ISSN:1306-3111
e-Journal of New World Sciences Academy
2011, Volume: 6, Number: 2, Article Number: 3A0033

PHYSICAL SCIENCES
Received: November 2010
Accepted: February 2011
Series : 3A
ISSN : 1308-7304
© 2010 www.newwsa.com

Elif Bulut[1]
Özlem Gürünlü Alma[2]
Ondokuz Mayıs University[1]
Mugla University[2]
bulut_elif@yahoo.com
Samsun-Turkey

DIMENSIONALITY REDUCTION METHODS: PCR, PLSR, RRR AND A HEALTH APPLICATION

ABSTRACT

Working with data set that has many variables or has fewer observation units than variables leads to difficulties in statistical analysis. In this situation dimension reduction is a necessary part of the data analysis. It is necessary because, it provides working with a subset of the existing features or to transform to a new reduced set of features and working with low dimensional data and simplify the data model by working with parsimonious model. There are some dimensionality reduction methods and all of them lean to use a linear combinations of *m* variables by reducing *m* dimensional data set to *a* dimensional data set (*a*<*m*) that explain the majority of the variability in the variables. This paper provides study of three dimension reduction techniques, namely Principal Component Regression (PCR), Partial Least squares Regression (PLSR), and Reduced Rank Regression (RRR), and they were illustrated on a data set that has PCOS disease to help to choose the efficient factors (latent variables) for modeling and predicting fsh and lh hormones when the data set has small number of observation unit.

Keywords: Dimension Reduction, Principal Component Regression, Partial Least Squares Regression, Polycystic Ovary Syndrome, Reduced Rank Regression

BOYUT İNDİRGEME TEKNİKLERİ: PCR, PLSR, RRR VE BİR SAĞLIK UYGULAMASI

ÖZET

Çok fazla değişkene sahip veya değişken sayısından daha az gözlem sayısına sahip veri seti ile çalışmak istatistiksel analizde bazı zorluklara yol açmaktadır. Böyle bir durumda boyut indirgemesi analizin önemli bir parçasıdır. Boyut indirgemesi, veri setinde var olan özelliklere sahip daha küçük bir veri seti ile çalışmayı mümkün kılmaktadır. Boyut indirgeme teknikleri m boyutlu veri setini, m değişkenlerdeki değişimin büyük bir kısmını açıklayacak ve bu değişkenlerin doğrusal birleşimi olacak şekilde a boyutlu veri setine indirgemektedir. Bu çalışmada, bu tekniklerden Temel bileşenler regresyonu, Kısmi en küçük kareler regresyonu ve İndirgenmiş rank regresyonu yöntemleri anlatılarak, sağlık verisi üzerinde uygulaması gösterilmiştir.

Anahtar Kelimeler: Boyut İndirgeme, Temel Bileşenler Regresyonu, Kısmi En Küçük Kareler Regresyonu, Polikistik Over Sendromu, Indirgenmiş Rank Regresyonu

1. INTRODUCTION (GİRİŞ)

Regression models the continuous relationship between two sets of variables, usually called explanatory and response variables (or inputs and outputs). The process of modeling entails finding the structure as well as the free parameters of a function such that it optimally describes a given set of input and output data. Regression is a generic and important statistical tool with a wide field of applications ranging from data mining, signal processing, chemometrics (Wold et al, 1984) and econometrics (Geweke, 1996) to adaptive learning control and robotics (Vijayakumar et al, 2002) (Hoffman et al., 2009).

In multivariate data analysis, existence of large number of variables sometimes fails to understand the data structure. To reduce the multivariate problems, dimension reduction is a necessary part of a statistical analysis. For this study, dimension reduction methods, PCR, PLSR and RRR, belong to the following three groups as defined by (Hoffman et al., 2009): "(1) reducing dimensionality only on the input data, (2) modeling the joint input-output data distribution, and (3) optimizing the correlation between projection directions and output data" were explained. Group 1 contains PCR, group 2 contains principal component analysis (PCA) in joint input and output space, factor analysis, and probabilistic PCA, and group 3 contains RRR and PLSR regression (Hoffman et al., 2009). The objective of this work is to use of dimension reduction methods such as PCR, RRR and PLSR on medical data, in order to make the selection of effective factors on PCOS disease. In this study, capital and bold letters represent matrix, lower case and bold letters represent vectors.

Section 3 contains summary of information about the PCR, PLSR, and RRR methods. Also, in section three, details of PLSR NIPALS algorithm was given. Section 4-5 give detailed information about application of PCR PLSR RRR, on PCOS disease and their results using SAS statistical program. Conclusions and comments are given in Sections 6.

