İstatistik Araştırma Dergisi
Journal of Statistical Research

ISSN: 2791-7616

(2024), 14 (2)
Başvuru / Received :13.09.2024
Kabul / Accepted :29.12.2024

https://dergipark.org.tr/tr/pub/jsstr

ARAŞTIRMA MAKALESİ **RESEARCH ARTICLE**

Performance Comparison of Least Squares, Ridge, Lasso and Principal Component Regression for Addressing Multicollinearity in Regression Analysis

Semih ERGİŞİ
Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı / Lisansüstü Öğrencisi
semihstat@gmail.com
Orcid No: 0009-0007-1364-1252

Beyza DOĞANAY
Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı / Doç. Dr.
bdoganay@medicine.ankara.edu.tr
Orcid No: 0000-0001-8845-2287

Yasemin YAVUZ
Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı / Prof. Dr.
dr.yasemin.yavuz@gmail.com
Orcid No: 0000-0003-1661-9468

Abstract

The purpose of this research was to evaluate the predictive accuracy of various regression methods in the context of multiple linear regression when multicollinearity invalidates the underlying assumptions of the least squares method. These methods included least squares regression (LS), ridge regression (RR), lasso regression (LR), and principal component regression (PCR). For this aim, the dataset including 6 variables simulated from normal with different sample of size from range of 50 to 1000. The performance was assessed using mean square error (MSE) and R square value. Despite the existence of multicollienarity among independent variables, research findings showed that LS method had the smallest MSE in the training dataset but RR had the smallest mse in the test dataset. When the sample size increases, the mse values increase for each methods in the training set but decrease in the test set. They are closer to each other. In terms of R square values, all methods showed similar performance both training and test data set.

Keywords: Lasso regression, Principal components, Multicollinearity, Least square regression.

Sorumlu Yazar / Corresponding Author: Semih ERGİŞİ, Ankara Üniversitesi Sağlık Bilimleri Enstitüsü, Biyoistatistik Anabilim Dalı.

Atıf / Citation: ERGİŞİ S., DOĞANAY B., YAVUZ Y. (2024). Performance Comparison of Least Squares, Ridge, Lasso and Principal Component Regression for Addressing Multicollinearity in Regression Analysis. *İstatistik Araştırma Dergisi*, 14 (2), 59-72.

Citation / Atıf: ERGİŞİ S., DOĞANAY B., YAVUZ Y. (2024). Performance Comparison of Least Squares, Ridge, Lasso and Principal Component Regression for Addressing Multicollinearity in Regression Analysis. *İstatistik Araştırma Dergisi*, 14 (2), 59-72.

Regresyon Analizinde Çoklu Doğrusallığın Ele Alınması için En Küçük Kareler, Ridge, Lasso ve Temel Bileşen Regresyonunun Performans Karşılaştırması

Özet

Bu araştırmanın amacı, çoklu doğrusal regresyon bağlamında çeşitli regresyon yöntemlerinin tahmin doğruluğunu değerlendirmektir. Bu bağlamda, en küçük kareler yönteminin (LS) temel varsayımlarını geçersiz kılabilecek çoklu bağlantı sorunları altında çeşitli yöntemler incelendi. İncelenen yöntemler arasında en küçük kareler regresyonu (LS), ridge regresyon (RR), lasso regresyon (LR) ve temel bileşen regresyonu (PCR) yer aldı. Bu amaçla, 50 ila 1000 arasında değişen örneklem boyutlarına sahip, normal dağılımdan simüle edilmiş 6 değişken içeren bir veri seti kullanıldı. Performans, hata kareler ortalaması (hko) ve R kare değeri kullanılarak değerlendirildi. Çoklu bağlantı sorunu olmasına rağmen, araştırma bulguları, LS yönteminin eğitim veri setinde en küçük hata kareler ortalamasına sahip olduğunu, ancak RR'nin test veri setinde en küçük hata kareler ortalamasına sahip olduğunu gösterdi. Örneklem boyutu arttıkça, her yöntem için eğitim setindeki hata kareler ortalaması değerleri arttı ancak test setinde azaldı ve yöntemler birbirine daha yakın hale geldi. R kare değerleri açısından, tüm yöntemler hem eğitim hem de test veri setlerinde benzer performans gösterdi.

Anahtar sözcükler: Lasso regresyon, Temel bileşenler, Çoklu bağlantı, En küçük kareler regresyon.

1. Introduction

Regression analysis is the frequently utilized statistical method for predicting the quantitative relationship between a dependent variable (Z) and one or more independent variables (Draper, 1966). The general applications of regression analysis are stated in the two main parts. One of this is to summarize data, relationship between dependent and independent variables while the other is to estimate the future values of dependent variable by using independent variables (Montgomery et al., 2021).

The most commonly utilized methodology in the regression analysis is the Least Squares (LS) method, which is dependent upon the satisfaction of certain necessary assumptions (Wooldridge, 2016). This approach is founded on the concept of minimizing the total sum of squares as the difference between observed y values and predicted y values from the regression equation (Altland, 1999). The dependability of the produced model is related to the fulfillment of the presumptions inherent in the LS method. If there is a considerable amount of multicollinearity exists among the independent variables, the regression coefficients obtained using the Least Squares method may lead to misinterpretation of the results (Alpar, 2017).

