View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

Doğuş Üniversitesi Dergisi, 20 (1) 2019, 31-47

The Comparison of Robust Partial Least Squares Regression Methods (RSIMPLS, PRM) with Robust Principal Component Regression for Predicting Tourist Arrivals to Turkey

Türkiye'ye Gelen Yabancı Turist Sayısını Kestirmek için Sağlam Kısmi En Küçük Kareler Regresyon Yöntemlerinin (RSIMPLS, PRM) Sağlam Temel Bileşenler Regresyon Yöntemi ile Karşılaştırılması

Esra POLAT [(1)]

ABSTRACT: Tourism is one of the most important component in the economic development strategy of many developing countries such as Turkey. The annual data set of Turkey (1986 - 2013), including the six factors affecting the tourist arrivals, is examined. The aim of this study is modelling the tourist arrivals to Turkey in cases of both multicollinearity and outlier existence in the data set by using a robust Principal Component Regression method: RPCR, two robust Partial Least Squares Regression methods: RSIMPLS and Partial Robust M-Regression (PRM). Hence, the best model giving the best predictions of tourist arrivals is selected and the most important factors are determined.

Keywords: multicollinearity, outliers, robust principal component regression, robust partial least squares regression, tourist arrivals

Jel Classification Code: *C52*

ÖZ: *Turizm, Türkiye gibi gelişmekte olan ülkelerin ekonomik kalkınma stratejilerinde anahtar bileşendir. Türkiye'nin 1986 - 2013 dönemi için, gelen yabancı turist sayısını etkileyen altı faktörün dâhil olduğu veri kümesi incelenir. Bu çalışmanın amacı, veri kümesinde hem çoklu bağlantı hem de uç değer olduğunda Türkiye'ye gelen yabancı turist sayısını bir sağlam Temel Bileşenler Regresyon yöntemi: RPCR, iki sağlam Kısmi En Küçük Kareler Regresyon yöntemleri: RSIMPLS ve Kısmi Sağlam M-Regresyon (PRM) kullanarak modellemektir. Böylece, yabancı turist sayısının en iyi kestirimlerini veren en iyi model seçilir ve en önemli faktörler belirlenir.*

Anahtar Kelimeler: *çoklu bağlantı, aykırı değerler, sağlam temel bileşenler regresyonu, sağlam kısmi en küçük kareler regresyonu, gelen turist sayısı*

1. Introduction

In existence of multicollinearity in the data set, Multiple Linear regression (MLR) analysis gives unreliable estimates for regression parameters and the variance of these parameters could be too large that leads to use biased methods: Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). Since they firstly reduce the dimensionality of the design matrix, they are the most popular regression techniques yielding better solutions. Straightforward Implementation of a Statistically Inspired Modification of the Partial Least Squares Method (SIMPLS) algorithm is the most popular PLSR algorithm as it is fast, efficient and the results of it are easily

[(1)]Hacettepe Üniversitesi, İstatistik Bölümü, espolat@hacettepe.edu.tr
Geliş/Received: 02-10-2017, Kabul/Accepted: 24-12-2018

interpreted. Since PCR is a combination of Principal Component Analysis (PCA) on the x-variables with Least Squares (LS) regression, in the case of outliers existence both steps of it are unreliable. Moreover, also the results of SIMPLS are affected by outliers in the data set as it is based on the empirical cross-covariance matrix between the y-variables and the x-variables and on linear LS regression. Hence, in Hubert and Verboven (2003) and Hubert and Vanden Branden (2003), two robust versions of these methods: RPCR and RSIMPLS have been suggested respectively. Another robust PLSR method 'Partial Robust M-Regression (PRM)' is conceptually different: instead of robust Partial Least Squares (PLS), Serneels et al. (2005) proposed a partial robust regression estimator.

World Travel & Tourism Council (WTTC) state clearly that both of travel and tourism are the top industries in the world on almost any economic measure, including gross output, value added, capital investment, employment and tax contributions (Aslan et al., 2008). Turkey and many developing countries utilize tourism as a key component in their economic development strategy. Turkey is a developing country which is both a candidate country for European Union membership and one of the attractive touristic places in the south of Europe. Since it contributes to Gross Domestic Product, tourism is one of the prominent industries in the Turkish economy. Since particularly from 1980's Turkey's active outer tourism started to show important development, tourism which contributes to the country's economy results in a very huge source of income. In 1982, forming of mass tourism investment is started. The bill on incentives for tourism introduced in 1982 (Tourism Intensive Law No. 2634) contributed to the development of the sector and the tourism actors included in tourism activities.

This law caused rapidly increment in tourism investments and increase the foreign number of tourists coming to Turkey and as a result the income of tourism increased within the share of Gross National Product. It seems that the number of foreign visitors has accelerated rapidly in last decade. In 2004, Turkey attracted 17.5 million foreign tourists, exceeding 41 million visitors in 2014.

There are many and various modelling and forecasting techniques for tourist arrivals. There isn't only one special model that exactly performs better than the other models in every situation. One of the forecasting method in tourism is predicting foreign tourist arrivals to particular countries. Different methods have been used in determining the determinants of demand for international tourism. It is clear that multiple regressions were used mostly in tourism demand researches. Approximately in 84% of tourism demand studies seemed to have used MLR (Zhang, et al., 2009).

The aim of this study is to model the tourist arrivals (number of foreign tourists) to Turkey by using three popular biased robust RPCR, RSIMPLS and PRM methods in existence of both multicollinearity and outlier in the data set. Therefore, the best model giving the best predictions of tourist arrivals is selected and the most important factors affecting the tourist arrivals to Turkey are determined for the examined period.