2. RESEARCH SIGNIFICANCE (ÇALIŞMANIN ÖNEMİ)

In this study, some dimensionality reduction methods were introduced and illustrated on a health study. These methods simplify the data model by working with a parsimonious model. They find new latent variables with different aims. PCR and RRR are interested in latent variables that capture most of the variation in explanatory variables and response variables, respectively. PLSR works with latent variables that model the relation between two blocks of variables also it overcomes multicollinearity and less number of observation unit problems. This study aims to introduce methods especially for the researchers in medicine by interpreting the results.

3. MATERIALS AND METHODS (MATERYAL VE YÖNTEM)

3.1. Principal Component Regression (Temel Bileşenler Regresyonu)

PCR is obtained by regressing $\mathbf{Y}_{N\times K}$ on the components (latent variables) **t** obtained from PCA. PCA select a new set of explanatory variables called components with the decreasing of variance within the explanatory variable matrix, $\mathbf{X}_{N\times M}$. These components are perpendicular to each other that there is no multicollinearity among them. It defines all the latent variables, $\mathbf{T}_{N\times A}=(\mathbf{t_1},\ldots,\mathbf{t}_a)$, $a=1,\ldots,A$, depends on the original descriptors therefore only deals with the variance-covariance matrix of explanatory variables, $\mathbf{X'X}$. The aim is to find the first *a* principal component of $\mathbf{X'X}$ starting with the largest eigenvalue λ_1 and down. PCR used principal component analysis of **X** to determine loadings $\mathbf{P}_{(PCA)M\times A}$ to be used in $\mathbf{T}_{N\times A} = \mathbf{X}_{N\times M}\mathbf{P}_{(PCA)M\times A}$. Here $\mathbf{T'T} = \mathrm{diag}(\lambda_a)$. The regression coefficients $\hat{\mathbf{b}}_{PCR}$ for each **y** can be written as $\hat{\mathbf{b}}_{PCR} = \mathbf{Pq}$. Here **q** are found by least squares regression of **y** on **T**. For further information look (Martens and Naes, 1989).

3.2. Partial Least Squares Regression (Kısmi En Küçük Kareler Regresyonu)

Herman O.A. Wold vigorously pursued the creation and construction of models and methods for the social sciences, where "soft models and soft data" were the rule rather than the exception and where approaches strongly oriented at prediction would be of great value. The author was fortunate to witness the development firsthand for a few years. Herman Wold (1977) suggested to write a PhD-thesis on Lisrel versus PLS in the context of latent variable models, more specifically of "the basic design" (Dijkstra, 2010).

The use of the PLS method for chemical applications was pioneered by the groups of S. Wold and H. Martens in the late seventies after an initial application by Kowalski (Geladi et al., 1986). Geladi (1988) was offered a review of historical development of PLS. PLS regression was studied and developed from the point of view of statisticians by Agnar Höskuldsson (1988). The book by Martens and Naes (1989) used statistical concepts that began to provide a theoretical basis for PLS. The recent investigations were provided by Inge Helland (1990), Paul Garthwaite (1994), Svante Wold (2001), Tobias (2003) and Abdi (2007, 2010).

PLSR enables working with small number of observation units and/or data set with multicollinearity and/or more than one response variable. PLSR involves information on both **X** and **Y** in the calculation of components and loadings by using singular value decomposition of $\mathbf{S} = \mathbf{X'Y}$ cross product matrix. **X** and **Y** data matrices can be modeled separately by these components as given below (1), (2):

$$\mathbf{X}_{N\times M} = \mathbf{T}_{N\times A}\mathbf{P}'_{A\times M} + \mathbf{E}_{N\times M} \qquad (1)$$

$$\mathbf{Y}_{N\times K} = \mathbf{U}_{N\times A}\mathbf{C}'_{A\times K} + \mathbf{F}_{N\times K} \qquad (2)$$

Here, $\mathbf{T}_{N\times A} = \mathbf{X}_{N\times M}\mathbf{W}_{M\times A}$, and $\mathbf{U}_{N\times A} = \mathbf{Y}_{N\times K}\mathbf{C}_{K\times A}$, summarize **X** and **Y** variables and $\mathbf{P}'$ and $\mathbf{C}'$ represent loading and weight matrices, respectively. Matrices of loading and weight are loading of the **t**'s on the **X** variables and the weight of the response variable on the latent vector of **X**, respectively. $\mathbf{W}_{M\times A}$ is the weight matrix of **X** variables obtained by regressing **X** on $\mathbf{u}_{N\times 1}$ that is, by using variability in the response variable. Here, $\mathbf{E}_{N\times M}$ and $\mathbf{F}_{N\times K}$ are matrices of residuals and show the unmodeled structure at all the **X** and **Y**. PLS is an iterated process. In each step, data matrices are deflated until a convergence between the latent variables obtained in current step and used in the previous step or a null-matrix of **X** variables are obtained. Then, regression coefficients for PLSR are obtained from $\mathbf{B}_{M\times K} = \mathbf{W}\,(\mathbf{P'W})^{-1}\mathbf{C}'$.

