserving the necessary information and to analyze the underlying structure of the data [20]. In sciences, FA proves to be highly valuable due to its ability to account for the influence of common underlying factors on multiple variables. These factors can range from socioeconomic status to political ideology. Identifying these factors allows researchers to streamline their models and concentrate on the critical aspects of the data.

2.5.1. Extraction Method

An essential initial step in FA is the careful selection of an extraction method that can effectively identify the factors. How the factors are derived from the data is determined by the extraction method. The researchers opted for the Maximum Likelihood (ML) method to extract factors in this study.

- Estimation using the Maximum Likelihood method: ML method was chosen because it offers statistically optimal estimates of the factor loadings, under the assumption of normal distribution of the data. The factors are estimated in this method by maximizing the likelihood that the observed correlation matrix could be produced by the model. When the sample size is large, ML proves to be highly effective. It enables hypothesis testing related to the number of factors and the adequacy of the model [21].
- Determining the Number of Factors: Various criteria were employed to determine the number of factors to extract. These criteria included the Kaiser criterion (eigenvalues greater than 1), the screen test, and parallel analysis. The methods employed in this study aim to retain only the factors that account for a significant portion of the variance, as described in the referenced publication [22].

The study utilized the Maximum Likelihood method for factor extraction due to its statistical efficiency and ability to facilitate hypothesis testing in large samples, complemented by rigorous criteria to determine the optimal number of factors to retain, ensuring a robust and reliable factor analysis.

2.5.2. Factor Rotation and Retention

After extracting the initial factors, an orthogonal rotation was applied to improve their interpretability. The underlying solution remains unchanged, but rotation simplifies the output and enhances its interpretability by creating a more straightforward and comprehensible structure.

- Varimax Rotation: We utilized the Varimax rotation method, which is widely recognized as the predominant form of orthogonal rotation. The columns of the factor matrix are simplified by maximizing the loadings of certain variables for each factor, while minimizing the loadings for others. The rotation technique employed in this study aims to enhance the identification of variables associated with each factor by maximizing the variance of squared loadings across variables [23].

- Analysis of Rotated Factors: The examination of factor loadings was conducted after rotation to interpret the factors. The factors were labeled according to the variables with the highest loadings, which indicate the dimensions they represent. For example, a factor that has high loadings on variables associated with economic indicators could be labeled as a "Economic Factor."

One important step in factor analysis is determining the appropriate number of factors to retain. If too few factors are retained, valuable information may be lost. On the other hand, retaining too many factors can result in overfitting and make interpretation more difficult.

- The Kaiser Criterion states that factors with eigenvalues greater than 1 were kept, as they explain more variance than a single observed variable, as previously stated. This criterion finds extensive usage in exploratory factor analysis [23].
- Scree Test: The visualization of eigenvalues was achieved using a scree plot. The number of factors to retain was determined by identifying the point at which the plot begins to level off, commonly referred to as the "elbow" point. This method allows for the differentiation of significant factors from those that have minimal impact on explaining the variance [19].
- Parallel Analysis: A comparison was made between the observed eigenvalues and those obtained from randomly generated data using parallel analysis. We only kept the factors that had eigenvalues higher than those obtained from the random data. It has been noted that this method is regarded as more reliable when compared to the Kaiser criterion and scree test alone, especially when dealing with complex datasets [24].

The study carefully determined the appropriate number of factors to retain in factor analysis by using the Kaiser Criterion, Scree Test, and Parallel Analysis, ensuring a balance between retaining sufficient information and avoiding overfitting, with parallel analysis offering the most reliability in complex datasets.

2.5.3. Interpretation and Naming of Factors

An essential step in factor analysis is interpreting and naming the factors, as this gives meaning to the latent variables identified through the analysis.

- Examining the factor loadings allowed for a careful determination of the variables that make the most significant contributions to each factor. A strong relationship between the variable and the factor is indicated by high loadings. The factor was determined to have significant contributors based on variables with loadings above 0.4 [25].
- Identifying Factors: The factors were named to reflect the underlying constructs they represent, based on the pattern of loadings. As an illustration, a factor that exhibits high loadings on variables such as education, income, and occupation could be referred to as "Socioeconomic Status."
- Calculating Factor Scores: Each observation was assigned factor scores to measure the extent to which they exhibit the characteristics of the factor. In subsequent regression models, the scores were used as predictor variables.

The study conducted a thorough evaluation of factor loadings in order to identify significant contributors. Factors were then named based on the loading patterns, and factor scores were calculated to quantify the influence of each factor on observations. These scores were subsequently utilized in regression models.

2.5.4. Application in Regression Modeling

Predictor variables in linear regression models were derived from the factors identified through FA. The objective of this approach was to enhance the model's ability to accurately predict and explain voter behavior by placing emphasis on the latent variables that have the greatest impact.

- Model Construction: The construction of regression models involved utilizing the factor scores as independent variables. By utilizing a reduced number of uncorrelated predictors, this approach effectively addresses the issue of multicollinearity and enhances the stability of the model.
- Model Evaluation: Metrics such as R-squared, Adjusted R-squared, and Root Mean Squared Error (RMSE) were used to evaluate the performance of the regression model based on FA. The effectiveness of FA in improving predictive accuracy was assessed by comparing these metrics with those of models using the original and PCA-transformed data [26].

Factor Analysis is a crucial tool in revealing the hidden structure of intricate datasets, especially in the field of social sciences. In this study, the predictive power and interpretability of regression models are enhanced by identifying and interpreting latent variables, which makes factor analysis an essential tool.

2.6. Regression Modeling

A statistical technique called regression modeling is used to analyze the relationship between a dependent variable and one or more independent variables. This study utilized regression modeling to predict voter behavior in the Croatian presidential election. Three sets of predictor variables were employed: the original dataset, the dataset transformed by Principal Component Analysis (PCA), and the dataset transformed by Factor Analysis (FA). In this section, a thorough explanation of the regression modeling process is presented, covering model construction, evaluation, and comparison.

2.6.1. Model Construction

Understanding the relationships between variables and making predictions based on those relationships is greatly facilitated by constructing regression models. This study involved the construction of three separate regression models:

2.6.2. Data in its Original Form

A model was constructed using the original predictor variables, without applying any dimension reduction techniques. The other two models were compared to this model, which served as a baseline. The original data was used to establish a reference point for evaluating the impact of PCA and FA on model performance. The model's linear regression equation is specified as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

In this equation, the dependent variable (voter behavior) is denoted as Y, while the predictor variables are represented as $X_1, X_2, \cdots, X_n$. The intercept is denoted as β_0, and the coefficients of the predictors are $\beta_1, \beta_2, \cdots, \beta_n$. The error term is represented as ϵ.

Original Data with Removed Highly Correlated Variables

In constructing the second model, highly correlated variables were first identified and subsequently removed from the original dataset. To reduce multicollinearity, we removed variables that had high Variance Inflation Factor (VIF) values, usually greater than 10. The objective of this study was to analyze the impact of manually eliminating variables with high correlation on the predictive accuracy and stability of the regression model, in comparison to using all the original variables. Identifying Correlated Variables: We assessed multicollinearity by calculating the Variance Inflation Factor (VIF) for each predictor variable. The presence of variables with a VIF greater than 10 indicated a high level of correlation and suggested that they should be considered for removal [14]. The specified regression model was formulated after eliminating variables that showed high correlation. The regression model after removing the highly correlated variables was specified as:

$$Y = \delta_0 + \delta_1 X'_1 + \delta_2 X'_2 + \cdots + \delta_m X'_{nm} + \epsilon$$

where $X'_1, X'_1, \ldots, X'_m$ are the remaining predictor variables after removing those with high multicollinearity, and $\delta_1, \delta_2, \ldots, \delta_m$ are their corresponding coefficients?