2. Robust Biased Estimation Methods: Rpcr, Rsimpls, Prm

PCR and PLSR methods assume that the p-dimensional independent x-variables and a set of q-dimensional dependent y-variables are associated by using a bilinear model. n is the number of observations and for i=1,…,n this bilinear model is shown as in (1) and (2). Here $\tilde{t}_i$ are scores with the dimension of $k<<p$, $P_{p,k}$ is the x-loadings matrix

and $A_{k,q}$ is the slope matrix for the regression model of y_i on $\tilde{t}_i$. f_i and g_i are error terms. This bilinear model could be written in terms of the original independent variables as in (3). PCR and PLSR construct the scores $\tilde{t}_i$ in a different way. PCR and PLSR differentiate mainly in the construction of the scores $\tilde{t}_i$. PCR method computes the scores by extracting the most related information in the x-variables by using a variance criterion (as a result of PCA on the independent variables). However, the PLSR scores are computed by maximizing a covariance criterion between the x- and y-variables (Hubert and Verboven, 2003; Hubert and Vanden Branden, 2003; Engelen et al., 2004).

$$x_i = \bar{x} + P_{p,k}\tilde{t}_i + g_i \quad (1)$$

$$y_i = \bar{y} + A'_{q,k}\tilde{t}_i + f_i \quad (2)$$

$$y_i = \beta_0 + B'_{q,p}x_i + e_i \quad (3)$$

Hubert and Verboven (2003) and Hubert and Vanden Branden (2003) have suggested two robust types of these methods: RPCR and RSIMPLS, respectively. Another robust PLSR method called PRM is proposed by Serneels et al. (2005). In PRM method, weights ranging between zero and one are computed iteratively in order to reduce the influence of outliers both in the y and x spaces. PRM is very efficient in terms of computational cost and statistical properties (Serneels et al., 2005; Liebmann et al., 2010; Polat and Turkan, 2016).

2.1. Robust Principal Component Regression: RPCR

Before starting the PCR analysis, the data is centered as $\tilde{x}_i = x_i - \bar{x}$ and $\tilde{y}_i = y_i - \bar{y}$. Afterwards, a PCA on the x-variables is performed in order to remove the effect of multicollinearity. The first *k* dominant eigenvectors of the covariance matrix $S_x = \frac{1}{n-1}\tilde{X}'_{p,n}\tilde{X}_{n,p}$ is contained in PCA loading matrix $\tilde{P}_{k,p} = (p_1, \cdots, p_k)'$ and the scores satisfy $\tilde{t}_i = \tilde{P}'_{k,p}\tilde{x}_i$. In the second step of PCR, the response variables $\tilde{y}_i$ are regressed onto $\tilde{t}_i$ as $\tilde{y}_i = A'\tilde{t}_i + \tilde{\varepsilon}_i$ using MLR. Then, the parameter estimates and fitted values are obtained as $\hat{A}_{k,q} = (T'T)^{-1}_{k,k}T'_{k,n}\tilde{Y}_{n,q}$ and $\hat{y}_i = \hat{A}'_{q,k}\tilde{t}_i + \bar{y}$, respectively. The unknown regression parameters in model (3) are then estimated as $\hat{B}_{p,q} = \tilde{P}_{p,k}\hat{A}_{k,q}$ and $\hat{\beta}_0 = \bar{y} - \hat{B}'_{q,p}\bar{x}$ (Hubert and Verboven, 2003).

Both steps of PCR is robustified and a robust PCR method is proposed by Hubert and Verboven (2003). In the first step, the highly robust Minimum Covariance Determinant (MCD) estimator is used as a robust estimator of the covariance matrix of the x_i in case of the data has a low-dimension ($p<n/2$), however, in case high-dimensional data the ROBPCA method chosen. ROBPCA, which combines projection pursuit ideas with MCD covariance estimation in lower dimensions, is a robust PCA method. In MCD estimator, the subsets of size *h* out of the whole data set

(of size *n*) is examined. Later the MCD estimator searches to find *h* subset for whom classical covariance matrix has minimal determinant. The robustness of the estimator is determined by the number 'h' that must be at least (n+p+1)/2. The MCD location estimate shown by $\bar{x}_h$ and the MCD scatter estimator shown by its covarince matrix $\hat{\Sigma}_h$. A tolerance ellipse, capturing the covariance structure of the majority of the data points, is yielded by a robust PCA method. The highly robust MCD estimator of location and scatter ($\hat{\mu}_{MCD}$ and $\hat{\Sigma}_{MCD}$) applied to the data and the points x of whose robust distance $D(x) = D(x, \hat{\mu}_{MCD}, \hat{\Sigma}_{MCD}) = \sqrt{(x-\hat{\mu}_{MCD})' \hat{\Sigma}_{MCD}^{-1} (x-\hat{\mu}_{MCD})}$ equals to $\sqrt{\chi^2_{2,0.975}}$ are plotted for the purpose of yielding a robust tolerance ellipse. In order to increase finite sample efficiency substantially, the raw MCD estimate can be reweighted. So that each data point belonging to the robust tolerance ellipse takes a weight of one and in other case a weight of zero. Therefore, the classical mean and covariance matrix of the data points having weight one gives reweighted MCD estimator. At last, robust loadings are obtained by the first k eigenvectors of the MCD estimator that ranked in descending order of the eigenvalues (Hubert and Verboven, 2003; Engelen et al., 2004).

In the second step of RPCR method, if there is only one y-variable the reweighted Least Trimmed Squares (LTS) regression is chosen for regressing y_i on t_i, otherwise the MCD regression is applied. Here, the regression model with intercept written as in (4) with $Cov(\breve{\varepsilon}) = \Sigma_{\breve{\varepsilon}}$. In case of one response variable (q=1), this model simplifies as in (5) with $\sigma_{\breve{\varepsilon}}$ scale of the errors. The parameters in (5) could be estimated by using the LTS estimator. The raw LTS estimator minimizes the sum of the h smallest squared residuals as shown in (6). Here, $r^2_{1:n} \le r^2_{2:n} \le \cdots \le r^2_{n:n}$ denote the ranked squared residuals. A starting estimate of the error dispersion is shown in (7). Here c_h is a consistency factor for normally distributed errors. Hence, the LS estimator performed on the observations whose absolute standardized residual is not too large corresponding to the reweighted LTS estimator. That means, if $|r_i(\hat{\alpha}, \hat{\alpha}_0)_{LTS} / \hat{\sigma}_0| > 2.5$ it is set $w_i = 0$ and otherwise, $w_i = 1$. Then, final estimates of $(\hat{\alpha}, \hat{\alpha}_0)$ are computed as the vector minimizing $\sum_{i=1}^{n} w_i (y_i - \alpha' t_i - \alpha_0)^2$.

$$y_i = \alpha_0 + A' t_i + \breve{\varepsilon}_i \tag{4}$$

$$y_i = \alpha_0 + \alpha' t_i + \breve{\varepsilon} \tag{5}$$

$$(\hat{\alpha}, \hat{\alpha}_0)_{LTS} = \arg\min_{\alpha, \alpha_0} \sum_{i=1}^{h} \left(r^2(\alpha, \alpha_0)\right)_{i:n} \tag{6}$$

$$\hat{\sigma}_0 = c_h \sqrt{\frac{1}{h} \sum_{i=1}^{h} \left(r^2(\hat{\alpha}, \hat{\alpha}_0)_{LTS}\right)_{i:n}} \tag{7}$$

