



Handling multicollinearity in quantile regression through the use of principal component regression

C. Davino¹  · R. Romano¹ · D. Vistocco^{1,2}

Received: 27 October 2020 / Accepted: 2 January 2022
© The Author(s) 2022

Abstract

In many fields of applications, linear regression is the most widely used statistical method to analyze the effect of a set of explanatory variables on a response variable of interest. Classical least squares regression focuses on the conditional mean of the response, while quantile regression extends the view to conditional quantiles. Quantile regression is very convenient, whereas classical parametric assumptions do not hold and/or when relevant information lies in the tails and therefore the interest is in modeling the conditional distribution of the response at locations different from the mean. A situation common to most regression applications is the presence of strong correlations between predictors. This leads to the well-known problem of collinearity. While the effects of collinearity on least squares estimates are well investigated, this is not the case for quantile regression estimates. This paper aims to explore the collinearity problem in quantile regression. First, a simulation study analyses the problem concerning different degrees of collinearity and various response distributions. Then the paper proposes using regression on latent components as a possible solution to collinearity in quantile regression. Finally, a case study shows the assessment of the quality of service in the presence of highly correlated predictors.

Keywords Collinearity · Least squares regression · Quantile regression · Principal component regression

1 Introduction

Regression is by far the most widely used statistical methodology in empirical applications in several fields, ranging from economic and social sciences to life sciences. Even if

✉ C. Davino
cristina.davino@unina.it

R. Romano
rosaroma@unina.it

D. Vistocco
domenico.vistocco@unina.it

¹ Department of Economics and Statistics, University of Naples Federico II, Naples, Italy

² Department of Political Science, University of Naples Federico II, Naples, Italy

plenty of variants have been proposed, encompassing parametric, semi- and non parametric approaches, conditional linear models retain their appeal for the ease interpretation and the availability of tools and strategies suitable to deal with different types of data and/or validity of theoretical assumptions. Today more than ever, in the era of big data, it is common practice to collect numerous variables to use as predictors of a response variable of interest. For instance, many of the customer satisfaction surveys conducted by businesses and industries collect numerous variables to predict how consumers perceive and evaluate the quality of a product/service. If, on the one hand, the use of numerous explanatory variables offers a more accurate view on the response variable, on the other engenders redundant information deriving from correlation among predictors. The *collinearity* among predictors is one of the main problems associated with multiple linear regression (MLR) [5, 45]. In particular, it affects least squares estimates, standard errors, computational accuracy, fitted values and predictions [19, 34]. Several are the methods generally used to diagnose collinearity. Among these, the condition index and the variance inflation factor [15]. Several are also the proposals to address the problem. They range from methods based on calculating various types of components (partial least squares regression (PLS) and principal component regression (PCR)) to a technique based on penalizing the solution using the L2 norm [24, 37]. The most well-known here is ridge regression [35]. Other examples are the more general Tikhonov regression [43] and the elastic net [50], which also penalizes the L1 norm (in addition to the L2 norm) to reduce the variables. In principle, any of these methods could be used since there is no consensus about which one is best in general: they are good in different situations.

Quantile regression (QR) extends least squares regression (LSR) beyond conditional mean. It fits conditional quantiles of the response variable with a general linear model that assumes no parametric form for the conditional distribution of the response [11, 17]. QR is a valuable tool when the conditions of linear regression are not met (i.e., linearity, homoscedasticity, independence, or normality) [29], as well as the interest is in modeling the effect of regressors at different locations of the response. This interest is present in several applications. In the decades since the introduction of quantile regression, the potential of this method has been appreciated more and more and in the most diverse fields of application (see [32] and [47] for a review of the research areas). In addition to the numerous applications in the economic framework that characterised the first years of the methodology's dissemination ([6, 16, 22]), contributions in social and behavioral sciences (among many [9, 12]) as well as in medicine and survival analysis ([23]), consumer and customer satisfaction analysis [8, 10] are also widespread.

The contributions in the literature dealing with possible solutions to the problem of multicollinearity in QR focus, essentially, on variants of the ridge regression or the proposal of variable selection techniques. While the former [3, 49] shrink the coefficients, the latter introduces many different types of penalties to achieve a proper variable selection (among many, see [1, 28, 46]). The present paper faces the multicollinearity problem from a different perspective, where the whole set of variables is preserved but eliminating any redundancy that is unnecessary and detrimental to the estimation process. The use of a method of synthesising the original variables before proceeding to a quantile regression has been proposed by other authors but with different methods and aims. On the one hand, Fan et al. [13] combine Principal Component Analysis and QR with a different purpose as the authors propose the use of principal component factors as input variables for a stepwise cluster analysis prior to a QR. On the other hand, Fang et al. [14], Ando and Tsay (2011) [2] and Giglio et al. [18] use very different dimensionality reduction methods and in different contexts, respectively proper orthogonal decomposition for time series, quantile regression model with factor-augmented predictors for panel data and factor models for time series quantile regression.

The principal component regression (PCR) [39] inspires the present work to solve the problem of collinearity. It is used a lot for standard regression and is well understood theoretically. PCR is a transparent technique concerning the effects involved and easy to implement. An additional advantage of PCR is the graphical representation of the results. The representation of the regression coefficients on the principal components (loading plot) allows understanding which are the most critical variables in constructing the principal components and, therefore, in predicting the response variable. Furthermore, the representation of the scores (coordinates of the observations on the principal components) allows the detection of similarities and differences between the different statistical units and can be used for outlier detection [39]. PCR essentially consists of two steps, the first of which applies a PCA on the predictor matrix, while the second regresses the response variable on the first principal components (those that explain most of the variability). The present paper transfers the principles of PCR to the context of quantile regression giving life to a new method called quantile principal component regression (QPCR). The two approaches share the first step of the analysis, while they differ in the second as the QR is used in QPCR instead of the classic LSR to regress the response variable on the principal components. QPCR is simple to implement and can use the same graphical outputs as PCR. The primary purpose of the new method is to apply quantile regression even in contexts where the model's predictors are strongly correlated with each other. To this end, the QPCR identifies components that summarise the information in the predictors. Like all other regression methods on latent components, the principle is to exploit the correlation structure between the predictors to precisely identify a few latent components that best summarise the set of predictors and leave aside the remaining components that would explain residual information. Unlike the other regression methods mentioned above, QPCR does not penalise the regression coefficients by selecting the observed variables. Instead, it exploits the correlations between these to identify new variables, synthesising all the observed variables that contribute to the construction of the components to different degrees. The role of each variable in the construction of the components lies in the loadings, generally displayed through appropriate graphical representations (loadings plots). On the other hand, the main components themselves can be represented by scatterplots that allow you to view the values of the statistical units on the new components (scores plots). A comparative study of different approaches will be presented in the case study to understand the different philosophies of QPCR compared to other treatment methods of multicollinearity. The two approaches are not competitors but follow different strategies with different relative results, to solve a common problem. QPCR is a good choice when the correlation structure between the variables assumes the existence of latent factors, which can be interpreted in terms of all the starting variables and used as new predictors in the regression model taken into account. The case study presented, based on the assessment of the quality of the service, represents an example of a context in which the QPCR is suitable: the set of starting predictors, highly correlated with each other, can be suitably synthesized into new variables, representing latent dimensions of observed data, and that can be effectively used to predict online purchases.

The paper is organized as follows. Section 2 offers an essential presentation of LSR and QR, discussing the problem of multicollinearity. Section 3 presents PCR, one of the best known and easiest to implement solutions to the problem. Furthermore, it presents QPCR as the extension of PCR to the context of QR. An in-depth evaluation of the effect of collinearity among predictors in QR with some insights on the use of PCR as a possible solution is addressed through a simulation study in Sect. 4. Section 5 concerns a case study on the application of QR in evaluating service quality in terms of consumer satisfaction buying

online and in-store products. Finally, concluding remarks are offered in Sect. 6 along with a short discussion of the implications of these findings for future research.

2 Least squares regression and quantile regression: a short recap

Regression analysis is widely used for modelling the relationship between a single dependent variable (response) and one or more independent variables (regressors or predictors). In the following subsections, we focus on the linear model, briefly summarizing LSR and QR, limiting the treatment to the basic notation and aspects concerning collinearity.