3.2.1. NIPALS (Non-Linear Iterative Partial Least Squares) Algorithm (Doğrusal Olmayan İteratif Kısmi En Küçük Kareler Regresyon Algoritması)

The NIPALS algorithm, also known as the classical algorithm, was developed by H. Wold, 1960s’. It was first used for PCA and later for PLS. It is the most commonly used method for calculating the principal components of a data set. It numerically gives more accurate results when compared with Singular Value Decomposition (SVD) of the covariance matrix, however is slower to calculate. The starting point of the algorithm is two data matrices **X** and **Y**. Algorithm based on deflating **X** and **Y** variables in each iteration that is PLS weights are iteratively estimated. In each iteration matrices are deflated as $\mathbf{X_{d+1}} \rightarrow \mathbf{X_d} - \mathbf{t_{d+1}p'_{d+1}}$ and $\mathbf{Y_{d+1}} \rightarrow \mathbf{Y_d} - \mathrm{b}\mathbf{t_{d+1}c'_{d+1}}$. Here, d represents the iteration number. $\mathbf{t_{d+1}p'_{d+1}}$ represents the predicted part obtained by algorithm in the (d+1)th iteration, $\mathbf{X_d}$ represents the matrix obtained by algorithm from the d)th iteration. $\mathbf{X_{d+1}}$ represents residual value that will use in the next iteration. Same comments are valid for $\mathbf{Y_{d+1}}$ only b is the regression coefficients of inner relation. Iteration continues with these deflated matrices until **X** becomes a null matrix. For further details look Höskuldsson (1988). SAS software uses NIPALS algorithm if you did not specify a different method.

3.3. Reduced Rank Regression (RRR) (İndirgenmiş Rank Regresyonu)

In the study of the experimental properties of mixtures, a linear model is often proposed to relate response to composition. The statistical technique of linear regression analysis is then appropriate and it is often applied severally when there are a number of responses of interest. Now, the responses are often inter-related, so that, for instance, it may be possible to use an empirical linear relationship to predict the approximate value of a certain response from knowledge of the others. The procedure for determining the regression coefficients of response on composition should, in these circumstances, is modified to reflect the known presence of such relationships (whose linearity is implied by the mutual linear dependence of responses on composition). This leads to consideration of the multivariate regression model with a constraint imposed on the rank of the matrix of coefficients, sometimes termed reduced-rank regression. Such models have been studied, e.g. by Izenman (1975) and also by Burket (1964), who used a factor analysis model (Davies et al., 1982).

The RRR model is a multivariate regression model with a coefficient matrix with reduced rank. The RRR algorithm is an estimation procedure, which estimates the RRR model. It is related to canonical correlations and involves calculating eigenvalues and eigenvectors (Johansen, 2008). The solution for the RRR analysis is related to the singular value decomposition of the full rank matrix. In RRR analysis, principal component analysis is first performed on **Y** followed by regressing **X** on the principal components. It is based on maximizing the covariance between the principal components and response variables. That is, k response variables are regressed separately on the explanatory variables.

As discussed in the preceding sections, partial least squares depends on selecting components $\mathbf{t} = \mathbf{Xw}$ of the explanatory variables and $\mathbf{u} = \mathbf{Yc}$ of the responses that have maximum covariance, whereas principal component regression effectively ignores **u** and selects **t** to have maximum variance, subject to orthogonality constraints. In contrast, reduced rank regression selects **u** to account for as much variation in the predicted responses as possible, effectively ignoring the explanatory variables for the purposes of factor extraction. In reduced rank regression, the **Y** weights $\mathbf{c_i}$ are the eigenvectors of the covariance matrix $\hat{\mathbf{Y}}'_{\mathrm{LS}}\hat{\mathbf{Y}}_{\mathrm{LS}}$ of the responses predicted by ordinary least squares regression; the **X** components are the projections of the **Y** components $\mathbf{Yc_i}$ onto the **X** space (SAS/STAT 9.1 User's Guide, 2004). RRR takes the first principal components of the ordinary regression matrix. These eigenvectors play the same role as the components **T** in PCR and PLS (Kiers et al., 2007). Coefficient matrix can be written as a product of two component matrices of lower dimension. It follows that the assumption of lower rank for the regression coefficient matrix leads to estimation results which take into account the interrelations among the multiple responses and use the entire set of explanatory variables in a systematic fashion (Reincell, 2006).