In case of q>1, the MCD regression estimator is used. First of all, the reweighted MCD estimator is calculated on the (t_i, y_i) jointly, hence, (k+q)-dimensional location estimate $\hat{\mu} = (\hat{\mu}_t, \hat{\mu}_y)$ and a scatter estimate $\hat{\Sigma}_{k+q,k+q}$ are obtained as shown in (8). Secondly, similar to the MLR estimates, which are based on the emprical covariance matrix of the joint (t_i, y_i) variables, robust parameter estimates are estimated as shown in (9). A reweighting step is done for increasing the efficiency of this robust regression estimator's efficiency. To apply this reweighting scheme, each data point receives a zero weight if it is initial residual distance is unsually large as shown in (10), with (11) and (12). All other observations have a weight $w_i = 1$. Later, the reweighted MCD regression parameters related to the MLR estimates based on observations having weight one. Updating the reweighted estimates for A and $\hat{\alpha}_0$ in (9), (11) and (12), the final residual distances are obtained. A different notation for the final estimates and residual distances is not used. The fitted values are obtained as in (13) and regression parameters derived as in (14). Finally, $\hat{\Sigma}_\varepsilon = \hat{\Sigma}_{\bar{\varepsilon}}$ is set (Hubert and Verboven, 2003)

$$\hat{\Sigma}_{MCD} = \begin{pmatrix} \hat{\Sigma}_t & \hat{\Sigma}_{ty} \\ \hat{\Sigma}_{yt} & \hat{\Sigma}_y \end{pmatrix} \tag{8}$$

$$\hat{A}_{k,q} = \hat{\Sigma}_t^{-1}\hat{\Sigma}_{ty} \quad \hat{\alpha}_0 = \hat{\mu}_y - \hat{A}'\hat{\mu}_t \quad \hat{\Sigma}_{\bar{\varepsilon}} = \hat{\Sigma}_y - \hat{A}'\hat{\Sigma}_t\hat{A} \tag{9}$$

$$w_i = 0 \text{ if } RD_i > \sqrt{\chi^2_{q,0.975}} \tag{10}$$

$$r_i = y_i - \hat{A}'t_i - \hat{\alpha}_0 \tag{11}$$

$$RD_i = D\left(r_i, 0, \hat{\Sigma}_{\bar{\varepsilon}}\right) = \sqrt{r_i\hat{\Sigma}_{\bar{\varepsilon}}^{-1}r_i} \tag{12}$$

$$\begin{aligned} \hat{y}_i &= \hat{A}'_{q,k}t_i + \hat{\alpha}_0 \\ &= \hat{A}'_{q,k}P'_{k,p}\left(x_i - \hat{\mu}_x\right) + \hat{\alpha}_0 \end{aligned} \tag{13}$$

$$\hat{B}_{p,q} = P_{p,k}\hat{A}_{k,q} \qquad \hat{\beta}_0 = \hat{\alpha}_0 - \hat{B}_{p,q}\hat{\mu}_x \tag{14}$$

2.2. Robust Partial Least Squares Regression: RSIMPLS

SIMPLS algorithm assuming that the x and y variables are related through a bilinear model as given in (1) and (2). After mean centering the data as $\tilde{X} = \{(x_i - \bar{x})\}_{i=1}^n$ and $\tilde{Y} = \{(y_i - \bar{y})\}_{i=1}^n$, firstly, SIMPLS will obtain k latent variables (LVs) $\tilde{T}'_{n,k} = (\tilde{t}_1, \cdots, \tilde{t}_n)'$ and after the response variables will be regressed on these k LVs. K components (the columns of $\tilde{T}_{n,k}$), which have maximum covariance with a certain linear combination of the y-variables, are constructed as a linear combination of the x-variables. In order to obtain k components, firstly, it is needed to calculate weight vectors. The first normalized PLSR weight vectors r_1 and q_1 are obtained as the first

left and right singular eigenvectors of $S'_{yx} = S_{xy} = \tilde{X}'_{p,n}\tilde{Y}_{n,q}/(n-1)$. The first coordinate of the score $\tilde{t}_i$ is computed as $\tilde{t}_{i1} = \tilde{x}'_i r_1$ for each observation. If we need that $\sum_{i=1}^{n} t_{ia} t_{ib} = 0$ and a≠b (that means orthogonality of components), other PLSR weight vectors are computed by deflating the S_{xy} matrix. Firstly, computing the x-loading $p_j = S_x r_j / (r'_j S_x r_j)$ with S_x then this deflation is made. Later $\{p_1,\ldots,p_a\}$ is orthonormalised as $\{v_1,\ldots,v_a\}$ and the deflation of S_{xy} is made as $S^a_{xy} = S^{a-1}_{xy} - v_a (v'_a S^{a-1}_{xy})$ with $S^1_{xy} = S_{xy}$. Then, $\tilde{t}_i$'s are defined as $\tilde{t}_{ia} = \tilde{x}'_i r_a$ or similarly as matrix form $\tilde{T}_{n,k} = \tilde{X}_{n,p} R_{p,k}$ with $R_{p,k} = (r_1, \cdots, r_k)$. Lastly, regressing the response variables y_i on these k-dimensional scores $\tilde{t}_i$ by using MLR, the formal regression model is obtained as in (15). Here, $E(f_i) = 0$ and $Cov(f_i) = \Sigma_f$. MLR yields estimates as in (16), (17) and (18). By inserting $\tilde{t}_i = R'_{k,p}(x_i - \bar{x})$ in (2), the parameters' estimators of the original model are obtained as in (19) (Hubert and Vanden Branden, 2003; Engelen et al., 2004; Polat and Turkan, 2016).

$$y_i = \alpha_0 + A'_{q,k}\tilde{t}_i + f_i \tag{15}$$

$$\hat{A}_{k,q} = (S_t)^{-1} S_{ty} = \left(R'_{k,p} S_x R_{p,k}\right)^{-1} R'_{k,p} S_{xy} \tag{16}$$

$$\hat{\alpha}_0 = \bar{y} - \hat{A}'_{q,k}\bar{\tilde{t}} \tag{17}$$

$$S_f = S_y - \hat{A}'_{q,k} S_t \hat{A}_{k,q} = Y'_{q,n} Y_{n,q} - \hat{A}'_{q,k} T'_{k,n} T_{n,k} \hat{A}_{k,q} \tag{18}$$

$$\hat{B}_{p,q} = R_{p,k}\hat{A}_{k,q} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{B}'_{q,p}\bar{x} \tag{19}$$

