FBED
Çanakkale Onsekiz Mart Üniversitesi
FEN BİLİMLERİ ENSTİTÜSÜ DERGİSİ
eISSN:2459-1580

Çanakkale Onsekiz Mart University
Journal of Graduate School of Natural and Applied Sciences
Open Access

doi.org/10.28979/comufbed.714111 2020, Vol. 6, Issue 2, Pages: 353-363 dergipark.org.tr/en/pub/comufbed

Identification of Leverage Points in Principal Component Regression and $r-k$ Class Estimators with AR(1) Error Structure

Tuğba Söküt Açar [1,*]

[1]Department of Statistics, Faculty of Arts and Sciences, Çanakkale Onsekiz Mart University, Çanakkale, Turkey

Article History

Received: 03.04.2020
Accepted: 05.11.2020
Published: 15.12.2020

Research Article

Abstract — The determination of leverage observations have been frequently investigated through ordinary least squares and some biased estimators proposed under the multicollinearity problem in the linear regression models. Recently, the identification of leverage and influential observations have been also popular on the general linear regression models with correlated error structure. This paper proposes a new projection matrix and a new quasi-projection matrix to determination of leverage observations for principal component regression and r-k class estimators, respectively, in general linear regression model with first-order autoregressive error structure. Some useful properties of these matrices are presented. Leverage observations obtained by generalized least squares and ridge regression estimators available in the literature have been compared with proposed principal component regression and r-k class estimators over a simulation study and a numerical example. In the literature, the first leverage is considered separately due to the first-order autoregressive error structure. Therefore, the behaviours of first leverages obtained by principal component regression and r-k class estimators has been also investigated according to the autocorrelation coefficient and biasing parameter through applications. The results showed that the leverage of the first observation obtained by principal component regression and r-k estimators is smaller than that obtained by generalized least squares and ridge regression estimators. In addition, as the autocorrelation coefficient goes to -1, the leverage of the first transformed observation decreases for PCR and r-k class estimators, while its increases while the autocorrelation coefficient goes to 1.

Keywords — *Autocorrelation, first-order autoregressive error, leverages, multicollinearity, biased estimators*

1. Introduction

Regression analysis is used to model the effect of one or more regressor on the response. Multicollinearity, defined as the correlation between the regressors, is a major problem to overcome in regression analysis. Applied data collection method, variable requirement at the model, or over fitted model may lead to multicollinearity problem. It may not be the correct approach to omit one of the regressors that cause the correlations. For example, there is a high correlation between the number of individuals in the family and family expenditures when the socio-economic level is determined, and it is expected that the contribution of these variables to the model will be high. In this instance omitting one of the regressors may reduce the explanatory percentage of the model. The ordinary least squares (OLS) estimation procedure leads to unreliable and unstable estimates of regression coefficients under this problem because the variance of the estimators are inflated. Some bias estimators that overcome to multicollinearity problem have been proposed in the statistics literature and are still a popular topic that is being proposed (Hoerl & Kennard, 1970; Liu, 1993; Marquardt, 1970). Most popular biased estimations are based on ridge and principal component procedures.

Another problem in the data set is that some observations may be located at different points in the regressor space. The concept of leverage is being used in regression diagnostics as a measure of differently located observations in the regressor space (Steece, 1986). The leverage points are determined by the diagonal entries of projection matrix which depend on only the regressor matrix. Myers (1990, p. 253) noted that which data

[1] t.sokut@comu.edu.tr
[*]Corresponding Author

points are high leverage may change if the model formulation is changed. However some authors have reviewed leverage points for different estimators and have indicated that the position of the leverage point has shrinked according to the used estimator procedure. For instance, Steece (1986) showed that the leverage values of relevant observation reduce with increasing biasing parameter in ridge regression (RR) under the $E(\varepsilon\varepsilon') = \sigma^2 I_n$ assumption. It means that RR estimator with homoscedastic errors yield smaller leverage values than least squares estimator does (Walker & Birch, 1988). Steece (1986) also noted that RR estimator can copes with the outliers by down weighting of relevant influential observations. But it should be noted that even if the leverage's are reduced as value, the important thing is the success in determining the leverage observations in same spaces. Leverages should be carefully investigated to see if a reason for their unusual behaviours can be found. Once the leverage has been determined, it can either be deleted from the dataset or be corrected (if possible). But, the leverage points may contain useful information or may be systematically leverage point as described in the following paragraph. In this case observation deletion may cause useful information loss.