2.1 Least squares regression

In formal notation, the MLR model can be expressed as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad (1)$$

where \mathbf{y} is a $(n \times 1)$ vector of observations on the dependent variable, \mathbf{X} is a $(n \times p)$ fixed matrix of observations on the independent variables, $\boldsymbol{\beta}$ is a $(p \times 1)$ 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.

Least squares (LS) method is commonly used to estimate the regression coefficient $\boldsymbol{\beta}$. It minimizes the least squares loss function $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ whose closed-form solution provides the LS estimates:

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

when $\mathbf{X}'\mathbf{X}$ is nonsingular. The covariance matrix of the LS estimator is:

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

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

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{a=1}^A \mathbf{p}_a(1/\lambda_a)\mathbf{p}_a' \quad (4)$$

where the \mathbf{p}_a s and λ_a s are the eigenvectors and the eigenvalues of $\mathbf{X}'\mathbf{X}$, respectively.

The problem of collinearity concerns the case where there are linear or near-linear relationships among the predictors. In the case of exact linear relationships, i.e. if one predictor is an exact linear combination of some others, $\mathbf{X}'\mathbf{X}$ becomes singular and no unique $\hat{\boldsymbol{\beta}}$ can be produced. If the predictors are nearly linearly dependent, $\mathbf{X}'\mathbf{X}$ is nearly singular, and the estimation equation for the regression parameters is ill-conditioned. Therefore, parameter estimates $\hat{\boldsymbol{\beta}}$ will be unstable, as evident from Eq. (3): the variances of the regression coefficients become very large, which implies that the confidence intervals tend to be larger. Equation 4 suggests the same conclusion, but in terms of the eigenvalues of $\mathbf{X}'\mathbf{X}$. If some eigenvalues are very small, as occurs in the presence of multicollinearity, the variances of the regression coefficients become very large.

Collinearity is relatively easy to detect by calculating the variance inflation factor (VIF) and the condition number (CN) [39]. The VIF is computable for each predictor \mathbf{x}_j ($j = 1, \dots, J$) as follows

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R_j^2 is the squared correlation coefficient obtained by predicting \mathbf{x}_j with the remaining explanatory variables. The VIF represents the factor of increase of the estimator's variance due to the correlation between \mathbf{x}_j and the other explanatory variables. Multicollinearity is present if one of the R_j^2 is close to 1. Generally, a VIF of 10 or above indicates that (multi) collinearity is a problem [20]. The condition number can be written as

$$CN = \left(\frac{\hat{\lambda}_1}{\hat{\lambda}_J} \right)^{1/2}$$

where $\hat{\lambda}_1$ and $\hat{\lambda}_J$ are respectively the largest and the smallest eigenvalue of the empirical covariance matrix of \mathbf{X} . An informal rule of thumb is that if the condition number is 15, multicollinearity is a concern, while multicollinearity is a severe concern if it is greater than 30 (these are just informal rules of thumb but have little theoretical basis).

2.2 Quantile regression

QR extends LSR replacing the classical estimate of the conditional mean (a single value) with estimates of conditional quantiles (several values). Therefore, it allows estimating the whole distribution of the conditional quantiles of the response variable. A typical QR model is formulated as:

$$Q_\theta(\hat{y}|\mathbf{X}) = \mathbf{X}\hat{\beta}(\theta), \tag{5}$$

where $Q_\theta(\cdot)$ is the conditional quantile function for the θ -th conditional quantile with $0 < \theta < 1$. QR provides separate models for each conditional quantile θ of interest. Although it is theoretically possible to estimate an infinite number of quantiles, a finite number is numerically distinct, the so-called quantile process. QR does not pose any parametric assumption for the error (and hence response) distribution. In line with classical linear models, each $\hat{\beta}_p(\theta)$ coefficient represents the rate of change in the θ -th conditional quantile of the dependent variable per unit change in the value of the p -th regressor ($p = 1, \dots, P$), holding the others constant.

QR is a modified version of the L_1 problem, the median regression, placing asymmetric weights on positive and negative residuals:

$$\hat{\beta}(\theta) = \underset{\beta(\theta)}{\operatorname{argmin}} \sum_{i=1}^n \rho_\theta \left(y_i - \mathbf{x}_i^\top \beta(\theta) \right) \tag{6}$$

where $\rho_\theta(\cdot)$ denotes the following asymmetric loss function:

$$\rho_\theta(y) = [\theta - I(y < 0)]y = [(1 - \theta)I(y < 0) + \theta I(y > 0)]|y| \tag{7}$$

Such loss function is a weighted sum of absolute deviations, where a $(1 - \theta)$ weight is assigned to the negative deviations and a θ weight is instead used for the positive deviations.

Therefore QR estimates of Eq. (6) can be formulated as:

$$\hat{\beta}(\theta) = \underset{\beta(\theta)}{\operatorname{argmin}} \sum_{y_i < \mathbf{x}_i^\top \beta(\theta)} (1 - \theta) |y_i - \mathbf{x}_i^\top \beta(\theta)| + \sum_{y_i \geq \mathbf{x}_i^\top \beta(\theta)} \theta |y_i - \mathbf{x}_i^\top \beta(\theta)| \quad (8)$$

QR allows the vector $\hat{\beta}(\theta)$ to vary on different θ , the median case ($\theta=0.5$) being equivalent to minimize the sum of absolute values of the residuals.

Asymptotic normality holds for the distribution of the estimators, where the form of the covariance matrix depends on the model assumptions [30, 31]. An alternative to asymptotic inference is provided by resampling methods [27] that do not require any assumption for the error distribution. Finally, the assessment of goodness of fit exploits the general idea leading to the typical R^2 goodness of fit index in classical regression analysis. The most common goodness of fit index in the QR framework is called pseudo- R^2 [33]. For an extensive discussion of QR methodological details, the reader is referred to the reference literature [11, 17, 29].

As the asymptotic distribution of the QR estimator depends on the inverse of the variance covariance matrix [29], the variance of the QR estimator increases with the degree of correlation among the predictors.

3 Principal component regression approach

3.1 Principal component regression

The basic idea of PCR is to find some linear combinations (components or factors) of the original variables and use them as regressors to predict \mathbf{y} . The identification of the main components takes place on the basis of dimensionality reduction techniques [24]. Specifically, principal components analysis (PCA) is applied to the matrix of predictors \mathbf{X} in order to extract the A most dominating principal components.

The model structure for PCR is given by the following two equations:

$$\begin{aligned} \mathbf{X} &= \mathbf{TP}' + \mathbf{E} \\ \mathbf{y} &= \mathbf{Tq} + \mathbf{f} \end{aligned} \quad (9)$$

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} relates to the original variables in \mathbf{X} and \mathbf{y} , respectively [39].

The estimated scores matrix $\hat{\mathbf{T}}$ of PCR is obtained by minimizing the loss function $\|\mathbf{X} - \mathbf{TP}\|^2$, whose solution is obtained through the singular value decomposition of \mathbf{X} . The estimated scores $\hat{\mathbf{T}}$ are then used in the regression equation in place of the original predictors:

$$\mathbf{y} = \hat{\mathbf{T}}\mathbf{q} + \mathbf{f}, \quad (10)$$

where LSR is used to estimate the regression coefficients in \mathbf{q} , and \mathbf{f} corresponds to the error term. Note that the PCR solution, i.e. the loadings $\hat{\mathbf{P}}$ and the regression coefficients $\hat{\mathbf{q}}$, can be combined to give the regression equation:

$$\hat{\mathbf{y}} = \bar{y} + \mathbf{X}\hat{\mathbf{P}}\hat{\mathbf{q}}, \quad (11)$$

which can be interpreted in the same way as a classical LSR and where the intercept is equal to the mean \bar{y} since the \mathbf{X} matrix is centred.