A robust RSIMPLS method starts by applying ROBPCA on the *x*- and *y*-variables with the aim of replacing S_{xy} and S_x, which are used in computing $\tilde{t}_i$, by robust counterparts and then continues similar to the SIMPLS algorithm. Similar to RPCR instead of MLR a robust regression method (ROBPCA regression) is performed in the second stage (Hubert and Vanden Branden, 2003; Engelen et al., 2004). To obtain robust scores, firstly, ROBPCA is applied on $Z_{n,m} = (X_{n,p}, Y_{n,q})$. ROBPCA is robust covariance estimator for high-dimensional data sets (m>n). The outlyingness of every observation is calculated and later the empirical covariance matrix of the h observations with smallest outlyingness is considered by ROBPCA using projection pursuit ideas. The data are then projected onto the subspace K_0 spanned by the $k_0 << m$ dominant eigenvectors of this covariance matrix. Later the MCD method is applied to estimate the center and scatter of the data in this low dimensional subspace. Finally, these estimates are back transformed to the original space and a robust estimate of the center $\hat{\mu}_z$ of $Z_{n,m}$ and of its scatter $\hat{\Sigma}_z$ are computed. This scatter matrix can be decomposed as $\hat{\Sigma}_z = P^z L^z (P^z)'$ with robust Z-eigenvectors P^z_{m,k_0} and Z-eigenvalues $diag(L_{k_0,k_0})$. Diagonal matrix L^z containing the k_0 largest eigenvalues of $\hat{\Sigma}_z$ in decreasing order. Then Z-scores T^z can be computed by

$T^z = (Z - 1_n \hat{\mu}'_z)P^z$. After the application of ROBPCA on $Z_{n,m}$, this yields robust estimates $\hat{\mu}_z = (\hat{\mu}'_x, \hat{\mu}'_y)'$ and $\hat{\Sigma}_z$. $\hat{\Sigma}_z$ can be decomposed as in (20). The cross-covariance matrix Σ_{xy} is estimated by $\hat{\Sigma}_{xy}$ and the PLS weight vectors r_a are computed as in the SIMPLS algorithm, but now starting with $\hat{\Sigma}_{xy}$ instead of S_{xy}. The x-loadings are defined as $p_j = (r'_j \hat{\Sigma}_x r_j)^{-1} \hat{\Sigma}_x r_j$. Then the deflation of the scatter matrix $\hat{\Sigma}^a_{xy}$ is performed as in SIMPLS. In each step the robust scores are calculated as in (21), where the $\breve{x}_i$ are the robustly centered observations (Hubert and Vanden Branden, 2003).

$$\hat{\Sigma}_z = \begin{pmatrix} \hat{\Sigma}_x & \hat{\Sigma}_{xy} \\ \hat{\Sigma}_{yx} & \hat{\Sigma}_y \end{pmatrix} \tag{20}$$

$$t_{ia} = \breve{x}'_i r_a = (x_i - \hat{\mu}_x)' r_a \tag{21}$$

After the robust scores are derived, a robust linear regression is performed. The regression model, based on robust scores, is written as in (22). In order to estimate parameters in this model a robust regression method called ROBPCA regression is used (Hubert and Vanden Branden, 2003).

$$y_i = \alpha_0 + A'_{q,k} t_i + \breve{f}_i \tag{22}$$

$r_{i(k)}$ is the residual for the i_{th} observation based on the initial estimates which were computed with k components and $\hat{\Sigma}_f$ is the initial estimate of the covariance matrix of the errors. The robust distance of the residuals is given as in (23). The weights $c_{i(k)}$ are computed as in (24). Here I shows the indicator function. Observations with weight $c_{i(k)}$ equal to one are used to compute the final regression estimates (similar to MLR method). The robust residual distances $RD_{i(k)}$ are recalculated as in (23) and at the same time the weights $c_{i(k)}$ are updated. Finally, robust parameter estimators of the original model (3) are obtained as in (25).

$$RD_{i(k)} = \left(r'_{i(k)} \hat{\Sigma}^{-1}_{\breve{f}} r_{i(k)} \right)^{1/2} \tag{23}$$

$$c_{i(k)} = I\left(RD^2_{i(k)} \le \chi^2_{q,0.975} \right) \tag{24}$$

$$\hat{\beta}_0 = \hat{\alpha}_0 - \hat{B}'_{q,p} \hat{\mu}_x \quad \hat{B}_{p,q} = R_{p,k} \hat{A}_{k,q} \tag{25}$$

2.3. Partial Robust M-Regression (PRM)

The latent regression model is then given by (26). Here T is a score matrix of size $n \times k$, having as rows the vectors t_i, with $1 \le i \le n$ (Serneels et al., 2005):

$$y_i = t_i \gamma + \varepsilon_i \tag{26}$$

Here, the vector $\gamma_{k\times 1}$ can be estimated by regressing the response variable on the LVs (t_i) by means of a robust M-estimator. The new model dimension is lower than as $k < p$ and it is a regression on the score vectors (t_i) that must be determined. Generally, leverage points and vertical outliers could be effective while estimating the regression coefficients, PRM gives robust parameter estimations. In PRM, a weight w_i^x is used to reduce the effect of leverage points, while a weight w_i^r is used for reducing the effect of vertical outliers. w_i^r are calculated from the residuals $r_i = y_i - t_i \gamma$ and w_i^x are obtained from the scores t_i (not from independent variables). In order to protect estimates against both vertical outliers and leverage points, weights need to be taken as in (27) and the obtained estimator called as the "PRM estimator" (Serneels et al., 2005).

$$w_i = w_i^r w_i^x \tag{27}$$

In order to compute the score matrix T, the following scheme is used. Loading vectors a_h, for $h = 1, \cdots, k$ are computed in a sequential manner as in (28), under the constraint in (29). $\mathrm{Cov}_w(y, u)$, in (29), with u another vector of length n, shows a weighted covariance as in (30) (Serneels et al., 2005).

$$a_k = \arg\max_a \mathrm{Cov}_w(y, X_a) \tag{28}$$

$$\|a\| = 1 \text{ and } \mathrm{Cov}_w(Xa, Xa_j) = 0 \text{ for } 1 \le j < k \tag{29}$$

$$\mathrm{Cov}_w(y, u) = \frac{1}{n}\sum_{i=1}^{n} w_i y_i u_i \tag{30}$$

Since $A_{p\times k}$ is the matrix of loading vectors, the score matrix is obtained as $T = XA$. The final estimate for β can be obtained as $\hat{\beta} = A\hat{\gamma}$ after the computation of $\hat{\gamma}$ (Serneels et al., 2005).

