THE JOURNAL OF
COGNITIVE SYSTEMS

Journal Homepage:https://dergipark.org.tr/en/pub/jcs

Vol.7, No.2, 2022

Comparison of Ordinary Least Squares and Principal Components Regression Analyses

Zeynep Küçükakçalı , Harika Gözükara Bağ

[1]Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey.(e-mail:zeynep.tunc@inonu.edu.tr).
[1]Inonu University Department of Biostatistics and Medical Informatics, Faculty of Medicine, Malatya, Turkey.(e-mail: harika.gozukara@inonu.edu.tr).

ARTICLE INFO

Received: March.,09.2022
Revised: April,19.2022
Accepted: May.,11.2022

Keywords:
Linear regression, multicollinearity, least squares, principal components regression, sample size.

Corresponding author: Zeynep Küçükakçalı
🖂zeynep.tunc@inonu.edu.tr
☎+90 536 424 32 06

ISSN:2548-0650

DOI:https://doi.org/10.52876/jcs.1117961

ABSTRACT

The aim of this study is to compare Least Squares Regression (LSR) and Principal Components Regression (PCR) results when multicollinearity is determined in a dataset.Inorder to examine the effect of the degree of multicollinearity in the study, 10 datasets with different levels of multicollinearity were derived. Each data set consists of three independent and one dependent variable, and the variables were derived from the standard normal distribution. The multicollinearity status in the derived data has been demonstrated by the commonly used metrics to determine multicollinearity. Least Squares and Principal Components Regression was applied to all datasets.

When generating multicollinearity, all relationships were defined as positive in data simulation. However, the sign of the regression coefficients for the second (X2) and third (X3) independent variables were obtained as reverse (negative) as one of the expected effects of multicollinearity in Least Squares analysis. In the analysis of the Principal Components Regression, the sign of coefficients was found to be in the right direction (positive). The sign of the coefficients obtained from OLS and PCR were different and they also differed in magnitude. In addition, the standard errors of the coefficients in PCR results were lower than OLS results.

The existence of multicollinearity must be examined while performing multiple linear regression analysis, and if multicollinearity is determined, one of the methods that can solve this situation should be used. Otherwise, the estimations to be made as a result of regression may lead to wrong results. In line with the results of this study, it is recommended to use Principal Components Regression instead of Least Squares regression in case of multicollinearity in the data.

1. INTRODUCTION

OFTEN a researcher or experimenter wants to find out if there is a relationship between two or more variables and show how this relationship can be expressed with an equation. In other words, what we really want to do is to try to explain the dependent variable by estimating it by the independent variable(s). When this process is desired to be done mathematically, the method to be applied is called regression. The equation showing the relationship between two (or more) variables not only shows the functional form of the relationship between the variables, but also enables predictions about the other when the value of one of the variables is known [1]. As a result, regression is used to reveal the functional relationship of the dependent variable Y with the independent variable(s) X and to control the asserted claims. If a single independent variable X is used to explain the dependent variable Y, this regression is called simple regression, while if two or more independent variable *X* is used to explain the dependent variable Y, the regression is called multiple regression.

If the assumption that the independent variables are not related to each other, which is one of the assumptions of the multiple linear regression analysis, is not fulfilled, the problem called the multicollinearity problem is encountered [2]. In this case, there will be some negative effects on the results to be obtained by regression [3]. In the most general form, the estimates will be very different from the actual results of the parameters to be estimated, and very different estimates and signs will be obtained for the parameters. There are many suggested methods to overcome this situation. Basically, it is recommended to be careful in the selection of the variables that will form the data at the beginning [4]. In addition, adding different observations to the data, rebuilding the model, or using some biased estimation methods are the most common methods. But instead of doing all these, it is better to use biased estimation methods instead of adding or removing data. The most effective way to eliminate the problems that occur in the presence of multicollinearity is to estimate the regression coefficients biasedly without adding or subtracting the variables in the model. Principal Component Regression is the most preferred method that uses biased estimation results [3].

The aim of this study is to show how the results of the least-squares method, which is frequently used in linear regression analysis, are affected in the case of multicollinearity between independent variables. Another aim is to compare the results of principal components regression (PCR) and the LSR method used in case of a multicollinearity problem.