If all the components are included in the regression ($A = P$), the resulting model is equivalent to the LSR model. It is worth highlighting that the variability of the estimates is larger for the last components, corresponding to the smaller eigenvalues. By considering only the first components, the ones associated with the larger eigenvalues, estimates are more stable. Equation (4) shows indeed that the directions with the smaller eigenvalues have a large impact on the variances if they are unstable, as shown in [38]. This is quite natural since they are dominated by noise. The variances of PCR estimates are the same as for LSR except that now the influence of the eigenvalues after component A is eliminated. This shows exactly that PCR gives more stable regression coefficients than LSR. For this reason, the PCR aims to use in the regression a reduced number of components that maximize the variability of \mathbf{X} , leaving aside those that include noise. PCR belongs to the class of regression estimators that are biased but can simultaneously greatly reduce any large variances for regression coefficient estimators caused by multicollinearity [25].

3.2 Quantile principal component regression

As anticipated in Sect. 1, the extension of the principal component regression to the context of the QR is straightforward. The extraction of the main components from the predictor matrix occurs in the same way. In contrast, the regression of the response variable on the extracted components uses the QR instead of the LSR.

The model structure for PCR is given by the following two equations:

$$\begin{aligned} \mathbf{X} &= \mathbf{TP}' + \mathbf{E} \\ Q_\theta(\hat{y}|\mathbf{T}) &= \mathbf{T}\hat{\beta}(\theta) \end{aligned} \tag{12}$$

where $Q_\theta(\cdot|\cdot)$ 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, with the only difference being that the results will be specific for each selected theta.

4 On the effects of collinearity in QR: a simulation study

4.1 Description of the simulation plan

The simulation presented in this section exploits the concept of relevant subspace and relevant predictors. Relevant subspace essentially consists of the subspace of the predictors, space that is relevant for the variation in the response variable. The approach is based on principal components and exploits, in particular, the degree of dependence among the predictors on the basis of the eigenvalue structure of their covariance matrix. More specifically, the relevant components derive from a subset of the eigenvectors of the covariance matrix of the predictors. They allow obtaining a set of relevant predictors preserving all relevant information for the prediction of the response variable. The set of relevant predictor variables will have truly non-zero regression coefficients [21, 36].

The simulation study considers different degrees of correlation among predictors and different types of response with the aim to compare LSR and QR performance. We carried out the analysis using the software R [40] and the *simrel* package [41] for linear model data simulations. We considered a sample size of 100 observations and 3 relevant predictors. The last might seem a small number, but it is sufficient to illustrate the presence of multicollinearity

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

	<i>Comp</i> ₁	<i>Comp</i> ₂	<i>Comp</i> 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

in QR. Taking into account the small number of predictors, we set only one component relevant for prediction. Without loss of generality, the relevant component was set at the first position. Data were simulated so that the correspondent population model explain 70% of the variation in the response, i.e. setting the theoretical R^2 equal to 0.7. Finally, we regulated the level of collinearity among predictors using a coefficient that control for the speed of decline in eigenvalues (i.e. variances) of the principal components. We denote this coefficient with γ in the following. The first eigenvalue was set equal to 1, and the subsequent decline according to an exponential model. 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 1 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). For each value of the γ grid, the standard errors of the classical linear regression and QR models were computed. We opted to compute LS standard errors using the bootstrap procedure in order to have a fair comparison with QR, where bootstrap is typically used. We carried out 1000 simulations for each value in the design grid. The illustrated simulation schema was replicated for different types for the following types of responses:

- classical normal i.i.d. errors, and hence the response
- normal heteroscedastic errors
- skewness in the response (errors).

4.2 Main simulation results

In the presentation of the simulation results, we focus on discussing the effects of multicollinearity in QR. A comparison with what happens in LSR is still presented in case the relevant components method is adopted, namely when the PCs are used in place of the original regressors. Since we considered only three regressors for our illustrative purposes, we retain all the three components in PCR. The expected result is a different pattern in correspondence of the last component, where the largest variance associated with the noise should provide less stable estimates. In summary, the main objective is to explore the effect of multicollinearity

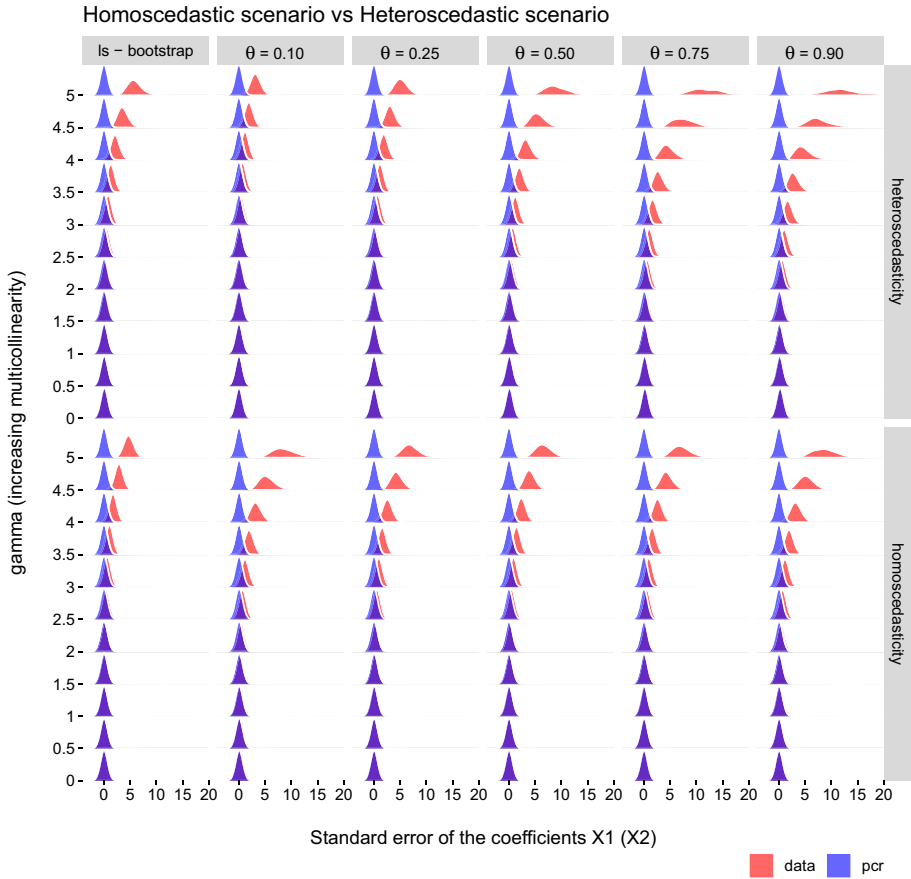


Fig. 1 Distributions of the standard errors of the first predictor (equivalent for the second predictor) in case of homoscedasticity and heteroscedasticity (rows) for LSR (first column) and QR (from the second to the last column, for the five considered quantiles). Colours refer to the results obtained using the original regressors (darker densities) or the principal components (lighter densities) in the model

in QR both when the original regressors are considered in the model and when PCs are used. Such results are compared with classical PCR for the different types of response considered.

The distribution of the standard errors of the coefficients for the estimated models are summarized in Figs. 1, 2 and 3. The three figures display all the settings of the simulation design, and in particular the data generation scheme (homogeneous vs heterogeneous errors in Figs. 1 and 2, skewed errors in Fig. 3), the type of regressors included in the model (original regressors versus PCs), and the estimated model (classical LS versus QR for several quantiles). The representations are organized in the following way: the horizontal panels refer to the type of errors while the columns display the different estimated models. At each intersection two different situations are compared: the standard errors estimated when the original regressors are considered in the model (darker histograms) and the ones when PCs are used (lighter histograms). The different values of γ (from 0, no collinearity, to 5, highest collinearity) are depicted on the vertical axis, the values of the standard errors being represented on the horizontal axis. The results corresponding to the first two variables (X_1

