Book of Short Papers SIS 2020

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

Published by Pearson

www.pearson.com

ISBN 9788891910776

Exploring multicollinearity in quantile regression

Analisi della multicollinearità nella regressione quantile

Cristina Davino, Tormod Naes, Rosaria Romano and Domenico Vistocco

Abstract The aim of the paper is to propose a simulation study to explore the multicollinearity problem in quantile regression, as compared to the classical linear regression. The simulation exploits the concept of a relevant subspace and relevant predictors, considering different degrees of collinearity among the involved predictors. The approach is based on principal components and consists in evaluating the degree of dependence between the predictors on the basis of the eigenvalue structure of their covariance matrix. It is well known that in case of highly intercorrelated predictors the least squares coefficients, although determinate, posses large standard errors, causing precision problems for their estimation. For this reason, results of the simulation study focus on standard errors estimated according to different collinearity levels. A possible solution based on regression on principal components is briefly presented.

Abstract *L'articolo presenta uno studio di simulazione che mira ad esplorare il problema della multicollinearità nella regressione quantile, offrendo al contempo un parallelo con la regressione lineare classica. Lo studio sfrutta il concetto di sottospazio rilevante e predittori rilevanti, prendendo in considerazione diversi gradi di collinearità tra i predittori. L'approccio, basato sulle componenti principali, consiste nel valutare il grado di dipendenza tra i predittori rispetto alla struttura degli autovalori della loro matrice di covarianza. In caso di predittori altamente intercorrelati, è infatti noto che i coefficienti dei minimi quadrati, sebbene determinati, presentano errori standard elevati, causando problemi di precisione delle stime. I*

Cristina Davino
University of Naples Federico II, Italy, e-mail: cristina.davino@unina.it

Tormod Næs
NOFIMA AS, Norway, e-mail: tormod.naes@nofima.no

Rosaria Romano
University of Naples Federico II, Italy, e-mail: rosaroma@unina.it

Domenico Vistocco
University of Naples Federico II, Italy, e-mail: domenico.vistocco@unina.it

risultati presentati si concentrano pertanto sugli errori standard dei coefficienti, stimati in base a diversi livelli di collinearità. Una possibile soluzione al problema è mostrata sfruttando la regressione sulle componenti principali, in luogo dei regressori originari, al fine di eliminare il problema di collinearità.

Key words: multicollinearity, least squares regression, quantile regression, principal component regression

1 Introduction

Regression is widely used in the analysis of socio-economic phenomena to study the dependence of a response variable on a set of predictors. One typical problem in such fields concerns the structure of relations among the predictors, engendering the well-known problem of collinearity [**?**]. Multicollinearity has been extensively explored for least square regression (LS). This is not the case for quantile regression (QR), a well established statistical method aimed to explore the whole conditional distribution of a response variable without posing any parametric assumption for the error (and hence response) distribution ([**?**]; [**?**]). QR estimates separate models for different quantiles $\tau \in [0,1]$, namely QR provides a regression model for each conditional quantile of interest [**?**]. Even if an infinite number of conditional quantiles can be estimated (the quantile process [**?**]), in practice the researcher defines few quantiles of interest, in most cases, the three quartiles, $\tau = [0.25, 0.50, 0.75]$, along with two extreme quantiles, typically $\tau = [0.10, 0.90]$. For each considered quantile, a regression model is estimated, providing a set of coefficients and a fitted response vector. This paper presents a simulation study to explore the collinearity in quantile regression, as compared to the classical linear regression model. Results of a possible solution based on principal component regression are presented. The simulation scheme considers classical normal i.i.d. errors. Further developments will include normal heteroscedastic errors, costant skewness in the response (error) and increasing skewness in the response (error).

2 Simulation study

2.1 Experimental design

The concept of a relevant subspace has been exploited to simulate different degrees of correlation among predictors. It essentially consists of the subspace of the predictor space that is relevant for the variation in the response variable. Principal components analysis allows to take into account of different degrees of correlation. We carried out the analysis using the software R [**?**] and the *simrel* package [**?**]

for linear model data simulations. In particular we set the following values for the simulations:

- number of observations: 100
- number of predictors: 3
- number of relevant principal components: 1
- theoretical R^2 for generating data: 0.7
- coefficient controlling the degree of collinearity: γ
- number of iterations: 1000

According to the concept of relevant subspace, and given the small number of selected predictors, we assume that only one component is relevant for prediction. With respect to the γ coefficient, it regulates the speed of decline in eigenvalues (variances) of the principal components. In particular, the eigenvalues are assumed to decline according to an exponential model and the first eigenvalue is set equal to 1. We considered a grid of values for γ ranging from 0 to 5, using an increment of 0.5. In case of low values of γ we expect no or very low collinearity among predictors, while high collinearity should be present by incrementing γ. As an example, Table **??** reports the percentage of cumulated explained variance for the three components ($Comp_1$, $Comp_2$ and $Comp_3$ on the columns) using one random sample for each level of γ (rows).