DOI: 10.55214/25768484.v8i6.2307

2.6.3. Data Transformed by PCA

The predictor variables for the third model were derived from PCA's principal components. The objective of this study was to evaluate the impact of dimension reduction using PCA on the predictive accuracy and stability of the regression model. An application of PCA was utilized to decrease multicollinearity and the dimensionality of the dataset, while still preserving the maximum variance. In the regression model, the predictors used were the uncorrelated resulting principal components. The regression model using PCA-transformed data was specified as:

$$Y = \alpha_0 + \alpha_1 PC_1 + \alpha_2 PC_{21} + \cdots + \alpha_m PC_m + \epsilon$$

where $PC_1, PC_2, \ldots, PC_m$ are the principal components, and $\alpha_1, \alpha_2, \ldots, \alpha_m$ are the coefficients of the principal components? The number of predictors in this model is usually reduced compared to the original model, as a result of dimensionality reduction [15].

2.6.3. Data Transformed Using the Fourier Analysis (FA)

The predictor variables for the fourth model were constructed using the factor scores derived from FA. The objective of this model was to assess the efficacy of utilizing latent variables (factors) identified through factor analysis in predicting voter behavior. The application of factor analysis allowed for the identification of latent variables that explain the correlations among the observed variables. The predictors in the regression model were based on the factors that represent the underlying structure of the data. The specified regression model utilizes FA-transformed data and is represented by the following formula: The regression model using FA-transformed data was specified as:

$$Y = \gamma_0 + \gamma_1 F_1 + \gamma_2 F_{21} + \cdots + \gamma_m F_m + \epsilon$$

where $F_1, F_2, \ldots, F_m$ are the factors, and $\gamma_1, \gamma_2, \ldots, \gamma_m$ are the coefficients of the factors. Like the PCA model, this model also uses fewer predictors due to dimensionality reduction.

2.6.4. Model Evaluation

Several metrics were used to evaluate the performance of the regression models, assessing their predictive accuracy, goodness of fit, and overall effectiveness.

The coefficient of determination, commonly referred to as R-squared (R^2), is a statistical measure that quantifies the proportion of the variance in the dependent variable that can be explained by the independent variable(s). The R-squared statistic quantifies the extent to which the independent variables can predict the variance in the dependent variable. The model performance is indicated by values ranging from 0 to 1, with higher values suggesting better performance. The explanatory power of the three models was compared using R-squared. The model's ability to explain the variance in voter behavior increases with a higher R-squared value. It is worth mentioning that evaluating model performance based solely on R-squared is not enough, as it fails to consider the complexity of the model [27].

The adjusted R-squared value takes into account the number of predictors in the model, providing a more accurate measure of the model's goodness of fit. The addition of irrelevant variables is penalized, resulting in a more accurate measure of model performance. This is particularly useful when comparing models with different numbers of predictors. The effectiveness of dimension reduction techniques (PCA and FA) in improving model performance was assessed using Adjusted R-squared. A higher adjusted R-squared value suggests a more optimal trade-off between the complexity of the model and its ability to explain the data [14].

The RMSE quantifies the average magnitude of prediction errors by taking the square root of the average squared differences between observed and predicted values. Greater model accuracy is indicated by lower RMSE values. The predictive accuracy of the models was evaluated using RMSE. An indication is given of how accurately the model's predictions align with the actual outcomes. The comparison of RMSE values among the three models allowed for the identification of the model that yielded the most precise predictions [28].

The Akaike Information Criterion (AIC) quantifies the relative quality of a statistical model for a specific dataset. The evaluation takes into account both the fit quality and model complexity, where lower AIC values suggest superior models. The three regression models were compared using AIC. The selection of a model that strikes the optimal balance between fit and complexity is crucial in order to avoid overfitting [29].

2.6.5. Comparison of Models

A comparison was made between four models, evaluating their performance based on different metrics. The aim was to determine how PCA, FA, and variable removal affect the models' performance. Insights were gained from the comparison regarding the effectiveness of these techniques in improving predictive accuracy and model stability.

The model that had highly correlated variables removed was anticipated to exhibit similar or slightly improved performance in terms of adjusted R-squared and AIC. This is because it reduces multicollinearity while retaining substantial explanatory power. However, it should be noted that the removal of variables could potentially lead to a decrease in the R-squared value. This is because when fewer predictors are included in the model, the overall explanatory power may be reduced.

4. Results and Discussion

The effectiveness and reliability of the regression models used in this study are determined by statistical inference, which is of utmost importance.

4.1. Goodness-of-Fit Measures

The goodness-of-fit of a regression model is typically evaluated through metrics such as the Total Sum of Squares (SST), the Residual Sum of Squares (SSR), and the Explained Sum of Squares (SSE). The relationship between these measures helps in understanding how well the model explains the variance in the dependent variable:

- Total Sum of Squares (SST): Represents the total variation in the dependent variable that would be observed if no predictors were included in the model.
- Residual Sum of Squares (SSR): Represents the variation in the dependent variable that is not explained by the model.
- Explained Sum of Squares (SSE): Measures the reduction in SST due to the inclusion of predictors in the model, indicating the amount of variance explained by the model.

The relationship is summarized as:

$$SST = SSE + SSR$$

This decomposition is fundamental in calculating the coefficient of determination (R^2), which indicates the proportion of the variance in the dependent variable explained by the model.

4.2. Significance Tests

Significance tests are used to assess whether the predictors in the model make a significant contribution to explaining the variance in the dependent variable. An assessment of the overall significance of the model is made using the F-test. The F-statistic compares the variance that can be explained by a model to the variance that cannot be explained, taking into account the degrees of freedom for each. The prediction of the dependent variable is significantly improved by a model with predictors, as indicated by a higher F-value compared to a model without predictors.

T-tests are used to test the significance of coefficients for individual predictors. The significance of the coefficients is assessed in these tests, indicating that the predictor has a meaningful impact on the dependent variable.

DOI: 10.55214/25768484.v8i6.2307

4.3. Model Selection and Optimization

Selection of the optimal model can be achieved by utilizing criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). The criteria in question strike a delicate balance between the quality of fit and the complexity of the model, discouraging the inclusion of excessive parameters in order to avoid overfitting. The most efficient model is determined by selecting the one with the lowest AIC or BIC, as it offers a good fit while minimizing complexity.

The significance of these measures and tests in validating the regression models is underscored in statistical inference, ensuring that they not only exhibit a good fit to the data but also demonstrate effective generalization to unseen data.

4.4. Data Sources and Preparation

The data sources utilized in this study played a crucial role in the development of a predictive model for the Croatian presidential election. The study utilized data from two primary sources: the Croatian Bureau of Statistics (DZS) and the State Election Commission (DIP), which were publicly available. The statistical data at the municipal level, provided by the Croatian Bureau of Statistics (DZS), played a crucial role in constructing the predictor variables utilized in the regression models. The variables encompassed a range of demographic, economic, and social indicators, including the proportion of the educated population, rates of unemployment, and levels of household computer and internet usage. The researchers meticulously collected the data, merged it using primary keys (municipality/city), and structured it into a cohesive dataset, which formed the foundation for analysis.

The data from the State Election Commission (DIP) was scraped and processed using custom scripts to extract results from the first round of the presidential elections. Although the data was not presented in a tabular form, it was successfully extracted. The dataset contained electoral outcomes by municipality, which played a crucial role in determining the dependent variable in the predictive models. In the study, predictors such as local government indicators and the percentage of votes in the marriage referendum were derived and utilized.