The model for reduced-rank regression may be written as

$$\mathbf{Y}_{\mathrm{N\times K}} = \mathbf{X}_{\mathrm{N\times M}}\mathbf{D}_{\mathrm{M\times K}} + \mathbf{E}^{*}{}_{\mathrm{N\times K}} \tag{3}$$

$$\text{rank } (\mathbf{D}) < s,$$

where s is an integer to be specified. The interpretation of (3) is as follows. $\mathbf{X}_{\mathrm{N\times M}}$ and $\mathbf{Y}_{\mathrm{N\times K}}$ are data matrices whose N rows contain measurements on M and K variables respectively for N individuals or experimental units. Assume that column means have been subtracted from each variable of **X** and **Y**. This corresponds to the situation that a constant term associated with each regression has been previously estimated (by maximum likelihood in the case of normality) and allows us to consider the homogeneous model (1) without loss of generality. The problem is to estimate the unknown matrix of regression coefficients $\mathbf{D}_{\mathrm{M\times K}}$ subject to the rank constraint rank (**D**)< s <min (M, K) given **X** and **Y**. (For convenience we shall assume that **X** and **Y** are full- rank matrices.) The rank restriction on **D** has the interpretation that fewer than min (M, K) linear combinations of the **x**-variables in fact enter into the prediction of the **y**-variables; thus for K < M it imposes the condition that the predictions shall be linearly dependent. Finally, $\mathbf{E}^{*}$ is the matrix of stochastic errors which are assumed to be uncorrelated row-wise, that is from unit to unit, but which may be correlated between variables measured on the same unit. We shall assume a zero-mean K-variate multi-normal distribution for the rows of $\mathbf{E}^{*}$, $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ is an unknown positive definite covariance matrix. It is natural to make explicit the reduced-rank nature of $\mathbf{D}_{\mathrm{M\times K}}$ by expression this matrix as the product of two matrices, $\mathbf{D}_{\mathrm{M\times K}} = \mathbf{Q}_{\mathrm{M\times s}}\mathbf{B}_{\mathrm{s\times K}}$, where $\mathbf{Q}_{\mathrm{M\times s}}$ is a matrix whose s columns are a set of linearly independent vectors representing a basis for the unknown subspace spanned by the columns of **D**, and $\mathbf{B}_{\mathrm{s\times K}}$ has K columns that define the appropriate linear combinations to represent the columns of **D**; i.e. the regression coefficients for each **y**-variable with respect to this basis. We shall choose **Q** to have the normalization $\mathbf{P}'\mathbf{P} = \mathbf{I_s}$ where $\mathbf{P}_{\mathrm{N\times s}} = \mathbf{X}_{\mathrm{N\times M}}\mathbf{Q}_{\mathrm{M\times s}}$; that is, the s columns of **P** are orthogonal linear combinations of the **x**-variables. This definition is consistent with canonical variate analysis and in fact it shall show that the columns of **Q** may be estimated as the s principal canonical linear combinations (Tso, 1981).

4. APPLICATION OF PCR PLSR AND RRR ON PCOS DISEASE (PCR PLSR VE RRR PCOS HASTALIĞI ÜZERİNE UYGULAMA)

The data used in this study was a part of a study that have done by (Çapoğlu et al., 2009) in Atatürk University School of Medicine. We used only the data of patients with PCOS disorder. In this study, the explanatory variables matrix $\mathbf{X}$=($\mathbf{x}_m$, m=1,…,M) consists age, body mass index and some hormones measured in different scales, whereas the response matrix $\mathbf{Y}$=($\mathbf{y}_k$, k=1,…,K) are formed from hormones fsh (follicle stimulating hormone) and lh (lutenizing hormone). The data sets in this study is about 15 females, ages between 18 and 26 and have Polycystic ovary syndrome disorder (PCOS). PCOS is one of the female endocrine disorders with one of characteristics increase in lh relative to fsh release, have long been recognized in PCOS (Yen, 1980). There are many studies in the literature about PCOS. Some of them are Escobar-Morreale et al., (2006), Barnard et al., (2007) and Brennan et al., (2009). In this study, PCR, PLSR and RRR methods were used to obtain components and then PLSR was used in modeling fsh and lh hormones in terms of explanatory variables. As mentioned above, fsh and lh hormones are one of the diagnostics of PCOS disorder. Variables are centered and scaled to have zero mean and one standard deviation because of having different measurement units. All of the analyses were performed by using SAS statistical software.