Ridge regression (RR) and Lasso regression (LR) are popular methods for handling multicollinearity in regression analysis. They are specifically designed to mitigate the adverse impact of multicollinearity on parameter predictions in MLR analysis. Principal Components Regression (PCR) is a regression technique that aims to elucidate original variables with intercorrelations by utilizing a reduced set of newly derived variables, which are linear combinations of the original variables (Alpar, 2017).

In disciplines like environmental science, public health, engineering, and finance, the utilization of the LS method for making predictions without adhering to its fundamental assumptions can lead to inaccurate deductions. In the finance sector, failing to consider multicollinearity among economic indicators may lead to untrustworthy assessments of investment risk (Gujarati & Porter, 2009). As a result, it is important to thoroughly examine the accuracy of the outcomes derived from the LS method when its assumptions are not sufficiently fulfilled.

Nevertheless, in research conducted within this framework, the existence of multicollinearity, a crucial factor in the field of statistics, and resolution methods are not afforded the requisite attention. In addition, despite the decreasing effectiveness of MLR analysis, the optimal method for addressing multicollinearity remains uncertain. A variety of approaches have been implemented to ensure the prevention of multicollinearity issues. Some examples of these techniques include lasso regression (LR), ridge regression (RR), and principal component regression (PCR). These three approaches are commonly employed to address issues of multicollinearity as they

are capable of producing effective estimators for predictors, making them a focus of interest for many researchers. This research paper examines and compares the performance of four selected regression techniques of LS, Lasso, Ridge, and PCR across different sample sizes. The Mean Squared Error (MSE) and R square metrics are used to evaluate the performance of various regression models.

2. Dataset and Method

In this study, Monte Carlo simulation techniques were utilized to create the dataset, which comprised six continuous variables, with one designated as the dependent variable and five as independent variables. To ensure the robustness of our findings across different sample sizes, datasets were simulated for sample sizes of 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000, with 1000 independent simulations conducted for each sample size. X1 was generated as random values drawn from a normal distribution with a mean of 3 and a standard deviation of 5. X2 was derived as a linear transformation of **X1**, specifically calculated as in the Equation 2.1. X3 was derived as linear combination of X1 and X2 and it was given in the Equation 2.2. X4 was independently generated as random values from a normal distribution with a mean of 3 and a standard deviation 3. X5 was created as linear combination of X1, X2, X3 and it was given in the Equation 2.3.

$$X_2 = 4*X_1 + e_2, \;\; e_2 \sim N(0.5,2) \tag{2.1}$$

$$X_3 = 2*X_1 + 0.6*X_2 + e_3, \; e_3 \sim N(5,2) \tag{2.2}$$

$$X_5 = X_1 + X_2 - X_3 + e_5, \;\; e_5 \sim N(0,0.1) \tag{2.3}$$

The dependent variable Y was then created as linear combination of independent variables using specified weights and a random error term was added to introduce variability. The simulation of the Y variable was given in the Equation 2.4. $\omega_1 = 1.5$, $\omega_2 = 2$, $\omega_3 = 0.5$, $\omega_4 = 0.8$ and $\omega_5 = 1.2$ were initial weights.

$$Y = \omega_1 X_1 + \omega_2 X_2 + \omega_3 X_3 + \omega_4 X_4 + \omega_5 X_5 + e, \; e \sim N(0,20) \tag{2.4}$$

While MLR models are often seen using matrix notation, Equation 2.5 provides a general outline of the regression analysis framework.

$$\gamma = \alpha_0 + \alpha_1 * \omega_1 + \alpha_2 * \omega_2 + \alpha_3 * \omega_3 + \cdots + \alpha_p * \omega_p + \varepsilon \tag{2.5}$$

In Equation 2.5, the dependent variable is on the left side, while the independent variables are on the right side. This equation mathematically expresses the portion of the variation in the dependent variable that can be explained by the independent variables. The error term in Equation 2.5 represents the portion of the variation that is not explained by the independent variables. In order to see the significance of multicollinearity, correlation matrices for selected sample of sizes (50, 200, 500, 900) are given in the Appendix.

2.1. Least Square Regression

The method of LS regression is frequently employed in statistical analysis to establish a relationship between the dependent variable and the predictor variables (Miles & Shevlin, 2000). This method is effective unless confronted with the issue of multicollinearity, which occurs when two or more explanatory variables are closely linearly correlated (Altland, 1999). The objective of this approach is to reduce the total of squared error values, under the assumption that the error values adhere to a normal distribution with consistent variance, and to enhance the model based on this (Weisberg, 1985).

