



COMPARATIVE ANALYSIS OF REGRESSION METHODS IN SMALL-SAMPLE AGRICULTURAL TIME SERIES: THE CASE OF TURKISH BEEKEEPING DATA

Mehmet BEYAZ^{1*}, Lütfi BAYYURT¹

¹Tokat Gaziosmanpaşa University, Faculty of Agriculture, Department of Animal Science, 60250, Tokat, Türkiye

Abstract: Accurate modeling of agricultural time series with a limited number of observations is a significant challenge, as multicollinearity and small sample sizes may reduce the reliability of conventional regression methods. In this study, the predictive performance of five different regression techniques was compared for modeling real honey prices using Turkish beekeeping data covering the period 2015–2025. Ordinary Least Squares (OLS) regression, Ridge Regression, Elastic Net Regression, Principal Component Regression (PCR), and Partial Least Squares Regression (PLSR) were evaluated using a nested leave-one-out cross-validation approach. Model performance was compared based on the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), and predicted residual sum of squares (PRESS). Among the evaluated models, PLSR demonstrated the highest predictive performance under cross-validation, with an R^2 of 0.972, RMSE of 1.937, MAE of 1.563, and PRESS of 41.290. Correlation analysis revealed substantial multicollinearity among the explanatory variables, supporting the use of latent-variable-based regression methods. The findings indicate that PLSR provides a robust and reliable approach for modeling small-sample agricultural time series and represents a potentially applicable method for agricultural price forecasting and decision-making processes in the beekeeping sector.

Keywords: Agricultural time series, Partial Least Squares Regression, Nested cross-validation, Multicollinearity, Small-sample regression

*Corresponding author: Tokat Gaziosmanpaşa University, Faculty of Agriculture, Department of Animal Science, 60250, Tokat, Türkiye

E mail: mehmet.beyaz6622@gop.edu.tr (M. BEYAZ)

Mehmet BEYAZ  <https://orcid.org/0009-0005-3352-4454>

Lütfi BAYYURT  <https://orcid.org/0000-0003-2613-9302>

Received: February 07, 2026

Accepted: April 19, 2026

Published: June 15, 2026

Cite as: Beyaz, M., & Bayyurt, L. (2026). Comparative analysis of regression methods in small-sample agricultural time series: The case of Turkish beekeeping data. *Black Sea Journal of Statistics*, 2(1), 15-27.

1. Introduction

With the increasing adoption of digital technologies in agricultural production systems, the volume of data obtained from sensor networks, remote sensing applications, farm management record systems, and official statistics has increased substantially. These developments have accelerated the use of data-driven modeling approaches in various fields, including production process monitoring, yield forecasting, disease management, and economic analysis (Guindani et al., 2024). In particular, reliable forecasting of agricultural commodity prices is of great importance for enabling producers to plan production, develop marketing strategies, mitigate market risks, and formulate policies aimed at ensuring food supply security (Jin et al., 2024). However, agricultural prices are shaped by the combined effects of numerous interrelated factors, such as production volume, climatic conditions, input costs, supply-demand dynamics, and macroeconomic indicators, making price forecasting a statistically complex problem (Avinash et al., 2024).

In recent years, a substantial proportion of studies on agricultural price forecasting have shifted toward

artificial intelligence, deep learning, and hybrid modeling approaches (Guindani et al., 2024). However, the vast majority of these studies rely on high-frequency time series, such as daily, weekly, or monthly data, and datasets containing thousands of observations. In contrast, as in many countries, key indicators of agricultural production in Türkiye are predominantly published as annual official statistics. Consequently, researchers, particularly in long-term economic analyses, are often required to work with datasets consisting of a limited number of observations. The overfitting of forecasting models to the training data, overestimation of model performance, and reduced generalizability to new observations under small sample sizes are among the major methodological challenges associated with such studies (Vabalas et al., 2019; Pavlou et al., 2024; Ghasemzadeh et al., 2024).

Another important problem encountered in small-sample agricultural datasets is the high degree of linear dependence among explanatory variables. Since variables such as the number of beekeeping operations, production volume, hive population, yield, and climatic indicators represent different components of the same



production system, strong correlations among these variables are expected. Such dependence among explanatory variables can lead to unstable coefficient estimates, inflated standard errors, and excessive sensitivity of the model to small changes in the data in ordinary least squares regression (Dormann et al., 2013; Schreiber-Gregory, 2018). The effects of multicollinearity become particularly pronounced in datasets where the number of observations approaches the number of explanatory variables, potentially leading to poorer out-of-sample predictive performance of conventional regression models (Pavlou et al., 2024).

As small sample sizes and multicollinearity limit the performance of conventional linear regression models, alternative regression approaches have been developed in recent years to mitigate these problems. Regularization-based methods, such as Ridge Regression and Elastic Net, aim to control regression coefficients, reduce prediction variance, and improve model generalizability (Hoerl & Kennard, 1970; Zou and Hastie, 2005). Similarly, dimensionality-reduction methods, including Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR), can mitigate the adverse effects of multicollinearity by representing highly correlated explanatory variables through a smaller number of components (Jolliffe, 1982; Wold et al., 1984). Although the performance of these methods across different data structures has been investigated in numerous studies, the relative effectiveness of each approach has been reported to vary depending on the data structure, sample size, and degree of association among the explanatory variables (James et al., 2021; Hastie et al., 2021). Therefore, comparing different regression methods under the same conditions is particularly important for small-sample agricultural datasets.

Reliable evaluation of model performance depends not only on the selection of an appropriate regression method but also on the use of an appropriate validation strategy. In small-sample datasets, conducting model development and model evaluation using the same data can result in an overestimation of predictive performance (Wainer and Cawley, 2021). Similarly, applying preprocessing procedures such as scaling, variable selection, or dimensionality reduction to the entire dataset before the validation process may lead to information leakage and consequently to a biased assessment of model performance (Parvande et al., 2020; Moscovich and Rosset, 2022). Therefore, particularly in the fields of biometry, bioinformatics, and applied statistics, the use of validation approaches that allow model development and performance evaluation to be conducted independently has been increasingly recommended in recent years (Yates et al., 2023; Pavlou et al., 2024).

Recent research on agricultural commodity price forecasting indicates that studies have largely focused on high-frequency time series or large-scale datasets. These

studies have primarily concentrated on the predictive performance of machine learning, deep learning, and hybrid modeling approaches (Guindani et al., 2024). In contrast, relatively few studies have systematically compared conventional linear regression with regularization- and dimensionality-reduction-based regression methods under the same validation strategy using small-sample agricultural datasets derived from annual official statistics. Furthermore, a substantial proportion of existing studies have evaluated model performance based on training data or single-level cross-validation results, without adequately accounting for the potentially optimistic estimates of predictive performance that may arise under small-sample conditions (Wainer and Cawley, 2021; Pavlou et al., 2024). This indicates that methodological uncertainty remains regarding which regression approach can provide more reliable and generalizable results for small-sample agricultural time series.

