THE USE OF PRINCIPAL COMPONENTS IN QUANTILE REGRESSION: A SIMULATION STUDY

Cristina Davino [1], Tormod Næs[2], Rosaria Romano[1] and Domenico Vistocco[3]

[1] Department of Economics and Statistics, University of Naples Federico II, (e-mail: c.davino@unina.it, rosaroma@unina.it)

[2] Nofima AS, Norway, (e-mail: tormod.næs@nofima.no)

[3] Department of Political Sciences, University of Naples Federico II, (e-mail: vistocco@unina.it)

ABSTRACT: Least squares regression is highly unreliable when a strong collinearity structure is present among the predictors. Among several proposals introduced in the literature, principal component regression is a straightforward method to overcome the problem, even if it introduces a slight bias in the parameter estimation. This paper presents a simulation study to evaluate the use of principal component regression in the context of quantile regression and, focusing on the variability of the estimates and the model's prediction ability.

KEYWORDS: mutlicollinearity, principal component regression, quantile regression.

1 Introduction

In classical multiple linear regression applications, multicollinearity occurs very often, i.e. whenever two or more predictors are strongly correlated with each other. Such an issue can affect least-squares (LS) regression coefficients, their standard deviation, and consequently the associated t-tests, fitted values, and predictions.

Although multicollinearity has been extensively covered in the linear regression literature (Weisberg, 2005, Martens & Næs, 1992), little attention has been devoted to its effects in the context of quantile regression (QR) (Koenker & Hallock, 2001, Davino *et al.*, 2013, Furno & Vistocco, 2018). Possible solutions to the problem have been proposed from the ridge regression viewpoint (Bager, 2018), or focusing on variable selection techniques (Zaikarina *et al.*, 2016), for instance. However, an alternative approach addresses the problem of multicollinearity from a different perspective: the entire set of variables is preserved but replaced by some synthetic variables defined as *principal components*. This alternative approach is known as regression on latent variables

(James *et al.*, 2013), the variants of which differ in how these latent variables are obtained. Among these, the best-known method is the principal component regression (PCR)(Massy, 1965), from which the technique of quantile on principal component regression (QPCR) (Davino *et al.*, (2022)) originated.

The contribution of this article is to investigate the multicollinearity issue in the QR by evaluating its effects and deepening the study of the QPCR method.

2 Methods

In formal notation, the multiple linear regression model can be expressed as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is the $(n \times 1)$ vector of the dependent variable, $\mathbf{X}$ is a $(n \times K)$ fixed matrix representing the independent variables, $\boldsymbol{\beta}$ is a $(K \times 1)$ vector of unknown regression coefficients, and $\mathbf{e}$ is a $(n \times 1)$ vector of errors assumed to be normally distributed, with $\mathbf{E}(\mathbf{e}) = \mathbf{0}$, and $\mathbf{E}(\mathbf{e}\mathbf{e}') = \sigma^2\mathbf{I}_n$. In the following, without loss of generality, we assume that $\mathbf{X}$ and $\mathbf{y}$ are centered columnwise. The LS estimator is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{2}$$

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is equal to

$$cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{3}$$

and can be also formulated in terms of the singular value decomposition of the $\mathbf{X}'\mathbf{X}$ matrix as

$$cov(\hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{k=1}^{K} \mathbf{p}_k(1/\lambda_k)\mathbf{p}_k', \tag{4}$$

where $\mathbf{p}$ and λ are the eigenvectors and the eigenvalues of $\mathbf{X}'\mathbf{X}$, respectively (Næs & Mevik, 2001). Equation (4) highlights how, in presence of collinearity among the predictors, i.e. when some eigenvalues are very small, the variance of the regression coefficients increases.

The LS predictor $\hat{y}$ is unbiased, and the related Mean Squared Error (MSE), written using the eigenvector and eigenvalue decomposition of $\mathbf{X}'\mathbf{X}$, is

$$MSE(\hat{y}) = \sigma^2/N + \sigma^2 \sum_{k=1}^{K} t_k^2/\lambda_k + \sigma^2, \tag{5}$$

where $t_k = \boldsymbol{x}'\mathbf{p}_k$ is the score of $\boldsymbol{x}$ along eigenvector k. Equation (5) shows that the MSE depends not only on the magnitude of the eigenvalue but also on the t-score, i.e., on how much the new observations fall within the range of variability of the observed data along the different axes.