The assumption of uncorrelated errors must be valid for the application of the OLS procedure to the linear regression model. However, violation of this assumption which is called autocorrelation can be encountered in practice. Working with time series data, the presence of some regressors which is not included in the model but should be in the model or non-random measurement errors at the response may cause autocorrelation problem. For example, if the change in one person's income affects the other's savings, autocorrelation exists. In time series data, if the observations show inter-correlation, especially in cases where the time intervals are little, the concept of autocorrelation occurs. Leverages in general linear regression model (GLRM) with autocorrelation problem has been considered by several authors (Özkale & Açar, 2015; Puterman, 1988; Roy & Guria, 2004; Stemann & Trenkler, 1993). Especially there are more studies under the autocorrelation problem from the first-order autoregressive errors, AR(1). Because of the AR(1) structure, first observation can be leverage based on the autocorrelation coefficient. For instance, Stemann & Trenkler (1993) noted that the leverages of first observation obtained by generalized least squares (GLS) increase with increasing autocorrelation coefficient. Açar & Özkale (2016) report that the leverages of first observation obtained by RR increase to 1 while the autocorrelation coefficient goes to 1. In this case, deleting technique will not be effective solution on the first observation leverage's. The first observation in the new data set will show the same behaviour versus the autocorrelation coefficient. In such a case, could different estimators can be a solution? In this paper, leverage concept which has been examined in general linear regression model with multicollinearity problem over the RR estimator will be investigated for the other biased estimators such as principle component regression (PCR) and $r-k$ class estimators. The contribution of this paper to the statistics literature as follows: The projection and quasi-projection matrix for PCR and $r-k$ class estimators, respectively, with AR(1) structure were obtained. Some useful properties of the this matrices were investigated. Another contribution of this paper is the examination of the behaviour of the first leverages obtained by PCR and r-k class estimators versus to the different autocorrelation coefficients and biasing parameters.

The continuation of the study is structured as follows. The model and the estimators used throughout the paper are presented as Materials and Methods. Also the proposed projection matrices for PCR and $r-k$ class estimators are given in this section. The applications are given as results and discussion. Finally, the results of this paper are presented.

2. Materials and Methods

2.1. The Model and Estimators

The GLRM can be written as

$$y = X\beta + \varepsilon,\ E(\varepsilon) = 0,\ E\left(\varepsilon\varepsilon'\right) = \sigma^2\Psi \tag{2.1}$$

where y is the response vector consisting of n lines, $X = [1\ x_1\ x_2 \ldots x_k]$ is an $n \times p, p = k+1$ non stochastic regressor matrix with $x_{j=}(x_{1j}\ x_{2j} \ldots x_{nj})'$, β is the p-vector containing the regression parameters to be estimated. Random errors from the normal distribution denoted by ε have zero mean and $\sigma^2\Psi$ variance.
The matrix of covariance for the errors can be given as

- $E(\varepsilon\varepsilon') = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I_n$

- $E(\varepsilon\varepsilon') = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_3^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix} = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_n^2)$

- $E(\varepsilon\varepsilon') = \begin{pmatrix} \sigma^2 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & \sigma^2 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & \sigma^2 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & \sigma^2 \end{pmatrix} = \Omega = \sigma^2\Psi.$

As it can be seen, there are 3 cases: In the first case, the errors have a constant variance and are not correlated to each other. The model which has the first case error covariance matrix is called as linear regression model. This state is optimal situation for applying the OLS procedure. The matrix with the second state shows a diagonal matrix whose elements on the diagonal are different, that is non constant variance. This state is known as heteroscedasticity. When the constant variance assumption is violated, the weighted least squares estimator is used. Final state shows the fundamental structure of this paper called as autocorrelation. That is, autocorrelation problem refers to the situation where the errors are correlated, $cov(\varepsilon_i, \varepsilon_j) \neq 0$ for $i \neq j$. There are several structures under this problem such as autoregressive (AR) processes and moving average (MA) processes, as well as a combination of both types, the so-called ARMA processes. A further assumption of this paper based on the errors are modelled by AR(1) process.

Durbin Watson (DW) statistics proposed by Durbin & Watson (1950) tests whether the errors are in the AR(1) structure in a model with a constant term ($dw = \frac{\sum_{t=2}^{n}(v_t - v_{t-1})^2}{\sum_{t=1}^{n} v_t^2}$, where v_t is the tth OLS residual).
The errors with AR(1) process are modelled as

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i, \ |\rho| < 1, \ u_i \sim WN(0, \sigma_u^2) \tag{2.2}$$

where ρ denotes the autocorrelation coefficient and WN denotes the white noise. If ρ is not known, the estimation procedure given by Judge et al. (1985) is used ($\hat{\rho} = \frac{\sum_{t=2}^{n} v_t v_{t-1}}{\sum_{t=1}^{n} v_t^2}$).
The variance-covariance matrix for AR(1) error structure is reported by Judge et al. (1985) as

$$E\left(\varepsilon\varepsilon'\right) = \sigma^2\Psi = \frac{\sigma_u^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{bmatrix}. \tag{2.3}$$

Ψ is a positive defined and symmetric matrix, so there is always a non-singular $n \times n$ matrix P with $\Psi = PP'$. The matrix P^{-1} provides $P^{-1}\Psi\left(P^{-1}\right)' = I_n$ which is as follows

$$P^{-1} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix}.$$

(2.3)

By multiplying the model 2.1 by P^{-1}, the transformed model can be written as

$$y^* = X^*\beta + \varepsilon^* \tag{2.4}$$

where $y^* = P^{-1}y = \left(\sqrt{1-\rho^2}y_1, y_2 - \rho y_1, ..., y_n - \rho y_{n-1}\right)'$,

$X^* = P^{-1}X = \begin{bmatrix} \sqrt{1-\rho^2} & \sqrt{1-\rho^2}x_{11} & \cdots & \sqrt{1-\rho^2}x_{1k} \\ 1-\rho & x_{21}-\rho x_{11} & \cdots & x_{2k}-\rho x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1-\rho & x_{n1}-\rho x_{(n-1)1} & \cdots & x_{nk}-\rho x_{(n-1)k} \end{bmatrix}$ and