In this study, the performance of Ordinary Least Squares (OLS) Regression, Ridge Regression, Elastic Net, Principal Component Regression (PCR), and Partial Least Squares Regression (PLSR) was compared for forecasting real honey prices using beekeeping indicators from Türkiye for the period 2015–2025. All models were evaluated under the same data preprocessing procedure and a nested Leave-One-Out Cross-Validation (Nested LOOCV) framework designed to prevent information leakage. Thus, the objective of the study was not only to compare the predictive accuracy of different regression methods but also to provide a methodological framework for reliable model evaluation under conditions of small sample size and strong dependence among explanatory variables. In this respect, the study is expected to contribute to the literature on the selection of appropriate regression methods for small-sample agricultural time series.

2. Materials and Methods

2.1. Materials

The data used in this study were obtained from official beekeeping and precipitation statistics published by the Turkish Statistical Institute (TÜİK) and the Turkish State Meteorological Service (MGM) for the period 2015–2025. The dataset includes annual observations on the number of beekeeping operations, the number of modern and traditional beehives, honey production, beeswax production, annual precipitation, honey yield, and honey prices. The nominal honey prices used as the dependent variable in the analyses were converted into real honey prices using the Consumer Price Index (CPI) to eliminate the effects of inflation across years. All data used in the study are publicly available and were obtained from the official TÜİK database. The variables used in the analysis and their units of measurement are presented in Table 1.

Table 1. Descriptive statistics of the variables included in the regression analyses

Variable	Unit	Mean	SD	CV (%)	Minimum	Maximum
Real honey price	TRY kg ⁻¹	49.17	12.21	24.84	37.32	73.14
Enterprises	Count	88,715.91	7,457.25	8.41	80,675	100,399
Modern hives	Count	8,206,061.91	497,584.87	6.06	7,525,652	8,969,387
Traditional hives	Count	228,272.00	25,757.18	11.28	194,406	277,089
Honey production	t	106,538.82	7,758.87	7.28	95,492.31	118,297.46
Beeswax production	t	3,989.36	464.66	11.65	3,258.09	4,755.53
Annual rainfall	mm	555.66	55.74	10.03	502.30	659.40
Honey yield	kg hive ⁻¹	12.68	1.25	9.86	10.66	14.32
Beeswax ratio	%	3.74	0.33	8.82	3.35	4.40

Descriptive statistics were calculated from annual observations collected between 2015 and 2025 (n=11). SD= standard deviation, CV= coefficient of variation, TRY= Turkish Lira.

2.2. Methods

2.2.1. Data preprocessing and descriptive statistics

Prior to the regression analyses, the dataset was examined for missing observations, duplicate records, and inconsistent data entries, and only complete and consistent observations were included in the analyses. The nominal honey prices used as the dependent variable were converted into real prices using the Consumer Price Index (CPI) to eliminate the effects of inflation across years. To assess the general characteristics of the dataset, the mean, standard deviation, minimum value, maximum value, and coefficient of variation (CV, %) were calculated for all variables. These descriptive statistics were used not only to characterize the central tendency and dispersion of the variables but also to facilitate comparisons of relative variability among variables measured on different scales.

Because the independent variables were measured on different scales, z-score standardization was applied prior to the analyses to eliminate the effects of scale differences in the regression models, as shown in equation 1.

$$z_i = \frac{x_i - \bar{x}}{s} \quad (1)$$

Here, x_i denotes the corresponding observation value, μ represents the variable mean, and σ denotes the standard deviation. Following standardization, all variables were rescaled to have a mean of zero and a standard deviation of one. This procedure contributes to obtaining more stable and comparable model estimates, particularly for scale-sensitive methods such as Ridge Regression, Elastic Net, PCR, and PLSR (Sharma et al., 2024; Tilevik, 2026). Finally, standardized boxplots were generated and visually examined to assess the overall structure of the data distributions and identify potential outliers.

2.2.2. Trend analysis

The non-parametric Mann-Kendall trend test was applied to determine the temporal trends of the variables included in the study. This test is widely used in environmental, agricultural, and hydrological studies because it does not require any assumption regarding the normality of the data distribution and provides reliable

results for small-sample time series (Meshram et al., 2020; Gadedjisso-Tossou et al., 2021; Yue and Pilon, 2004). The Mann-Kendall test statistic was calculated as shown in equation 2:

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^n \text{sgn}(x_j - x_i) \quad (2)$$

Here, n denotes the number of observations, while x_i and x_j represent the i th and j th observation values, respectively, and sgn denotes the sign function. A positive S value indicates an increasing trend, whereas a negative S value indicates a decreasing trend. In addition to determining the direction of the trend, the Theil-Sen slope estimator was used to quantify its magnitude. Since this method is based on the median of the slope estimates calculated from all pairs of observations, it is less sensitive to outliers and provides reliable slope estimates, particularly for short time series (Sen, 1968; Bal, 2024). The slope coefficient was calculated using equation 3:

$$\beta = \text{Median} \left(\frac{x_j - x_i}{j - i} \right), i < j \quad (3)$$

Here, β represents the median rate of change of the variable per unit of time. Positive values indicate an increasing trend, whereas negative values indicate a decreasing trend. In this study, the statistical significance of the trend was assessed using the Mann-Kendall test, while the direction and magnitude of the trend were jointly interpreted using the Theil-Sen slope estimator. Thus, the temporal variation characteristics of all variables included in the study were evaluated independently of parametric assumptions.

2.2.3. Correlation analysis and assessment of multicollinearity

Prior to applying the regression models, Pearson correlation analysis was performed to determine the linear relationships among the variables. The Pearson correlation coefficient is one of the most widely used parametric methods for assessing the direction and strength of the linear relationship between two

continuous variables (Šverko et al., 2022). The correlation coefficient was calculated using equation 4:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (4)$$

Here, r denotes the Pearson correlation coefficient, x_i and y_i represent the observation values, and $\bar{x}$ and $\bar{y}$ denote the means of the corresponding variables. The correlation coefficient ranges from -1 to $+1$, with increasing absolute values indicating a stronger linear relationship between the variables. A high degree of linear association among the independent variables may reduce the stability of regression coefficient estimates and complicate model interpretation. Therefore, the problem of multicollinearity was assessed using the Variance Inflation Factor (VIF) (Kyriazos and Poga., 2023; Faraz et al., 2023; Şahin 2026). The VIF was calculated using equation 5:

$$VIF_i = \frac{1}{1 - R_i^2} \quad (5)$$

Here, R_i^2 denotes the coefficient of determination of the auxiliary regression model constructed for the corresponding independent variable using the other independent variables. An increase in the VIF value indicates a higher degree of linear association between the corresponding variable and the other explanatory variables. In this study, VIF values were used to assess multicollinearity among the explanatory variables prior to the application of the regression models. Correlation analysis and VIF assessment were conducted solely to examine the structure of the data, and no variables were excluded from the analyses. The primary reason for this approach is that the methods compared in the study, including Ridge Regression, Elastic Net, Principal Component Regression (PCR), and Partial Least Squares Regression (PLSR), are regression techniques specifically developed to address multicollinearity. Thus, all models were compared using the same set of explanatory variables, ensuring a fair evaluation across the methods.