PCR finds some linear combinations of the original variables and use them as regressors to predict $\mathbf{y}$. Specifically, principal components analysis is applied to the matrix of predictors $\mathbf{X}$ to extract the A most dominating principal components. The PCR model structure is given by the following two equations

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}' + \mathbf{E}, \\ \mathbf{y} &= \mathbf{T}\mathbf{q} + \mathbf{f}, \end{aligned} \tag{6}$$

where $\mathbf{T}$ is called scores matrix and collects the A dimensions responsible for the systematic variation in $\mathbf{X}$, $\mathbf{P}$ and $\mathbf{q}$ are called loadings and describe how the variables in $\mathbf{T}$ are related to the original variables in $\mathbf{X}$ and $\mathbf{y}$, respectively. The PCR estimator is no longer unbiased since only the main dimensions are retained, while the less relevant ones are discarded. The MSE of the predictor $\hat{y}_{PCR}$ is

$$MSE(\hat{y}) = \sigma^2/N + \sigma^2 \sum_{k=1}^{A} t_k^2/\lambda_k + \left(- \sum_{k=A+1}^{K} (t_k/\sqrt{\lambda_k})\alpha_k \right)^2 + \sigma^2. \tag{7}$$

It has been empirically demonstrated (Næs & Mevik, 2001) that in situations of collinearity among the predictors, the PCR predictor performs better than the LS predictor in terms of MSE. Equation (7) suggests that a more considerable contribution of the variance along the eigenvectors with small eigenvalues ($a = A+1, \ldots, K$) for the LS predictor is replaced in the case of the PCR predictor by a more negligible bias contribution.

The extension of the PCR to the context of the QR is straightforward, as shown in Davino *et al.*, (2022). The model structure for the so-called QPCR is given by the following two equations:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}' + \mathbf{E} \\ Q_\theta &(\hat{\mathbf{y}}|\mathbf{T}) = \mathbf{T}\hat{\beta}(\theta) \end{aligned} \tag{8}$$

where $Q_\theta(.|.)$ is the conditional quantile function for the θ–th conditional quantile with $0 < \theta < 1$. It is worth noting that QPCR can produce the same numerical and graphical outputs as PCR, for each selected θ.

3 Simulation study

The simulation study aims to investigate the QPCR properties assessing:

- the variability of the regression coefficients in terms of MSE, given that the PCR estimator is biased;
- the prediction ability of the model both in the case of new cases within the range of the sampled data (i.e. to interpolate) and in the case of new data outside such a range (i.e. to extrapolate).

References

BAGER, AS. 2018. Ridge parameter in quantile regression models: An application in biostatistics. *International Journal of Statistics and Applications*, **8**(2), 72–78.

DAVINO, C, FURNO, M, & VISTOCCO, D. 2013. *Quantile regression: theory and applications*. Vol. 988. John Wiley & Sons.

DAVINO, C, ROMANO, R, & VISTOCCO, D. (2022). Handling multicollinearity in quantile regression through the use of principal component regression. *METRON*, **80**(2), 153–174.

FURNO, M, & VISTOCCO, D. 2018. *Quantile regression: estimation and simulation, Volume 2*. Vol. 216. John Wiley & Sons.

JAMES, G, WITTEN, D, HASTIE, T, & TIBSHIRANI, R. 2013. *An introduction to statistical learning*. Vol. 112. Springer.

KOENKER, R, & HALLOCK, KF. 2001. Quantile regression. *Journal of economic perspectives*, **15**(4), 143–156.

MARTENS, H, & NÆS, T. 1992. *Multivariate calibration*. John Wiley & Sons.

MASSY, WF. 1965. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, **60**(309), 234–256.

MCCULLAGH, P, & NELDER, JA. 1989. Binary data. *Pages 98–148 of: Generalized linear models*. Springer.

NÆS, T, & MEVIK, B-H. 2001. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, **15**(4), 413–426.

WEISBERG, S. 2005. *Applied linear regression*. Vol. 528. John Wiley & Sons.

ZAIKARINA, H, DJURAIDAH, A, & WIGENA, AH. 2016. Lasso and ridge quantile regression using cross validation to estimate extreme rainfall. *Global Journal of Pure and Applied Mathematics*, **12**(3), 3305–3314.