By leveraging the power of the least squares regression method, designed to minimize residuals, meticulous researchers and astute analysts can achieve an unquestionably reliable estimation of coefficients. This increases statistical inference which allows more investigation and interpretation of complex data sets (Draper, 1966). Furthermore, the LS regression method stands out as an important aspect of many regression techniques. Simplicity and efficiency of LS regression makes its status as a necessary data analysis tools which allow researchers and analysts to go deeper into the field's ever-increasing depth and complexity (Alpar, 2017). The regression equation is written in the form of weighted averages of independent variables and it is given in Equation 2.6. In the Equation 2.6, ε is normally distributed with zero mean and σ standard deviation. In the Equation 2.6, ω_i, $i = 0,1,..k$ are weights and learned from training data. These parameters can be estimated by minimizing the Equation 2.7 that minimizes the sum squared difference between observed Y values and estimated $\hat{Y}$ values.

$$Y = \omega_0 + \omega_1 * X_1 + \omega_2 * X_2 + ... + \omega_k * X_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma) \tag{2.6}$$

$$\hat{Y} = \hat{\omega}_0 + \hat{\omega}_1 * X_1 + \hat{\omega}_2 * X_2 + ... + \hat{\omega}_k * X_k \tag{2.7}$$

$$\min_{\omega} \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 \tag{2.8}$$

Minimizing the Equation 2.8, the regression parameters can be found by taking derivative with respect to ω and equating it zero. In contrast, despite the benefits of the LS regression method, its application may be hampered by failure to meet certain basic assumptions (Kutner et al.,2004). A vital assumption is that the independent variables should not be significantly linearly correlated with each other (Kutner et al.,2004). Under the existence of multicollinearity, this can lead to increased variance in the LS estimates, potentially giving biased results that deviate from the true values (Kutner et al.,2004). Therefore, if there is a multicollinearity in a regression model, using different techniques instead of the LS method can reduce the variance and produce more reliable results (Kutner et al.,2004).

2.2. Ridge Regression

Ridge regression (RR) is a technique that aims to apply the L2 norm shrinkage penalty to the regression coefficients and thereby reduce the effects of multicollinearity in regression analysis containing multicollinear data (Hoerl & Kennard, 1970). The negative effects of multicollinearity can be reduced by constraining the coefficients from becoming too large. The penalty term applied in the RR method reduces the coefficient estimates towards zero by penalizing them without making them completely zero. As a result, some of the highly correlated predictors are allowed to remain in the model, albeit with reduced magnitudes. (Zoe & Hastie, 2005). While it is not aimed to select variable in the scope of RR, it is aimed to reduce the effect of coefficients and by this way increase the performance of model on the test data set. In the training data set, the performance of the model is not as good as the LS model (Müller & Guido, 2016).

Ridge regression excels in this aspect by permitting the retention of all predictors in the model without elimination. However, in addition to preserving the relevant predictors, the L2-norm reduction method forces the important coefficients to become smaller, bringing them closer to zero. This algorithm balances the coefficients by adding a penalty term to the residual sum of squares (Müller & Guido, 2016). It has a tuning parameter that provides a balance between reducing the effect of the coefficients and reducing the variance of the model(s) (Tibshirani, 1996). In fact, the form of this penalty term can be expressed in the Equation 2.9.

$$\min_{\omega} \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 + \lambda \sum_{m=1}^{k} \omega_m^2 \tag{2.9}$$

In the Equation 2.9, the right hand side of the plus sign is a regularization term to the least squares objective function, which serves to get rid of overfitting and enhance the model's generalization performance (Müller & Guido, 2016). The inclusion of this penalty component pushes the model to strike a compromise between precisely fitting the data and keeping coefficients low. The coefficients are biased toward zero by this regularization procedure, which also keeps any one predictor from governing the model (Müller & Guido, 2016). Ridge regression is therefore especially remedy for datasets with multicollinearity. Ridge regression stabilizes the model's performance and produces more accurate coefficient estimates by lessening the effect of multicollinearity. Ridge regression is an effective method for enhancing a linear regression model's stability and predictive precision when multicollinearity is present (Hoerl & Kennard, 1970).

2.3. Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression (LR) has been widely used to address the problem of multicollinearity in repressors in high-dimensional datasets (Müller & Guido, 2016). This is because the L1 penalty used in lasso regression allows for the automatic reduction and deletion of coefficients associated with weak or irrelevant variables, substantially limiting the influence of multicollinearity (Müller & Guido, 2016). Consequently, lasso regression increases model interpretability while also increasing predicted accuracy by saving only the most significant factors. The feature-selective behavior of Lasso regression is particularly useful when the effects of individual predictors that are truly important to the model are obscured due to multicollinearity. By penalizing the coefficients of variables which leading to multicollinearity, their influences are reduced effectively by Lasso regression, thereby enhancing the interpretability and robustness of the model (Fu ,1998). In the mathematical logic under the lasso regression is that it applies an L1 regularization technique, which encourages sparsity in the model. As a result, some coefficients being shrunk to zero, effectively performing variable selection and allowing for a simpler model that includes only the most important variables. The logic behind LR is that it adds a penalty term to the loss function, which discourages complexity and helps mitigate the impact of

multicollinearity. In Equation 2.10, the mathematical expression of this model is given. In the Equation 2.10, the right hand side of the plus sign is a regularization term to the least squares objective function.

$$\min_{\omega} \sum_{j=1}^{n} (Y_j - \hat{Y}_j)^2 + \lambda \sum_{m=1}^{k} |\omega_m| \tag{2.10}$$

By decreasing the detrimental effect of multicollinearity, the coefficients of the regression model are stabilized, resulting in more credible and interpretable results. This strategy essentially decreases some coefficients to zero, thus conducting variable selection and increasing the model's predictive power (Müller & Guido, 2016).