2.2.4. Regression models

To compare the performance of different regression approaches in forecasting real honey prices, a total of five regression methods were evaluated, comprising one conventional and four modern approaches. These included Ordinary Least Squares (OLS) Regression, Ridge Regression, Elastic Net Regression, Principal Component Regression (PCR), and Partial Least Squares Regression (PLSR). These methods differ in their approaches to addressing multicollinearity among explanatory variables. While the OLS model represents the conventional linear regression approach, Ridge and Elastic Net are regularization-based methods. In contrast, PCR and PLSR are dimensionality-reduction-based regression methods that perform modeling by transforming explanatory variables into lower-dimensional components (Önder et al., 2025; Tirink and

Önder., 2022). The same dependent and independent variables were used in all regression models to enable direct comparison among the methods. Model hyperparameters were optimized using only the training data, and all performance evaluations were conducted under the same validation strategy. Thus, differences among the methods were attributable solely to their respective modeling approaches. The following subsections provide detailed descriptions of each regression method in terms of its theoretical foundations, mathematical formulation, and implementation in the present study.

Ordinary Least Squares Regression

Ordinary Least Squares (OLS) Regression is the most fundamental regression method used to model the linear relationship between a dependent variable and one or more independent variables. The OLS approach estimates the regression coefficients by minimizing the sum of squared errors between the observed values and the values predicted by the model (Burton., 2021; Mahaboob et al., 2018). The general linear regression model is expressed by equation 6.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i \quad (6)$$

Here, Y denotes the dependent variable, X represents the independent variables, β_0 denotes the intercept, β represents the regression coefficients, and ε denotes the error term. In the OLS method, the regression coefficients are estimated as the values that minimize the Residual Sum of Squares (RSS), as calculated in equation 7.

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (7)$$

Minimizing this objective function yields the regression coefficients in matrix form, as shown in equation 8.

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (8)$$

Here, X denotes the matrix of explanatory variables, Y represents the dependent variable vector, and $\hat{\beta}$ denotes the estimated regression coefficients. For OLS regression to provide reliable results, several fundamental assumptions are expected to be satisfied, including linearity, independence of error terms, constant variance (homoscedasticity), approximately normally distributed error terms, and the absence of severe multicollinearity among the explanatory variables (Aslam., 2014). Particularly in datasets with high multicollinearity, the variance of the regression coefficients may increase, thereby reducing the predictive stability of the model. In this study, the OLS model was used as the baseline reference model without regularization or dimensionality reduction. The results were evaluated by comparison with the Ridge Regression, Elastic Net, Principal Component Regression (PCR), and Partial Least Squares Regression (PLSR) models under the same validation strategy.

Ridge Regression

Ridge Regression is a regularization-based regression method developed to improve the stability of OLS regression in situations where a high degree of multicollinearity exists among explanatory variables. The Ridge approach shrinks coefficient estimates by adding a penalty term based on the squared regression coefficients (L_2 penalty term) to the residual sum of squares, thereby reducing coefficient variance and improving model generalizability (Owen 2007; Hastie et al., 2022). The objective function for Ridge Regression is presented in equation 9.

$$\min_{\beta} \left\{ \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\} \quad (9)$$

Here, λ denotes the regularization parameter. When $\lambda=0$, the model is equivalent to the conventional OLS regression, whereas as the value of λ increases, the regression coefficients are increasingly shrunk toward zero. However, because Ridge Regression does not set coefficients exactly to zero, all explanatory variables included in the model are retained. In Ridge Regression, the coefficient estimates are obtained in matrix form as shown in equation 10.

$$\hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T y \quad (10)$$

Here, I denotes the identity matrix. The addition of the penalty term reduces numerical instability that may arise when taking the inverse of the $X^T X$ matrix and substantially mitigates the adverse effects of multicollinearity on regression coefficients (Wang et al., 2021). The performance of Ridge Regression depends largely on the selection of the regularization parameter (λ). While very small λ values produce results similar to those of the OLS model, excessively large values may cause the coefficients to be shrunk excessively, resulting in underfitting. Therefore, in this study, the optimal λ value was determined exclusively using the training data within the inner loop of the Nested Leave-One-Out Cross-Validation (Nested LOOCV) procedure. Ridge Regression was applied in this study particularly to mitigate the effects of the multicollinearity identified among the explanatory variables on model performance and to evaluate the predictive performance of regularization-based approaches. Model performance was compared with that of the other regression methods under the same validation procedure.

Elastic Net Regression

Elastic Net Regression is a hybrid regularization method that combines the advantages of Ridge and LASSO regression. By combining the stability of Ridge Regression against multicollinearity with the variable selection capability of the LASSO method, Elastic Net can provide effective results, particularly in datasets containing highly correlated explanatory variables (Zou and Hastie, 2005). Therefore, Elastic Net is widely preferred in regression problems where strong

relationships exist among explanatory variables. The objective function of Elastic Net is expressed in equation 11.

$$\min_{\beta} \left\{ \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \left[(1 - \alpha) \sum_{j=1}^p \beta_j^2 + \alpha \sum_{j=1}^p |\beta_j| \right] \right\} \quad (11)$$

Here, λ denotes the penalty parameter that determines the level of regularization, while α represents the mixing parameter that controls the contribution of the L_1 and L_2 penalty terms to the model. When $\alpha=0$, the model is equivalent to Ridge Regression, whereas when $\alpha=1$, it becomes equivalent to LASSO Regression. One of the key advantages of Elastic Net is its ability to jointly account for groups of highly correlated variables. While the LASSO method often selects only one variable from a group of highly correlated variables, Elastic Net can retain these variables simultaneously in the model and produce more stable coefficient estimates (Zou and Hastie, 2005; Hastie et al., 2022). This characteristic provides an important advantage in studies involving explanatory variables that may exhibit strong interrelationships, such as agricultural production data. In this study, the Elastic Net hyperparameters, λ and α , were optimized exclusively using the training data within the inner loop of the Nested LOOCV approach. The model fitted using the optimal parameters was evaluated in the outer validation loop, and its performance was compared with that of the other regression methods using the same evaluation metrics.

Principal Component Regression

Principal Component Regression (PCR) is a regression method that combines Principal Component Analysis (PCA) with linear regression to reduce the effects of multicollinearity among explanatory variables. In PCR, highly correlated explanatory variables are first transformed into new components that are linearly uncorrelated with one another, and the regression model is subsequently fitted using these components (Lee et al., 2015; Xu and Hsu., 2019).

The principal components are expressed as linear combinations of the explanatory variables, as shown in equation 12.

$$PC_k = a_{1k}X_1 + a_{2k}X_2 + \dots + a_{pk}X_p \quad (12)$$

Here, PC_j denotes the j th principal component, p_{jk} represents the loading coefficients associated with the corresponding component, and X_k denotes the explanatory variables. The principal components are orthogonal to one another and are ordered according to the proportion of the total variance in the explanatory variables that they explain. In the PCR method, the regression model is fitted using the selected principal components rather than the original explanatory variables, as expressed in equation 13.

$$y = \beta_0 + \beta_1 PC_1 + \beta_2 PC_2 + \dots + \beta_m PC_m + \varepsilon \quad (13)$$

Here, k denotes the number of principal components included in the regression model. Thus, the problem of multicollinearity arising from highly correlated variables is substantially reduced. The performance of PCR depends largely on the number of principal components included in the model. Using too few components may result in the loss of important information, whereas including too many components may increase model complexity (Boggy and Woolf, 2010). Therefore, in this study, the optimal number of components was determined exclusively using the training data within the inner loop of the Nested Leave-One-Out Cross-Validation approach. A major limitation of PCR is that the principal components are constructed solely based on the variance in the explanatory variables. Consequently, components with high variance but weak associations with the dependent variable may be included in the model, whereas some components containing important information about the dependent variable may be overlooked (Agarwal et al., 2019). Therefore, the performance of PCR was compared with that of the PLSR method, which also incorporates information from the dependent variable.