The final dataset included variables at both the municipal and regional levels, after data from DZS and DIP were collected and merged. Before developing the model, the comprehensive dataset underwent necessary preprocessing steps, including the removal of outliers and conducting correlation analyses, using statistical software (SPSS and Orange).

4.5. Data Verification

After completing the process of collecting and integrating the data, the next important step was to ensure the quality and suitability of the dataset for analysis. The section on data verification discusses the important steps taken to ensure the suitability of the data for constructing accurate predictive models.

4.5.1. Detecting Outliers

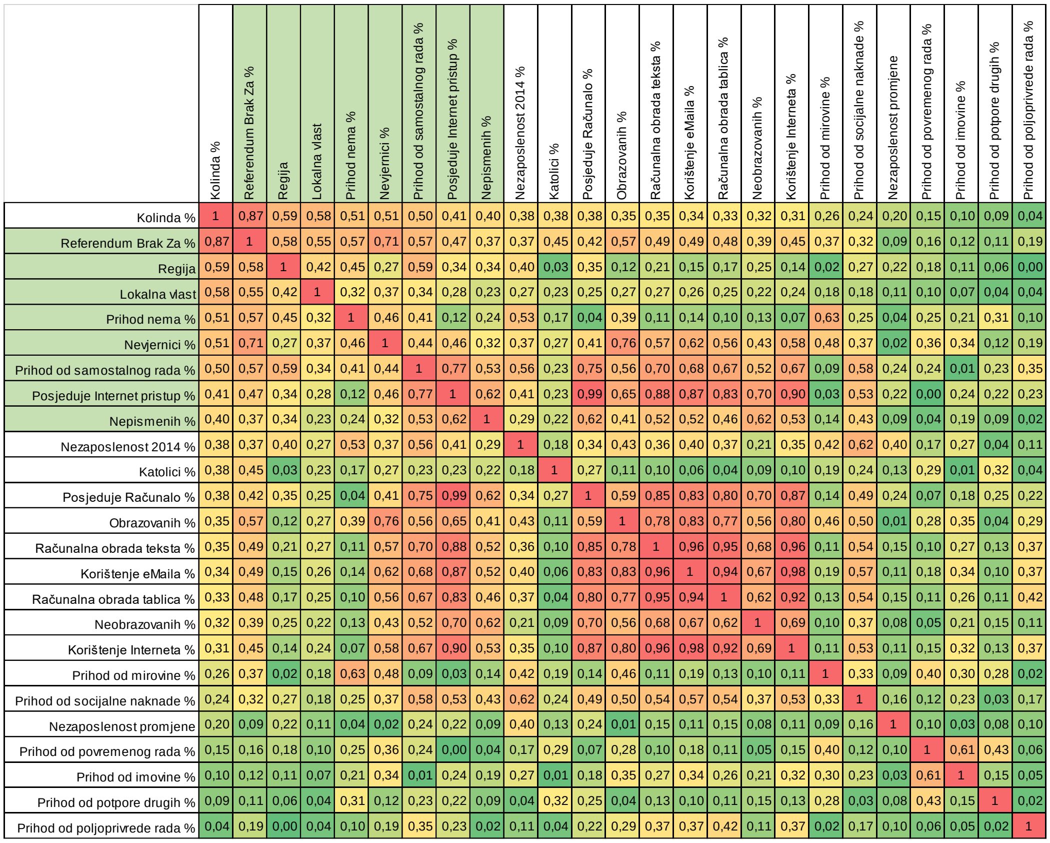
The presence of outliers in a dataset can have a distorting effect on the results of regression models, as they represent data points that deviate significantly from the norm. The study utilized the statistical software Orange to identify outliers within the dataset. This was achieved by calculating the Z-scores for each observation. Ervenik and Civljane, two municipalities, were found to be outliers with remarkably high Z-scores. Excluding observations with small population sizes and minimal impact on the overall results was necessary to ensure the accuracy of the model outcomes.

4.5.2. Relevance Analysis of Predictor Variables

Information Gain (IG) was used to assess the relevance of each predictor variable to the target variable, voter behavior. This measure aids in the identification of the variables that have the greatest impact on the accuracy of the prediction. Only the most informative predictors were included in the regression models by retaining variables with an IG greater than 0.2 for further analysis. The number of predictor variables was reduced in this step, with a focus on those that had significant predictive value.

4.5.3. Correlation Analysis

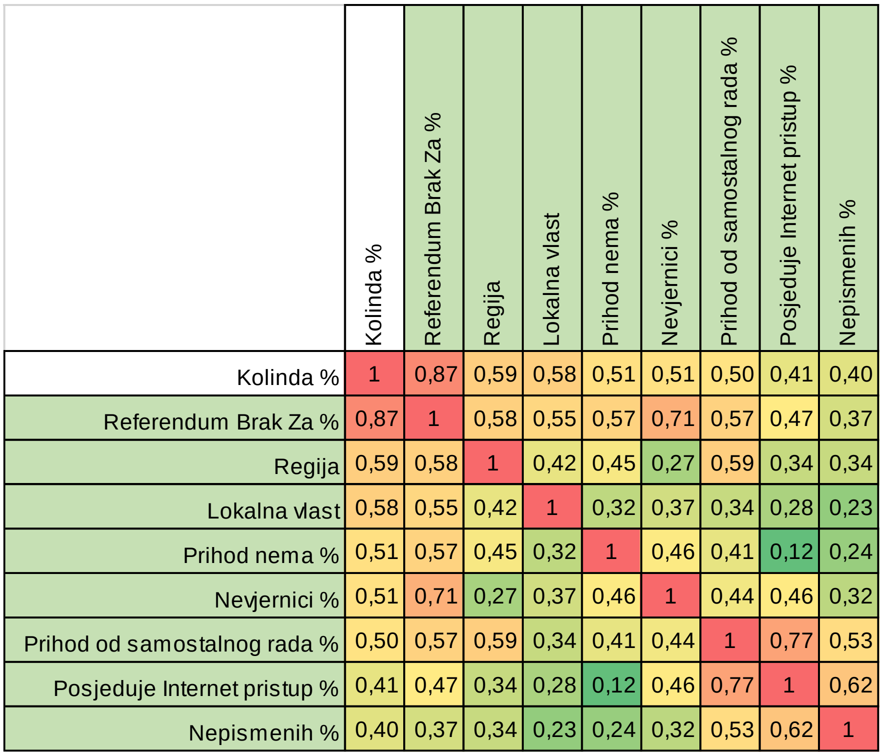
Correlation analysis was conducted to identify any multicollinearity among the predictor variables. Unstable regression coefficients can be a consequence of high multicollinearity, which can undermine the reliability of the model.