2.4. Principal Component Regression

The Principal Component Regression (PCR) method is a statistical technique that allows researchers to estimate the coefficients of independent variables by performing multiple linear regression analysis between the observed dependent variable and the new variables obtained using the PCR method, without having to remove the independent variables in the presence of multicollinearity (Alpar,2017). In this kind of study used to deal with multicollinearity, researchers who choose to utilize this model need first acquire independent components using principal component analysis. They should then retrieve the coefficients of the original variables by applying the back transformation procedure described in Equations 2.11 and 2.12. Unscaled regression coefficients in terms of MLR is given in left hand side of the Equation 2.11, P is the matrix of Eigen values and ψ_{PCR} is the coefficients of principal components.

$$\omega' = P * \psi_{PCR} \tag{2.11}$$

$$\omega = \omega' * \frac{\sigma_Y}{\sigma_{X_i}} \tag{2.12}$$

The coefficients of independent variables are given in the left hand side of the Equation 2.12. In fact, this method not only provides a summary of how independent variables are related through linear combinations of a set of separate variables, but also overcomes the misinterpretation of original variables under the effect of multicollinearity by providing the opportunity to return to the original variables (Johnson, 1998). This method converts the original regression problem into a condensed regression structure in which the number of independent variables is reduced to the maximum number of principal components used in the PCA sequence to explain the variance while providing the benefit of dimensionality reduction (Hintze, 2007). The goal of this method is to apply the least squares method to a set of derived variables known as principle components, which are produced from the correlation matrix using a specific process (Göktaş, 2010). In the context of PCR, the multicollinearity situation can be addressed by estimating the regression coefficients through the application of the Least Squares regression method on a new set of variables (Topal, 2010). These fresh variables are acquired through conducting orthogonal transformations on the initial variables in the dataset (Ortabaş, 2001).

3. Result and Discussion

In this study, the performance metrics of four different models were compared, presented in four distinct tables and one visual, with separate results for the training and test data. The tables sequentially display the MSE values for various sample sizes across different models. The same structure applies to the R-squared values. Each table consists of 5 columns and 11 rows, where the first column contains the sample sizes, while the remaining columns provide the average MSE and R-squared values for each model.

When examining Table 1, the MSE values obtained from different models for various sample sizes in the training data are presented. According to the results, the MSE values obtained through LS analysis are the lowest across all sample sizes, while the highest average MSE is observed in PCR analysis.

When examining Table 2, the changes in average MSE values in the test set can be observed. In all sample sizes, the lowest average MSE value is obtained from the RR and LR analysis, while the highest average MSE value is obtained from the PCR analysis. The most notable observation in this table is that, for the dataset with a sample size of 50, the highest average MSE is recorded in the LS analysis.

From the results presented in Table 1 and Table 2 for the training and test sets, it is observed that as the sample size increases, the MSE values in the training set increase and converge across the three models, while in the test set, the MSE values decrease and converge Furthermore, in terms of generalization performance, the RR and LR model demonstrates the best results, though with increasing sample size, the performance of LS, RR, and LR models also converges. Figure 1 visualizes the results provided in Table 1 and Table 2, showing the changes in average MSE values for each model in both the training and test sets as the sample size increases.

Table 1. Average MSE for training datasets.

Sample Size	LS	RR	LR	PCR
50	338,32	352,55	352,22	393,02
100	373,91	379,98	380,25	414,02
200	386,52	389,24	389,53	420,01
300	390,08	391,91	392,14	421,7
400	391,8	393,07	393,26	421,65
500	393,95	395,01	395,16	422,77
600	394,2	395,05	395,14	423,36
700	396	396,72	396,82	424,44
800	396,87	397,49	397,56	425,47
900	396,62	397,17	397,24	424,93
1000	397,39	397,88	397,94	425,45

Table 2. Average MSE for test datasets.

Sample Size	LS	RR	LR	PCR
50	475,29	446,96	447,2	461,48
100	433,02	424,1	423,46	445,18
200	414,95	411,6	411,37	433,06
300	413,92	411,42	411,53	434,8
400	408,11	406,43	406,48	430,79
500	403,36	402,49	402,61	427,44
600	408,34	407,33	407,36	432,81
700	406,36	405,62	405,7	430,6
800	401,63	400,82	400,81	426,4
900	403,19	402,59	402,62	428,27
1000	402,58	402,03	402,08	427,6

Table 3. Average R-squared for training datasets.

Sample Size	LS	RR	LR	PCR
50	0,92	0,91	0,91	0,90
100	0,91	0,91	0,91	0,90
200	0,91	0,91	0,91	0,90
300	0,91	0,91	0,91	0,90
400	0,91	0,91	0,91	0,90
500	0,91	0,91	0,91	0,90
600	0,91	0,91	0,91	0,90
700	0,91	0,91	0,91	0,90
800	0,91	0,91	0,91	0,90
900	0,91	0,91	0,91	0,90
1000	0,91	0,91	0,91	0,90

Table 4. Average R-squared for test datasets.