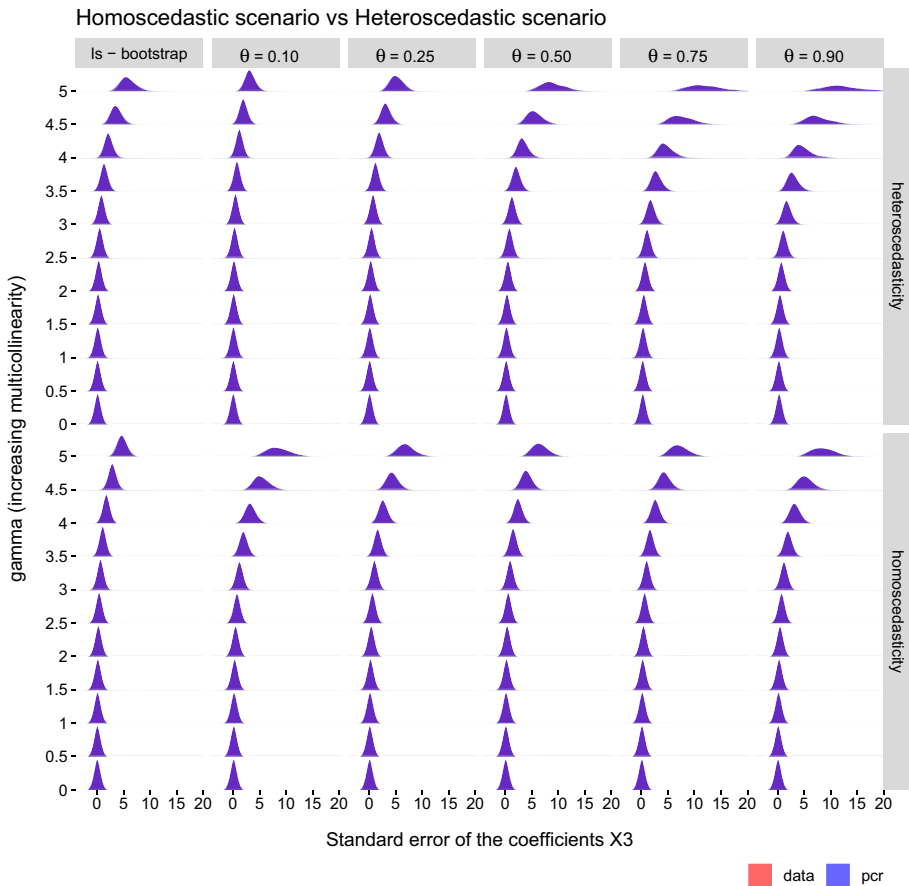


Fig. 2 Distributions of the standard errors of the third predictor in case of homoscedasticity and heteroscedasticity (rows) for LSR (first column) and QR (from the second column to the last column, for the five considered quantiles). Colours refer to the results obtained using the original regressors (darker densities) or the principal components (lighter densities) in the model

and X_2), as well as the ones associated to the first two components are identical and therefore Fig. 1 expresses both the two coefficients (darker densities) and the two components (lighter densities). Figure 2 depicts instead the standard errors for the regressor X_3 and the third PC.

From the analysis of the plots the following findings seem relevant:

- when original regressors are used in the model, standard errors increase in value and variability as the collinearity increases. This is more marked in QR than in LSR. The effect is more pronounced in the extreme parts of the distribution ($\theta = 0.1$ and $\theta = 0.9$)
- the previous consideration is amplified in the heteroscedastic case
- when PCs are used in place of the original regressors, multicollinearity does not affect the variability of the estimates both in the homoscedastic and heteroscedastic case. This is true for all the values of γ regulating the level of collinearity among predictors
- the distributions of the standard errors for X_3 and the third PC (Fig. 2) coincide and the densities related to the two cases perfectly overlap

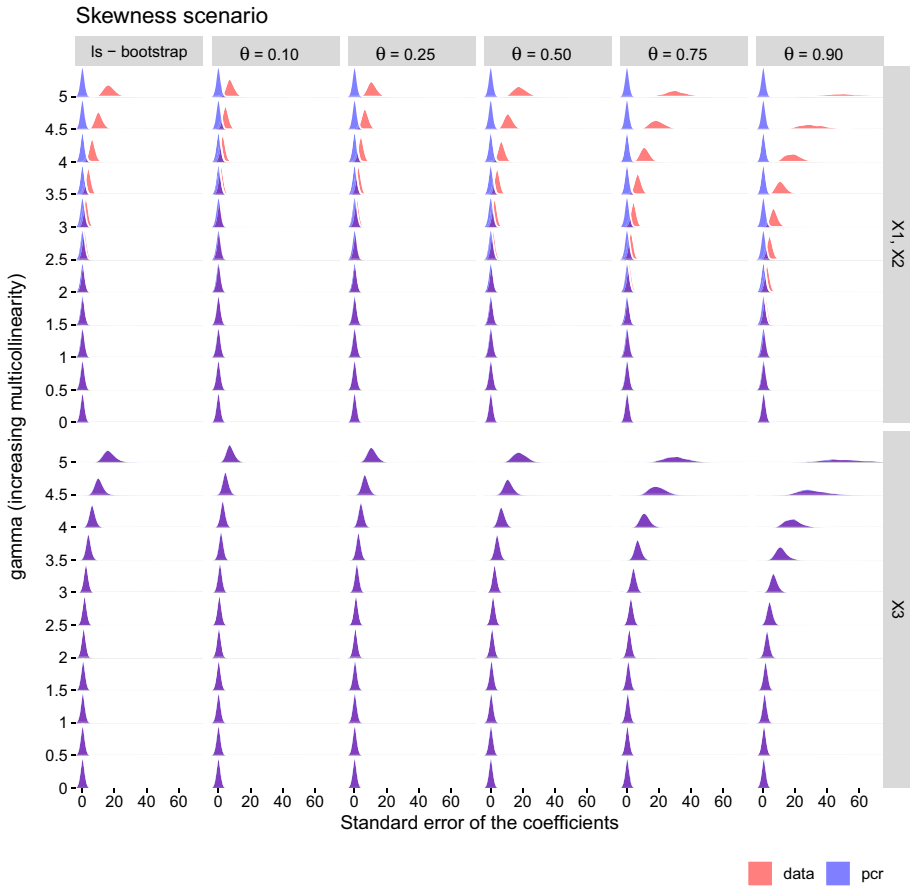


Fig. 3 Distributions of the standard errors of the three predictors (rows) in case of skewness for LSR (first column) and QR (from the second to the last column, for the five considered quantiles). Colours refer to the results obtained using the original regressors (darker densities) or the principal components (lighter densities) in the model

- the variability of the estimators for X_3 and the third PC is more pronounced than shown in Fig. 1, especially in the heteroscedastic case. This result was expected considering that the third component explains a residual part of the variability. In PCR, in fact, it is appropriate to use only a subset of components that explain a sufficient part of variability, or even the first component that is the most relevant
- PCR estimator, even if biased, is a reduced variance estimator, as highlighted in Sect. 2
- in the skewness scenario (Fig. 3), QR is less affected by the multicollinearity on the left side of the distribution ($\theta = 0.1$ and $\theta = 0.25$) while standard errors increase in size and variability on the right side. It is worthy of notice that this scenario is quite different from the others: the range of the standard errors is much wider, as evident comparing the values on the horizontal axis in Fig. 3 with the values in Figs. 1 and 2. Using the PCs in place of the regressors, any multicollinearity does not affect the variability of the estimates.

5 A case study for the analysis of consumer purchasing behaviour of a multi-channel retailer

The aim of the study presented in this section is to describe the effects of the presence of multicollinearity on the results of LSR and QR through a practical case study. The empirical analysis will show that strong linear relationships among regressors provide unstable estimated regression coefficients and inadequate statistical measures. Dropping one or more of the highly collinear regressors could be a possible solution to the collinearity problem even if the improvement in the efficiency of the estimates not always balances the loss of information triggered from the deletion of some variables. Moreover, the omission of relevant regressor(s) from the model may result in a specification error. A different option, as described in Sect. 2, consists of eliminating only the redundant information in the data by identifying a subset of new variables through a PCA. The resulting PCs are indeed linear combinations of the original variables but orthogonal and therefore uncorrelated. The use of a regression on the PCs will lead to satisfactory results both for LSR and QR, allowing a reduction in the standard errors of the estimators and, therefore, showing the real and significant contribution of the regressors.