$\varepsilon^* = P^{-1}\varepsilon = \left(\sqrt{1-\rho^2}\varepsilon_1, \varepsilon_2 - \rho\varepsilon_1, ..., \varepsilon_n - \rho\varepsilon_{n-1}\right)'$

are the transformed response vector, regressor matrix and the error vector, respectively. It can be easily seen that the first row of y^*, X^* and ε^* exhibit different structures due to the AR(1). From this point of view, the first observation makes important differences in the examination against autocorrelation coefficient. In the equation 2.4, the expected value of the error is $E(\varepsilon^*) = 0$ and the variance-covariance matrix is $E(\varepsilon^*) = \sigma_\varepsilon^2 I_n$.
Since the transformed model has homoscedastic and uncorrelated errors, the least squares method can be applied. As a result, the application of least squares to the transformed observations yield

$$\widehat{\beta}^{GLS} = \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}y^* \tag{2.5}$$

which is known as the GLS estimator (Aitken, 1935). Under the multicollinearity problem, $X^{*\prime}X^*$ matrix is ill-conditioned. Although the GLS estimator is the best linear unbiased estimator (BLUE) in the GLRM, it causes a large total variance when multicollinearity exists and in this case yields estimates should not be trusted Trenkler (1984).

In GLRM model, the detection of multicollinearity techniques are similar to linear regression model. The most famous technique is condition number (CN) which depends on a ratio of maximum to minimum eigenvalues of $X^{*\prime}X^*$. Montgomery et al. (2001) noted that there is no serious multicollinearity problem when CN is less then 100. Also, they noted that if $100 < CN < 1000$ then there is a moderate multicollinearity problem and if $CN > 1000$ then there is severe multicollinearity problem among the regressors. Trenkler (1984) has addressed Hoerl & Kennard (1970)'s RR estimator which is well-known technique for reducing the variance under multicollinearity and autocorrelation problems as follows

$$\widehat{\beta}^{RR} = \left(X^{*\prime}X^* + kI_p\right)^{-1} X^{*\prime}y^*, k > 0 \tag{2.6}$$

with $bias(\widehat{\beta}^{RR}) = -k\left(X^{*\prime}X^* + kI_p\right)^{-1}\beta$. Here k represents the biasing parameter. Under the model 2.1, k can be estimated by $k = \frac{p\hat{\sigma}^2}{\hat{\beta}^{GLS\prime}\hat{\beta}^{GLS}}$ where $\hat{\sigma}^2$ is the unbiased estimator of σ^2 which is the most widely used estimation method of k based on Hoerl et al. (1975)'s procedure. However, a major problem in RR is based on choosing of k since the $MSE(\widehat{\beta}^{RR})$ and $bias(\widehat{\beta}^{RR})$ are also depends on the k (Kibria, 2003). Recommended estimated procedures for the selection of k in the literature are quite a lot.

PCR is the another popular biased estimator used in the presence of multicollinearity. In PCR, the matrix X^* is expressed in terms of its principal components. The singular value decomposition of $X^{*\prime}X^*$ is expressed as $X^{*\prime}X^* = \Phi\Lambda\Phi'$ where $\Lambda = \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix}$ is a diagonal matrix with the entries are the eigenvalues of $X^{*\prime}X^*$, Λ_1 and Λ_2 are the $r \times r$ and $p-r \times p-r$ diagonal matrices respectively such that the main diagonal entries of Λ_1 are the $r \leq p$ largest eigenvalues of $X^{*\prime}X^*$, while the main diagonal entries of Λ_2 are the remaining $p-r$ eigenvalues. $\Phi = (\Phi_1\ \Phi_2)$ is a orthogonal matrix corresponding to the kth column is the kth eigenvector of $X^{*\prime}X^*$. $\Phi_1 = (\phi_1, ..., \phi_r)$ is an $p \times r$ matrix formed by eigenvectors which is corresponding to the largest eigenvalues than 1. $\Phi_2 = (\phi_{r+1}, ..., \phi_p)$ is the remaining eigenvectors, so that $\Phi_1'\Phi_1 = I_r$, $\Phi_2'\Phi_2 = I_{p-r}$, $\Phi_1'\Phi_2 =$

0, $X^{*\prime}X^{*}=\Phi_1\Lambda_1\Phi_1'+\Phi_2\Lambda_2\Phi_2'$. By following Marquardt (1970), Trenkler (1984) gives the PCR estimator for model (2.1) as

$$\hat{\beta}^{PCR}=\Phi_1\left(\Phi_1'X^{*\prime}X^{*}\Phi_1\right)^{-1}\Phi_1'X^{*\prime}y^{*}. \tag{2.7}$$

$\hat{\beta}^{PCR}$ is biased estimator with $bias(\hat{\beta}^{PCR})=[\Phi_1\Phi_1'-I_p]\beta$.
Şiray et al. (2014) proposed the $r-k$ class estimator as a combination of ridge and PCR estimators under model (2.1) as below

$$\hat{\beta}^{r-k}=\Phi_1\left(\Phi_1'X^{*\prime}X^{*}\Phi_1+kI_r\right)^{-1}\Phi_1'X^{*\prime}y^{*},k>0. \tag{2.8}$$

$\hat{\beta}^{r-k}$ is a biased estimator with $bias(\hat{\beta}^{r-k})=\left(k\Phi_1\left(\Lambda_1+kI_r\right)^{-1}\Phi_1'+\Phi_2\Phi_2'\right)\beta$.
Şiray et al. (2014) proposed the selection of k, which makes the $r-k$ class estimator better than GLS according to the mean of error squares as $k=\frac{\hat{\sigma}^2}{\max\left|\hat{\alpha}_i^2-\frac{\hat{\sigma}^2}{\lambda_i}\right|}$.