Sample Size	LS	RR	LR	PCR
50	0,85	0,86	0,86	0,85
100	0,88	0,88	0,88	0,88
200	0,90	0,90	0,90	0,89
300	0,90	0,90	0,90	0,89
400	0,90	0,90	0,90	0,90
500	0,90	0,90	0,90	0,90
600	0,90	0,90	0,90	0,90
700	0,90	0,90	0,90	0,90
800	0,91	0,91	0,91	0,90
900	0,91	0,91	0,91	0,90
1000	0,91	0,91	0,91	0,90

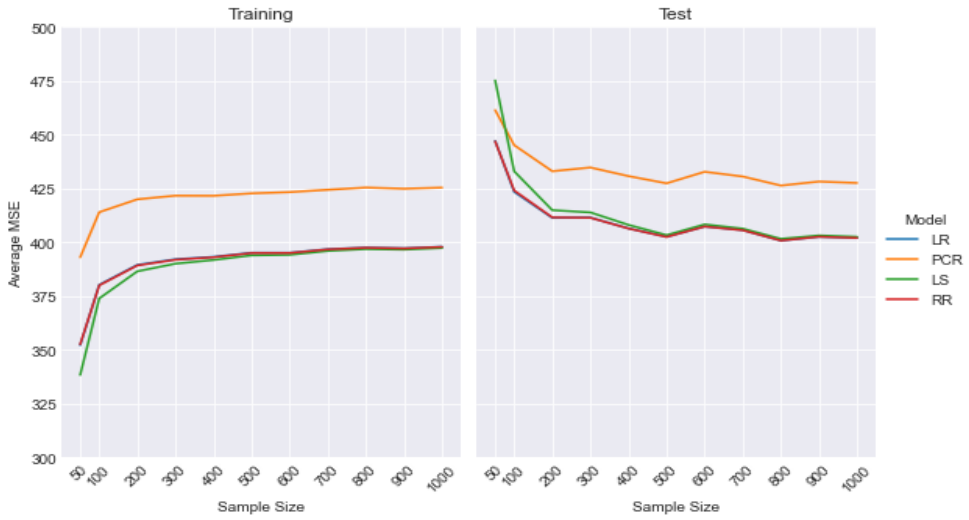

Figure 1. The change of MSE according to the sample size.

There are two different columns in the Figure 1 and they represent the change of MSE values for training and test data set respectively. X axes represent the sample size and y axis represents the average MSE values for this study. Each model is represented by line with different colored. Blue lines represent Lasso regression, orange lines represent Principal Component regression, green lines represent Least square regression and red lines represent Ridge regression. When the sample size increases in the training data, average MSE values increase. However, when the sample size increases in test data, average MSE values decrease.

It can be said that PCR method in terms of average MSE values are separated from other methods both in training and test data set. When the sample size is equal and above from 500, average MSE values are overlapped for LR, LS and RR in the training data set. On the other hand, when the sample size is below from 500, average MSE values are lowest for LS method and it is expected in the literature because the performance of LS method under the multicollinearity is not affected for the training data set.

When the sample size is equal 50, average MSE values are highest for LS in the test data set. When the sample size is equal and above from 500, average MSE values of LS are getting closer to LR and RR. On the other hand, when the sample size is below from 500, average MSE values are lowest for LR and RR methods and it is expected in the literature because the generalization performance of RR and LR methods under the multicollinearity is better than LS method.

From the Figure 1, it can be said that average MSE values are getting closer for each model both in training and test data set. The first one increases with increasing sample size and up to certain point and the second one decreases with increasing sample size up to certain point.

When examining Table 3 and Table 4, the R-square values for the models across different sample sizes are presented for both the training and test datasets. In the training set, despite the increase in sample size, all models achieve high R-square values, indicating that the increase in sample size has not led to significant changes in R-square values. On the other hand, in the test set, the R-square values show a slight increase as the sample size grows, but even for smaller sample sizes, the R-square values remain relatively high.

Table 5. Changing of coefficients across different sample size and model.