A completely different approach to deal with the problem of multicollinearity is the use of penalty methods such as LASSO regression [42]. The empirical analysis discussed in this paper is further refined by the results obtained by applying a penalty term to the regression model. The objective is not a comparison between the LASSO regression and the proposed quantile regression on principal components as the logic followed by the two approaches are completely different, although in both cases it leads to overcoming the multicollinearity problem. The results provided by the LASSO approach, however, make it possible to show empirically the different philosophies followed by the two approaches. In the first, principal component regression, the original space is reduced by identifying components that are linear combinations of all variables. Penalty term methods, on the other hand, are in effect variable selection techniques whereby they suggest a subset of the observed variables.

The empirical analysis is carried out on a data set regarding customers of a retail who offers products both online and in-store. Data have been simulated using the guidelines provided in [7] to reproduce a typical situation of customer relationship management where personal information of the customers are related to their purchase habits and to the evaluation of the quality of service/product. We aim to assess whether and to what extent the purchasing behaviour, the level of satisfaction with the seller and the personal characteristics of customers influence purchases made online and whether this impact changes according to the amount of money spent (for example, according to categories of customers who make low, medium or high amount purchases). The use of artificial data guarantees the transparency of the analysis process and allows the interested reader to replicate the procedure, in line with the standard of reproducible research [26]. In this section, the simulation process proposed by [7] has been completely pursued, limiting the analysis to consumers with at least one online purchase in the year. Data deals with 632 customers, which are supposed to represent a random sample from a company's customer relationship management system.

The variables in the dataset are listed below, along with quick comments (labels in parentheses are used for tables and graphs):

- Age (*age*): the distribution is quite symmetric and ranges from 48 to 50.
- Credit score (*credit.score*): a typical measure that reflects the propensity of a customer to pay the credit back. This variable has been generated as a function of age, assuming

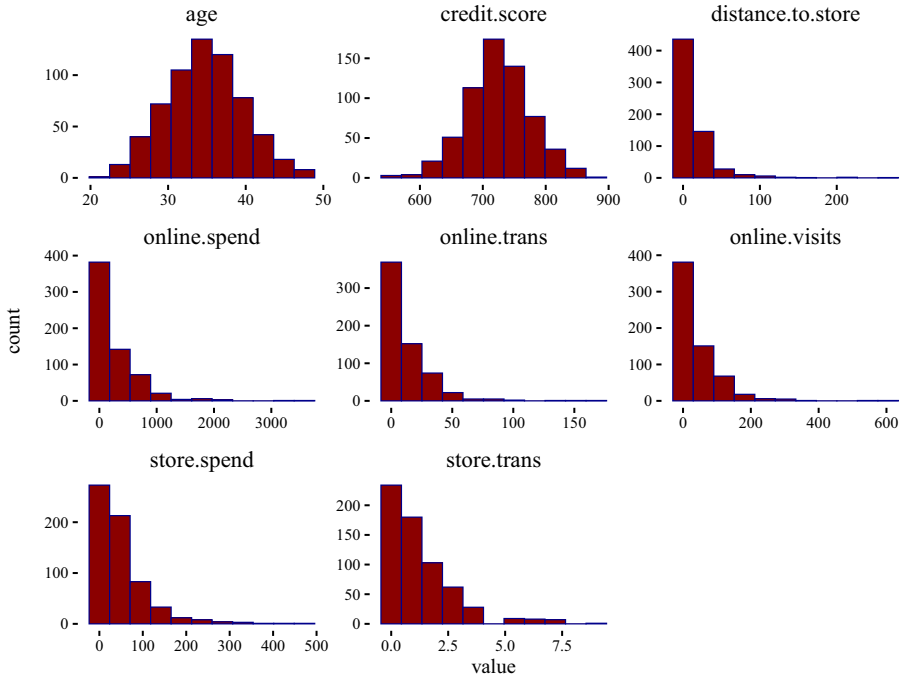


Fig. 4 Distribution of the quantitative variables

that older customers have higher credit scores on average. For this reasons the shape of the distribution is very similar to age, centred on an average value of €725.

- Distance from the store (*distance.to.store*): most of customers lives very close to the store and very few of them (less than 25% of the sample) very far. The variable is expressed in meters.
- On line visits (*online.visits*), transactions (*online.trans*) and total spending (*online.spend*) during a year: online activity is rather limited both in terms of frequency and amounts spent. Half of the consumers make a maximum of 20 visits for a total of 20 transactions on the company website. The average amount spent is equal to €269, and very high amounts of expenditure are very limited.
- Store transactions (*store.trans*) and total spending (*store.spend*) during a year: the attendance of the stores is more limited compared to the website both in terms of the transaction and purchase amounts. Also in this case, the distributions are rather asymmetrical.
- Level of satisfaction with service (*sat.service*) and with the selection of products (*sat.selection*): the evaluation of the satisfaction, recorded through a 5-points Likert scale, shows a quite symmetric distribution with respect to the service while the group of customers providing a negative evaluation about the selection of products (below the central point 3) prevails.

Distributions and univariate statistics are reported in Figs. 4 and 5 and Table 2.

Some remarks on the bivariate relationships are necessary before proceeding with the estimation of a model that involves the simultaneous analysis of all the variables. Figures 6 and 7 show the correlation coefficients and scatter plots (respectively above and below the

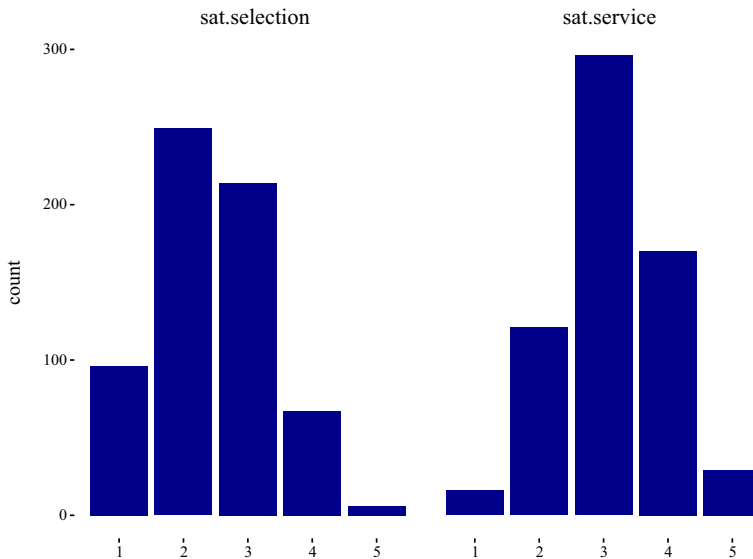


Fig. 5 Distribution of the ordinal satisfaction variables

Table 2 Central tendency statistics

	Age	Credit.score	Distance.to.store	Online.visits	Online.trans
Min.	22.13	543.00	0.21	1.00	1.00
1st Qu.	31.31	691.80	3.45	7.00	2.00
Median	34.87	725.00	7.58	20.00	6.00
Mean	34.73	725.00	15.48	44.43	13.27
3rd Qu.	37.91	757.00	17.14	59.00	18.00
Max.	48.56	869.10	267.09	606.00	169.00
	Online.spend	Store.trans	Store.spend	Sat.service	Sat.selection
Min.	14.73	0.00	0.00	1.00	1.00
1st Qu.	42.45	0.00	0.00	3.00	2.00
Median	122.84	1.00	31.58	3.00	2.00
Mean	269.49	1.32	47.04	3.129	2.43
3rd Qu.	361.18	2.00	66.44	4.00	3.00
Max.	3593.03	9.00	472.20	5.00	5.00

main diagonal) for the two groups of variables with very high correlations and thus herald of multicollinearity problems.