2.2. Proposed Projection Matrices for PCR and $r-k$ Class Estimators

Several authors noted that one of the popular diagnostics methods to measure the impact of a particular observation is based on the projection matrix (Puterman, 1988; Steece, 1986). Dodge & Hadi (2010) state that this matrix is called as projection or called more commonly as hat matrix. The diagonal entries of the projection matrix give the leverage of the relevant observation. The projection matrix maps y into the fitted y. Therefore, the leverage points in the regressor space will differ according to the estimation procedure used by fitted y. For example, Puterman (1988) stated that the diagonal entries of the projection matrix in OLS are dependent solely on independent variables or regressor matrices as well as autocorrelation coefficient is effective on the diagonal entries of the projection matrix when GLS is used. In addition, Steece (1986) noted that the biasing parameter k is effective on leverage values when the RR estimator is used. The diagonal entries of projection matrix for GLS estimator are given by Puterman (1988) as

$$h_{i,i}^{GLS}=x_i^{*}\left(X^{*'}X^{*}\right)^{-1}x_i^{*\prime} \tag{2.9}$$

where x_i^{*} is the ith row of X^{*}. For RR estimator the diagonal entries of projection matrix which is called as quasi-projection matrix is given by Açar & Özkale (2016) as

$$h_{i,i}^{RR}=x_i^{*}\left(X^{*'}X^{*}+kI_p^{*}\right)^{-1}x_i^{*\prime}. \tag{2.10}$$

To determine the leverages with the PCR and $r-k$ class estimators for the model (2.1) is aimed in this paper. Let us investigate the projection matrix for PCR and $r-k$ class estimators. The response vector fitted by PCR estimator is

$$\begin{aligned}\hat{y}^{PCR} &= X^{*}\hat{\beta}^{PCR}\\ &= H^{PCR}y^{*}\end{aligned} \tag{2.11}$$

where $H^{PCR}=X^{*}\Phi_1\left(\Phi_1'X^{*\prime}X^{*}\Phi_1\right)^{-1}\Phi_1'X^{*\prime}$ is the projection matrix for PCR estimator. $h_{i,i}^{PCR}$ denotes the ith diagonal entries of H^{PCR} and $h_{i,j}^{PCR}$ denotes the off-diagonal entries of H^{PCR}. The diagnostics entries of the projection matrix for the PCR estimator is then

$$h_{i,i}^{PCR}=x_i^{*}\Phi_1\Lambda_1^{-1}\Phi_1'x_i^{*\prime}. \tag{2.12}$$

The projection matrix H^{PCR} has some important properties like the projection matrix of OLS (Cook& Weisberg, 1982) and GLS (Puterman, 1988; Stemann & Trenkler, 1993) estimators. We introduced this properties as follow

(i) H^{PCR} and (I_n-H^{PCR}) are symmetric matrices

(ii) H^{PCR} is an idempotent matrix, (I_n-H^{PCR}) is also idempotent

(iii) $\sum_{i=1}^{n} h_{i,i}^{PCR} = trace(H^{PCR}) = rank(H^{PCR}) = r$

(iv) $trace(I_n - H^{PCR}) = n - r$

where r is the number of eigenvalues equal to or higher than 1.

When the $r-k$ class estimator is used, the fitted values can be written as

$$\begin{aligned} \hat{y}^{r-k} &= X^* \hat{\beta}^{r-k} \\ &= H^{r-k} y^* \end{aligned} \tag{2.13}$$

where $H^{r-k} = X^* \Phi_1 \left(\Phi_1' X^{*\prime} X^* \Phi_1 + kI_r\right)^{-1} \Phi_1' X^{*\prime}$. H^{r-k} is symmetric but is not idempotent matrix so it is called quasi-projection matrix.

$$h_{i,i}^{r-k} = x_i^* \Phi_1 \left(\Lambda_1 + kI_r\right)^{-1} \Phi_1' x_i^{*\prime} \tag{2.14}$$

is the diagonal entries of H^{r-k}.

3. Results and Discussion

All computations were performed using MATLAB R2013a.