Table 1. Explanatory variables
(Tablo 1. Açıklayıcı değişkenler)

$\mathbf{x}_1$	İnsulin (insu)
$\mathbf{x}_2$	Cpeptide (cpep)
$\mathbf{x}_3$	Dehydroepiandrosterone sulfate (dhs)
$\mathbf{x}_4$	Thyroid stimulating hormone (ths)
$\mathbf{x}_5$	Ferritin
$\mathbf{x}_6$	Resistin (resi)
$\mathbf{x}_7$	Testesterone (tes)
$\mathbf{x}_8$	Androgen (andr)
$\mathbf{x}_9$	Age
$\mathbf{x}_{10}$	Body mass index (bmi)
$\mathbf{x}_{11}$	Hemoglobin (hb)
$\mathbf{x}_{12}$	Endorphin (endo)
$\mathbf{x}_{13}$	Erythrosin (erit)
$\mathbf{x}_{14}$	Vascular endothelial growth factor (vegf).
$\mathbf{x}_{15}$	Adiponektin (adipo)
$\mathbf{x}_{16}$	C-reactive protein (crp)

5. FINDINGS AND DISCUSSION (BULGULAR VE TARTIŞMALAR)

The PCR and PLSR analysis results showed that 7 components explain most of the variability on both explanatory and response variables while RRR works with 2 components. So, analysis was carried on 7 components for PCR and PLSR.

Table 2. Percent variation accounted for PCR
(Tablo 2. PCR ile açıklanan değişim yüzdesi)

Percent variation accounted for by principal components				
Number of extracted factors	Model effects		Response variables	
	Current (X variance)	Total (summary for model)	Current (Y variance)	Total (summary for responses)
1	25.2133	25.2133	1.6055	1.6055
2	19.0046	44.2179	2.0781	3.6836
3	16.7312	60.9491	4.3231	8.0067
4	10.5514	71.5006	14.4033	22.4100
5	7.2449	78.7455	4.2312	26.6411
6	6.4187	85.1642	1.5976	28.2387
7	4.3115	89.4757	13.8213	42.0600

Table 2 expresses the percentage of variance explanation over 7 components. Percentages of the explained variances are 89.48% for explanatory variables and 42.06% for response variables. In current block, values show the percentages of explained variance for each explanatory variables and total block shows the cumulative total of the percentage of explained variance for the model. The same explanations are valid for the response variables block.

Table 3. Percent variation accounted for components for PLSR
(Tablo 3. PLSR bileşenleri tarafından açıklanan değişim yüzdesi)

Percent variation accounted for by partial least squares factors						
Number of extracted factors	Model effects		Response variables		fsh	lh
	Current (X variance)	Total (summary for model)	Current (Y variance)	Total (summary for responses)	R-Sq	R-Sq
1	15.3886	15.3886	42.4811	42.4811	43.2122	41.7503
2	15.2228	30.6115	19.2578	61.7389	61.8756	61.6026
3	12.6001	43.2116	13.2686	75.0075	67.7759	82.2393
4	12.1295	55.3411	5.5905	80.5979	68.1255	93.0706
5	7.3759	62.7170	3.9935	84.5914	75.6870	93.4960
6	7.1210	69.8380	4.1743	88.7658	80.7268	96.8049
7	13.7375	83.5755	1.3976	90.1634	83.1242	97.2027

Table 3 gives the results for PLS regression. The total percentages of the explained variations are 83.58% for model and 90.2% for responses. Unlike PCR, PLSR also gives the percentages of explained variation for each response variable. 7 components explain the total variation as 83.12% for fsh and 97.20% for lh. For both response variables, components explain the variation as 90.16%.

Table 4. Percent variation accounted for RRR
(Tablo 4. RRR için açıklanan değişim yüzdesi)

Percent variation accounted for by reduced rank regression factors				
Number of extracted factors	Model effects		Response variables	
	Current	Total	Current	total
1	5.1380	5.1380	91.0128	91.0128
2	3.5400	8.6780	8.9872	100.000

In RRR analysis, first factor explains maximum part of the variability in the response variables. Second factor alone explains only 9% of the response variation. Same factors explain less amount of variation in the explanatory variables.