2. MATERIAL and METHODS

2.1. Data set

Our aim in this study is to compare the results of Least Squares and Principal Component Regression analyses by making analyzes with datasets containing different degrees of multicollinearity. For this purpose, 10 data sets were derived using the simulation technique with IBM SPSS Statistics version 25.0 [5]. All derived data sets contained 3 independent (x1, x2, and x3) and 1 dependent (y) variable, and each variable was derived according to standard normal distribution.

In order to obtain 10 data sets that will contain multicollinearity at different degrees, data sets consisting of 1000 observations were derived, provided that the correlation values between the dependent variable and independent variables were kept constant and the correlation values between the independent variables were increased. The correlation values between the properties and variables of these data sets are given in Table 1.

TABLE I

Correlation representation between variables	The value of the correlation between the variable									
x1-x2	0.75	0.8	0.85	0.9	0.95	0.95	0.95	0.95	0.95	0.95
x1-x3	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90
x2-x3	0.50	0.50	0.50	0.50	0.50	0.55	0.6	0.65	0.7	0.75
y-x1	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
y-x2	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65
y-x3	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55

2.2. Regression

Regression analysis can be defined as "making correlations by mathematically modeling and explaining the relationship between the dependent variable and the independent variable(s)". The obtained relations are expressed with the help of equations (regression equations). Regression equations vary according to the linearity of the relationship between two variables, the data type structure of the dependent variable, and the number of independent variables [6]. The main purpose is to make estimations with the obtained regression equations. In other words, it is estimating a variable that is difficult to obtain with some easily obtainable variables. Variables are affected negatively or positively by some factor or factors. In addition, some of the factors may have a great effect, while others may have a very low effect [7].

In the regression model, if there is a dependent and an independent variable and the relationship between these variables shows a linear structure, the regression is called simple linear regression. The method used to explain the cause-effect relationships between two or more independent variables affecting a variable with a linear model and to determine the effect levels of these independent variables is called multiple linear regression analysis [7].

The most common method used to estimate the coefficients in the regression equation is the Least Squares method. Apart from many possible regression methods, the Least Squares method is generally used as the most appropriate estimation method due to its simplicity in mathematical calculations. [8]. The aim of this method is to optimize the model by minimizing the optimum results, in other words, the sum of the squares of the error terms, if the variances of the error terms are homogeneous and normally distributed [9].

2.3. Multicollinearity Problem

One of the most common problems in multiple linear regression analysis is the existence of a linear relationship between independent variables, that is, multicollinearity. If one of the independent variables is expressed as a linear function of the other independent variable or variables, the existence of a linear relationship between the independent variables is mentioned. Multicollinearity can lead to incorrect estimation of the regression coefficients, and exaggeration of the standard errors of the regression coefficients, resulting in an increase in the confidence intervals and a decrease in the t-test value. The enlargement of the standard error may cause statistically significant regression coefficients to be insignificant, thus resulting in incorrect results.

The correlation matrix between the independent variables is used to determine whether there is multicollinearity or not. If the absolute value of the correlation coefficient between the variables is close to 1, it is stated that there is multicollinearity between those independent variables. As these values get larger (VIF values ≥ 10) for variance swelling values (VIF), it can be mentioned that there is multicollinearity between the relevant independent variables. The larger the VIF value, the greater the multicollinearity between the variables. In general, when the VIF value is above 10, it is accepted that there is multicollinearity between those variables. The determinant of the correlation matrix of the independent variables ranges from 0 to 1 and can be used to determine multicollinearity. When the determinant of the correlation matrix is 1, it is stated that there is no linear dependence between the variables, but when the determinant of the correlation matrix is 0, there is multicollinearity between the variables. The smaller the determinant value, the higher the degree of multicollinearity. If one or more characteristic roots (eigenvalues) of the correlation matrix are zero or close to zero, it is said to be multicollinear. The ratio of the largest to the smallest eigenvalues (λmax / λmin) is also used as a measure of multicollinearity, and the larger this ratio, the higher the degree of multicollinearity. In general, when the ratio (λmax / λmin) is less than 10, it is stated that there is no serious multicollinearity in the data set. In addition, Although none of the t values of the regression coefficients are significant, the fact that the F statistic belonging to the regression model is significant is an indication of the existence of multicollinearity in the data set [10-12].