3.1. Monte Carlo simulation

To determine the effect of ρ and k on the first leverages obtained by GLS, RR, PCR and $r-k$ class estimators a Monte Carlo simulation study was conducted. The following steps were followed;
Step 1: The sample size was taken as $n = 100$ and regressors number was fixed to $p = 4$.
Step 2: Following McDonald & Galarneau (1975), the correlated regressor variables are generated from

$$x_{ij} = \left(1 - \gamma^2\right)^{1/2} w_{ij} + \gamma w_{ip}, \quad j = 1, ..., p, \; i = 1, ..., n \tag{3.1}$$

where w_{ij} are independent standard normal pseudo-random numbers and γ^2 is the correlation between any two regressor variables. γ^2 were fixed to be as 0.99. The regressor matrix are centralized and standardized after x_{ij} was produced, so that the $X'X$ becomes the correlation form.
Step 3: The responses were generated from $y_i = \beta_0 + \sum_{j=1}^{r} \beta_j x_{ij} + \varepsilon_i$. ε_i''s are generated by Eq.(2.2) where ρ is taken as $|\rho| = 0.99, 0.90, 0.70, 0.50$ and $u_i \sim IN(0,1)$.
Step 4: Following Kibria (2003), β was determined as the normalized eigenvector corresponding to the largest eigenvalue of the $X^{*\prime}X^*$ matrix.
Step 5: P^{-1} matrix was created for each different value of ρ and the corresponding transformed response vector and the regressor matrix were constructed.
Step 6: Three different values of the biasing parameter were used as $k = 0.1, 0.5, 1$.
Step 7: The experiment was replicated 1000 times. For each of them, the first leverages were obtained then, averages and standard errors of leverage values were calculated for 1000 replications.
The average and standard error values of each state are given in Table 1.

Table 1
Leverage values of first transformed observation according to different ρ and k

		$h_{1,1}^{GLS}$		$h_{1,1}^{RR}$		$h_{1,1}^{PCR}$		$h_{1,1}^{r-k}$	
ρ	k	Mean	Std Er.	Mean	Std Er.	Mean	Std Er.	Mean	Std Er.
-0.99	0.1	0.0008	0.0005	0.0007	0.0003	0.0008	0.0006	0.0007	0.0000
	0.5	0.0008	0.0005	0.0006	0.0008	0.0008	0.0006	0.0006	0.0004
	1	0.0008	0.0005	0.0004	0.0007	0.0008	0.0006	0.0004	0.0002
-0.90	0.1	0.0081	0.0087	0.0058	0.0048	0.0006	0.0011	0.0006	0.0002
	0.5	0.0081	0.0087	0.0030	0.0021	0.0006	0.0011	0.0006	0.0009
	1	0.0081	0.0087	0.0020	0.0036	0.0006	0.0011	0.0006	0.0002
-0.70	0.1	0.0263	0.0625	0.0178	0.0222	0.0020	0.0047	0.0020	0.0031
	0.5	0.0263	0.0625	0.0087	0.0151	0.0020	0.0047	0.0019	0.0012
	1	0.0263	0.0625	0.0059	0.0038	0.0020	0.0047	0.0019	0.0041
-0.50	0.1	0.0457	0.0569	0.0295	0.0378	0.0037	0.0079	0.0037	0.0039
	0.5	0.0457	0.0569	0.0140	0.0205	0.0037	0.0079	0.0036	0.0046
	1	0.0457	0.0569	0.0095	0.0137	0.0037	0.0079	0.0036	0.0041
0.50	0.1	0.0691	0.0625	0.0539	0.0236	0.0297	0.0191	0.0296	0.0430
	0.5	0.0691	0.0625	0.0390	0.0521	0.0297	0.0191	0.0291	0.0774
	1	0.0691	0.0625	0.0342	0.0750	0.0297	0.0191	0.0286	0.0156
0.70	0.1	0.0764	0.1083	0.0682	0.0819	0.0543	0.1000	0.0537	0.1208
	0.5	0.0764	0.1083	0.0578	0.0548	0.0543	0.1000	0.0516	0.0771
	1	0.0764	0.1083	0.0528	0.1083	0.0573	0.1000	0.0491	0.0590
0.90	0.1	0.1673	0.3888	0.1529	0.0417	0.1612	0.1361	0.1468	0.0305
	0.5	0.1673	0.3888	0.1152	0.1333	0.1612	0.1361	0.1133	0.1722
	1	0.1673	0.3888	0.0885	0.1555	0.1612	0.1361	0.0873	0.0264
0.99	0.1	0.6682	0.5998	0.1539	0.1611	0.0007	0.0016	0.0006	0.0008
	0.5	0.6682	0.5998	0.0380	0.0062	0.0007	0.0016	0.0005	0.0009
	1	0.6682	0.5998	0.0197	0.0197	0.0007	0.0016	0.0004	0.0004

The comments obtained from Table 1 are as follows: The reason of the decrease in the leverage values of first transformed observation when PCR and $r-k$ estimators are used is the change in r. As also observed in the numerical example, if r does not change and the autocorrelation coefficient positively increases the leverage value of first transformed observation increases. The attitude towards autocorrelation coefficient of the first transformed observation obtained with GLS and RR was in parallel with the literature so that the first leverages increased when $\rho \to 1$. And also the leverage values of first transformed observation decreased with increasing k. This decrease was more evident on the RR estimator. In all cases, the first leverages obtained by PCR is smaller than GLS. Moreover, the first leverages obtained by $r-k$ class is smaller than RR.