Sample Size	Coefficient	LR	RR	PCR	LS
50	Intercept	37.33 (0.34)	37.33 (0.34)	37.33 (0.34)	37.33 (0.34)
	X1	15.41 (0.7)	11.42 (0.62)	16.26 (0.06)	2.46 (5.98)
	X2	36.31 (0.77)	22.23 (0.41)	16.33 (0.06)	19.2 (23.22)
	X3	5.81 (0.37)	21.7 (0.25)	16.12 (0.06)	32.99 (25.61)
	X4	2.14 (0.1)	2.31 (0.1)	2.27 (0.11)	2.33 (0.11)
	X5	3.81 (0.15)	6.76 (0.17)	14.32 (0.07)	7.96 (4.27)
100	Intercept	37.89 (0.23)	37.89 (0.23)	37.89 (0.23)	37.89 (0.23)
	X1	14.36 (0.53)	10.43 (0.52)	16.51 (0.04)	2.58 (3.77)
	X2	41.28 (0.58)	23.47 (0.34)	16.58 (0.04)	17.68 (14.98)
	X3	3.4 (0.23)	22.58 (0.21)	16.36 (0.04)	34.91 (16.45)
	X4	2.2 (0.07)	2.41 (0.07)	2.39 (0.08)	2.42 (0.07)
	X5	3.23 (0.11)	6.68 (0.12)	14.63 (0.05)	8.33 (2.74)
200	Intercept	37.59 (0.16)	37.59 (0.16)	37.59 (0.16)	37.59 (0.16)
	X1	12.98 (0.42)	8.85 (0.45)	16.55 (0.03)	7.37 (2.59)
	X2	44.96 (0.45)	24.61 (0.29)	16.62 (0.03)	35.95 (10.08)
	X3	1.56 (0.12)	23.17 (0.19)	16.39 (0.03)	14.71 (11.12)
	X4	2.17 (0.05)	2.39 (0.05)	2.36 (0.05)	2.39 (0.05)
	X5	2.87 (0.08)	6.7 (0.09)	14.67 (0.03)	5.06 (1.86)
300	Intercept	37.74 (0.13)	37.74 (0.13)	37.74 (0.13)	37.74 (0.13)
	X1	11.72 (0.34)	7.39 (0.39)	16.54 (0.02)	6.26 (2.14)
	X2	47.06 (0.37)	25.59 (0.26)	16.61 (0.02)	35.37 (8.37)
	X3	0.97 (0.1)	23.73 (0.18)	16.38 (0.02)	16.3 (9.22)
	X4	2.21 (0.04)	2.39 (0.04)	2.4 (0.04)	2.4 (0.04)
	X5	2.64 (0.07)	6.63 (0.07)	14.69 (0.03)	5.15 (1.54)
400	Intercept	37.63 (0.12)	37.63 (0.12)	37.63 (0.12)	37.63 (0.12)
	X1	10.71 (0.31)	6.28 (0.35)	16.54 (0.02)	5.18 (1.79)
	X2	48.47 (0.32)	26.45 (0.27)	16.62 (0.02)	34.42 (7.08)
	X3	0.59 (0.06)	23.92 (0.18)	16.38 (0.02)	18.05 (7.78)
	X4	2.19 (0.04)	2.35 (0.04)	2.33 (0.04)	2.35 (0.04)
	X5	2.66 (0.06)	6.73 (0.06)	14.7 (0.02)	5.52 (1.3)
500	Intercept	37.62 (0.11)	37.62 (0.11)	37.62 (0.11)	37.62 (0.11)
	X1	10.7 (0.29)	6.21 (0.32)	16.58 (0.02)	4.21 (1.66)
	X2	48.8 (0.3)	26.35 (0.25)	16.65 (0.02)	28.72 (6.49)
	X3	0.36 (0.05)	24.13 (0.2)	16.42 (0.02)	23.91 (7.15)
	X4	2.2 (0.03)	2.35 (0.03)	2.35 (0.03)	2.35 (0.03)
	X5	2.73 (0.06)	6.84 (0.06)	14.73 (0.02)	6.66 (1.19)

Table 5. Changing of coefficients across different sample size and model(cont.).

Sample Size	Coefficient	LR	RR	PCR	LS
600	Intercept	37.64 (0.1)	37.64 (0.1)	37.64 (0.1)	37.64 (0.1)
	X1	10.89 (0.26)	6.43 (0.29)	16.56 (0.02)	6.85 (1.48)
	X2	48.67 (0.28)	26.39 (0.25)	16.63 (0.02)	37.07 (5.86)
	X3	0.38 (0.05)	23.97 (0.21)	16.39 (0.02)	14.32 (6.44)
	X4	2.25 (0.03)	2.38 (0.03)	2.38 (0.03)	2.38 (0.03)
	X5	2.55 (0.05)	6.62 (0.06)	14.71 (0.02)	4.87 (1.08)
700	Intercept	37.69 (0.09)	37.69 (0.09)	37.69 (0.09)	37.69 (0.09)
	X1	11.09 (0.25)	6.43 (0.28)	16.57 (0.02)	8.04 (1.4)
	X2	48.64 (0.27)	26.58 (0.26)	16.64 (0.02)	40.63 (5.46)
	X3	0.23 (0.04)	23.79 (0.24)	16.4 (0.02)	10.04 (6.02)
	X4	2.23 (0.03)	2.35 (0.03)	2.35 (0.03)	2.35 (0.03)
	X5	2.61 (0.05)	6.66 (0.06)	14.73 (0.02)	4.3 (1.0)
800	Intercept	37.64 (0.08)	37.64 (0.08)	37.64 (0.08)	37.64 (0.08)
	X1	10.8 (0.24)	6.2 (0.26)	16.57 (0.02)	8.54 (1.27)
	X2	49.1 (0.25)	27.16 (0.26)	16.64 (0.02)	43.89 (5.03)
	X3	0.14 (0.03)	23.55 (0.23)	16.4 (0.02)	6.82 (5.52)
	X4	2.32 (0.03)	2.43 (0.02)	2.43 (0.03)	2.43 (0.02)
	X5	2.52 (0.04)	6.53 (0.05)	14.72 (0.02)	3.65 (0.92)
900	Intercept	37.6 (0.08)	37.6 (0.08)	37.6 (0.08)	37.6 (0.08)
	X1	10.52 (0.22)	5.74 (0.25)	16.59 (0.01)	7.24 (1.23)
	X2	49.45 (0.24)	27.09 (0.28)	16.66 (0.01)	39.6 (4.84)
	X3	0.15 (0.03)	24.08 (0.27)	16.42 (0.01)	11.82 (5.33)
	X4	2.29 (0.02)	2.4 (0.02)	2.39 (0.03)	2.4 (0.02)
	X5	2.53 (0.04)	6.63 (0.06)	14.75 (0.02)	4.5 (0.89)
1000	Intercept	37.61 (0.07)	37.61 (0.07)	37.61 (0.07)	37.61 (0.07)
	X1	10.7 (0.21)	5.89 (0.24)	16.59 (0.01)	8.32 (1.18)
	X2	49.39 (0.22)	27.23 (0.28)	16.66 (0.01)	42.78 (4.56)
	X3	0.04 (0.01)	23.82 (0.29)	16.42 (0.01)	8.04 (5.04)
	X4	2.29 (0.02)	2.4 (0.02)	2.38 (0.02)	2.4 (0.02)
	X5	2.53 (0.04)	6.6 (0.06)	14.75 (0.02)	3.88 (0.84)