Partial Least Squares Regression

Partial Least Squares Regression (PLSR) is a dimensionality-reduction-based regression method widely used for data structures characterized by multicollinearity and limited sample sizes. In addition to summarizing the variance in the explanatory variables, PLSR aims to maximize the covariance between the latent components and the dependent variable (Mevik and Wehrens, 2007; Mehmood et al., 2012). Owing to this characteristic, PLSR can provide more effective results than PCR for predicting the dependent variable. In the PLSR method, the explanatory variable matrix X and the dependent variable vector Y are simultaneously decomposed into lower-dimensional latent variables, as shown in equation 14.

$$\begin{aligned} X &= TP^T + E \\ y &= Uq^T + F \end{aligned} \quad (14)$$

Here, T and U denote the score matrices of the explanatory and dependent variables, respectively; P and Q represent the loading matrices; and E and F denote the residual matrices. The primary objective of the PLSR algorithm is to identify linear components that maximize the covariance between the latent components derived from the explanatory and dependent variables. Thus, each component incorporates not only the variance in the explanatory variables but also information relevant to explaining the dependent variable (Helland, 1990). This characteristic can improve model performance, particularly in datasets with highly correlated explanatory variables. In the PLSR model, the dependent variable is predicted using the linear structure given in equation 15.

$$\hat{y} = XB \quad (15)$$

Here, B denotes the matrix of regression coefficients estimated by the PLSR algorithm. In PCR, the principal components are constructed solely according to the variance in the explanatory variables, whereas in PLSR, information from the dependent variable is also incorporated into the component construction process. Therefore, PLSR may provide superior predictive performance, particularly in studies characterized by highly correlated explanatory variables and a limited number of observations (Mehmood et al., 2012; Cheng and Sun, 2017). In this study, the optimal number of components to be used in the PLSR model was determined exclusively using the training data within the inner loop of the Nested Leave-One-Out Cross-Validation approach. The resulting optimal model was then evaluated in the outer validation loop, and its performance was compared with that of the other regression methods using the same evaluation metrics.

2.2.5. Nested Leave-One-Out Cross-Validation

To obtain an unbiased assessment of the generalizable predictive performance of the regression models, a Nested Leave-One-Out Cross-Validation (Nested LOOCV) approach was used for all analyses. This method is considered one of the most reliable resampling approaches because it separates model development from model evaluation and reduces optimistic estimates of performance arising from hyperparameter optimization (Varma and Simon, 2006; Cawley and Talbot, 2006). Nested LOOCV consists of a two-stage validation structure. The outer loop is used to evaluate the model's independent test performance, whereas the inner loop is used exclusively for hyperparameter optimization. Thus, the test data were not involved in the model development process at any stage, thereby preventing data leakage. In this study, because the dataset consisted of 11 annual observations, one observation was held out as the test set in each iteration of the outer loop, while the remaining observations constituted the training set. A second Leave-One-Out Cross-Validation procedure was then applied to the training set to determine the optimal hyperparameters for each model. The regularization parameter (λ) in Ridge Regression, λ and the mixing parameter (α) in the Elastic Net model, and the optimal number of components in the PCR and PLSR models were optimized exclusively using the training data. The OLS model, which does not involve hyperparameters, was fitted directly to the training data. At the end of the inner loop, the optimal model was refitted using the entire training set from the outer loop, and predictions were generated for the previously unseen test observation. This process was repeated for each observation so that all predictions were obtained from independent test observations. Thus, model selection and performance evaluation were completely separated, allowing the actual generalization performance of the models to be assessed (Cawley and

Talbot, 2006). The Nested LOOCV approach was preferred because the dataset contained a limited number of observations. While the Leave-One-Out method uses the largest possible training set in each iteration, the nested validation structure reduces bias arising from model selection and provides more reliable estimates of predictive performance, particularly in small-sample studies (Vabalas ve ark., 2019).

2.2.6. Model performance metrics

The predictive performance of the regression models was compared using the coefficient of determination (R^2), root mean squared error (RMSE), mean absolute error (MAE), prediction error sum of squares (PRESS), and predicted coefficient of determination (Predicted R^2) to jointly assess explanatory power and prediction error. The coefficient of determination (R^2) was calculated to evaluate the explanatory power of the models. To assess the magnitude of prediction error, the root mean squared error (RMSE) was calculated using the following equation. The mean absolute error (MAE) was calculated to determine the average absolute deviation of the model predictions. The prediction error sum of squares (PRESS) was used to assess the total prediction error obtained during the cross-validation process. Finally, the predicted coefficient of determination (Predicted R^2) was calculated to evaluate the explanatory power of the models for observations outside the training sample. In this study, all regression models were evaluated under the same Nested LOOCV procedure, and performance comparisons were based exclusively on the R^2 , RMSE, MAE, PRESS, and Predicted R^2 values obtained from out-of-sample predictions. The best-performing model was identified as the method that achieved the highest R^2 and Predicted R^2 values while providing the lowest RMSE, MAE, and PRESS values. All data preprocessing, statistical analyses, and regression modeling were performed using the Python 3.12 programming language. Pandas and NumPy were used for data manipulation, SciPy for statistical analyses, Scikit-learn for regression modeling and cross-validation procedures, and Matplotlib for graphical visualization. All analyses were conducted within the same computational environment, and all regression models were evaluated using the same data preprocessing steps and validation strategy.

3. Results

3.1. Descriptive Statistics and Data Distribution

The descriptive statistics for the variables used in the study are presented in Table 1. During the study period, the mean real honey price was 49.17 TL kg⁻¹, with a standard deviation of 12.21 and a coefficient of variation (CV) of 24.84%. Among the variables examined, the highest variability was observed in real honey prices. This finding suggests that honey prices may be influenced not only by production-related indicators but also by market conditions and economic factors. Evaluating real prices after adjusting for inflation allows

actual changes in prices across years to be compared more reliably.