3.2. An Example: Macroeconomics Data

To demonstrate the leverage points in PCR and $r-k$ class estimators, Macroeconomics Data given by Gujarati (2004, p. 794) was used. And the data was used to compare the results of PCR and $r-k$ class estimators with the outputs of GLS and RR estimators. The variables as follows: y is the quarterly US data on gross domestic product (GDP) growth, x_1 is personal disposable income (PDI), x_2 is personal consumption expenditure (PCE), x_3 is corporate profits after tax (Profits) and x_4 is net corporate dividend payments. All data are in billions of 1987 dollars and are for the quarterly periods of 1970–1991, for a total of 88 quarterly observations. The main model was constructed as follows:

$$y = 1\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + \varepsilon = X\beta + \varepsilon. \tag{3.2}$$

The regressor matrix X was centralized and standardized; that is, $\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} x_{ij}^2 = 1$, $j = 1,2,3,4$ as the $X'X$ is in correlation form. The relations between the regressors were obtained as follows,

$$X'X = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{pmatrix} 1 & & & \\ 0.9970 & 1 & & \\ 0.8469 & 0.8521 & 1 & \\ 0.9823 & 0.9842 & 0.7894 & 1 \end{pmatrix}.$$

Durbin-Watson test as dw =0.4784 which indicated that the positive autocorrelation existed in data. The autocorrelation coefficient was obtained as $\hat{\rho} = 0.7596$. Figure 1 shows the graphs of autocorrelation function (ACF) and partial autocorrelation function (PACF). It is observed that the first lag is statistically significant and all the others are not significant. This indicates a possible AR(1) model for this data set.

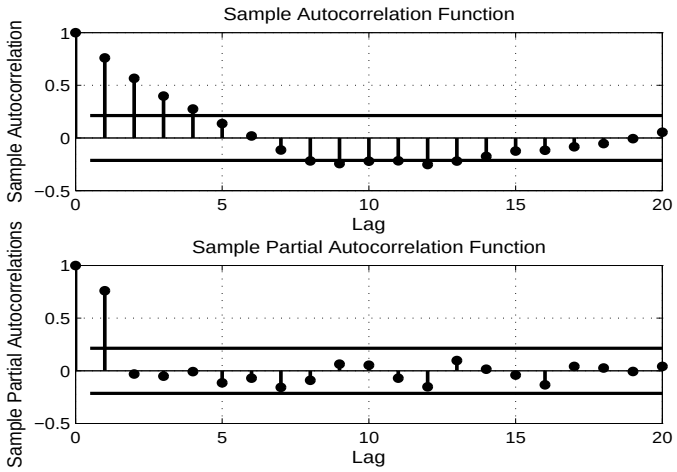

Figure 1. ACF and PACF for the Macroeconomics data

The strong relations between regressors drew attention to multicollinearity problem. The CN was obtained as $CN = 32592.5$ ($\lambda_1 = 88$, $\lambda_2 = 3.7516$, $\lambda_3 = 0.2514$, $\lambda_4 = 0.0142$, $\lambda_5 = 0.0027$) which indicated that there was a serious collinearity problem in the data set.

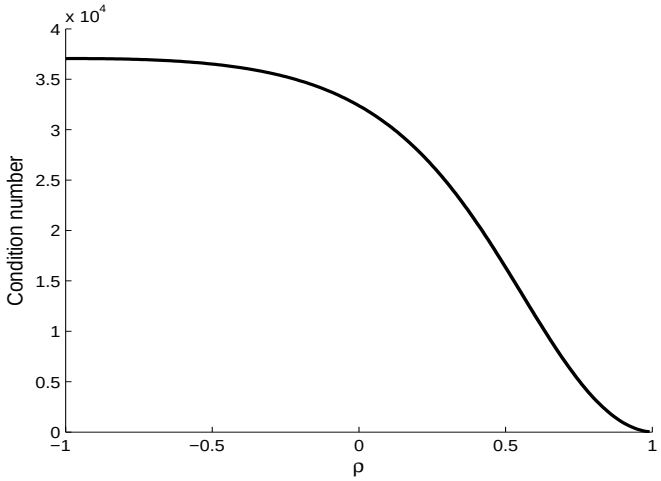

Figure 2. Condition number versus ρ for the Macroeconomics data

It can be seen from the Figure 2 that the CN decreased when the autocorrelation coefficient increased in positive direction. In the negative autocorrelation zone, the CN increased when $\rho \to -1$. It can be inferred from the graphs that the eigenvalues are significantly affected by the autocorrelation coefficient. Leverage values of all observation over the different estimators are given in Figure 3. Note that these results were only taken into the estimated $\hat{\rho} = 0.7596$ and $k = \frac{p\hat{\sigma}^2}{\hat{\beta}^{GLS'}\hat{\beta}^{GLS}} = 7.5806e05$ for RR estimator and $k = \frac{\hat{\sigma}^2}{\max\left|\hat{\alpha}_i^2 - \frac{\hat{\sigma}^2}{\lambda_i}\right|} = 6.1112e05$ for

$r-k$ class estimator.