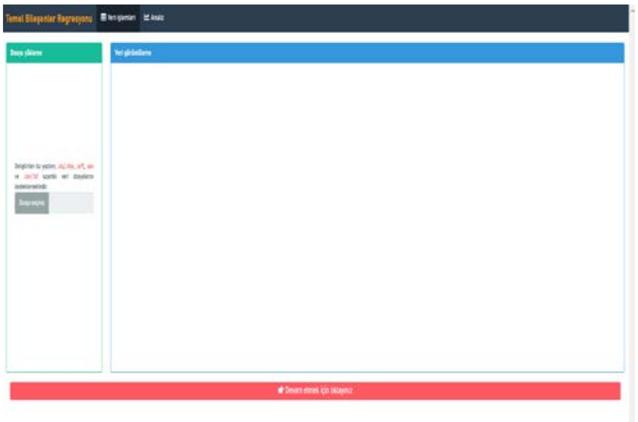
In the case of multicollinearity, new variables can be added to the data or one or more of the multilinked variables in the dataset can be removed from the model. However, by removing the variables from the model, a variable that really contributes to the model may be removed, which causes information loss. Methods that are less sensitive to multicollinearity than the least-squares method are generally used to estimate the regression coefficients without removing the variables in the model. At the beginning of these methods, it is recommended to use ridge regression and principal component regression methods, which give biased estimates.

2.4. Principal Components Regression

Principal Components Regression is one of the biased estimation methods used in the presence of multicollinearity in the data. It was first discussed by Hotelling in 1933. In this method, the aim is to apply the Least Squares method on a set of artificial variables obtained from the correlation matrix and called principal components, that is, obtained by operation. In Principal Components Regression, which can be used to eliminate the multicollinearity situation, the regression coefficients are estimated by applying the Least Squares regression method on a set of new, that is, artificial variables obtained with the help of their orthogonal transformations instead of the original variables in the data set. In the estimation using Principal Component Regression, it is expected that the mean squared error value will take a smaller value than the estimation made by the Least Squares regression method [13-16].

2.5. Data Analyses

The suitability of the derived data to the multiple normal distributions was examined with the "Normal Distribution Analysis Software" developed by the Biostatistics and Medical Informatics Department of İnönü University Faculty of Medicine [17]. First of all, multiple linear regression analysis was applied to the sets obtained by simulation technique with IBM SPSS Statistics version 25.0 package program [18]. As a result of the analysis, the existence of multiple connections in the data set was examined. The existence of multicollinearity was supported by looking at the VIF, tolerance values, eigenvalues, and the condition index. R programming language was used for Principal components regression, which will be compared with least squares regression results. For principal components regression, the pcr function available in the pls package in the R programming language was used. The functions written in order to be able to analyze are visualized with the Shiny library in the R studio environment. The screenshot of the image is given in figure 1.



3. RESULTS

The information on the measures used in determining the presence of multicollinearity for 10 data sets with different degrees of multicollinearity used in the study is presented in Table 2

TABLE II
TABLE OF CRITERIA FOR DETERMINING MULTICOLLINEARITY FOR DATASETS WITH DIFFERENT DEGREES OF MULTICOLLINEARITY

NUMBER OF OBSERVATION	INDEPENDENT VARIABLES	VIF	TOLERANCE VALUE	EIGENVALUE	CONDITION. INDEX
n=1000	x1	16.87850	0.05925	2.47096	1.00
	x2	4.04280	0.24735	0.49355	5.01
	x3	9.67960	0.10331	0.03550	69.61
n=1000	x1	44.18220	0.02263	2.50335	1.00
	x2	10.89020	0.09183	0.48303	5.18
	x3	20.83430	0.04800	0.01362	183.84
n=1000	x1	165.55410	0.00604	2.52092	1.00
	x2	46.75720	0.02139	0.47543	5.30
	x3	64.36160	0.01554	0.00365	691.17
n=1000	x1	235.73770	0.00424	2.50118	1.00
	x2	79.79830	0.01253	0.49627	5.04
	x3	78.98630	0.01266	0.00255	980.73
n=1000	x1	441.51720	0.00226	2.53002	1.00
	x2	168.38760	0.00594	0.46861	5.40
	x3	123.11860	0.00812	0.00137	1848.17
n=1000	x1	249.43130	0.00401	2.57379	1.00
	x2	94.99260	0.01053	0.42376	6.07
	x3	66.44860	0.01505	0.00245	1050.42
n=1000	x1	158.99470	0.00629	2.61782	1.00
	x2	60.96010	0.01640	0.37830	6.92
	x3	40.47790	0.02470	0.00389	673.85
n=1000	x1	109.18950	0.00916	2.66186	1.00
	x2	42.65640	0.02344	0.33243	8.01
	x3	26.57870	0.03762	0.00571	465.93
n=1000	x1	79.36310	0.01260	2.70579	1.00