The analysis of the VIF values (Table 3, second column) confirms that these strong correlations will also condition the joint analysis of all variables. The results of the LSR model but also of the QR for the median (Table 3, from the third column) obviously reveal a strange behaviour of the regressors in terms of rather high error standards. The effect seems even amplified in QR where only *online.trans* has a significant effect among the five variables related to online and in-store behaviour. The effect of the multicollinearity on the estimate of

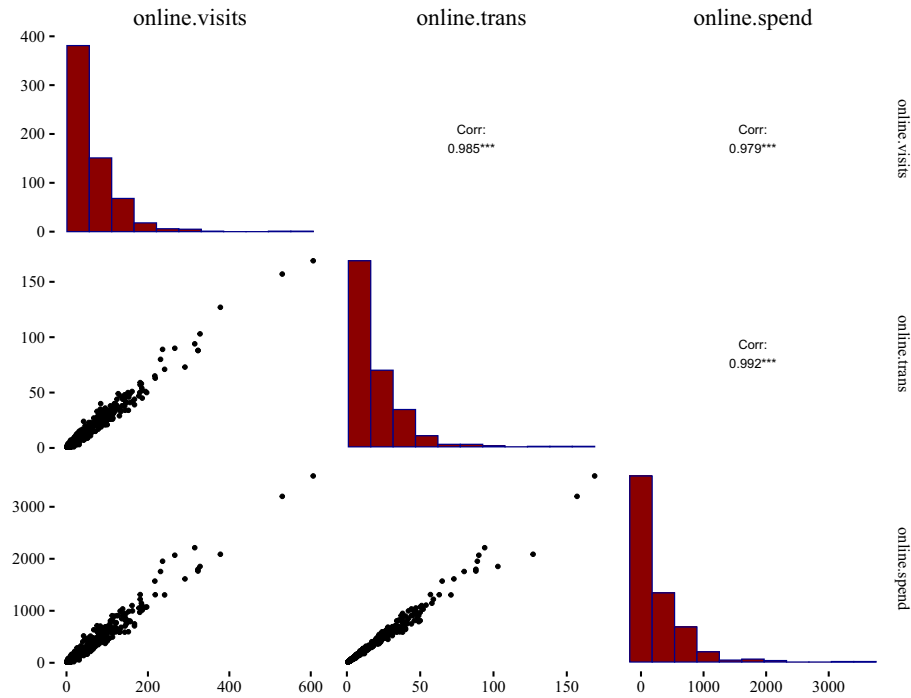


Fig. 6 Scatterplot matrix of the variables related to the online behavior

the conditioned median of the dependent variable is evident so that the analysis of the effects on more extreme parts of the distribution would lead to a predictable result. As it is known, regression results for more extreme θ values often lead to higher error standards.

Using PCA and inserting PCs in place of the original regressors allows eliminating redundant information in the data. By choosing the first four PCs as a synthesis of the original variables, it is possible to maintain an interesting share of variability (76.14%) as well as respecting the criterion of eigenvalue one (Table 4). These PCs, orthogonal to each other, represent different aspects of the phenomenon. From the correlations between variables and PCs (Table 5), it is in fact possible to interpret the first PC (F1.store) as the one linked to consumers who prefer the shop for their purchases and transaction, the second to customers who prefer online sales (F2.online). The third PC, on the other hand, contrasts customers with respect to the level of satisfaction for the seller (F3.sat), while the fourth summarises the personal aspects of the consumer, age and credit score (F4.personal). It therefore contrasts mature customers, both in a personal and economic sense, with those who are younger and have less credit credibility.

The four PCs can provide a valuable contribution to the explanation of expenses made online. Table 6 shows the results of LSR and QR for the median considering the four synthetic factors as regressors. It results that all the regressors provide significant contributions to the prediction of *online.spend*. In both models, the main impact is exerted by online visits and transactions (F2.online) followed, with a negative coefficient, by shopping behaviour at the stores. Median regression does not add interesting information to the results of classical LSR, but the estimation of QR for a denser grid of quantiles allows differentiating the impact of the regressors on the dependent variable. Figure 8 shows, for each regressor, the coefficient plot

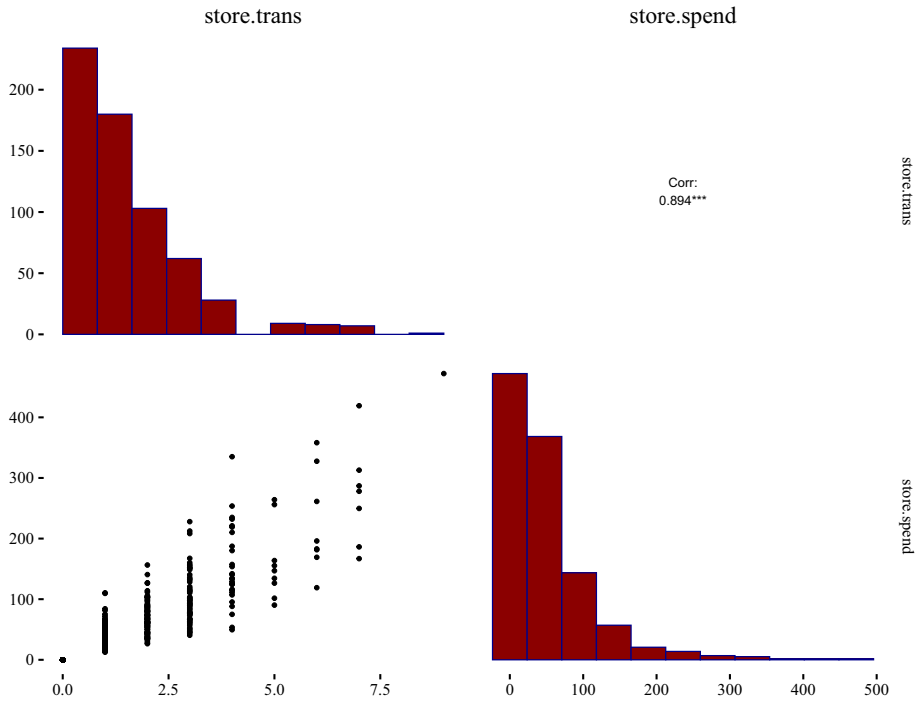


Fig. 7 Scatterplot matrix of the variables related to the in store behavior

typically used in QR. The horizontal axis displays the different quantiles, and the vertical axis the values of regression coefficients estimated at different quantiles (from 0.01 to 0.99 with a step equal to 0.01). Moving from lower to higher quantiles, the sign of the coefficients does not change while the size changes, in some cases of relevant amount. Results limited to the average effect flattens the different effects that can occur when the amounts spent online vary. For example, for large online purchases, it is very important how the consumer behaves in terms of how she/he accesses the seller's products (online or in-store). This means, for example, that being able to increase access to the site and also online transactions involves a greater impact among those who make very expensive online purchases than those who spend small amounts.

The analysis of the data presented in this section now extends to a description of the results obtained by applying the LASSO approach to the median regression model presented in Table 3. It is known that the output and performance of a LASSO regression depends on a tuning parameter, named lambda: when lambda is small, the result is essentially the least squares estimate while as lambda increases, the method shrinks the coefficient estimates towards zero. Figure 9 shows the distribution of the p-values associated with the regressors of the median regression model obtained by considering a dense grid of lambdas (from 0.01 to 10 with step 0.01). It is interesting to note that the method highlights a predominant, indeed unique, role for the variable *online.trans*, which is the only one to have significant coefficients. In other words, this means that, regardless of the performance of the model, the results obtained lead to a complete loss of the information provided by the other regressors. The combined use of PCA-regression allows, on the other hand, to preserve part of the contribution of all

Table 3 VIF and OLS and median regression results (coefficients, standard errors and p-value)

	VIF	OLS regression			QR regression		
		Estimate	Std. error	p-value	Estimate	Std. error	p-value
(Intercept)		- 11.98	29.64	0.69	- 13.23	6.62	0.04
Age	1.07	0.26	0.39	0.50	0.07	0.07	0.33
Credit.score	1.07	- 0.01	0.04	0.81	0.01	0.01	0.14
Distance.to.store	1.09	- 0.04	0.08	0.65	- 0.02	0.02	0.46
Online.visits	34.06	0.41	0.18	0.02	0.11	0.14	0.43
Online.trans	34.13	19.02	0.60	$\leq 2e-16$	20.01	0.53	0.00
Store.trans	5.11	- 2.31	2.84	0.42	- 0.35	0.56	0.53
Store.spend	4.99	0.05	0.07	0.46	0.01	0.01	0.42
Sat.service	1.53	2.74	2.73	0.32	0.51	0.64	0.43
Sat.selection	1.52	0.52	2.58	0.84	0.01	0.63	0.99