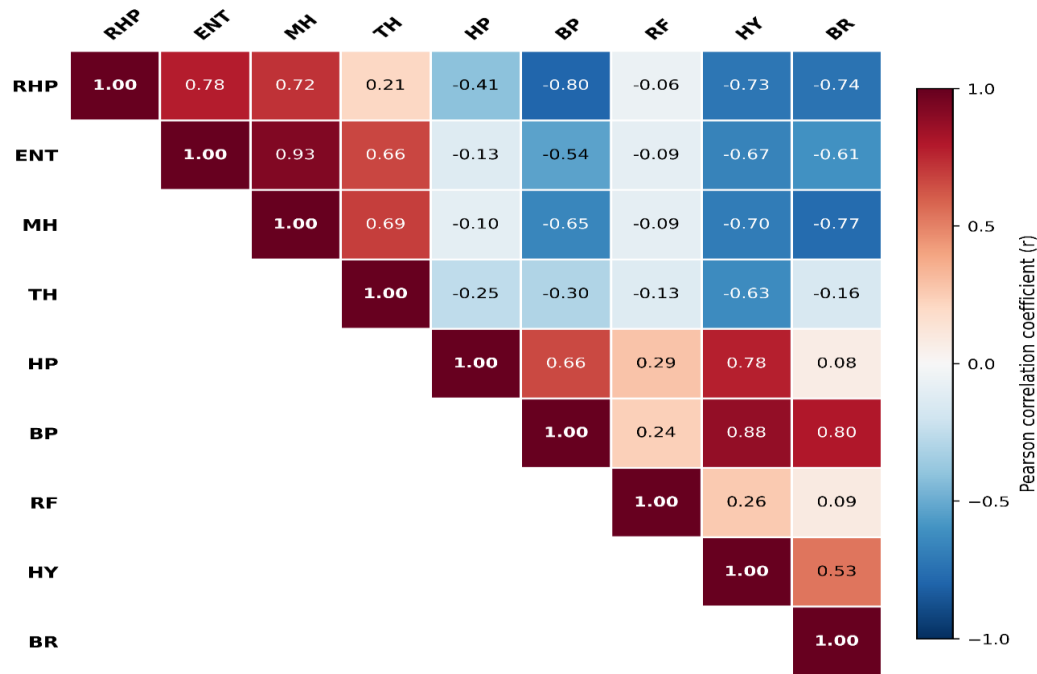
The mean number of beekeeping operations was 88.715, with a coefficient of variation of 8.41%, indicating relatively stable variation in the number of operations throughout the study period. The mean number of modern beehives was 8.206.061, with a coefficient of variation of 6.06%, representing one of the lowest levels of variability among all variables. In contrast, the mean number of traditional beehives was 228.272, with a coefficient of variation of 11.28%, indicating a higher degree of variability compared with modern beehives.

The mean honey production was 106.538.82 tons, with a coefficient of variation of 7.28%. Similarly, the mean honey yield per hive was 12.68 kg hive⁻¹, with a coefficient of variation of 9.86%, indicating that production efficiency remained relatively stable throughout the study period. The mean beeswax production was 3.989.36 tons, with a coefficient of variation of 11.65%, while the mean beeswax ratio was 3.74%, with a coefficient of variation of 8.82%. The mean annual precipitation was 555.66 mm, with a coefficient of variation of 10.03%. These findings indicate that production and environmental indicators generally exhibited low to moderate levels of variability.

Overall, the dataset consisted of explanatory variables exhibiting varying levels of variability. In particular, the relatively low variability observed in variables such as the number of modern and traditional beehives, production volume, and honey yield indicates a certain degree of stability across annual observations. In contrast, the higher variability of real honey prices compared with the other variables suggests that, in addition to production-related indicators, economic and sectoral factors may also influence price formation. The standardized distributions of the variables are presented in Figure 1. Examination of the boxplots based on z-scores indicates that the overall distributions of the variables were balanced and that there were no apparent outlying observations likely to substantially affect the results of the analyses. Furthermore, standardization enabled variables measured in different units to be evaluated on a common scale, thereby eliminating potential effects arising from differences in scale in the regression models.

3.2. Assessment of Temporal Trends

The annual changes in all variables examined in the study are presented in Figure 2. Overall, the graphs indicate that the number of beekeeping operations in the sector increased over the study period. The number of operations generally exhibited an upward trend throughout the years. This increase was relatively continuous and consistent, suggesting a stable expansion of the production capacity of the sector. Similarly, a clear increase in the number of modern beehives was observed throughout the study period.



RHP, real honey price; ENT, enterprises; MH, modern hives; TH, traditional hives; HP, honey production; BP, beeswax production; RF, annual rainfall; HY, honey yield; BR, beeswax ratio.

Figure 1. Upper-triangular Pearson correlation matrix among the variables included in the regression analyses.

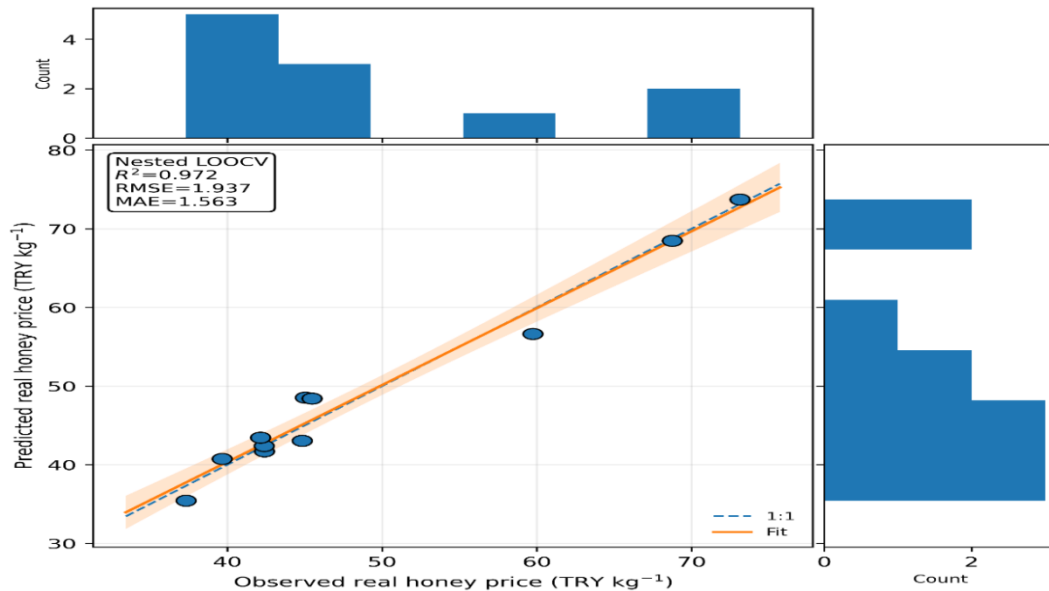


Figure 2. Observed versus nested leave-one-out cross-validated predictions of real honey price obtained from the partial least squares regression model.

The more consistent increase in the number of modern beehives compared with the growth in the number of beekeeping operations suggests that the use of modern beehives became increasingly widespread during the study period. In contrast, a pronounced decreasing trend was observed in the number of traditional beehives. This finding suggests a shift from traditional production practices toward modern beehive systems in Turkish beekeeping. When the trends in the numbers of modern and traditional beehives are considered together, they indicate a structural transformation in the production system.

Honey production remained relatively stable throughout the study period, despite limited year-to-year fluctuations. Honey yield per hive similarly did not exhibit substantial fluctuations. This indicates that variations in production volume are influenced not only by the number of beehives but also by environmental conditions and management practices. In contrast, real honey prices exhibited a more volatile pattern than the other variables. In particular, the increases observed in some years suggest that, in addition to production-related indicators, market conditions and economic factors may also play an important role in price

formation. The real price series obtained after adjusting for inflation provides a more comparable basis for evaluating price changes alongside production-related indicators.