	x2	32.09370	0.03116	0.28628	9.45
	x3	18.52200	0.05399	0.00793	341.24
n=1000	x1	59.96010	0.01668	2.74993	1.00
	x2	25.75090	0.03883	0.23951	11.48
	x3	13.46050	0.07429	0.01057	260.23

When Table 2 was examined, it was observed that VIF values were above 10, and above 30 when the degree of multicollinearity increased. Similarly, the existence of multicollinearity was proved by obtaining eigenvalues close to 0 and tolerance values approaching 0. Condition indexes, another multicollinearity indicator, also supported the existence of multicollinearity. The least-squares regression and principal components regression results applied to these datasets are presented in Table 3.

TABLE III
TABLE OF RESULTS OF LSR AND PCR FOR DATASETS WITH DIFFERENT DEGREES OF MULTICOLLINEARITY

NUMBER OF OBSERVATION	VARIABLES	LSR REGRESSION COEFFICIENTS	LSR STANDARD ERROR	LSR R2	LSR sigma	PCR COEFFICIENTS	PCR STANDARD ERROR	PCR R2	PCR sigma
n=1000	sabit	-0.00586		0.668512711	0.580366307	-0.01716		0.556861024	0.67102504
	x1	159.867	0.07344			0.26096	0.00896		
	x2	-0.12979	0.03644			0.41149	0.02465		
	x3	-0.83471	0.05551			0.13588	0.01912		
n=1000	sabit	-0.005685		0.744456972	0.507873575	-0.019505		0.546636028	0.676467757
	x1	3.148.592	0.103986			0.266733	0.008574		
	x2	-0.962421	0.052201			0.396218	0.024229		
	x3	-1.809.292	0.071357			0.125027	0.020592		
n=1000	sabit	0.00100		0.998867995	0.03305884	-0.02118		0.550601570	0.658687034
	x1	844.752	0.01303			0.26831	0.00806		
	x2	-396.968	0.00702			0.37778	0.02287		
	x3	-497.492	0.00818			0.11844	0.02120		
n=1000	sabit	0.00187		0.998215812	0.041392357	-0.02712		0.502951218	0.690874776
	x1	1.083.777	0.02010			0.27357	0.00873		
	x2	-581.548	0.01182			0.35418	0.02342		
	x3	-599.535	0.01172			0.12453	0.02341		
n=1000	sabit	0.00114		0.996843317	0.05504378	-0.02112		0.528508731	0.672712648
	x1	1.361.870	0.03474			0.26849	0.00804		
	x2	-806.414	0.02194			0.34123	0.02174		
	x3	-707.178	0.01881			0.12313	0.02330		
n=1000	sabit	-0.00087		0.998415603	0.038985305	-0.02157		0.525494947	0.674666883
	x1	1.039.442	0.01858			0.26555	0.00800		
	x2	-598.341	0.01169			0.34862	0.02274		
	x3	-516.923	0.00979			0.11132	0.02453		
n=1000	sabit	0.00134		0.998830334	0.03347448	-0.02214		0.522450897	0.676391255
	x1	841.439	0.01280			0.26314	0.00798		
	x2	-471.336	0.00805			0.35727	0.02389		
	x3	-399.810	0.00656			0.09738	0.02600		
n=1000	sabit	-0.00087		0.999285843	0.026152056	-0.02305		0.519883684	0.678081934
	x1	706.912	0.00833			0.26190	0.00800		
	x2	-386.264	0.00527			0.36932	0.02525		
	x3	-319.804	0.00415			0.07926	0.02780		
n=1000	sabit	0.00143		0.999466839	0.022590281	-0.02388		0.518435234	0.678921543
	x1	610.250	0.00617			0.26251	0.00811		
	x2	-326.198	0.00395			0.38502	0.02684		
	x3	-261.594	0.00299			0.05615	0.03004		
n=1000	sabit	-0.00137		0.999603574	0.019479355	-0.02465		0.517939963	0.679272706
	x1	536.803	0.00464			0.26650	0.00843		
	x2	-282.543	0.00305			0.40754	0.02876		
	x3	-215.966	0.00220			0.02282	0.03301		