Table 4 Eigenvalues, percentage and cumulative percentage of variability

Component	Eigenvalue	% of variability	Cumulative % of variability
1	2.12	23.58	23.58
2	1.95	21.66	45.25
3	1.58	17.51	62.76
4	1.20	13.38	76.14
5	0.86	9.53	85.67
6	0.76	8.40	94.07
7	0.41	4.60	98.67
8	0.11	1.17	99.84
9	0.01	0.16	100.00

Table 5 Correlations between variables and factors

	F1	F2	F3	F4
Age	0.12	- 0.03	- 0.28	0.72
Credit.score	0.11	0.07	- 0.22	0.75
Distance.to.store	- 0.31	- 0.35	- 0.20	0.08
Online.visits	- 0.62	0.76	0.13	0.08
Online.trans	- 0.62	0.76	0.13	0.08
Store.trans	0.78	0.51	0.20	0.00
Store.spend	0.78	0.48	0.22	0.00
Sat.service	- 0.13	- 0.30	0.80	0.22
Sat.selection	- 0.08	- 0.28	0.81	0.23

Table 6 OLS and median regression results (coefficients, standard errors, p-values)

	OLS regression			QR regression		
	Coefficient	Std. error	p-value	Coefficient	Std. error	p-value
(Intercept)	269.49	2.34	$< 2e-16$	265.39	2.32	0.00
F1.store	-160.266	1.60	$< 2e-16$	-157.79	2.11	0.00
F2.online	204.64	1.66	$< 2e-16$	203.53	2.35	0.00
F3.satisfaction	39.54	1.77	$< 2e-16$	40.80	1.27	0.00
F4.personal	28.69	1.89	$< 2e-16$	27.11	1.13	0.00

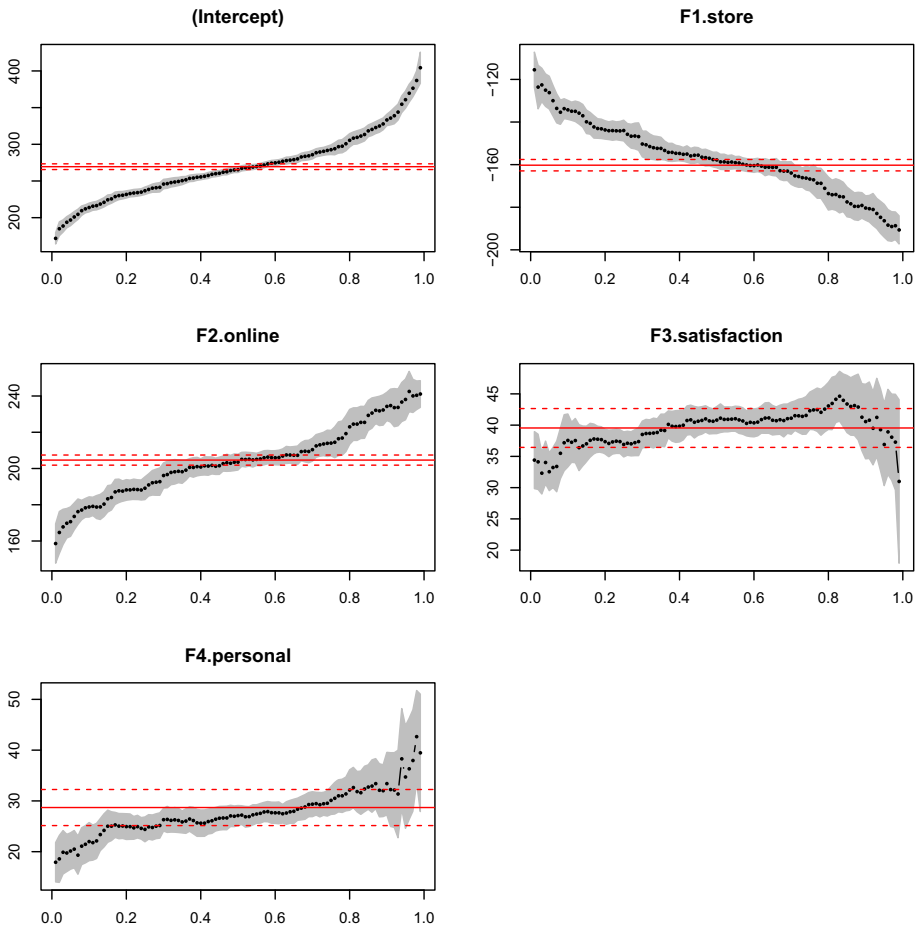


Fig. 8 Quantile regression coefficients

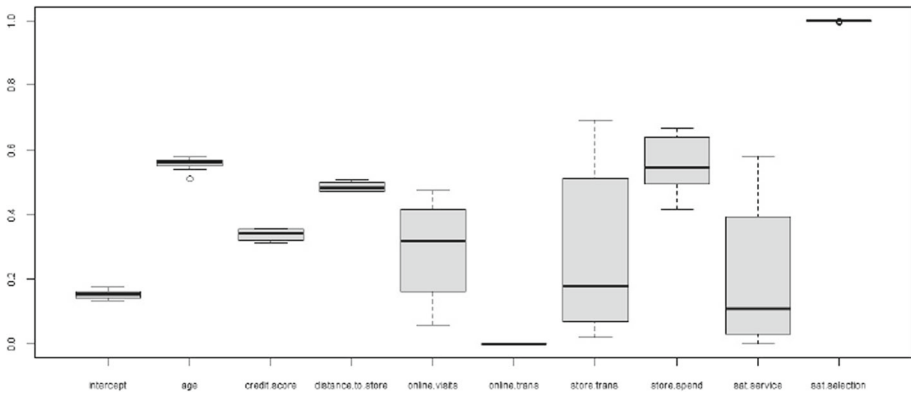


Fig. 9 Results obtained by applying the LASSO approach to the median regression model

Table 7 Results by a cross validated quantile regression with quantile equal to the median

	Estimate	Std. error	t value	$Pr(> t)$
(Intercept)	- 13.196	9.748	- 1.354	0.176
Age	0.072	0.125	0.576	0.565
Credit.score	0.0116	0.0126	0.926	0.355
Distance.to.store	- 0.015	0.023	- 0.660	0.509
Online.visits	0.112	0.157	0.717	0.474
Online.trans	20.010	0.557	35.892	<2e-16
Store.trans	- 0.352	0.882	- 0.399	0.690
Store.spend	0.010	0.022	0.461	0.645
Sat.service	0.509	0.911	0.559	0.576
Sat.selection	0.004	0.909	0.005	0.996

regressors. In fact, the factors extracted by PCA represent a linear combination of all the initial regressors, albeit with different weights.

Table 7 provides results obtained estimating a cross-validated quantile regression with quantile equal to the median. The use of cross-validation to identify the best lambda is quite common in LASSO regression [48], and it has been extended in quantile regression setting by Wang et al. [44]. Results in Table 7 shows a very penalised model where just the regressor *online.trans* is significantly different from zero.

6 Conclusion

This paper presents a thorough study of the effect of collinearity in QR using both artificial and empirical data.

Simulations results suggest different effects of collinearity in case of the different settings considered in the simulation design, and in particular several degrees of collinearity and different distributions of the response. Empirical findings show that as collinearity increases, standard errors in LS increase, but those in QR increase more. The larger increase is even more

evident in the heteroscedastic case. In case PCR is adopted, the stability of the results was confirmed in the different scenarios. Multicollinearity is properly solved using as regressors only the components that maximize the variability. The effect of the collinearity is similar in LSR and QR from the case study on the evaluation of the quality of services. The case study on empirical data also shows how to estimate a quantile regression model appropriately when some explanatory variables correlate highly with each other. Therefore, the steps to follow in the analysis should be as follows:

1. Descriptive bivariate analysis to analyze the correlations between pairs of variables;
2. Use of traditional multicollinearity diagnostic methods, such as VIF and NC, to identify the predictors responsible for multicollinearity;
3. Implementation of the QPCR, first identifying the relevant principal components and then regressing the explanatory variable on them;
4. Analysis of the correlations between original variables and principal components (loadings) to trace the most relevant explanatory variables.