Table 1 Percentage of cumulated explained variance on the three principal components (columns) for the considered scenarios (different values of γ on the rows)).

	$Comp_1$	$Comp_2$	$Comp_3$
$\gamma = 0.0$	36.33	70.20	100.00
$\gamma = 0.5$	45.44	79.60	100.00
$\gamma = 1.0$	55.19	91.56	100.00
$\gamma = 1.5$	65.17	95.88	100.00
$\gamma = 2.0$	66.17	98.10	100.00
$\gamma = 2.5$	72.16	99.19	100.00
$\gamma = 3.0$	85.03	99.77	100.00
$\gamma = 3.5$	91.79	99.91	100.00
$\gamma = 4.0$	95.43	99.97	100.00
$\gamma = 4.5$	97.47	99.99	100.00
$\gamma = 5.0$	97.87	100.00	100.00

For each value of the γ grid, the standard errors of the classical linear regression and QR models were computed. We decided to compute LS standard errors using the bootstrap procedure in order to have a fair comparison with QR, where bootstrap is typically used to this end.

2.2 Main results

Simulation results for LS and the three quartiles of QR are shown in Figure **??**. In particular, the different models (LS bootstrap, QR bootstrap for $\tau \in (0.25, 0.5, 0.75)$, where τ denotes the conditional quantile) are depicted on the columns, while rows refer to the regression coefficient ($X1$, $X2$, $X3$). The different boxplot in each panel represent the distribution of the standard error (vertical axis) for the different values of the γ coefficient (horizontal axis). includes several panels. Each panel corresponds to one of the estimated parameters by row ($X1$, $X2$, $X3$).

Results highlight how increasing the degree of collinearity, the distribution of the standard errors increases both in size and variability for each predictor. This trend can be found both in the LS and in the QR. In case of QR the inaccuracy of the estimates is even larger and no particular pattern stands out with respect to the considered quantile. This effect is higher for extreme quantiles (results not shown).

Principal component regression (PC–R) exploits principal component analysis to overcome multicollinearity [**?**]. It exploits the principal components of regressors as explanatory variables. Typically, only a subset of the principal components is used, and in particular the ones with higher variances. Therefore PC–R is considered also a shrinkage method [**?**]. Since in this first simulation study only three regressors have been considered, we present PC–R results on all the three principal components. Figure **??** presents the same structure of the previous Figure **??**, the two columns representing results for the two extreme conditional quantiles ($\tau = 0.1$ and $\tau = 0.9$) being the only exception. It is safe to say that rows in Figure **??** refers to the principal components (and not to the origigiinal regressors). It is evident how PC–R represents a possible solution also for QR in case of multicollinearity: the first two regression coefficients become really stable, the only variability remains on the third coefficients.

2.3 Concluding remarks

The simple simulation study shows that QR coefficients suffer the same problem of LS coefficients in case of strong collinearity among regressors. In fact, they become unstable, and this in particular for quantiles far from the median. PC–R is a possible solution to deal with such problem, providing more stable solutions.

This early study will be expanded considering different error (response) distribution, that is normal heteroscedastic response, constant skewness in the response and increasing skewness in the response. Moreover, the influence of prediction performances will be studied in the various scenarios, both for prediction inside the sample, i.e. inside the typical range of the regressor(s), and for prediction outside the sample. We expect that prediction performances will not be necessarily influenced from collinearity when the input is inside the range of the data used for model fitting. As soon as one moves outside, collinearity can have a huge effect on the performance. The small eigenvalue directions are the most susceptible to this because