Models and coefficients for all data sets were statistically significant. In order to obtain multiple connections, all relations were defined positively in the data generation. However, as one of the expected effects of multicollinearity in the least-squares analysis, the regression coefficients for the independent variables x2 and x3 were obtained with inverse (negative) signs. In the analysis of Principal Components Regression, the sign of the coefficients was found to be in the right direction (positive). The coefficients obtained in the LSR analysis and the coefficients obtained as a result of the PCR analysis are different from each other in terms of magnitude, although they are different in sign. In addition, the standard errors of the coefficients in PCR results are lower than in the LSR results.

4. DISCUSSION

If the assumption that the independent variables are not related to each other, which is one of the assumptions of the multiple linear regression analysis, is not fulfilled, the multicollinearity problem is encountered. This situation may cause the real values of the parameters to be estimated cannot be obtained, the absolute values of the estimations to be large and the signs of the estimations to change. Due to the negative effects of the multicollinearity situation in the data on the obtained regression model, this situation should be eliminated or its effect should be reduced [19]. There are some suggested methods for eliminating the multi-connection situation. The first thing to do is to select the variable appropriately during the creation of the regression model. In addition, adding new observations to the data, rebuilding the model, or using some biased estimation methods are also methods used in the process of eliminating multicollinearity. Each method may have its own application area and drawbacks. For example, it is recommended to add new observations appropriate to the data in a multicollinearity situation that arises because the sample created does not represent the universe from which it was selected very well. However, it may not always be possible to add sampling units. One or more independent variables may need to be dropped from the model. This process is called model redefinition. In this process, which variables will be removed from the model may be a problem and this approach may cause us to define the model incorrectly [3, 20].

Another approach that can be used to cope with the negative effects of multicollinearity on the model is the use of biased estimation methods. The most preferred methods using biased estimation results are Principal Component Regression, Ridge Regression, and Partial Least Squares Regression.

In this study, it was aimed to remove the effects of multicollinearity by applying PCR to the datasets in which the effects of multicollinearity were observed and to determine in which cases PCR can be used instead of LSR regression. For this reason, both methods were applied to the 10 data sets in the data group that was derived to have different degrees of multicollinearity for our purpose. Although the relationships between all variables were defined to be positive while generating the data, as one of the expected effects of multicollinearity, the sign of the regression coefficients for x2 and x3 variables was obtained as negative in the LSR analysis. In PCR results, the signs of the coefficients are positive. LSR regression coefficients are also different as a numerical size from PCR results, and the coefficient value tends to increase as multicollinearity increases. At the same time, the standard errors of the LSR coefficients are also larger than the PCR results. Although it seems more explanatory for the LSR analysis, the estimations to be made with the help of this model will not be correct since the assumptions are not provided.

As a result, the existence of multicollinearity should be examined while performing multivariate models and one of the regression methods using biased estimation results should be used as a solution to this situation. Otherwise, the estimations to be made may lead to wrong and biased results.

REFERENCES

[1] F. Akdeniz, Olasılık ve istatistik: Akademisyen Kitabevi, 2015.

[2] S. Şahinler, "En küçük kareler yöntemi ile doğrusal regresyon modeli oluşturmanın temel prensipleri," Mustafa Kemal Üniversitesi Ziraat Fakültesi Dergisi, vol. 5, pp. 57-73, 2000.

[3] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to linear regression analysis vol. 821: John Wiley & Sons, 2012.

[4] R. Özkale, "Çoklu İç İlişki ile İlgili Yöntemler," Doktora, Fen Bilimleri Enstitüsü, Çukurova Üniversitesi, 2007.