The main findings of the study is the relevance of collinearity also in QR and the use of PCR as a possible solution, as already well experienced in LSR. The opportunity of having more stable results even for the most extreme quantiles can be a really significant advantage considering the QR feature of modelling the tails of the distribution of the response. The present study focused on the effect of collinearity on standard errors. Future research will focus on the effect of collinearity also in terms of prediction ability, considering both in-sample and out of sample prediction and in terms of estimate bias. Moreover, a comparison between QR on the principal components and the use of penalized regression will be also explored.

Acknowledgements The authors would like to thank Professor Tormod Naes for his intellectual contribution.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Alhamzawi, R., Yu, K.: Variable selection in quantile regression via Gibbs sampling. *J. Appl. Stat.* **39**(4), 799–813 (2012)
2. Ando, T., Tsay, R.S.: Quantile regression models with factor-augmented predictors and information criterion. *Econom. J.* **14**(1), 1–24 (2011)
3. Bager, A.S.M.: Ridge parameter in quantile regression models. An application in biostatistics. *Int. J. Stat. Appl.* **8**(2), 72–78 (2018)
4. Bayer, S.: Combining value-at-risk forecasts using penalized quantile regressions. *Econom. Stat.* **8**, 56–77 (2018)
5. Bowerman, B.L., O'Connell, R.T.: *Linear Statistical Models: An Applied Approach*. PWS-KENT Publishing Co, Boston (1990)
6. Buchinsky, M.: Changes in the U.S. wage structure 1963–1987. *Econometrica* **62**, 405–458 (1994)
7. Chapman, C., Feit, E.M.: *R for Marketing Research and Analytics*. Springer, New York (2015)
8. Davino, C., Vistocco, D.: Quantile regression for the evaluation of student satisfaction. *Ital. J. Appl. Stat.* **20**(3–4), 179–196 (2008)

9. Davino, C., Vistocco, D.: The evaluation of University educational processes: a quantile regression approach. *Statistica* **67**, 281–292 (2007)
10. Davino, C., Romano, R., Næs, T.: The use of quantile regression in consumer studies. *Food Qual. Preference* **40**, 230–239 (2015)
11. Davino, C., Furno, M., Vistocco, D.: *Quantile Regression: Theory and Applications*, vol. 988. Wiley, New York (2013)
12. Eide, E., Showalter, M.H.: The effect of school quality on student performance: a quantile regression approach. *Econ. Lett.* **58**(3), 345–350 (1998)
13. Fan, Y., Huang, G., Li, Y., Wang, X., Li, Z., Jin, L.: Development of PCA-based cluster quantile regression (PCA-CQR) framework for streamflow prediction: application to the Xiangxi river watershed, China. *Appl. Soft Comput.* **51**, 280–293 (2017)
14. Fang, W., Chen, Y., Cheng, D., Zhang, H., Li, L.: Improved quantile regression analysis on small sample multicollinear time series measured data. *IOP Conf. Ser. Earth Environ. Sci.* **304**(3) (2011)
15. Faraway, J.J.: *Linear Models with R*. CRC Press, Boca Raton (2014)
16. Fitzenberger, B., Koenker, R., Machado, J.A.F.: *Economic Applications of Quantile Regression*, Series: Studies in Empirical Economics. Physica, Heidelberg (2002)
17. Furno, M., Vistocco, D.: *Quantile Regression: Estimation and Simulation*, vol. 216. Wiley, New York (2018)
18. Giglio, S., Kelly, B., Pruitt, S.: Systemic risk and the macroeconomy: an empirical evaluation. *J. Financ. Econ.* **119**(3), 457–471 (2016)
19. Gunst, R.F.: Regression analysis with multicollinear predictor variables: definition, detection, and effects. *Commun. Stat. Theory Methods* **12**(19), 2217–2260 (1983)
20. Hair, J.F., Jr., Black, W.C., Babin, B.J., Anderson, R.E.: *Multivariate Data Analysis*, 8th edn. Cengage, Boston (2019)
21. Helland, I.S., Almøy, T.: Comparison of prediction methods when only a few components are relevant. *J. Am. Stat. Assoc.* **89**(426), 583–591 (1994)
22. Hendricks, W., Koenker, R.: Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Am. Stat. Assoc.* **87**, 58–68 (1992)
23. Huang, Q., Zhang, H., Chen, J., He, M.: Quantile regression models and their applications: a review. *J. Biom. Biostat.* **8**, 354 (2017)
24. James, G., Witten, D., Hastie, T., Tibshirani, R.: *An Introduction to Statistical Learning*. Springer, New York (2013)
25. Jolliffe, I.T.: *Principal Component Analysis*. Springer, New York (1986)
26. King, G.: Replication, Replication. *PS Political Sci. Politics* **28**(3), 444–452 (1995)
27. Kocherginsky, M., He, H., Mu, Y.: Practical confidence intervals for regression quantiles. *J. Comput. Graph. Stat.* **14**(1), 41–55 (2005)
28. Koenker, R.: Quantile regression for longitudinal data. *J. Multivar. Anal.* **91**(1), 74–89 (2004)
29. Koenker, R., Bassett, G.W.: Regression quantiles. *Econometrica* **46**, 33–50 (1978)
30. Koenker, R., Bassett, G.W.: Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* **50**, 43–61 (1982)
31. Koenker, R., Bassett, G.W.: Tests for linear hypotheses and L1 estimation. *Econometrica* **46**, 33–50 (1982)
32. Koenker, R., Chernozhukov, V., He, X., Peng, L.: *Handbook of Quantile Regression*. Chapman and Hall/CRC (2017)
33. Koenker, R., Machado, J.: Goodness of Fit and related inference processes for quantile regression. *J. Am. Stat. Assoc.* **94**, 1296–1310 (1999)
34. Martens, H., Næs, T.: *Multivariate Calibration*. Wiley, Chichester (1989)
35. McDonald, G.C.: Ridge regression. *Wiley Interdiscip. Rev. Comput. Stat.* **1**(1), 93–100 (2009)
36. Næs, T., Helland, I.S.: Relevant components in regression. *Scand. J. Stat.* **20**, 239–250 (1993)
37. Næs, T., Indahl, U.: A unified description of classical classification methods for multicollinear data. *J. Chemom.* **12**(3), 205–220 (1998)
38. Næs, T., Mevik, B.H.: Understanding the collinearity problem in regression and discriminant analysis. *J. Chemom.* **15**, 413–426 (2001)
39. Næs, T., Isaksson, T., Fearn, T., Davies, T.: *A User-Friendly Guide to Multivariate Calibration and Classification*. NIR Publications, Chichester (2004)
40. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. <https://www.R-project.org/> (2020)
41. 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. *Chemom. Intell. Lab. Syst.* **146**, 128–135 (2015)
42. Tibshirani, R.: Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **58**(1), 267–288 (1996)

43. Tychonoff, A.N.: Solution of incorrectly formulated problems and the regularization method. *Sov. Math.* **4**, 1035–1038 (1963)
44. Wang, L., Wu, Y., Li, R.: Quantile regression of analyzing heterogeneity in ultra-high dimension. *J. Am. Stat. Assoc.* **107**, 214–222 (2008)
45. Weisberg, S.: *Applied Regression Analysis*. Wiley, New York (1985)
46. Wu, Y., Liu, Y.: Variable selection in quantile regression. *Stat. Sin.* **19**(2), 801–817 (2009)
47. Yu, K., Lu, Z., Stander, J.: Quantile regression: applications and current research areas. *J. R. Stat. Soc. Ser. D (The Statistician)* **52**(3), 331–350 (2003)
48. Zou, H., Li, R.: One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat.* **36**, 1509–1533 (2008)
49. Zaikarina, H., Djuraidah, A., Wigena, A.H.: Lasso and ridge quantile regression using cross validation to estimate extreme rainfall. *Glob. J. Pure Appl. Math.* **12**(3), 3305–3314 (2016)
50. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* **67**(2), 301–320 (2005)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.