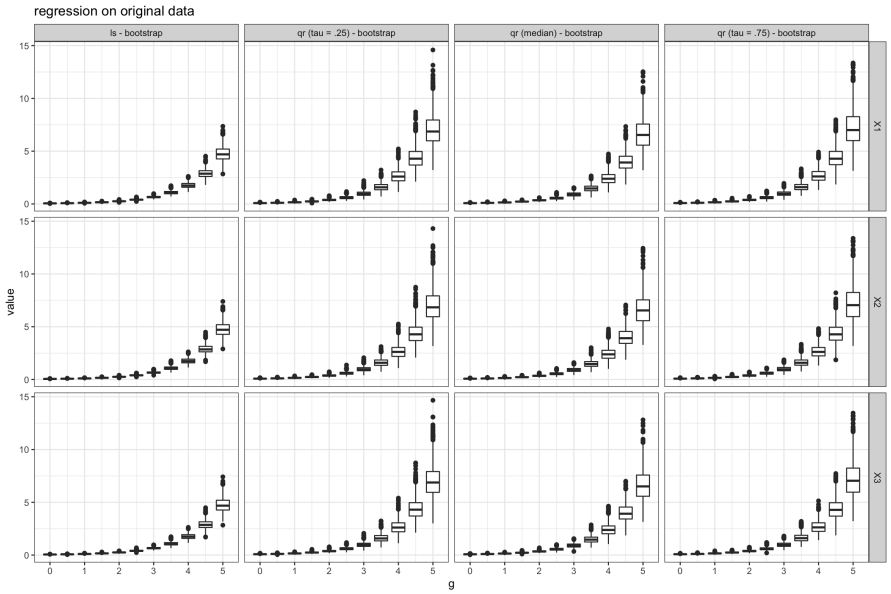

Fig. 1 Results from the simulation study: standard errors of the involved coefficients (different boxplots) for LS and QR models (columns) for three selected values of the γ coefficient (rows), regulating the degree of collinearity.

they are more unstable. Also in this case, we expect that PC–R may help to overcome the problem.

References

1. Davino, C., Furno, M., Vistocco, D.: *Quantile Regression. Theory and applications*. Wiley Series in Probability and Statistics, John Wiley & Sons (2013)
2. Furno, M., Vistocco, D.: *Quantile Regression. Estimation and Simulation*. Wiley Series in Probability and Statistics, John Wiley & Sons (2018)
3. Hastie, T., Tibshirani, R., Friedman, J: *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd edition. Springer Series in Statistics, Springer (2009)
4. Koenker, R., Bassett, G.: Regression quantiles. *Econometrica. Journal of the Econometric Society*. 33–50 (1978)
5. Koenker, R.: *Quantile Regression* (Econometric Society monographs; no. 38). Cambridge university press (2005)
6. Næs, T., Isaksson, T.: Data Compression by PLS/PCR. *NIR News*. **3**(1), 10–11 (1992)
7. Næs, T., Mevik, B.H.: Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*. **15**(4), 413–426 (2001)
8. R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
9. Sæbø, S., Almøy, T., Helland, I.S.: simrel—A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. Chemometrics and Intelligent Laboratory Systems. **146**, 128–135 (2015)

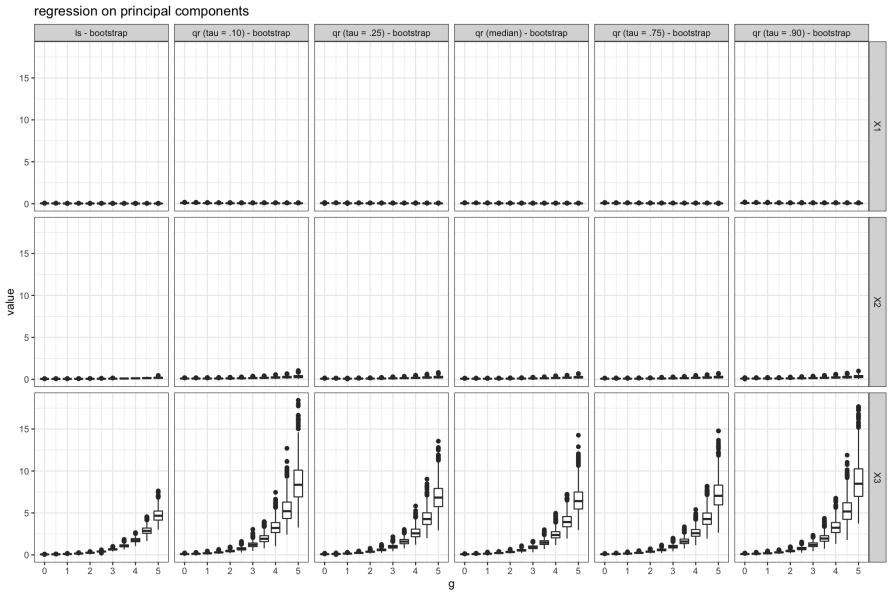

Fig. 2 Results from the simulation study: standard errors of the involved coefficients (different boxplots) for LS and QR models (columns) for three selected values of the γ coefficient (rows), regulating the degree of collinearity, when regressions are carried out on the principal components of the regressors.