	Kolinda %	Referendum Brak Za %	Regija	Lokalna vlast	Prihod nema %	Nevjernici %	Prihod od samostalnog rada %	Posjeduje Internet pristup %	Nepismenih %	Nezaposlenost 2014 %	Katolici %	Posjeduje Računalo %	Obrazovanih %	Računalna obrada teksta %	Korištenje eMaila %	Računalna obrada tablica %	Neobrazovanih %	Korištenje Interneta %	Prihod od mirovine %	Prihod od socijalne naknade %	Nezaposlenost promjene	Prihod od povremenog rada %	Prihod od imovine %	Prihod od potpore drugih %	Prihod od poljoprivrede rada %
Kolinda %	1	0,87	0,59	0,58	0,51	0,51	0,50	0,41	0,40	0,38	0,38	0,38	0,35	0,35	0,34	0,33	0,32	0,31	0,26	0,24	0,20	0,15	0,10	0,09	0,04
Referendum Brak Za %	0,87	1	0,58	0,55	0,57	0,71	0,57	0,47	0,37	0,37	0,45	0,42	0,57	0,49	0,49	0,48	0,39	0,45	0,37	0,32	0,09	0,16	0,12	0,11	0,19
Regija	0,59	0,58	1	0,42	0,45	0,27	0,59	0,34	0,34	0,40	0,03	0,35	0,12	0,21	0,15	0,17	0,25	0,14	0,02	0,27	0,22	0,18	0,11	0,06	0,00
Lokalna vlast	0,58	0,55	0,42	1	0,32	0,37	0,34	0,28	0,23	0,27	0,23	0,25	0,27	0,27	0,26	0,25	0,22	0,24	0,18	0,18	0,11	0,10	0,07	0,04	0,04
Prihod nema %	0,51	0,57	0,45	0,32	1	0,46	0,41	0,12	0,24	0,53	0,17	0,04	0,39	0,11	0,14	0,10	0,13	0,07	0,63	0,25	0,04	0,25	0,21	0,31	0,10
Nevjernici %	0,51	0,71	0,27	0,37	0,46	1	0,44	0,46	0,32	0,37	0,27	0,41	0,76	0,57	0,62	0,56	0,43	0,58	0,48	0,37	0,02	0,36	0,34	0,12	0,19
Prihod od samostalnog rada %	0,50	0,57	0,59	0,34	0,41	0,44	1	0,77	0,53	0,56	0,23	0,75	0,56	0,70	0,68	0,67	0,52	0,67	0,09	0,58	0,24	0,24	0,01	0,23	0,35
Posjeduje Internet pristup %	0,41	0,47	0,34	0,28	0,12	0,46	0,77	1	0,62	0,41	0,23	0,99	0,65	0,88	0,87	0,83	0,70	0,90	0,03	0,53	0,22	0,00	0,24	0,22	0,23
Nepismenih %	0,40	0,37	0,34	0,23	0,24	0,32	0,53	0,62	1	0,29	0,22	0,62	0,41	0,52	0,52	0,46	0,62	0,53	0,14	0,43	0,09	0,04	0,19	0,09	0,02
Nezaposlenost 2014 %	0,38	0,37	0,40	0,27	0,53	0,37	0,56	0,41	0,29	1	0,18	0,34	0,43	0,36	0,40	0,37	0,21	0,35	0,42	0,62	0,40	0,17	0,27	0,04	0,11
Katolici %	0,38	0,45	0,03	0,23	0,17	0,27	0,23	0,23	0,22	0,18	1	0,27	0,11	0,10	0,06	0,04	0,09	0,10	0,19	0,24	0,13	0,29	0,01	0,32	0,04
Posjeduje Računalo %	0,38	0,42	0,35	0,25	0,04	0,41	0,75	0,99	0,62	0,34	0,27	1	0,59	0,85	0,83	0,80	0,70	0,87	0,14	0,49	0,24	0,07	0,18	0,25	0,22
Obrazovanih %	0,35	0,57	0,12	0,27	0,39	0,76	0,56	0,65	0,41	0,43	0,11	0,59	1	0,78	0,83	0,77	0,56	0,80	0,46	0,50	0,01	0,28	0,35	0,04	0,29
Računalna obrada teksta %	0,35	0,49	0,21	0,27	0,11	0,57	0,70	0,88	0,52	0,36	0,10	0,85	0,78	1	0,96	0,95	0,68	0,96	0,11	0,54	0,15	0,10	0,27	0,13	0,37
Korištenje eMaila %	0,34	0,49	0,15	0,26	0,14	0,62	0,68	0,87	0,52	0,40	0,06	0,83	0,83	0,96	1	0,94	0,67	0,98	0,19	0,57	0,11	0,18	0,34	0,10	0,37
Računalna obrada tablica %	0,33	0,48	0,17	0,25	0,10	0,56	0,67	0,83	0,46	0,37	0,04	0,80	0,77	0,95	0,94	1	0,62	0,92	0,13	0,54	0,15	0,11	0,26	0,11	0,42
Neobrazovanih %	0,32	0,39	0,25	0,22	0,13	0,43	0,52	0,70	0,62	0,21	0,09	0,70	0,56	0,68	0,67	0,62	1	0,69	0,10	0,37	0,08	0,05	0,21	0,15	0,11
Korištenje Interneta %	0,31	0,45	0,14	0,24	0,07	0,58	0,67	0,90	0,53	0,35	0,10	0,87	0,80	0,96	0,98	0,92	0,69	1	0,11	0,53	0,11	0,15	0,32	0,13	0,37
Prihod od mirovine %	0,26	0,37	0,02	0,18	0,63	0,48	0,09	0,03	0,14	0,42	0,19	0,14	0,46	0,11	0,19	0,13	0,10	0,11	1	0,33	0,09	0,40	0,30	0,28	0,02
Prihod od socijalne naknade %	0,24	0,32	0,27	0,18	0,25	0,37	0,58	0,53	0,43	0,62	0,24	0,49	0,50	0,54	0,57	0,54	0,37	0,53	0,33	1	0,16	0,12	0,23	0,03	0,17
Nezaposlenost promjene	0,20	0,09	0,22	0,11	0,04	0,02	0,24	0,22	0,09	0,40	0,13	0,24	0,01	0,15	0,11	0,15	0,08	0,11	0,09	0,16	1	0,10	0,03	0,08	0,10
Prihod od povremenog rada %	0,15	0,16	0,18	0,10	0,25	0,36	0,24	0,00	0,04	0,17	0,29	0,07	0,28	0,10	0,18	0,11	0,05	0,15	0,40	0,12	0,10	1	0,61	0,43	0,06
Prihod od imovine %	0,10	0,12	0,11	0,07	0,21	0,34	0,01	0,24	0,19	0,27	0,01	0,18	0,35	0,27	0,34	0,26	0,21	0,32	0,30	0,23	0,03	0,61	1	0,15	0,05
Prihod od potpore drugih %	0,09	0,11	0,06	0,04	0,31	0,12	0,23	0,22	0,09	0,04	0,32	0,25	0,04	0,13	0,10	0,11	0,15	0,13	0,28	0,03	0,08	0,43	0,15	1	0,02
Prihod od poljoprivrede rada %	0,04	0,19	0,00	0,04	0,10	0,19	0,35	0,23	0,02	0,11	0,04	0,22	0,29	0,37	0,37	0,42	0,11	0,37	0,02	0,17	0,10	0,06	0,05	0,02	1

Figure 1.
Correlation matrix with all predictor variables.

Strong correlations were found among several predictor variables as shown on Figure 1, particularly between the percentage of votes for the marriage referendum (*Referendum Brak Za %*) and other variables, according to the analysis.



	Kolinda %	Referendum Brak Za %	Regija	Lokalna vlast	Prihod nema %	Nevjernici %	Prihod od samostalnog rada %	Posjeduje Internet pristup %	Nepismenih %
Kolinda %	1	0,87	0,59	0,58	0,51	0,51	0,50	0,41	0,40
Referendum Brak Za %	0,87	1	0,58	0,55	0,57	0,71	0,57	0,47	0,37
Regija	0,59	0,58	1	0,42	0,45	0,27	0,59	0,34	0,34
Lokalna vlast	0,58	0,55	0,42	1	0,32	0,37	0,34	0,28	0,23
Prihod nema %	0,51	0,57	0,45	0,32	1	0,46	0,41	0,12	0,24
Nevjernici %	0,51	0,71	0,27	0,37	0,46	1	0,44	0,46	0,32
Prihod od samostalnog rada %	0,50	0,57	0,59	0,34	0,41	0,44	1	0,77	0,53
Posjeduje Internet pristup %	0,41	0,47	0,34	0,28	0,12	0,46	0,77	1	0,62
Nepismenih %	0,40	0,37	0,34	0,23	0,24	0,32	0,53	0,62	1

Figure 2.
Correlation matrix with predictor variables having IG greater than 0.2.