The weights in the above definitions are unknown and they are not fixed. First approximation of the estimator $\hat{\gamma}$ is computed by using an appropriate initial value for the weights. Then, the weights are recomputed using the preliminary parameter estimates and a second approximation of $\hat{\gamma}$ is obtained by again applying weighted PLS. After that the weights w_i are recomputed and the iteration process continues. Hence, the Iterative Reweighted Partial Least Squares (IRPLS) algorithm can be used

to compute $\hat{\gamma}$. These continuous weights are iteratively executed for each observation, in order to minimise the negative influence of outliers in the regression model (Serneels et al., 2005).

2.3.1. PRM Algorithm

Since PRM can be calculated with a change in an algorithm proposed by Cummins and Andrews (1995) called as Iterative Reweighted Partial Least Squares (IRPLS) regression, the implementation of it is easy. PRM is entirely robust and also practical for high-dimensional data sets. It is significant to use robust initial values and relevant weights. The weights also have to depend on the scores for PRM, thus correcting for leverage points if presenting in the predictor space (Serneels et al., 2005; Liebmann et al., 2010).

The weights w_i^r have been computed as in (31) with $\hat{\sigma}$ an estimate of residual scale and the function in (32) (Serneels et al, 2005).

$$w_i^r = f\left(\frac{r_i}{\hat{\sigma}}, c\right) \tag{31}$$

$$f(z,c) = \frac{1}{\left(1+\left|\frac{z}{c}\right|\right)^2} \tag{32}$$

In (32) c is a tuning constant, used as c = 4. f is "Fair" weight function. Other weight functions could be used and Serneels et al. (2005) stated that it is not claimed any optimality properties for c=4. However, many numerical experiments revealed that the fair function used with c = 4 is a good compromise between robustness and statistical efficiency. If the tuning constant c increases to infinity, then the weight function becomes more and more flat, as a result, the PRM-estimator look likes more and more PLS (Serneels et al., 2005).

By using standardized residuals, the weights in (33) are calculated. A simple and robust choice for $\hat{\sigma}$ s the Median Absolute Deviation: $\hat{\sigma} = \mathrm{MAD}(r_1,\cdots,r_n) = \underset{i}{\mathrm{median}}\left|r_i - \underset{j}{\mathrm{median}}\, r_j\right|$. The weights w_i^x measuring the leverage of each score vector t_i are computed as in (33) (Serneels et al, 2005).

$$w_i^x = f\left(\frac{\left\|t_i - \mathrm{med}_{L_1}(T)\right\|}{\mathrm{median}_i \left\|t_i - \mathrm{med}_{L_1}(T)\right\|}, c\right) \tag{33}$$

Here $\|.\|$ used for the Euclidean norm and $\mathrm{med}_{L_1}(T)$ shows the L1-median computed from the collection of score vectors $\{t_1,\cdots,t_n\}$; it is a robust estimator of

the center of k-dimensional score vectors. This L_1-median is a multivariate version of the sample median, also known as a spatial median and it could be computed very quickly. Coordinate-wise or component-wise median also could be used for estimating the multivariate median (Serneels et al, 2005).

The PRM steps could be given briefly as in the following (Serneels et al, 2005):

1. Robust starting values for the weights $w_i = w_i^r w_i^x$ are computed. The formula in (31) is used with $r_i = y_i - \text{median}_j y_j$ for the residual weights and formula in (33) is used with the score vectors replaced by x_i, *for* $1 \le i \le n$ for the leverage weights.
2. PLSR analysis is performed by using SIMPLS algorithm on the (re)weighted data matrices $\breve{X}$ and $\breve{y}$ computed by multiplying each row of X and y with $\sqrt{w_i}$. This PLS analysis results then in an update of $\hat{\gamma}$ and of the score matrix T. By dividing each row of T by $\sqrt{w_i}$, score matrix T is updated.
3. The residuals $r_i = y_i - t_i\hat{\gamma}$ are recomputed and the weights $w_i = w_i^r w_i^x$ are updated using (31) and (33).
4. Go back to step (2) until $\hat{\gamma}$ converges. Whenever the relative difference in norm between two consecutive approximations of $\hat{\gamma}$ s is smaller than a specified threshold, e.g. 10^{-2}, then convergence is achieved.
5. The final estimate $\hat{\beta}$ is directly obtained from the last weighted PLS step.

Many numerical computations revealed that this iterative procedure is stable and converges quite quickly. If software for computing standard PLS is available, then it is easy and quick to program the above algorithm (Serneels et al, 2005).

3. Application and Results

Tourism is one of the most quickly growing sectors in the world. Global tourism flows and tourism receipts show a stable increase in recent years. Hence, as an effective tool, significance of tourism on economic growth and development of a country increases. For most of the countries, tourism constitutes a prominent source of additional income, foreign exchange, employment and tax revenue. Turkey is one of the popular destinations in the world and today, tourism has become an important sector in the Turkish economy.

The tourism demand literature shows that there are several measurements for international tourism demand such as: the number of the tourist arrivals, the number of nights spent by tourist or the receipts from tourism. The number of tourist arrivals is still the most popular measurement in tourism demand studies. The main reason for this choice is the availability of tourist arrivals data. In this study, tourism demand is measured in terms of number of tourist arrivals to Turkey. Therefore, in order to develop the sector in a most planned and controlled manner it is important to determine the factors which have impact on Turkey's tourist arrivals. In this paper, it

is aimed to investigate some of these effective factors based on robust biased methods since the data set contains both multicollinearity and outliers.