Note: Estimate (standard error) given in the Table 6.

Tables 5 and 6 present the regression coefficients along with their standard errors. It is important to highlight that, even with varying sample sizes, the coefficients derived from the LR, RR, and PCR methods are fairly consistent. In contrast, the coefficients produced by the LS method show considerable fluctuations in their values as the sample size varies. This indicates that the LS method may not provide reliable coefficient estimates when multicollinearity is present. Also, both multicollinearity exist and small sample size from the Table 5 and Table 6, coefficients have higher standard error and they are not statistically significant.

In addition, an analysis of the standard errors for the coefficients indicates that the LR, RR, and PCR approaches consistently produce exceedingly low standard errors, which exhibit a slight decrease as the dataset's number of observations grows. Conversely, the LS method consistently demonstrates high standard errors across all sample sizes.

According to the findings of this study, the LS regression method yielded the lowest MSE value in the training dataset for various sample sizes. The size of our sample had a significant impact on our results. The Ridge regression and Lasso regression methods performed best in terms of the Mean Squared Error (MSE) in the test dataset. Conversely, the PCR method showed the highest average MSE value in both the training and test datasets. In terms of MSE performance, these findings are consistent with those reported by (Çankaya et al. 2019).

Moreover, as the size of the sample grows, the average mean squared error (MSE) values also rise for the least squares (LS), ridge regression (RR), and Lasso regression (LR) methods in the training data set. In relation to the validity of the models, it was found that the models generated with RR and Lasso regression yielded the most favorable outcomes for the test data especially for the small sample sizes. Additionally, it was noted that as the sample size increases, the average MSE values for LS, RR, and LR methods tend to converge.

There are actually two main goals of the regression analysis in the literature. One of this is to summarize data set for the purposes of finding relationships among the variables, finding which independent variables have highest effect on dependent variable and getting adjusted coefficients of each independent variables. The other goal of this is to predict future value of the dependent variables by using independent variable. These two goals can be used under two main title as explorative model and predictive model respectively.

This study yields two main conclusions. Initially, the performance metrics, such as Mean Squared Error (MSE) and R-squared values, are not able to effectively differentiate between models in the presence of multicollinearity. However, the generalization of the LS method is low in the scope of the predictive model when the sample size small and multicollinearity exists. In simpler terms, the LS method does not exhibit a noticeable decline compared to other models when assessing model performance under the high sample size. On the other hand, when it comes to interpreting coefficients in the scope of summarizing the data set, the LS method clearly falls short, resulting in misleading conclusions and inaccurate interpretations due to high standard error of coefficients. Conversely, the LR, RR, and PCR methods yield reliable outcomes in scenarios where dataset summarization and coefficient interpretation are essential. Despite of the increasing sample size, the stability of the confidents for LS method cannot be guaranteed.

References

Alpar, R., (2017). Uygulamalı Çok Değişkenli İstatistiksel Yöntemler. Detay Yayıncılık.

Altland H. W. (1999). Regression analysis: statistical modeling of a response variable.

Çankaya, S., Eker, S., & Abacı, S. H. (2019). Comparison of least squares, ridge regression and principal component approaches in the presence of multicollinearity in regression analysis. *Turkish Journal of Agriculture - Food Science and Technology, 7*(3), 180-190. https://doi.org/10.24925/turjaf.v7i3.180-190.2019

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example*. John Wiley & Sons.

Draper, N. R. and Smith H. (1998). Applied Regression Analysis. New York, John Wiley and Sons, Inc.

Fu, W. J. (1998). Penalized regression: the Bridge versus the Lasso, Journal of Computation and Graphical Statistics, 7, 397-416

Göktaş, A., & Öznur, İ. (2010). Türkiye'de işsizlik oranının temel bileşenli regresyon analizi ile belirlenmesi. *Sosyal Ekonomik Araştırmalar Dergisi, 10*, 279-294.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). McGraw-Hill.

Hintze, J. L. (2007). *NCSS User's Guide III - Regression and Curve Fitting, Chapter 340 - Principal Components Regression*. Kaysville/Utah: NCSS Statistical System.