Figure 2 demonstrates the selection of the predictor with the highest Information Gain and the lowest correlation with other variables, which guarantees a regression model that is both stable and interpretable.

4.5.4. Preparation for Regression Modeling

After conducting relevance and correlation analyses, the dataset was refined to only include the predictor variables that were deemed most relevant and independent. The dataset underwent a thorough refinement process to ensure the resulting regression models would be strong, dependable, and unaffected by outliers and multicollinearity.

4.5.5. Model Construction Using Linear Regression

In this section, we will discuss the construction of a linear regression model that can be used to predict voter behavior in the Croatian presidential election. The LR1 model was constructed by incorporating key predictor variables that were identified through previous analysis, with a specific emphasis on their relevance and independence.

After conducting an analysis in earlier sections, we have selected two predictor variables to construct the linear regression model:

- The variable "*Referendum Brak Za %*" indicates the percentage of votes supporting the marriage referendum and was discovered to have a strong correlation with the dependent variable (percentage of votes for the candidate). Retaining it as a predictor was justified by its strong correlation and high Information Gain.
- The variable "*Nepismenih %*" represents the proportion of individuals who are unable to read or write within a municipality. The selection was based on its significant Information Gain and

lower correlation with other variables, which establishes it as a strong and independent predictor.

These two predictors were used to construct the linear regression model (LR1). The model sought to elucidate the variability in the dependent variable, specifically the percentage of votes for the presidential candidate, by considering the chosen predictors. Several key steps were involved in the construction and evaluation of the model:

The model was constructed utilizing the ordinary least squares (OLS) technique. As shown in Table 1 based on the analysis, it was determined that the R^2 value of the model is 0.874, suggesting that the model can account for 87.4% of the variability in the dependent variable. The high R^2 indicates that the selected predictors successfully capture most of the variance in voter behavior.

Table 1.
Coefficients of determination for model LR1.

Model	R	R^2	Adjusted R^2	Standard error
LR1	0.935	0.874	0.872	0.127

A significance test was conducted using Analysis of Variance (ANOVA) to evaluate the regression model. The model exhibited an exceptionally high F-statistic (894.472) and a p-value of less than 0.001, suggesting that the model is statistically significant at all levels of significance. The results of the ANOVA are presented in Table 2, which demonstrates the significance of the model.

Table 2.
Coefficients of determination for model LR1.

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F-value	Significance (p-value)
Regression	144.890	2	72.445	894.472	< 0.001
Residual	20.855	258	0.081		
Total	165.745	260			

Significance testing was conducted on the estimated coefficients of the predictor variables. The coefficients of both predictors, *Referendum Brak Za %* and *Nepismenih %*, were found to be highly significant ($p < 0.001$), indicating their strong influence on the dependent variable. The estimated coefficients and their significance levels are provided in Table 3.

Table 3.
Variable analysis for model LR1.

Model	Unstandardized coefficients (B)	Standard error (Std. error)	t-value	Significance (Sig.)
(Constant)	-26.001	1.952	-13.318	0.000
Referendum Brak Za %	1.013	0.027	37.607	0.000
Nepismenih %	1.103	0.277	3.981	0.000

4.5.6. Regression Modeling

Valuable insights into voter behavior during the Croatian presidential election are provided by the LR1 model. Municipalities that showed greater support for the marriage referendum also tended to have a higher percentage of votes for right-wing politicians, as indicated by the positive coefficient. Similarly, it can be observed that there is a positive correlation between higher illiteracy rates and greater support for this candidate.

Once the initial linear regression model (LR1) was constructed, it became essential to confirm the satisfaction of the Gauss-Markov assumptions. These assumptions play a crucial role in ensuring the reliability and efficiency of the Ordinary Least Squares (OLS) estimator. One crucial requirement is homoscedasticity, which necessitates a constant variance of the residuals (errors) across all levels of the independent variables.

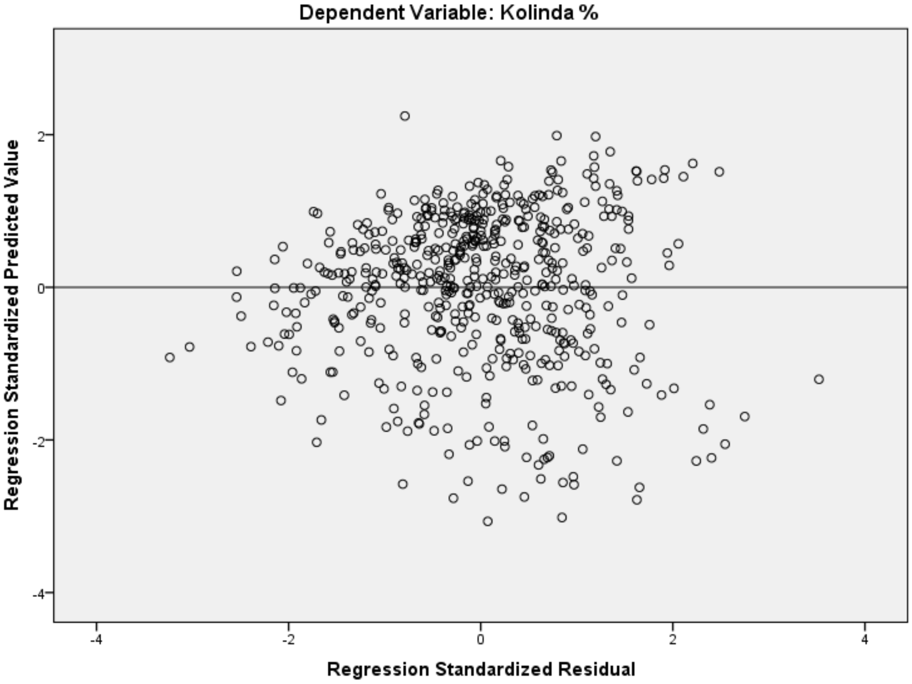


Figure 3.
Homoscedasticity conditions for the LR1 model.

Upon examination, it was determined that the LR1 model did not meet the homoscedasticity condition. In Figure 3, it can be observed that the residuals do not exhibit symmetrical dispersion around the line Y=0. This is evident from the residuals versus fitted values plot. Heteroscedasticity is indicated by the lack of symmetry, which suggests that the variance of residuals increases as the fitted values increase. It appears that the model may be underestimating in areas with high residuals, which could result in unreliable predictions in those regions.

Although the LR1 model had a high coefficient of determination (R^2), its rejection was necessary due to the poor dispersion of residuals. It is suggested that the presence of heteroscedasticity may lead to a potential decrease in the accuracy of outcome prediction, particularly when dealing with large residuals.

4.5.7. New Regression Modeling

A new model (LR2) was developed to tackle this issue. In order to address the problem of heteroscedasticity, the predictor variable Referendum Brak Za %, which had a high information gain but was highly correlated with the dependent variable, was removed. The decision was made based on the observation that models with fewer predictor variables could potentially result in inaccurate predictions. This is especially true if a strong variable experiences a significant change, which could excessively impact the outcome. It was anticipated that the LR2 model would evenly distribute the impact of changes by incorporating additional predictor variables, thereby creating a more robust model.