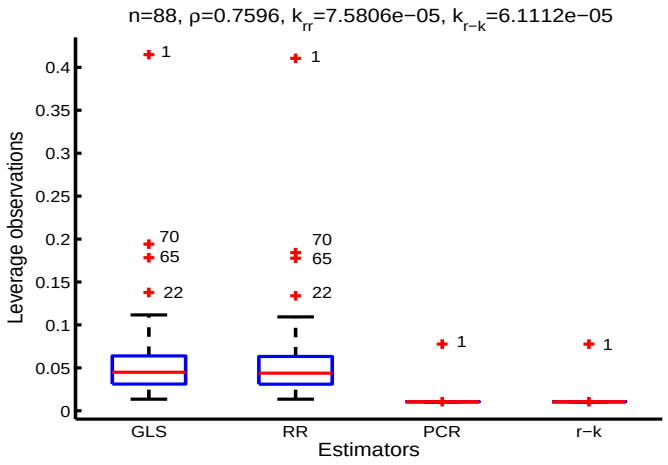

Figure 3. Leverage values at estimated ρ and k for the Macroeconomics data

In Figure 3, k_{r_r} represent the biasing parameter for RR estimator and k_{r-k} represent the biasing parameter for $r-k$ class estimator. As it can be seen in Figure 3, the first transformed observation was leverage point according to all the estimators for model 2.1. In addition, when the GLS and RR estimator were used 22, 65 and 70th observations were leverage points according to other observations at estimated ρ and k. Whereas, when PCR and $r-k$ estimators were used, there were no leverage points except 1 in the regressor spaces. That is, PCR and $r-k$ estimators shrank the leverages of 22, 65 and 70th observations on the regression spaces. The focus of this paper is the first transformed observation and to observe the behaviour of this observation against the autocorrelation coefficient and the biasing parameter. Many authors have studied the behaviour of first observation on the model with AR(1) error structure. For instance, Puterman (1988) showed that the first observation in a constant mean model ($y=\beta_0\mathbf{1}+\varepsilon$) and a regression through the origin model ($y=\beta_0 x_1+\varepsilon$) with AR(1) errors could have a large influence on regressor space. When the GLS is used, the first leverage goes to 1 at $\rho\to 1$ for a constant mean model while it goes to 0 at $\rho\to -1$ for a regression through the origin model. Stemann & Trenkler (1993) expanded Puterman (1988)'s investigation to the model which contained multiple regressors and they noted that if the model had constant term, then the first leverages close to 1 at $|\rho|\to 1$. Let's examine the behaviours of first leverages for different ρ and k.

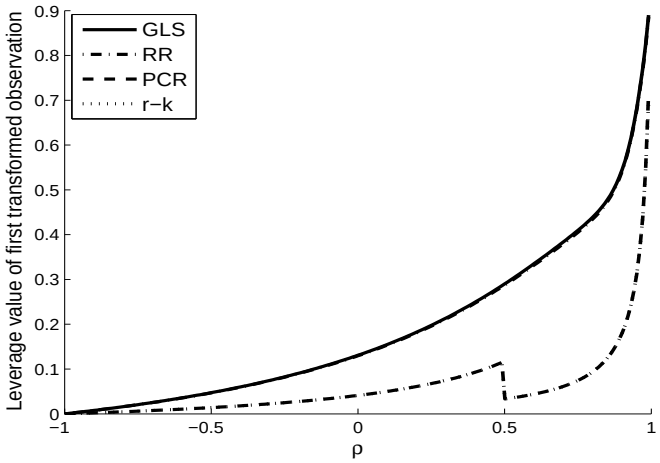

Figure 4. Leverage values of first transformed observation versus ρ according to different estimators for the Macroeconomics data

It can be realized in Figure 4 that the first transformed observation has a large leverage value as $|\rho| \to 1$. The first leverages decreases to 0 as $\rho \to -1$. That is, the first observation is not leverage on regressor space no matter which estimators are used while ρ goes to -1. GLS and RR estimators gave close leverage values in all cases and PCR and $r-k$ estimators too produced close values in all cases. But as we can see, GLS and RR estimator pairs offer higher leverage values than the pairs of PCR and $r-k$ estimators. The reason for the rapid decline of the first leverages obtained by PCR and $r-k$ estimators at $\rho \cong 0.4992$ is that $r=2$ at the $-1<\rho<0.4992$ while $r=1$ at the $0.4992<\rho<1$. Let's examine the first leverages against the biasing parameter, k. Because the GLS and PCR estimators do not contain k, these estimators were not included in the graphical representation.

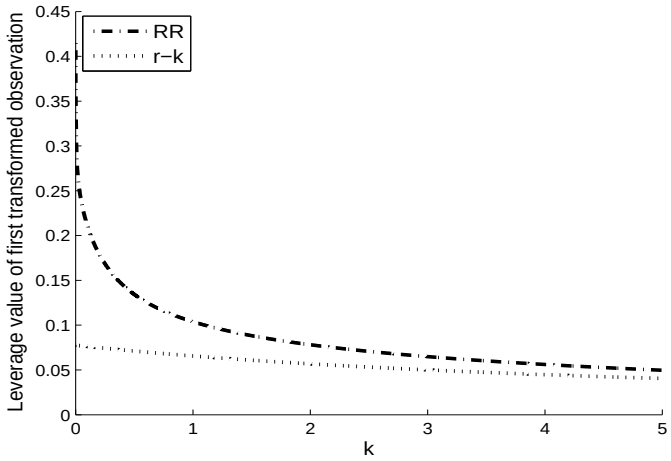

Figure 5. Leverage values of first transformed observation versus k according to RR and $r-k$ estimators for the Macroeconomics data

It is evident in Figure 5 that the first leverages decrease with increasing k.