Overall, the trend analyses revealed an increasing trend in modern production indicators and a decreasing trend in traditional production indicators. In contrast, production and yield indicators exhibited more limited changes. Although the findings indicate trends in different directions across production-related indicators during the study period, they suggest that real honey prices cannot be explained by a single production indicator alone. Rather, production capacity, environmental conditions, and structural changes within the sector should be considered jointly. Therefore, in the subsequent stage of the study, the predictive performance of different regression methods was compared.

3.3. Comparison of Regression Models

The Nested Leave-One-Out Cross-Validation results for the five regression methods are presented in Table 2. Overall, the models exhibited substantial differences in predictive performance. Although the conventional Ordinary Least Squares (OLS) model produced a relatively high R^2 value in terms of apparent fit, it showed lower performance than the other methods in out-of-sample validation. This finding suggests that the small sample size and high correlation among the explanatory variables may have limited the out-of-sample predictive performance of the conventional linear regression model. Ridge Regression produced more balanced predictions than the OLS model by shrinking the regression coefficients through the L_2 penalty term. This regularization approach particularly reduced the effects of multicollinearity and resulted in a marked improvement in cross-validation performance. However, because the Ridge model retains all variables in the model, it may not completely eliminate the shared information structure among highly correlated variables in some cases.

Although the Elastic Net model combines the Ridge and LASSO approaches, it did not provide a clear advantage over the Ridge model for this dataset. This may be because most of the explanatory variables in the dataset carry shared information, and variable selection may not have substantially improved predictive performance. The

PCR method largely eliminated the problem of multicollinearity by transforming the explanatory variables into orthogonal principal components. Consequently, it achieved better predictive performance than the OLS model. However, because the principal components in PCR are constructed solely based on the variance in the explanatory variables, some information that is important for explaining the dependent variable may not have been fully captured by the model.

The highest predictive performance was achieved by the PLSR model. The PLSR model yielded the highest Predicted R^2 (0.972) along with the lowest RMSE (1.937), MAE (1.563), and PRESS (41.290) values. These results indicate that PLSR achieved superior out-of-sample predictive performance for this dataset by simultaneously accounting for the shared variance among the explanatory variables and their covariance with the dependent variable. The superior performance of the PLSR model may be particularly attributable to the highly correlated structure of the dataset. Variables such as the number of modern beehives, number of beekeeping operations, honey production, and honey yield exhibited close relationships with one another. By effectively summarizing this shared information structure through latent components, PLSR provided higher out-of-sample predictive accuracy than the other methods. The observed and predicted real honey prices and the distribution of residuals for the best-performing PLSR model are presented in Figure 3. The close agreement between the observed and predicted values indicates that the model successfully represented the underlying structure of the data. The random distribution of residuals around zero further suggests that there was no pronounced systematic error in the model and that the prediction errors did not exhibit any clear systematic pattern.

Overall, the regularization- and dimensionality-reduction-based methods were found to achieve better out-of-sample predictive performance than conventional linear regression for this dataset, which was characterized by a small sample size and substantial multicollinearity. In particular, the PLSR method emerged as the best-performing approach among the methods evaluated, both in terms of predictive accuracy and model stability.

Table 2. Predictive performance of the evaluated regression models under nested leave-one-out cross-validation (LOOCV)

Model	Apparent R^2	LOOCV R^2 (95% CI)	RMSE (95% CI)	MAE (95% CI)	PRESS
PLSR	0.996	0.972 (0.606–0.992)	1.937 (1.204–2.513)	1.563 (0.898–2.255)	41.290
Ridge	0.996	0.962 (0.315–0.982)	2.259 (1.706–2.722)	2.040 (1.449–2.605)	56.121
PCR	0.996	0.961 (0.299–0.987)	2.304 (1.566–2.910)	1.970 (1.285–2.670)	58.403
Elastic Net	0.996	0.960 (0.367–0.981)	2.315 (1.769–2.753)	2.061 (1.403–2.652)	58.940
OLS	0.998	0.950 (0.140–0.973)	2.604 (1.844–3.201)	2.277 (1.526–3.009)	74.568

Apparent R^2 was calculated using the complete dataset. LOOCV R^2 , RMSE, and MAE are reported together with bootstrap 95% confidence intervals. PRESS = Prediction Residual Error Sum of Squares. Higher LOOCV R^2 and lower RMSE, MAE, and PRESS values indicate better predictive performance.

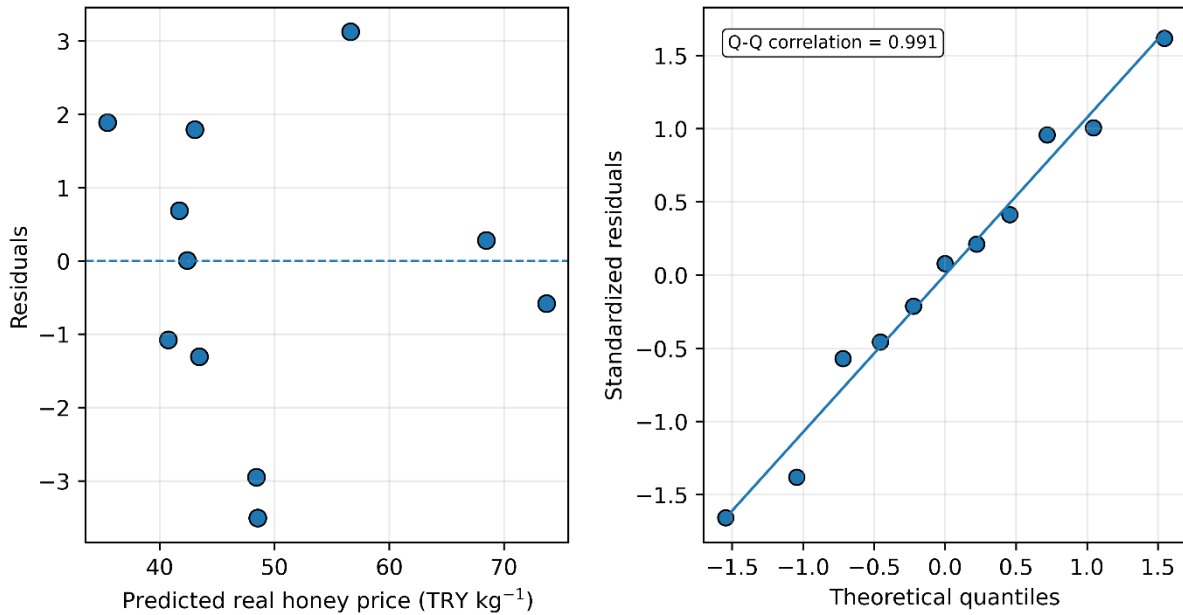


Figure 3. Residual diagnostics of the optimal partial least squares regression model based on nested leave-one-out cross-validation: (A) Residuals versus predicted values illustrating the absence of systematic prediction errors. (B) Normal Q-Q plot of standardized residuals for assessing residual normality.