Other predictor variables with an information gain greater than 0.1 were used to construct the LR2 model. The researcher utilized a stepwise selection method to determine the predictor variables that would be incorporated into the model. Variables are added to the model in the stepwise method if their significance level falls below 5%, while they are removed if their significance level exceeds 10%. The reduced model was achieved by ensuring that each parameter's significance was at least 10%.

The LR2 model showed remarkably high values of the coefficient of determination, suggesting that a substantial percentage of the variability in the dependent variable was explained by the model. As an illustration, the model with nine predictor variables yielded a R^2 value of 0.722, indicating that it explained 72.2% of the variance in the percentage of votes for the candidate.

We examined the residuals of the LR2 model to confirm that we had successfully addressed the issue of heteroscedasticity. Figure 4 demonstrates a notable enhancement in the dispersion of residuals, resulting in a more uniform distribution. This suggests a commendable standardization of errors. The LR2 model's improved accuracy and reliability in predicting voter behavior is evident.

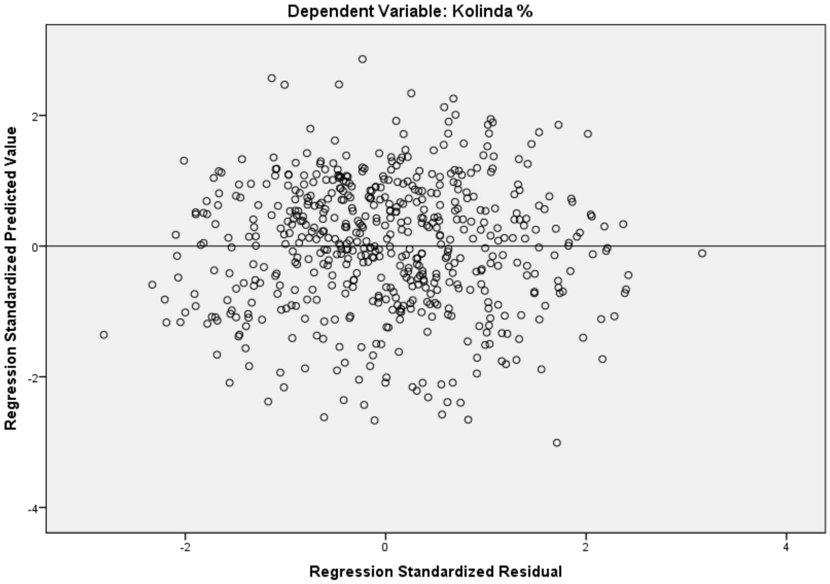


Figure 4.
Homoscedasticity conditions for the LR2 model.

Several metrics were used to evaluate the performance of the four regression models. These included R-squared (R^2), Adjusted R-squared, Root Mean Squared Error (RMSE), and Akaike Information Criterion (AIC). Below, we present the findings of these evaluations.

4.6. Model Construction Using Principal Component Analysis (PCA)

The original set of predictor variables was transformed using PCA to create a smaller set of uncorrelated components that effectively capture the majority of the variance in the data. The eigenvalues and eigenvectors of the correlation matrix of the predictor variables were calculated as part of the process. A substantial portion of the variance was explained by selecting the significant principal components.

A new regression model was constructed using the significant principal components as predictors instead of the original variables, once they were selected. The objective of this approach was to tackle multicollinearity and simplify the model without compromising its predictive capability.

The number of predictors (principal components) used in the PCA model was reduced compared to the original model, which led to a more streamlined model. The use of principal components as uncorrelated predictors helps to mitigate the risk of overfitting and enhances the robustness of the model. The evaluation of the PCA-transformed regression model involved the utilization of identical metrics as the original model, as well as the LR1 and LR2 models. These metrics encompassed R-squared, Adjusted R-squared, and Root Mean Squared Error (RMSE). Although the PCA model showed

DOI: 10.55214/25768484.v8i6.2307

improvement in reducing multicollinearity compared to the initial LR1 model, it did not match the performance of the LR2 model.

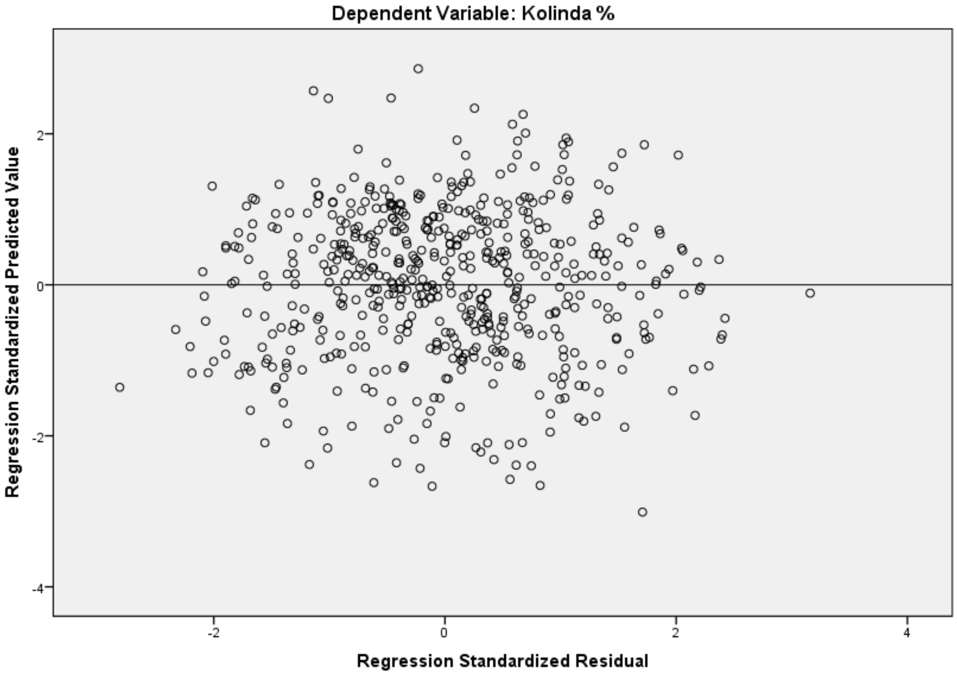


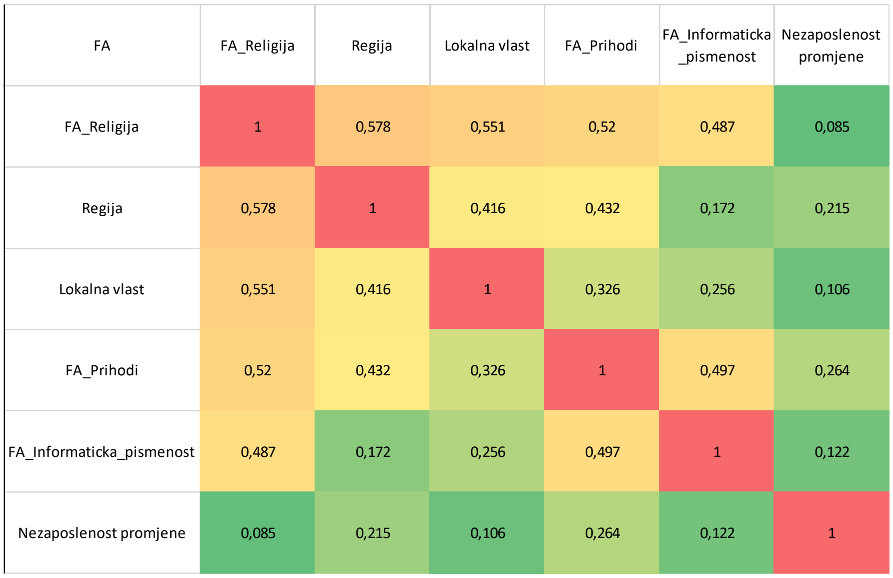
Figure 5.
Homoscedasticity conditions for the PCA model.