The purpose of this study is to model the tourist arrivals (number of foreign tourists) to Turkey for the period of 1986-2013 by using three popular biased robust RPCR, RSIMPLS and PRM methods in existence of both multicollinearity and outliers in the data set. The model giving the best predictions of tourist arrivals is selected and the most effective variables on the tourist arrivals to Turkey are found. Considering the studies of Alpu et al. (2010), Samkar et al. (2011) and Ispir et al. (2015) six independent variables are determined and a trend variable is also added to analysis. The variables in the models are given in below:

Y: Number of Foreign Tourists,
T: Trend
X1: Number of Incoming Airplanes,
X2: Number of Rooms in Tourism Facilities,
X3: Number of Rest Areas,
X4: Number of Licensed Operation Yachts,
X5: Total Bed Amount of Tourism Facilities,
X6: Number of Tourism Agencies,

Firstly, classical MLR model is applied and found to be significant with a probability of 95% (F=396.95; p=0.000). According the MLR analysis, 99.3% of variation occurs in the variable of number of foreign tourists is explained by these six independent variables. Even though the MLR model fits the data well, multicollinearity may severely prohibit quality of the prediction. Table 1 shows that all independent variables with the exception of X1 and X3 are not significant as an indicator of multicollinearity problem. Firstly, it is investigated whether there is multicollinearity or not in the dataset. For this purpose, the condition number is calculated as $\lambda max/\lambda min=7.240/0.006=1206.6$. The condition number greater than 30 means that there is multicollinearity. The other multicollinearity measure is Variance Inflation Factor (VIF) that is one of the most common techniques in statistics for detecting multicollinearity. In practice, if any of the VIF values is equal or larger than 10, there is a near collinearity. In this case, the regression coefficients are not reliable. As the results of MLR the VIF values for T, X1, X2, X5 and X6 are found as 234.950, 15.450, 5314.895, 5129.155 and 68.604. Hence, there is a near-collinearity problem for this dataset.

Table 1. The estimated regression coefficients for the MLR model.

Model	Coefficients	Standart Error of Coefficients	T	P
Constant	96832	1471796	0.66	0.518
T	-237803	351877	-0.68	0.507
X1	58.32	15.11	3.86	0.001
X2	150.7	142.1	1.06	0.302
X3	-9193	1052	-8.74	0.000
X4	3217	1849	1.74	0.097
X5	-5.19	64.54	-0.08	0.937
X6	-342.9	748.8	-0.46	0.652

Secondly, whether outliers exist or not is examined using normal Q-Q plot of the MLR residuals given in Figure 1. As seen from Figure 1, there is an outlier in the data.

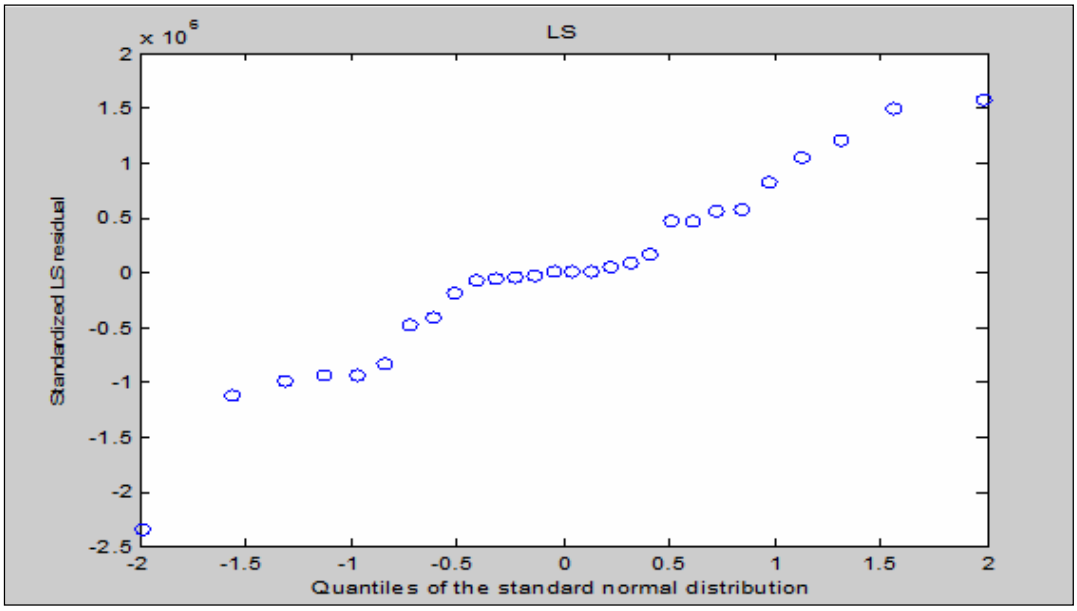
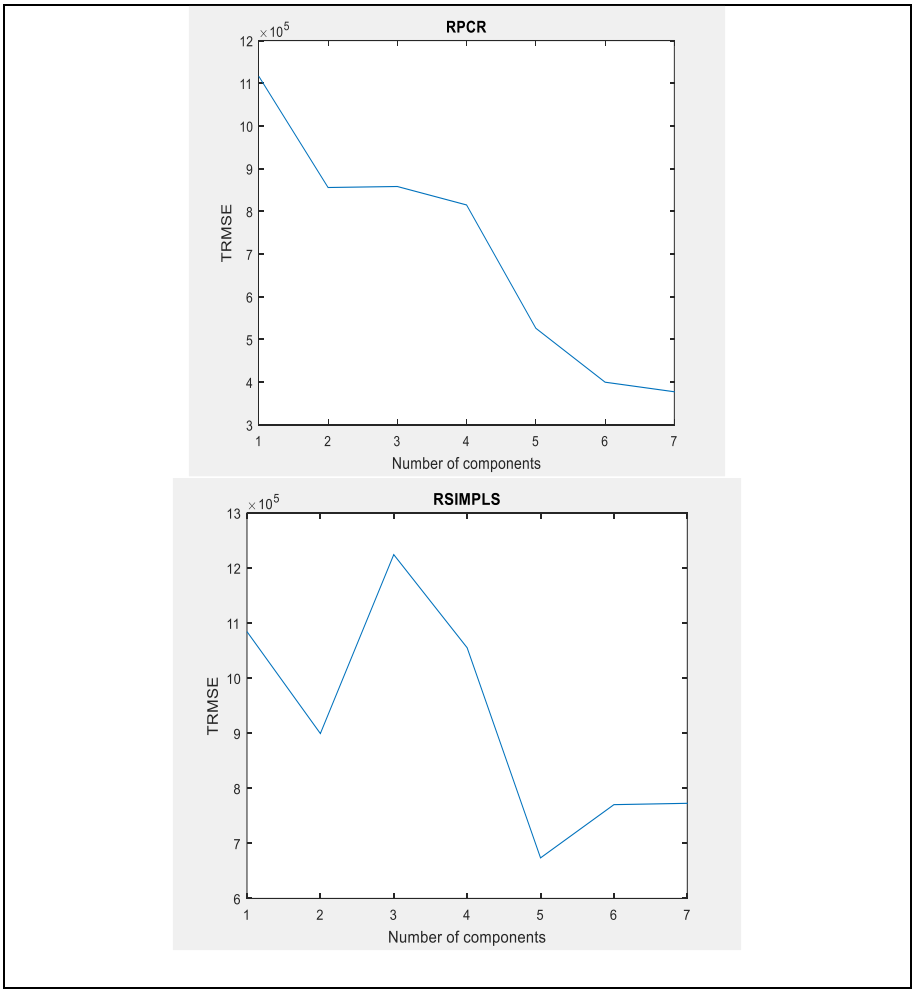