4. Conclusion

In this paper, a new projection matrix and a new quasi-projection matrix obtained by PCR and $r-k$ class estimators in the linear regression model with first-order autoregressive errors have been proposed. It has been emphasized that the projection matrix obtained by PCR have some important features like the projection matrices obtained with GLS and OLS. That is, the projection matrix of PCR estimator is symmetric and idempotent. Since the first transformed observation has required special investigation due to the structure of AR (1), the first leverages obtained by PCR and $r-k$ class estimators have been compared with GLS and RR estimators over a Monte Carlo simulation and numerical example. The first leverages obtained by PCR and $r-k$ class estimators were smaller than those obtained by GLS and RR. This paper revealed that the first leverages obtained by PCR and $r-k$ class estimators increased by the autocorrelation coefficient in AR(1) structure. Also, the first leverages obtained by $r-k$ class estimators decrease with increasing k.

Acknowledgement

No financial support has been received from any institution or organization.

Author Contributions

Author Tuğba Söküt Açar completed and wrote all the stages of the paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

Açar, T.S., & Özkale M.R. (2016). Influence measures in ridge regression when the error terms follow an Ar(1) process. *Computational Statistics, 31*(3), 879-898. https://doi.org/10.1007/s00180-015-0615-5

Aitken, A.C. (1935). IV.— On least square and linear combinations of observations. *Proceedings of Royal Statistical Society, 55*, 42-48. https://doi.org/10.1017/S0370164600014346

Cook, R.D., & Weisberg, S.(1982). *Residuals and influence in regression*. Chapman and Hall, New York, pp. 11.

Dodge, Y., & Hadi, A.S. (2010). Simple graphs and bounds for the elements of the hat matrix. *Journal of Applied Statistics, 26*(7), 817-823. https://doi.org/10.1080/02664769922052

Durbin, J., & Watson, G.S. (1950). Testing for serial correlation in least squares regression I, *Biometrica, 37*(3/4), 409-428. https://doi.org/10.2307/2332391

Gujarati, D.N. (2004). *Basic Econometrics, 4th ed.*, McGraw-Hill, New Jersey.

Hoerl, A.E., & Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics, 12*(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

Hoerl, A.E., Kennard, R.W., & Baldwin, K.F. (1975). Ridge regression: some simulation. *Communications in Statistics, 4*(2), 105-123. https://doi.org/10.1080/03610927508827232

Judge, G.G., Grifths, W.E., Hill, R.C., Lütkepohl, H., & Lee, T.C. (1985). *The Theory and Practice of Econometrics, 2nd ed.* John Wiley & Sons Inc, New York.

Kibria, B.M.G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation, 32*(2), 419-435. https://doi.org/10.1081/SAC-120017499

Liu, K. (1993). A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods, 22*(2), 393-402. https://doi.org/10.1080/03610929308831027

Marquardt, D.W. (1970). Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics, 12*(3), 591–612. https://doi.org/10.2307/1267205

McDonald, G.C.,& Galarneau, D.I. (1975). A Monte Carlo evaluation of some ridge type estimators. *Journal of the American Statistical Association, 70*(350), 407–416. https://doi.org/10.2307/2285832

Montgomery, D.C., Peck, E.A., & Vining, G.G. (2001). *Introduction to Linear Regression Analysis*. John Wiley & Sons, New York.

Myers, R.H. (1990). *Classical and Modern Regression with Applications*. Duxbury Press, California, 1990.

Özkale, M.R., & Açar, T.S. (2015). Leverages and influential observations in regression model with autocorrelated errors. *Communications in Statistics -Theory and Methods, 44*(11), 2267-2290. https://doi.org/10.1080/03610926.2013.781646

Puterman, M.L. (1988). Leverage and influence in autocorrelated regression models. *Journal of the Royal Statistical Society, 37*(1), 76-86. https://doi.org/10.2307/2347495

Roy, S.S., & Guria, S. (2004). Regression diagnostics in an autocorrelated model. *Brazilian Journal of Probability and Statistics, 18*(2), 103-112.

Steece, B.M. (1986). Regresor space outliers in ridge regression. *Communications in Statistics-Theory and Methods, 15*(12), 3599-3605. https://doi.org/10.1080/03610928608829333

Stemann, D., & Trenkler, G. (1993). Leverage and cochrane-orcutt estimation in linear regression. *Communications in Statistics-Theory and Methods, 22*(5), 1315-1333. https://doi.org/10.1080/03610929308831088

Şıray, G.Ü., Kaçıranlar, S., & Sakallıoğlu, S. (2014). $r-k$ class estimator in linear regression model with correlated errors. *Statistical Papers, 55*(2), 393–407. https://doi.org/10.1007/s00362-012-0484-8

Trenkler, G. (1984). On the performance of biased estimators in the linear regression model with correlated or heteroscedastic errors. *Journal of Econometrics, 25*(1/2), 179-190. https://doi.org/10.1016/0304-4076(84)90045-9

Tripp, R.E. (1983). *Nonstochastic ridge regression and effective rank of the regressors matrix*, unpublished Ph.D. dissertation, Virginia Polytechnic Institute and State University, Dept. of Statistics.

Walker, E. & Birch, J.B. (1988). Influence measures in ridge regression. *Technometrics, 30*(2), 221-227. https://doi.org/10.2307/1270168