The distribution of residuals for the PCA model is depicted in Figure 5. While the residuals show a more even distribution compared to the LR1 model, the homoscedasticity condition is not as well met as in the LR2 model. It can be inferred that although PCA was useful in tackling certain concerns, it did not completely solve the issue of heteroscedasticity. In this study, it was found that the model's performance can be improved by using PCA, especially through the reduction of multicol-linearity. This improvement was observed when comparing the initial LR1 model with dimension reduction. Nevertheless, the PCA model continued to demonstrate problems with homoscedasticity, which resulted in it being less reliable compared to the LR2 model. Although PCA had its advantages, the LR2 model outperformed it by carefully selecting predictor variables and addressing homoscedasticity.

4.7. Model Construction Using Factor Analysis (FA)

In this section, we explore into the utilization of Factor Analysis (FA) for the development of regression models with the purpose of predicting voter behavior in the Croatian presidential election. In contrast to Principal Component Analysis (PCA), which places emphasis on maximizing variance, Factor Analysis (FA) aims to uncover latent variables, or factors, that account for the correlations among observed predictor variables. Model construction can be based on these factors, which provide a basis that is both interpretable and theoretically meaningful.

The process started by identifying relevant factors from the set of predictor variables. Factors were created by grouping variables together based on their correlations and theoretical relevance. An example of a factor called Informatička pismenost (Information Literacy) was created by grouping variables related to internet usage, computer ownership, and literacy rates.



FA	FA_Religija	Regija	Lokalna vlast	FA_Prihodi	FA_Informaticka _pismenost	Nezaposlenost promjene
FA_Religija	1	0,578	0,551	0,52	0,487	0,085
Regija	0,578	1	0,416	0,432	0,172	0,215
Lokalna vlast	0,551	0,416	1	0,326	0,256	0,106
FA_Prihodi	0,52	0,432	0,326	1	0,497	0,264
FA_Informaticka_pismenost	0,487	0,172	0,256	0,497	1	0,122
Nezaposlenost promjene	0,085	0,215	0,106	0,264	0,122	1

Figure 6.
Correlation matrix with selected factors and predictor variables.

The correlation matrix for the selected factors and predictor variables is depicted in Figure 6. This figure demonstrates the correlation between variables within each factor, indicating that each factor represents a cohesive concept.

Two regression models were constructed using the identified factors.

- FA1 Model: The first model included a limited set of factors that were most strongly correlated with the dependent variable (percentage of votes for candidate).
- FA2 Model: The second model expanded on FA1 by including additional factors to explore whether a broader set of predictors would improve model accuracy.

Both models sought to utilize the interpretability of factors to make predictions about voter behavior while striking a balance between model complexity and predictive power.

The performance of the FA models was assessed using the same metrics as the other models, including R-squared, Adjusted R-squared, and Root Mean Squared Error. Although both FA models showed impressive predictive capabilities, they encountered difficulties with homoscedasticity, which was also observed in the PCA model.

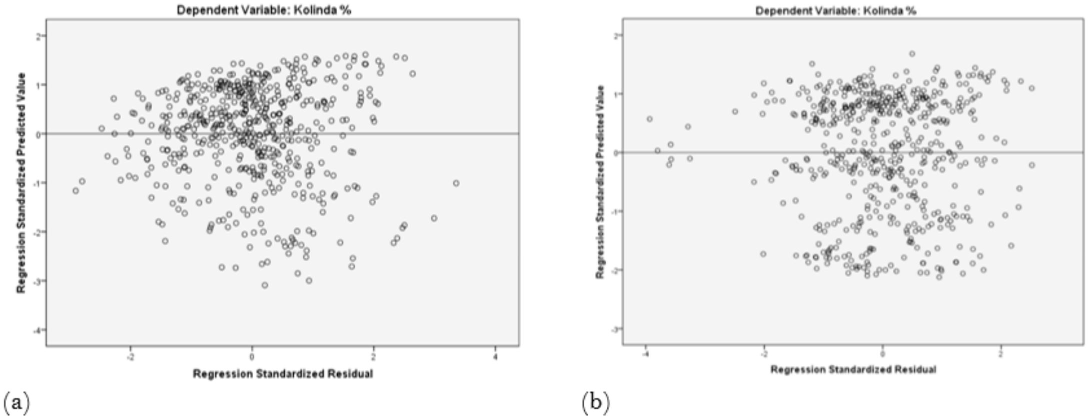


Figure 7.
The homoscedasticity of variance conditions for the FA1 (a) and FA2 (b) models.

DOI: 10.55214/25768484.v8i6.2307

Figure 7 presents a comparison of the homoscedasticity checks for the FA1 and FA2 models. The figure shows that there was an improvement in the dispersion of residuals when comparing to the LR1 model. However, the FA models still showed some heteroscedasticity, especially in areas where there were large residuals.

This study gained valuable insights into voter behavior by utilizing Factor Analysis, which focused on latent variables that hold theoretical significance. Despite the initial LR1 model being outperformed by the FA models, the LR2 model proved to be slightly more effective. This can be attributed to ongoing issues with homoscedasticity. Despite the limitations in predictive accuracy compared to LR2, the FA models' ability to offer interpretable and meaningful factors makes them a valuable tool for understanding the underlying dynamics in voter behavior.

5. Future Work

This study demonstrates the effectiveness of dimension reduction techniques like Principal Component Analysis (PCA) and Factor Analysis (FA) in improving the predictive accuracy and interpretability of linear regression models for electoral outcomes. Future research could explore other techniques, such as Independent Component Analysis (ICA) or t-Distributed Stochastic Neighbor Embedding (t-SNE), to assess their impact on electoral data modeling. Additionally, non-linear models like decision trees, random forests, or neural networks could be investigated in combination with dimension reduction methods. Expanding this research to other electoral contexts and incorporating temporal aspects, as well as improving data collection and preprocessing, would further enhance model accuracy. Understanding causal connections between socio-demographic factors and voting patterns through methods like structural equation modeling (SEM) and exploring the practical applications of these models in electoral strategies could provide valuable insights.

6. Conclusions

This study investigates the use of dimension reduction techniques, such as Principal Component Analysis (PCA) and Factor Analysis (FA), to improve the accuracy and interpretability of linear regression models in the context of the Croatian presidential election. These techniques were used to predict voter behavior by applying them to socio-demographic and political variables, while also addressing common challenges like multicollinearity and model complexity. The model that underwent PCA transformation showed a noteworthy decrease in multicollinearity, resulting in a more stable and easily understandable model when compared to the initial linear regression model (LR1). Although PCA enhanced the model's performance compared to LR1, it did not completely resolve the homoscedasticity issues. As a result, the model was found to be less robust than the optimized LR2 model. Factor Analysis has been found to be a highly effective tool in revealing latent variables that can explain underlying patterns in the data. The FA-based models (FA1 and FA2) demonstrated superior performance compared to the original LR1 model by offering a more significant and easily understandable set of predictors. However, like PCA, the FA models still faced difficulties with homoscedasticity, resulting in slightly lower predictive accuracy compared to the LR2 model. Out of the various models that were created, the LR2 model stood out as the most effective. This model was refined by carefully selecting predictor variables and addressing homoscedasticity. The study showcased the superior levels of predictive accuracy and model stability, surpassing the performance of both the PCA and FA models.