Figure 1. Normal Q-Q plot of MLR residuals

Table 1 shows that the significant variable X3 (Number of Rest Areas) has a negative effect on "Number of Foreign Tourists" variable which conflicts with both theoretical and logical expectations. Since the presence of both multicollinearity and outlier, the MLR results could not be reliable. In order to overcome both multicollinearity and outlier, biased robust RPCR and RSIMPLS, PRM methods (the robust counterparts of classical biased PCR and PLSR methods) are applied on the data set by using the functions given in MATLAB Toolboxes of 'LIBRA Toolbox' (Verboven and Hubert, 2005) and 'TOMCAT Toolbox' (Daszykowski et al., 2007).

The performance of the methods are evaluated by using the Root Mean Square Error (RMSE), $RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$ with upper % 20 trimming (TRMSE (0.8)), which is considered to be safer in the presence of outliers. Since we attend to assess the robust model's performance in fitting the data but not the outliers, a robust RMSE measure is necessary. The exclusion of a certain percentage of unusually large (absolute) residuals leads to an acceptable robust performance criterion. As mentioned in Daszykowski et al. (2007), the obtained values of RMSE are trimmed according to the assumed fraction of data contamination.

Firstly, the optimal number of components (showed by k_{opt}) could be selected for robust RPCR and robust PLSR methods (RSIMPLS and PRM) by taking the value for which TRMSE value is sufficiently small.



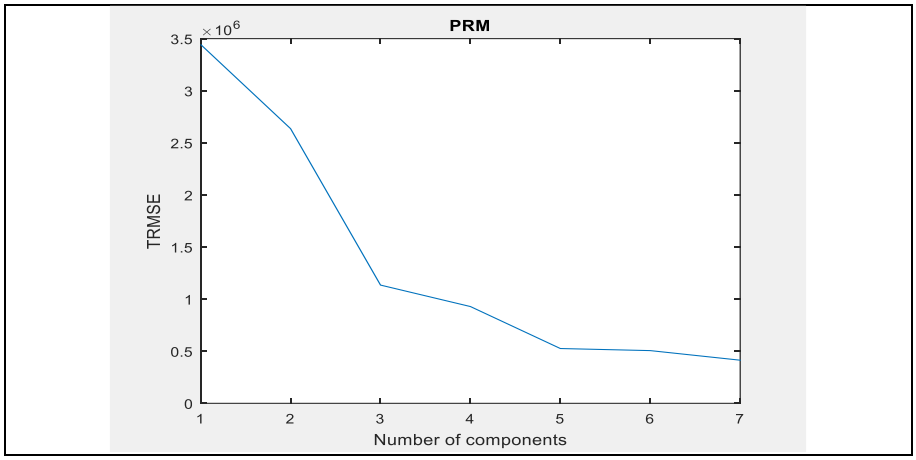


Figure 2. The plots of TRMSE (0.80) values of tourist arrival data set for RPCR, RSIMPLS, PRM

Since the model having sufficiently small TRMSE (0.80) value is always preferred, as seen from Figure 2 both of the RPCR and RSIMPLS models with two components (k_{opt}=2) and PRM model with three components (k_{opt}=3) is chosen.

Table 2. TRMSE values of three models for the tourist arrival dataset

	RPCR (k_{opt}=2)	**RSIMPLS (k_{opt}=2)**	**PRM (k_{opt}=3)**
TRMSE (0.8)	8.5582e+05	8.9937e+05	1.1348e+06

As seen from Table 2, RPCR is the model giving the best prediction of number of foreign tourists, hence, the estimated coefficients for RPCR given as shown in below.

The final model of RPCR is presented in terms of original variables:

Number of Foreign Tourists = -2.0969e+07 + 0.0037 **trend** + 40.1142 **airplanes** + 17.4722 **rooms** + 0.0407 **restareas** – 0.0175 **yachts** + 40.0919 **bedamount** + 1.0286 **agencies**

For the best model selected (RPCR) it is possible to detect outliers by using regression diagnostic plot and score diagnostic plot as shown in Figure 3. The first plot allows us to distinguish three types of outliers; good leverage points, bad leverage points and vertical outliers. The second one detects three types of outliers; good PCA leverage points, bad PCA leverage points and orthogonal outliers. The orthogonal outliers do not influence the computation of the regression parameters, but they might influence the loadings.