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

Johnson, R., A. (1998). Applied multivariate statistical analysis,Prentice Hall, 458-498.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). McGraw-Hill.

Miles J. & Shevlin M. (2000). Applying Regression and Correlation-A Guide for Students and Researchers. Sage Publication, Lodon.

Müller, A. C., & Guido, S. (2016). *Introduction to machine learning with Python: A guide for data scientists*. O'Reilly Media.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis*. sut.ac.th

Ortabaş, N. (2001). Principal components in the problem of multicollinearity (Master's thesis). DEÜ Fen Bilimleri Enstitüsü.

Topal, M., Eyduran, E., Yağanoğlu, A. M., Sönmez, A., & Keskin, S. (2010). Çoklu doğrusal bağlantı durumunda ridge ve temel bileşenler regresyon analiz yöntemlerinin kullanımı. *Atatürk Üniversitesi Ziraat Fakültesi Dergisi, 41*, 53-57.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological), 58*, 267-288.

Weisberg, S. (1985). Applied Linear Regression, Wiley.

Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach* (6th ed.). Cengage Learning.

Zou, H., & Hastie, T. (2005). The use of ridge regression in the analysis of high-dimensional data. *Journal of Computational and Graphical Statistics, 14*(3), 814-828.

Appendix

Table 6. Correlation matrix for selected sample size (50).

		X1	X2	X3	X4	X5	Y
X1	**P.C**	1	,998**	,997**	0.229	,854**	,965**
	Sig		0.000	0.000	0.109	0.000	0.000
X2	**P.C**	,998**	1	,997**	0.235	,870**	,967**
	Sig	0.000		0.000	0.100	0.000	0.000
X3	**P.C**	,997**	,997**	1	0.236	,827**	,962**
	Sig	0.000	0.000		0.099	0.000	0.000
X4	**P.C**	0.229	0.235	0.236	1	0.182	0.254
	Sig	0.109	0.100	0.099		0.205	0.075
X5	**P.C**	,854**	,870**	,827**	0.182	1	,855**
	Sig	0.000	0.000	0.000	0.205		0.000
Y	**P.C**	,965**	,967**	,962**	0.254	,855**	1
	Sig	0.000	0.000	0.000	0.075	0.000	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 7. Correlation matrix for selected sample size (200).

		X1	X2	X3	X4	X5	Y
X1	**P.C**	1	,996**	,996**	0.012	,814**	,950**
	Sig		0.000	0.000	0.866	0.000	0.000
X2	**P.C**	,996**	1	,995**	0.014	,835**	,952**
	Sig	0.000		0.000	0.845	0.000	0.000
X3	**P.C**	,996**	,995**	1	0.016	,780**	,945**
	Sig	0.000	0.000		0.822	0.000	0.000
X4	**P.C**	0.012	0.014	0.016	1	-0.006	0.040
	Sig	0.866	0.845	0.822		0.931	0.570
X5	**P.C**	,814**	,835**	,780**	-0.006	1	,818**
	Sig	0.000	0.000	0.000	0.931		0.000
Y	**P.C**	,950**	,952**	,945**	0.040	,818**	1
	Sig	0.000	0.000	0.000	0.570	0.000	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 8. Correlation matrix for selected sample size (500).

		X1	X2	X3	X4	X5	Y
X1	**P.C**	1	,995**	,995**	0.014	,822**	,956**
	Sig		0.000	0.000	0.756	0.000	0.000
X2	**P.C**	,995**	1	,995**	0.011	,839**	,958**
	Sig	0.000		0.000	0.803	0.000	0.000
X3	**P.C**	,995**	,995**	1	0.008	,784**	,953**
	Sig	0.000	0.000		0.854	0.000	0.000
X4	**P.C**	0.014	0.011	0.008	1	0.030	0.033
	Sig	0.756	0.803	0.854		0.508	0.460
X5	**P.C**	,822**	,839**	,784**	0.030	1	,815**
	Sig	0.000	0.000	0.000	0.508		0.000
Y	**P.C**	,956**	,958**	,953**	0.033	,815**	1
	Sig	0.000	0.000	0.000	0.460	0.000	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 9. Correlation matrix for selected sample size (900).

		X1	X2	X3	X4	X5	Y
X1	**P.C**	1	,995**	,995**	0.000	,820**	,954**
	Sig		0.000	0.000	0.990	0.000	0.000
X2	**P.C**	,995**	1	,995**	-0.003	,835**	,958**
	Sig	0.000		0.000	0.924	0.000	0.000
X3	**P.C**	,995**	,995**	1	-0.004	,779**	,951**
	Sig	0.000	0.000		0.898	0.000	0.000
X4	**P.C**	0.000	-0.003	-0.004	1	0.009	0.040
	Sig	0.990	0.924	0.898		0.797	0.229
X5	**P.C**	,820**	,835**	,779**	0.009	1	,814**
	Sig	0.000	0.000	0.000	0.797		0.000
Y	**P.C**	,954**	,958**	,951**	0.040	,814**	1
	Sig	0.000	0.000	0.000	0.229	0.000	

**. Correlation is significant at the 0.01 level (2-tailed).