Subsequent analysis was carried on PLSR since it has the maximum cumulative percentage of explained variance. PLSR used both response and explanatory variables in analysis. The VIP scores and the beta coefficients are obtained by PLS regression can be used to select the most influential variables (Chong and Jun 2005). The VIP score can be estimated for jth explanatory variable by:

$$\mathrm{VIP_j} = \sqrt{\mathrm{M} \times \frac{\sum_{a} \mathrm{w}_{ja}^2 \mathrm{b}_a^2 \mathbf{t}_a' \mathbf{t}_a}{\sum_{a} \mathrm{b}_a^2 \mathbf{t}_a' \mathbf{t}_a}} \qquad (4)$$

where w_{ja} is a weight of the jth **X**-variable to the ath latent variable which is obtained by NIPALS algorithm Jun et al. (2009), and b_a is the regression coefficients of inner relation.). w_{ja} $j=1,\dots,16$ and $a=1,\dots,7$ values are given in the following table. Weight values can be interpreted as the contribution of the jth explanatory variable to the ath latent variable.

Table 5. The matrix **W**
(Tablo 5. **W** matrisi)

	$\mathbf{w}_1$	$\mathbf{w}_2$	$\mathbf{w}_3$	$\mathbf{w}_4$	$\mathbf{w}_5$	$\mathbf{w}_6$	$\mathbf{w}_7$
Insu	0.34139	0.01498	0.29812	-0.3572	0.11899	-0.4196	-0.4052
Cpep	-0.2146	-0.1642	-0.2737	0.33091	0.14722	-0.4196	-0.2653
Dhs	0.02946	0.37145	-0.3913	0.19853	0.32498	0.1530	-0.0419
Ths	0.12330	0.25932	-0.0938	-0.1538	0.47243	-0.1820	0.48307
Ferrirtin	-0.2649	0.56607	0.09460	-0.2969	-0.1969	-0.2409	-0.0513
Resi	0.30816	0.41389	0.03404	-0.0802	-0.2134	0.56261	-0.4926
Tes	-0.2256	0.26107	-0.2549	0.18034	-0.0249	-0.1393	-0.5193
Andr	-0.1763	-0.0812	-0.6502	-0.3157	0.02761	0.40849	-0.1669
Age	-0.7652	0.02785	0.06177	0.06380	0.16825	0.29281	-0.0299
Bmi	-0.1463	0.17843	0.21003	-0.3942	0.17344	-0.3548	-0.1893
Hb	-0.3237	0.50381	0.39352	0.14846	0.12260	-0.0954	0.01096
Endo	-0.4538	-0.4841	0.05496	-0.4898	-0.2887	-0.0202	-0.0102
Erit	0.05601	0.12580	-0.5227	0.04278	-0.4974	-0.7504	0.08397
Vegf	-0.1437	0.20041	-0.1799	-0.1895	0.28242	-0.1448	0.21114
Adipo	0.10522	-0.3692	-0.1199	-0.1411	0.65249	-0.1479	-0.3574
crp	0.14008	0.25040	-0.3473	-0.4335	-0.0104	0.20821	0.18789

Table 6. The matrix **C**
(Tablo 6. **C** matrisi)

	$\mathbf{c}_1$	$\mathbf{c}_2$	$\mathbf{c}_3$	$\mathbf{c}_4$	$\mathbf{c}_5$	$\mathbf{c}_6$	$\mathbf{c}_7$
fsh	0.71316	0.69610	0.47153	0.17681	0.97299	0.77697	0.92611
lh	0.70099	0.71793	0.88184	0.98425	-0.2309	0.62954	0.37724

In Table 6, c_{ka} gives the weight of the kth **y**-variable k=1,2 to the ath latent variable. The importance of the ath latent variable in modeling $\mathbf{y}_k$ is measured with c_{ka}.

Table 7. Regression coefficients and VIP values
(Tablo 7. Regresyon katsayıları ve VIP değerleri)

Explanatory variable	Regression coefficients		VIP
	Fsh	Lh	
Insu	0.2056265	0.2473931	0.92802
Cpep	-0.3973759	-0.4273216	0.64815
Dhs	0.3172521	0.1832351	0.64525
Ths	0.3931684	0.1943243	0.5586
Ferritin	-0.0800293	0.0357948	1.21258
Resi	0.4841168	0.6084864	1.10008
Tes	-0.194544	-0.1467852	0.76045
Andr	-0.3394625	-0.5803727	0.49923
Age	-0.3983012	-0.5112621	2.07967
Bmi	-0.0582910	-0.0941500	0.50259
Hb	0.2067884	0.3086763	1.23595
Endo	-0.8250317	-0.8714472	1.48901
Erit	-0.4870774	-0.3178983	0.26494
Vegf	0.0125276	-0.1491177	0.52127
Adipo	0.0299710	-0.2653342	0.69772
crp	0.1348927	0.0050763	0.57547

Table 7 gives the regression coefficients of explanatory variables for different responses. Explanatory variables with small coefficients make a small contribution to the response prediction. According to the regression coefficients, the contribution of dhs to fsh is bigger than the contribution to lh and also the contributions of ferritin, erit, vegf and adipo show a contrast in terms of response variables. Adipo has a great contribution to a negative sign in predicting lh while it has less contribution to predicting fsh. The contribution of endorphin to response variables is the biggest of those.