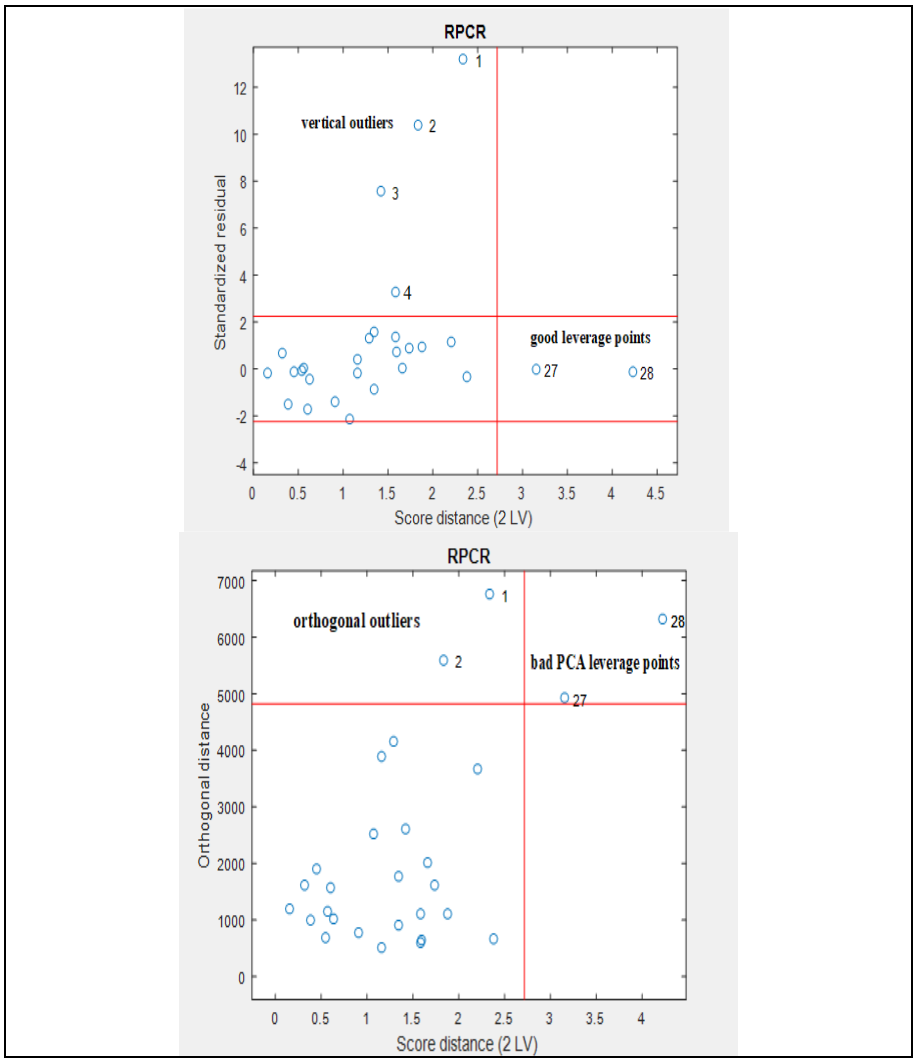


Figure 3. (a) Regression diagnostic plot (b) score diagnostic plot for RPCR (k_{opt}=2)

Figure 3 gives the order numbers of the observations, which are outliers and detected by RPCR (k_{opt}=2). It is seen that observations 1 and 2 are both vertical and orthogonal outliers, observations 3 and 4 are only vertical outliers, observations 27 and 28 are both good leverage and bad PCA leverage points.

4.Conclusion

The sector of tourism creates employment opportunities, decreasing unemployment and has an important role on providing the country with foreign currency income. Since it is a source of income and a supply of foreign currency input, it's eliminating instability between regions, farming, transportation, services and other tourisms

concerning direct and indirect commercial activities gaining motion, tourism is very important for a country's economy.

In this study, robust biased RPCR, RSIMPLS and PRM methods are applied to a real tourist arrival dataset of Turkey with both multicollinearity and outlier. They have been compared in order to determine which of them gives the best predictions of tourist arrivals. For the tourist arrival data set, RPCR model is chosen as the best model according to a robust RMSE performance criterion, TRMSE(0.8). The results obtained from RPCR robust biased estimation method showed that the most important independent variables affecting the number of foreign tourists are "Number of Incoming Airplanes" and "Total Bed Amount of Tourism Facilities". The least important variables affecting the number of foreign tourists are "Number of Licensed Operation Yachts" and "Number of Rest Areas". Hence, any increment in "Number of Incoming Airplanes" and "Total Bed Amount of Tourism Facilities" cause an important increment in number of foreign tourists. In this study, also it is observed that the addition or omission of the trend variable does not affect the results. Whether the trend variable present or not in the model, the parameters of independent variables remained same.
In conclusion, it could be declared that for the chosen best model RPCR, the directions of relationships between these six independent variables and the number of foreign tourists are consistent with the results obtained by Alpu et al. (2010), Samkar et al. (2011) and Ispir et al. (2015). Studies and meet the theoretical expectations. Moreover, in this study, different from other studies in literature about forecasting number of foreign tourists, three biased robust estimation methods RPCR, RSIMPLS and PRM are applied for the first time in the case of both multicollinearity and outlier existence.

5. References

Alpu, O., Samkar, H. and Altan, E. (2010). Saglam ridge regresyon analizi ve bir uygulama. *Dokuz Eylul Universitesi İktisadi ve İdari Bilimler Fakultesi Dergisi*, 25 (2), 137-148.

Aslan, A., Kaplan, M. and Kula, F. (2008). International tourism demand for Turkey: a dynamic panel data approach. Avaliable: https://mpra.ub.uni-muenchen.de/10601/1/MPRA_paper_10601.pdf.

Daszykowski, M., Serneels, S., Kaczmarek, K.., Van Espen, P., Croux, C. and Walczak, B. (2007). TOMCAT: A MATLAB toolbox for multivariate calibration techniques. *Chemometrics and Intelligent Laboratory Systems*, 85, 269–277.

Engelen, S., Hubert, M., Vanden Branden, K. and Verboven, S. (2004). Robust PCR and robust PLSR: a comparative study. M. Hubert, G. Pison, A. Struyf and S. V. Aelst (Ed.). In *Theory and Applications of Recent Robust Methods* (pp. 105–117). Birkhäuser; Basel.

Hubert, M. and Verboven, S. (2003). A robust PCR method for high-dimensional regressors. *Journal of Chemometrics*, 17, 438–452.

Hubert, M. and Vanden Branden, K. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics*, 17, 537-549.

Ispir, D., Ergul, B. and Yavuz Altın, A. (2015). Examining the ridge regression analysis of the number of foreign tourists coming to Turkey, *in Proceedings of the 2nd International Congress of Tourism & Management Researches* (pp. 242).

Liebmann, B. Filzmoser, P. and Varmuza, K. (2010). Robust and classical PLS regression compared. *Journal of Chemometrics*, 24 (3-4), 111-120.

Polat, E. and Turkan, S. (2016). The comparison of classical and robust biased regression methods for determining unemployment rate in Turkey: period of 1985-2012. *Journal of Data Science*, 14 (4), 739-768.

Samkar, H., Alpu, O. and Altan, E. (2011). Ridge regresyonda M tahmin edicilerinin kullanımı üzerine bir uygulama. *Dokuz Eylul Universitesi İktisadi ve İdari Bilimler Fakultesi Dergisi*, 26 (1), 67-77.

Serneels, S., Croux, C., Filzmoser, P. and Van Espen, P. J. (2005). Partial robust M-regression. *Chemometrics and Intelligent Laboratory Systems*, 79, pp. 55-64.

Verboven, S. and Hubert, M. (2005). LIBRA: a MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory System*, 75, 127–136.

Zhang, Y., Qu, H. and Tavitiyaman, P. (2009). The determinants of the travel demand on international tourist arrivals to Thailand. *Asia Pacific Journal of Tourism Research*, 14 (1), 77-92.