4. Discussion

In this study, the predictive performance of OLS, Ridge Regression, Elastic Net, PCR, and PLSR models was compared using small-sample annual data from the Turkish beekeeping sector covering the 2015–2025 period. The Nested Leave-One-Out Cross-Validation results showed that the PLSR model provided the highest out-of-sample explanatory power while also achieving the lowest prediction errors. This finding indicates that conventional linear regression may not always be the most appropriate predictive approach for data structures characterized by substantial multicollinearity among explanatory variables and a limited number of observations. Similarly, methodological studies have reported that regularization- and dimensionality-reduction-based regression methods can provide more stable and generalizable predictions under such data conditions (Dormann et al., 2013; Hastie et al., 2021; James et al., 2021; Mehmood et al., 2012).

The superior performance of the PLSR model compared with the other methods in this study is consistent with its theoretical properties. Whereas PCR constructs latent components solely based on the variance in the explanatory variables, PLSR also accounts for the covariance between the explanatory and dependent variables during the component construction process (Wold et al., 1984; Helland, 1990). Consequently, the resulting components are not intended solely for dimensionality reduction but also represent linear combinations that contain the greatest amount of information relevant to prediction. This property can enhance the predictive performance of PLSR, particularly in data structures characterized by high correlations among explanatory variables, and helps explain its

widespread use in multivariate regression problems (Mehmood et al., 2012; Hastie et al., 2021). The highest out-of-sample explanatory power achieved by PLSR in the present study is also consistent with this theoretical expectation.

Although the dataset used in this study consisted of only 11 annual observations, the high out-of-sample performance of the PLSR model is noteworthy. In datasets with small sample sizes, increasing model complexity raises the risk of overfitting and may cause performance metrics based solely on the training data to overestimate model performance (Varma and Simon, 2006; Cawley and Talbot, 2006). In the present study, the use of Nested Leave-One-Out Cross-Validation separated the model selection and model performance evaluation processes, thereby reducing selection bias and enabling a more reliable assessment of out-of-sample predictive performance (Cawley and Talbot, 2006; Yates et al., 2023). Similarly, Pavlou et al. (2024) emphasized that the use of appropriate validation strategies in small-sample datasets is critical for preventing overly optimistic estimates of model performance. Therefore, the validation approach employed in this study is considered to have enhanced the reliability of the reported performance results.

Although the OLS model showed a high degree of fit on the training data, it performed worse than the PLSR and Ridge models in out-of-sample validation. This finding can be explained by the fact that multicollinearity increases the variance of regression coefficient estimates and reduces model stability in data structures with a limited number of observations (Dormann et al., 2013). Ridge Regression, in contrast, produced more stable results than the OLS model by reducing prediction

variance through the application of L_2 regularization to the coefficients (Hoerl and Kennard, 1970). However, because the Ridge model retains all explanatory variables in the model, it does not eliminate the shared information structure among highly correlated variables but rather stabilizes the coefficient estimates (Hastie et al., 2021). Indeed, both methodological studies and comparative applications have shown that Ridge Regression can provide more reliable and lower-variance predictions than conventional linear regression in data structures characterized by multicollinearity (James et al., 2021; Hastie et al., 2021).

Although the Elastic Net model combines the characteristics of Ridge and LASSO methods, it underperformed relative to PLSR in this study. This outcome suggests that since a large majority of the explanatory variables in the dataset are highly correlated, preserving the shared information structure for predictive performance may be more crucial than variable selection. Elastic Net yields particularly effective results in high-dimensional data structures and problems containing numerous irrelevant or weak explanatory variables (Zou and Hastie, 2005; Hastie et al., 2021). Conversely, in data structures where the number of explanatory variables is limited and all variables are biologically or economically meaningful, PLSR's ability to construct components that account for the shared information structure between the explanatory variables and the dependent variable can enhance predictive performance (Helland, 1990; Mehmood et al., 2012). The findings obtained in our study are aligned with this methodological expectation.

5. Conclusion

In this study, the performances of OLS, Ridge Regression, Elastic Net, PCR, and PLSR models in predicting real honey prices were compared using small-sample annual time-series data covering the 2015–2025 period for the Turkish beekeeping sector. All models were evaluated under the same data preprocessing workflow and a Nested Leave-One-Out Cross-Validation approach. The results demonstrated that the PLSR model was the most successful approach among the evaluated methods, providing the highest out-of-sample explanatory power along with the lowest prediction errors. In contrast, the OLS model exhibited lower generalizability, while Ridge Regression, Elastic Net, and PCR methods yielded better results than classical linear regression but fell short of PLSR's performance. The findings reveal that the PLSR method is a reliable and effective alternative for agricultural time series characterized by small sample sizes and multicollinearity. Furthermore, the Nested Leave-One-Out Cross-Validation approach utilized in this study allowed for an unbiased evaluation of the models' out-of-sample performance. In future studies, it is considered that predictive performance could be further enhanced by incorporating longer time series, regional data, and climatic variables into the models.

Author Contributions

The percentage of the authors' contributions are presented below. All authors reviewed and approved the final version of the manuscript.

	M.B.	L.B.
C	50	50
D	50	50
S	50	50
DCP	50	50
DAI	50	50
L	50	50
W	50	50
CR	50	50
SR	50	50
PM	50	50
FA	50	50

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

Conflict of Interest

The authors declare that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

References

- Agarwal, A., Shah, D., Shen, D., & Song, D. (2019). On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32.
- Aslam, M. (2014). Using heteroscedasticity-consistent standard errors for the linear regression model with correlated regressors. *Communications in Statistics—Simulation and Computation*, 43(10), 2353–2373.
- Avinash, G., Ramasubramanian, V., Ray, M., Paul, R. K., Godara, S., Nayak, G. H. H., Kumar, R. R., Manjunatha, B., Dahiya, S., & Iquebal, M. A. (2024). Hidden Markov guided deep learning models for forecasting highly volatile agricultural commodity prices. *Applied Soft Computing*, 158, 111557. <https://doi.org/10.1016/j.asoc.2024.111557>
- Bal, A. (2024). Improving the robustness of the Theil-Sen estimator using a simple heuristic-based modification. *Symmetry*, 16(6), 698. <https://doi.org/10.3390/sym16060698>
- Boggy, G. J., & Woolf, P. J. (2010). A mechanistic model of PCR for accurate quantification of quantitative PCR data. *PLOS One*, 5(8), e12355.
- Burton, A. L. (2021). *OLS (linear) regression*. In *The Encyclopedia of Research Methods in Criminology and Criminal Justice* (Vol. 2, pp. 509–514).
- Cawley, G. C., & Talbot, N. L. C. (2006). Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. *Bioinformatics*, 22(19), 2348–2355. <https://doi.org/10.1093/bioinformatics/btl386>
- Cheng, J. H., & Sun, D. W. (2017). Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle. *Food*