It is suggested by the findings that dimension reduction techniques, such as PCA and FA, have value in enhancing model interpretability and decreasing multicollinearity. However, it is important to still consider the assumptions that underlie linear regression models, such as homoscedasticity. It is crucial to combine these techniques with thorough model validation and refinement to achieve predictions that are robust and reliable.

Ensuring the reliability and accuracy of linear regression models is dependent on the crucial assumption of homoscedasticity. This assumption states that the variance of residuals remains constant across all levels of independent variables. When this condition is satisfied, the model's predictions exhibit consistency and lack bias, enabling accurate estimation of the relationships between predictors and the dependent variable. When homoscedasticity is not upheld, there is a possibility for standard errors to be biased, which can result in inaccurate conclusions regarding the significance of predictors. In addition, the efficiency of estimators may be affected, and the use of goodness-of-fit measures such as R-squared can lead to potential misinterpretations. The LR2 model was found to be the most reliable and accurate in making predictions. This highlights the significance of addressing homoscedasticity in order to develop robust regression models.

This study adds to the ongoing discussion on the use of dimension reduction in predictive modeling. This research provides valuable insights into the potential of these techniques to improve the accuracy and interpretability of models used in predicting voter behavior. These findings offer valuable tools for both academic research and practical applications in political strategy.

Ultimately, after considering the advantages of PCA and FA in handling high-dimensional data and enhancing model performance, it is evident that the LR2 model emerges as the most dependable approach in this study due to its exceptional ability to manage homoscedasticity. Further exploration of alternative dimension reduction techniques and their integration with both linear and non-linear modeling approaches is recommended for future research to enhance predictive accuracy in electoral studies.

Data Availability Statement: The data supporting this study's findings are available on public services on government sites in Croatia.

Copyright:

References

[1] Shang, H.L. Dynamic Principal Component Regression: Application to Age-Specific Mortality Forecasting. ASTIN Bull 2019, 49, 619–645. doi:10.1017/asb.2019.20.

[2] García-Santillán, A.; Escalera-Chávez, M.; Venegas-Martínez, F. Principal Components Analysis and Factorial Analysis to Measure Latent Variables in a Quantitative Research: A Mathematical Theoretical Approach. BSMaSS 2013, 7, 3–12. doi: 10.18052/WWW.SCIPRESS.COM/BSMASS.7.3.

[3] Wang, F. Factor Analysis and Principal-Components Analysis. International Encyclopedia of Human Geography 2009, 1–7., doi: 10.1016/B978-008044910-4.00434-X

[4] Hansen, B.E. Reply to: Comment on "A Modern Gauss–Markov Theorem". ECTA 2024, 92, 925–928. doi:10.3982/ECTA22362.

[5] Brown, J.D. Generalized Least Squares Estimation. Advanced Statistics for the Behavioral Sciences 2018, 189–217. doi: 10.1007/978-3-319-93549-2_6.

[6] Larocca, R. Reconciling Conflicting Gauss-Markov Conditions in the Classical Linear Regression Model. Political Analysis 2005, 13, 188–207. doi: 10.1093/pan/mpi011.

[7] White, H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica 1980, 48, 817. doi: 10.2307/1912934

[8] Wolfe, S.J. A Note on the Complete Convergence of Stable Distribution Functions. The Annals of Mathematical Statistics 1972, 43, 363–364. doi: 10.1214/aoms/1177692733

[9] Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning; Springer New York 2009; doi: 10.1007/978-0-387-84858-7.

[10] Little, R.J.A.; Rubin, D.B. Statistical Analysis with Missing Data. Wiley Series in Probability and Statistics 2002. doi: 10.1002/9781119013563.

[11] Takahashi, M. Statistical Inference in Missing Data by MCMC and Non-MCMC Multiple Imputation Algorithms: Assessing the Effects of Between-Imputation Iterations. CODATA 2017, 16, 37. doi: 10.5334/DSJ-2017-037.

[12] Martin, E. Normalizing Data. Experiments of the Mind 2022, 84–108. doi: 10.23943/princeton/9780691230719.003.0005.

[13] Schreiber-Gregory, D.N. Ridge Regression and Multicollinearity: An In-Depth Review. MAS 2018, 13, 359–365. doi: 10.3233/MAS-180446.

[14] James, G.; Witten, D.; Hastie, T.; Tibshirani, R. An Introduction to Statistical Learning. Springer New York 2013. doi: 10.1007/978-1-4614-7138-7.

[15] Jolliffe, I.T. Principal component analysis. Springer-Verlag, New York 2002. doi: 10.1007/b98835.

[16] Abdi, H.; Williams, L.J. Principal Component Analysis. WIREs Computational Stats 2010, 2, 433–459. doi: 10.1002/wics.101.

[17] Bialkowski, S.E. Data Analysis in the Shot Noise Limit. Part III: Adaptive Data Smoothing Filters. Journal of Chemometrics 1990, 4, 271–289. doi: 10.1002/cem.1180040403.

[18] Kaiser, H.F. The Application of Electronic Computers to Factor Analysis. Educational and Psychological Measurement 1960, 20, 141–151. doi: 10.1177/001316446002000116.

[19] Cattell, R.B. The Scree Test for The Number of Factors. Multivariate Behavioral Research 1966, 1, 245–276. doi: 10.1207/s15327906mbr0102_10.

[20] Bartholomew, D.J. Factor Analysis and Latent Variable Modelling. International Encyclopedia of Statistical Science 2011, 501–503. doi: 10.1007/978-3-642-04898-2_247.

[21] Hirose, K.; Kim, S.; Kano, Y.; Imada, M.; Yoshida, M.; Matsuo, M. Full Information Maximum Likelihood Estimation in Factor Analysis with a Large Number of Missing Values. Journal of Statistical Computation and Simulation 2015, 86, 91–104. doi: 10.1080/00949655.2014.995656.

[22] Cattell, R.B.; Burdsal Jr., C.A. The Radial Parcel Double Factoring Design: A Solution to the Item-Vs-Parcel Controversy. Multivariate Behavioral Research 1975, 10, 165–179. doi: 10.1207/s15327906mbr1002_3.

[23] Izard, C.E.; Rosenberg, N. Effectiveness of a Forced-Choice Leadership Test Under Varied Experimental Conditions. Educational and Psychological Measurement 1958, 18, 57–62. doi: 10.1177/001316445801800105.

[24] Horn, J.L. A Rationale and Test for the Number of Factors in Factor Analysis. Psychometrika 1965, 30, 179–185. doi: 10.1007/BF02289447.

[25] Lucas, X. Review of The Art of Art Therapy. Contemporary Psychology 1985, 30, 995–995. doi: 10.1037/023426.

[26] Spence, I.; Graef, J. The Determination of The Underlying Dimensionality of An Empirically Obtained Matrix of Proximities. Multivariate Behavioral Research 1974, 9, 331–341. doi: 10.1207/s15327906mbr0903_8.

[27] Onyutha, C. From R-squared to coefficient of model accuracy for assessing "goodness-of-fits". Geoscientific Model Development Discussions 2020, 1-25. doi: 10.5194/gmd-2020-51.

[28] Hyndman, R.J.; Koehler, A.B. Another Look at Measures of Forecast Accuracy. International Journal of Forecasting 2006, 22, 679–688. doi: 10.1016/j.ijforecast.2006.03.001.

[29] Yang, X.-S.; Forbes, A.B. Model and Feature Selection in Metrology Data Approximation. Springer Proceedings in Mathematics 2010, 293–307. doi: 10.1007/978-3-642-16876-5_14.