VIP (variable importance in the projection) is a statistic of Wold (1994), summarizing the contribution that a variable makes to the model. VIP block gives the value of each explanatory variable in fitting the PLS model for both explanatory and response variables. If an explanatory variable has a relatively small coefficient (in absolute value) and a small value of VIP, then it is a prime candidate for deletion. Wold(1994) considers a value less than 0.8 to be “small” for the VIP(SAS/STAT 9.1 User’s Guide, 2004).

From VIP block, it can be easily seen that those explanatory variables: insu, ferritin, resi, age, hb and endo have VIP value bigger than 0.8. These variables are the most relevant ones modeling and predicting response variables. According to Palermo et al. (2009), “because of its definition a VIP score derived by multivariate PLS regression would not allow to separate the contribution of each explanatory variable to different responses” we need to give VIP values for each response variable separately. The following table gives VIP values for explanatory variables that exceed threshold value.

Table 8. VIP values for each response variable
(Tablo 8. Herbir bağımlı değişken için VIP değerleri)

Explanatory variable	fsh	Explanatory variable	lh
Insu	1.00083	Insu	0.79970≈0.80
Ths	0.88920	Ferritin	1.23858
Ferritin	1.16870	Resi	1.19726
Resi	0.90351	Tes	0.81854
Age	2.05781	Andr	0.80888
Hb	1.26217	Age	1.97222
Endo	1.44924	Hb	1.21129
Adipo	0.83664	Endo	1.40358
crp	0.80488	Adipo	0.85117

As seen from Table 7 and Table 8, insu, ferritin, resi, age, hb and endo have contributions to predicting both response variables simultaneously and predicting responses separately.

6. CONCLUSION (SONUÇ)

There are many multivariate statistical methods that overcome multicollinearity and multiple variables problem. These methods serve different aims according to the study’s purpose. In this study, we briefly summarized three of them, RRR, PCR and PLSR, and then interpret the results. Methods were explained on the application about medical PCOS data. There are so many studies about PCOS in the literature. In the literature, it is stated that clinical findings such as increasing and/or decreasing in hormones; insu, fsh, lh, resi, hb and endo are used as a diagnostic in PCOS. But in this study, we try to find the contributions of hormones to fsh and lh. As seen from the results, insu, ferritin, resi, age, hb and endo have the maximum contributions in responses. Also Table 7 shows that the variables with high VIP value have high coefficients except ferritin. Statistically, this is the expected situation.

According to Table 7, there is an interaction between insulin and responses. Insulin has a positive contribution to fsh and lh measurement levels. Coefficients show that ferritin has a negative contribution to fsh and positive contribution to lh, respectively. Resistin has a positive contribution to both responses with a highly coefficients. The contribution of age to responses is in a negative way. Because of PCOS being a disease in women of reproductive age, and also fsh and lh measurement levels are already different from normal, therefore being in small age has a negative contribution to responses. Finally, the contributions of hormones; hemoglobin and endorphin to responses are in a positive way and in a negative way with high coefficients, respectively.

All this comments are the results of PLSR analysis. We choose it because of better describing of variability on both responses and explanatory variables than PCR and RRR. As a further study the effects of age and body-mass index have been investigated with RRR analysis by taking them as explanatory variables and hormones as response variables.

REFERENCES (KAYNAKLAR)

1. Abdi, H., (2007). Partial Least Square Regression (PLS regression).In: Salkind. N. J. (ed.). Encylopedia of Measurement and Statistics.

2. Abdi, H., (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). Wiley Interdisciplinary Reviews: Computational Statistics. 2. Issue 1, pp: 97–106.

3. Barnard, L., Ferriday, D., Guenther, N., Strauss, B., Balen, A.H., and Dye, L., (2007). Quality of life and psychological well being in polycystic ovary syndrome. Human Reproduction 22, 8, pp: 2279–2286.

4. Brennan, K., Huang, A., and Ricardo, A.R., (2009). Dehydroepiandrosterone sulfate and insulin resistance in patients with polycystic ovary syndrome. Fertil Steril. 2009 May ; 91(5), pp: 1848–1852.