[5] I. Corp, "IBM SPSS statistics for windows, version 22.0," Armonk, NY: IBM Corp, 2013.

[6] R. Alpar, Spor, sağlık ve eğitim bilimlerinden örneklerle uygulamalı istatistik ve geçerlik-güvenirlik: Detay Yayıncılık, 2010.

[7] K. Özdamar, "SPSS ile Biyoistatistik. Kaan Kitabevi.. 4," Baskı. Eskişehir, pp. 315-368, 2001.

[8] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, Applied linear statistical models vol. 4: Irwin Chicago, 1996.

[9] S. Cankaya, "A comparative study of some estimation methods for parameters and effects of outliers in simple regression model for research on small ruminants," Tropical animal health and production, vol. 41, pp. 35-41, 2009.

[10] W. W. Hines, D. C. Montgomery, and D. M. G. C. M. Borror, Probability and statistics in engineering: John Wiley & Sons, 2008.

[11] R. Alpar, Uygulamalı çok değişkenli istatistiksel yöntemlere giriş-I: Bağırgan Yayımevi, 1997.

[12] M. Topal, E. Eyduran, A. M. Yağanoğlu, A. Sönmez, and S. Keskin, "Çoklu doğrusal bağlantı durumunda ridge ve temel bileşenler regresyon analiz yöntemlerinin kullanımı," Atatürk Üniversitesi Ziraat Fakültesi Dergisi, vol. 41, pp. 53-57, 2010.

[13] A. İşi, "Yanlı Tahmin Ediciler ve Kombinasyonları," Yüksek Lisans Tezi. Ankara: Gazi Üniversitesi Fen Bilimleri Enstitüsü, 2002.

[14] A. Göktaş and İ. Öznur, "Türkiye'de İşsizlik Oranının Temel Bileşenli Reegresyon Analizi ile Belirlenmesi " Sosyal Ekonomik Araştırmalar Dergisi, vol. 10, pp. 279-294, 2010.

[15] F. Aşkın, "Ortalama Artelyel Kan Basıncını Etkileyen Faktörlerin Temel Bileşenler Regresyonu İle Belirlenmesi," Yüksek Lisans, Fen Bilimleri Enstitüsü, Fırat Üniversitesi, 2011.

[16] N. Ortabaş, "Principal components in the problem of multicollineartity," DEÜ Fen Bilimleri Enstitüsü, 2001.

[17] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in explainable AI," arXiv preprint arXiv:1810.00184, 2018.

[18] S. IBMCorp Ibm, "statistics for windows, version 25.0," Armonk, NY: IBM Corp, 2017.

[19] A. S. Albayrak, "Çoklu doğrusal bağlanti halİnde enküçük kareler teknİğİnİn alternatİfİ yanli tahmİn teknİklerİ ve bİr uygulama," Uluslararası Yönetim İktisat ve İşletme Dergisi, vol. 1, pp. 105-126, 2005.

[20] K. Orçanli, B. Bİrgören, and E. Oktay, "Çok Değişkenli Kalite Kontrolünde Süreç Tabanlı Temel Gösterimleri Yönteminin Hata Teriminde Kovaryansın Etkileri," SosyalBilimler Araştırma Dergisi, vol. 6, pp. 20-40.

BIOGRAPHIES

Zeynep Küçükakçalı obtained her BSc. degree in mathematics from Çukurova University in 2010. She received MSc. degree in biostatistics and medical informatics from the Inonu University in 2018. She currently continues Ph.D. degrees in biostatistics and medical informatics from the Inonu University. In 2014, she joined the Department of Biostatistics and Medical Informatics at Inonu University as a researcher assistant. Her research interests are cognitive systems, data mining, machine learning, deep learning.

Harika Gözükara Bağ obtained her BSc. degree in Agricultural Engineering from Ankara University in 2000. She received MSc. degree in Biostatistics from the Inonu University in 2004, and a PhD degree in Biostatistics from Hacettepe University in 2011. She is currently working as an Associate Professor at the Department of Biostatistics and Medical Informatics at Inonu University since 2019. She is active in teaching and her research interests are genetics, survival analysis and meta-analysis.