- Engineering Reviews*, 9(1), 36–49. <https://doi.org/10.1007/s12393-016-9147-1>
- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., García Marquéz, J. R., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27–46. <https://doi.org/10.1111/j.1600-0587.2012.07348.x>
- Faraz, A., Tirink, C., Önder, H., Şen, U., Ishaq, H. M., Tauqir, N. A., Waheed, A., & Nabeel, M. S. (2023). Usage of the XGBoost and MARS algorithms for predicting body weight in Kajli sheep breed. *Tropical Animal Health and Production*, 55(4), 276. <https://doi.org/10.1007/s11250-023-03700-6>
- Gadedjisso-Tossou, A., Adjegan, K. I., & Kablan, A. K. M. (2021). Rainfall and temperature trend analysis by Mann–Kendall test and significance for rainfed cereal yields in Northern Togo. *Sci*, 3(1), 17. <https://doi.org/10.3390/sci3010017>
- Ghasemzadeh, H., Hillman, R. E., & Mehta, D. D. (2024). Toward generalizable machine learning models in speech, language, and hearing sciences: Estimating sample size and reducing overfitting. *Journal of Speech, Language, and Hearing Research*, 67(3), 753–781. <https://doi.org/10.1044/2023.JSLHR-23-00273>
- Guindani, L. G., Oliveira, G. A., & Ribeiro, M. H. D. M. (2024). Exploring current trends in agricultural commodities forecasting methods through text mining: Developments in statistical and artificial intelligence methods. *Heliyon*, 10(23), e40568. <https://doi.org/10.1016/j.heliyon.2024.e40568>
- Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2), 949–986.
- Hastie, T., Tibshirani, R., & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
- Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, 17(2), 97–114.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67. <https://doi.org/10.1080/00401706.1970.10488634>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer.
- Jin, Y., Li, W., & Gil, J. M. (2024). Forecasting fish prices with an artificial neural network model during the tuna fraud. *Journal of Agriculture and Food Research*, 18, 101340. <https://doi.org/10.1016/j.jafr.2024.101340>
- Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, 31(3), 300–303. <https://doi.org/10.2307/2348005>
- Klein, N. (2024). Distributional regression for data analysis. *Annual Review of Statistics and Its Application*, 11, 321–346. <https://doi.org/10.1146/annurev-statistics-040722-053607>
- Kyriazos, T., & Poga, M. (2023). Dealing with multicollinearity in factor analysis: The problem, detections, and solutions. *Open Journal of Statistics*, 13(3), 404–424.
- Lee, H., Park, Y. M., & Lee, S. (2015). Principal component regression by principal component selection. *Communications for Statistical Applications and Methods*, 22(2), 173–180.
- Mahaboob, B., Venkateswarlu, B., Narayana, C., & Balasiddamuni, P. (2018). A treatise on ordinary least squares estimation of parameters of linear model. *International Journal of Engineering & Technology*, 7(4.10), 518.
- Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118, 62–69. <https://doi.org/10.1016/j.chemolab.2012.07.010>
- Meshram, S. G., Kahya, E., Meshram, C., Ghorbani, M. A., Ambade, B., & Mirabbasi, R. (2020). Long-term temperature trend analysis associated with agriculture crops. *Theoretical and Applied Climatology*, 140(3–4), 1139–1159. <https://doi.org/10.1007/s00704-020-03137-z>
- Mevik, B. H., & Wehrens, R. (2007). The pls package: Principal component and partial least squares regression in R. *Journal of Statistical Software*, 18(2), 1–23. <https://doi.org/10.18637/jss.v018.i02>
- Moscovich, A., & Rosset, S. (2022). On the cross-validation bias due to unsupervised preprocessing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(4), 1474–1502. <https://doi.org/10.1111/rssb.12537>
- Önder, H., Tirink, C., Yakubets, T., Getya, A., Matvieiev, M., Kononenko, R., Şen, U., Özkan, Ç. Ö., Tolun, T., & Kaya, F. (2025). Predicting live weight for female rabbits of meat crosses from body measurements using LightGBM, XGBoost and support vector machine algorithms. *Veterinary Medicine and Science*, 11(1), e70149. <https://doi.org/10.1002/vms3.70149>
- Owen, A. B. (2007). A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443, 59–72.
- Parvande, S., Yeh, H. W., Paulus, M. P., & McKinney, B. A. (2020). Consensus features nested cross-validation. *Bioinformatics*, 36(10), 3093–3098. <https://doi.org/10.1093/bioinformatics/btaa046>
- Pavlou, M., Omar, R. Z., & Ambler, G. (2024). Penalized regression methods with modified cross-validation and bootstrap tuning produce better prediction models. *Biometrical Journal*, 66(5), e202300245. <https://doi.org/10.1002/bimj.202300245>
- Şahin, M. (2026). Estimate of albumen height in Atak-S hen eggs using ridge regression. *Black Sea Journal of Agriculture*, 9(4), 520–525.
- Schreiber-Gregory, D. N. (2018). Ridge regression and multicollinearity: An in-depth review. *Model Assisted Statistics and Applications*, 13(4), 359–365. <https://doi.org/10.3233/MAS-180446>
- Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63(324), 1379–1389. <https://doi.org/10.1080/01621459.1968.10480934>
- Sharma, P., Villegas-Diaz, R., & Fennell, A. (2024). Predicting grapevine physiological parameters using hyperspectral remote sensing integrated with hybrid convolutional neural network and ensemble stacked regression. *Remote Sensing*, 16(14), 2626. <https://doi.org/10.3390/rs16142626>
- Šverko, Z., Vrankić, M., Vlahinić, S., & Rogelj, P. (2022). Complex Pearson correlation coefficient for EEG connectivity analysis. *Sensors*, 22(4), 1477. <https://doi.org/10.3390/s22041477>
- Tilevik, A. (2026). *PCR, PLS, and Lasso Regression*. In *Multivariate Statistics and Machine Learning in R for Beginners: With Applications in Biology and Medicine* (pp. 279–297). Springer Nature Switzerland.
- Tirink, C., & Önder, H. (2022). Comparison of M, MM and LTS estimators in linear regression in the presence of outlier. *Turkish Journal of Veterinary & Animal Sciences*, 46(3), 420–

- 428.
- Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS One*, 14(11), e0224365. <https://doi.org/10.1371/journal.pone.0224365>
- Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, 91. <https://doi.org/10.1186/1471-2105-7-91>
- Wainer, J., & Cawley, G. C. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, 182, 115222. <https://doi.org/10.1016/j.eswa.2021.115222>
- Wang, X., Wang, X., Ma, B., Li, Q., & Shi, Y. Q. (2021). High precision error prediction algorithm based on ridge regression predictor for reversible data hiding. *IEEE Signal Processing Letters*, 28, 1125–1129.
- Wold, S., Ruhe, A., Wold, H., & Dunn, W. J., III. (1984). The collinearity problem in linear regression: The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3), 735–743. <https://doi.org/10.1137/0905052>
- Xu, J., & Hsu, D. J. (2019). On the number of variables to use in principal component regression. *Advances in Neural Information Processing Systems*, 32.
- Yates, L. A., Aandahl, Z., Richards, S. A., & Brook, B. W. (2023). Cross validation for model selection: A review with examples from ecology. *Ecological Monographs*, 93(1), e1557. <https://doi.org/10.1002/ecm.1557>
- Yue, S., & Pilon, P. (2004). A comparison of the power of the t test, Mann-Kendall and bootstrap tests for trend detection. *Hydrological Sciences Journal*, 49(1), 21–37. <https://doi.org/10.1623/hysj.49.1.21.53996>
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>