5. Burket, G.R., (1964). A study of reduced-rank models for multiple prediction. Psychometric Monograph No. 12 (Psychometrika)

6. Chong, I.G. and Jun, C.H., (2005). Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78, pp: 103–112.

7. Çapoğlu, İ., Erdem, F., Uyanık, A., and Turhan, H., (2009). Serum levels of resistin and hsCRP in women with PCOS. Central European Journal of Medicine, 4[4], pp: 428-432.

8. Davies, P.T., and Tso, M.K.S., (1982). Procedures for Reduced-rank Regression. Appl. Statist. 31. No. 3, pp: 244-255.

9. Dijkstra, T.K., (2010). Handbook of Partial Least Squares. Concepts. Methods and Applications. Springer.

10. Escobar-Morreale, H.F., Villuendas, G., Botella-Carretero, J.I., Álvarez-Blasco, F., Sanchón, R., Luque-Ramírez, M., and San Millán, J.L. (2006). Adiponectin and resistin in PCOS: a clinical, biochemical and molecular genetics study. Human Reproduction 21, 9 pp: 2257–2265.

11. Garthwaite, P.H., (1994). An Interpretation of Partial Least Squares. Journal of the American Statistical Association;89, pp: 122-127.

12. Geladi, P. and Kowalski, R., (1986). Partial Least Squares Regression: A Tutorial. Analytica Chimica Acta 185, pp: 1-17.

13. Geladi, P., (1988). Notes on the History and Nature of Partial Least Squares (PLS) Modelling. Journal of Chemometrics;2, pp: 231-246.

14. Geweke, J., (1996). Bayesian reduced rank regression in econometrics. J Econometr 75(1), pp: 121–146.

15. Helland, I.S., (1990). Partial Least Squares Regression and Statistical Models. Scandinavian Journal of Statistics 17, pp: 97-114.

16. Hoffmann, H., Schaal S., and Vijayakumar, S., (2009). Local Dimensionality Reduction for Non-Parametric Regression. Neural Process Lett. 29, pp: 109–131.

17. Höskuldson, A., (1988). PLS Regression Methods. Journal of Chemometrics;2, pp: 211-228.

18. Izenman, A.J., (1975). Reduced-rank regression for the multivariate linear model. J. Multiv. Anal. 5, pp: 248-264

19. Johansen, S., (2008). "Reduced Rank Regression". New Palgrave Dictionary of Economics. Basingstoke. Palgrave Macmillan. 7 pp.

20. Jun, C.H., Lee, S.H., Park, H.S., and Lee, J.H., (2009). Use of Partial Least Squares Regression for Variable Selection and Quality Prediction. This paper appears in: Computers & Industrial Engineering. 2009. CIE 2009. International Conference on 6-9 July 2009.

21. Kiers, H.A.L. and Smilde, A.K., (2007). A comparison of various methods for multivariate regression with highly collinear variables. Stat. Meth. & Appl., 16, pp: 193-228.

22. Martens, H. and Naes, T., (1989). Multivariate Calibration. John Wiley & Sons.

23. Palermo, G., Piraino, P., and Zucht, H.D., (2009). Performance of PLS regression coefficients in selecting variables for each response of a multivariate PLS for omics-type data. Advances and Applications in Bioinformatics and Chemistry 2.

24. Reinsel, G., (2006). University of Wisconsin. Encylopedia of statistical sciences, Volume 11. Page 7015-7028. Wiley-Interscience.

25. SAS/STAT 9.1 User’s Guide, SAS Institue Inc., Cary, NC, USA, 2004.

26. Tobias, R.D., (2003). An Introduction to Partial Least Squares Regression from: www.ats.ucla.edu/stat/sas/library/pls.pdf.

27. Tso, M.K.S., (1981). Reduced-Rank Regression and Canonical Analysis. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 43, No. 2, pp: 183-189.

28. Vijayakumar, S. and Schaal, S., (2000b). Locally weighted projection regression: an O(n) algorithm for incremental real time learning in high dimensional space. In: Proceedings of the 17th international conference on machine learning. Montreal. Canada. pp 1079–1086.

29. Wold, S., Ruhe, A., Wold, H., and Dunn III, W.J., (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J Sci Stat Comput 5(3), pp: 735–743.

30. Wold, S., (1994). PLS for multivariate linear modelling, QSAR: Chemometric metods in molecular design. Methods and principles in medicinal chemistry. (Ed. H. Van de Waterbeemd), Weinheim, Germany: Verlag-Chemie.

31. Wold, S., Sjöström M., and Eriksson, L., (2001). PLS-regression:a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 2001;58, pp: 109-130.

32. Yen, S.S., (1980). The polycystic ovary syndrome. Clin Endocrinol, 12, pp: 177-207.