https://doi.org/10.22319/rmcp.v14i1.6182

Article

Quantile regression for prediction of complex traits in Braunvieh cattle using SNP markers and pedigree

Jonathan Emanuel Valerio-Hernández [a]

Paulino Pérez-Rodríguez [b]*

Agustín Ruíz-Flores [a]

[a] Universidad Autónoma Chapingo. Posgrado en Producción Animal. Carretera Federal México-Texcoco Km 38.5, 56227, Texcoco, Estado de México, México.

[b] Colegio de Postgraduados. Socio Economía Estadística e Informática. Carretera Federal México-Texcoco Km 36.5, 56230, Texcoco, Estado de México.

*Corresponding author: perpdgo@colpos.mx

Abstract:

Genomic prediction models generally assume that errors are distributed as normal, independent, and identically distributed random variables with zero mean and equal variance. This is not always true, in addition there may be phenotypes distant from the population mean, which are usually removed when making the prediction. Quantile regression (QR) deals with statistical aspects such as high dimensionality, multicollinearity and phenotypic distribution different from the normal one. The objective of this work was to compare QR using marker and pedigree information with alternative methods such as genomic best linear unbiased prediction (GBLUP) and single-step genomic best linear unbiased prediction (ssGBLUP) to analyze the birth (BW), weaning (WW) and yearling (YW) weights of Braunvieh cattle and simulated data with different degrees of asymmetry and proportion of outliers. The predictive capacity of the models was assessed by cross-validation. The predictive performance of QR both with marker information alone and with information of markers plus pedigree, with the actual dataset, was better than the GBLUP and ssGBLUP methodologies for WW and YW. For BW, GBLUP and ssGBLUP were better, however, only quantiles 0.25, 0.50 and 0.75 were used, and the BW distribution was not asymmetric. In the simulated data experiment, correlations between “true” marker effects and estimated effects, as well as “true” and estimated signal correlations were higher when QR was used compared to GBLUP. The advantages of QR were more noticeable with asymmetric distribution of phenotypes and with a higher proportion of outliers, as was the case with the simulated dataset.

Key words: Quantile regression, GBLUP, ssGBLUP.

Received: 30/03/2022

Accepted: 04/08/2022

Introduction

The main motivation of the quantile regression (QR) method is that most models for genetic evaluation assume normality, which is not always true. Another problem is that sometimes phenotypic records very far from the population average are considered as recording errors or outliers and therefore removed from the analyses, seen from the genomic point of view, valuable information of markers associated with certain regions of DNA with strong influence on characteristics of interest is being lost.

With the QR method, robust results and a broad vision of the explanatory variables on the dependent ones are obtained[1]. The data generated from omics experiments are often complex and large, so there is a statistical challenge to extract relevant biological information from the large volume of data[2,3]. Using a robust approach such as QR makes inference less biased and less subject to false positives[2]. Recent studies using QR describe various applications such as etic association studies[4], population genetics[5], gene expression[6,7], and genomic selection[8–10].

One of the first studies where QR was used to predict individual genetic merit was presented by Nascimento *et al*[11], who used simulated data, finding advantages when using QR compared to conventional methodologies. In the same year[12], results using QR to adjust growth curves with data from pigs and molecular markers were published; not only did they successfully adjust the growth curves, but they identified important markers associated with the studied characteristic. Another similar work by the same team of researchers was presented by Nascimento *et al*[13], but with bean data. Recently, Pérez-Rodríguez *et al*[10] extended the quantile regression model to include pedigree information through the use of the additive genetic relationship matrix, further improving the predictive ability of the models and at the same time identifying the proportions of the variances attributed to markers, relationships between individuals and the residual, which allows a precise partitioning of the phenotypic variance to be obtained.

The objective of the present study was to study the predictive power of the quantile regression model using simulated data and actual data (birth, weaning and yearling weights) from Braunvieh cattle and the following models were considered: 1) QR with information of SNP molecular markers (QRM), 2) QR simultaneously including molecular marker information and genealogical information derived from pedigree (QRH); 3) GBLUP which, like QRM, only included molecular marker information, and 4) single-step genomic evaluation (ssGBLUP) which included marker and pedigree information.

Material and methods

Genotypes

The information used was from 300 animals (236 females, 64 males) born from 2001 to 2016 in eight herds located in Eastern, Central and Western Mexico. Hair samples were collected for genotyping by the company GeneSeek (Lincoln, https://www.neogen.com/, NE, USA), using the GeneSeek® Genomic Profiler Bovine LDv.4 panel, with 30,000 and 50,000 SNP markers, 150 animals with each Chip. Genotyping was performed on two separate occasions, initially 150 individuals with the 30K Chip and later another group of 150 individuals with the 50K Chip since the 30K Chip was not available at the time. The SNPs in common between the 30K and 50K chips (12,835 SNPs) were used. The proportions of missing values were calculated for each marker and for each individual. The average of missing values per individual was 2.09 % with a standard deviation of 7.50 %. The average call rate (not missing proportion for each marker) was 97.90 % with a standard deviation of 4.66 %. Markers with a call rate of less than 95 % were removed. The genotypes were recoded as AA= 0, AB= 1 and BB= 2, from which a matrix with 300 rows (individuals) and 12,835 columns (markers) was obtained, whose cells take values in the set $\{0,1,2,-\}$, where “$-$” denotes a missing value. For the 12,835 common markers of the two chips, the missing values were randomly imputed, generating samples of the $Binomial(2,\hat{p})$ distribution, where $\hat{p}$ is the frequency of the major allele, calculated from the observed marker genotypes. Monomorphic markers or those with minor allele frequency (MAF) less than 0.04 were removed. After quality control, 9,628 of the 12,835 SNPs in common between the two chips were available for further analyses.

Phenotypes

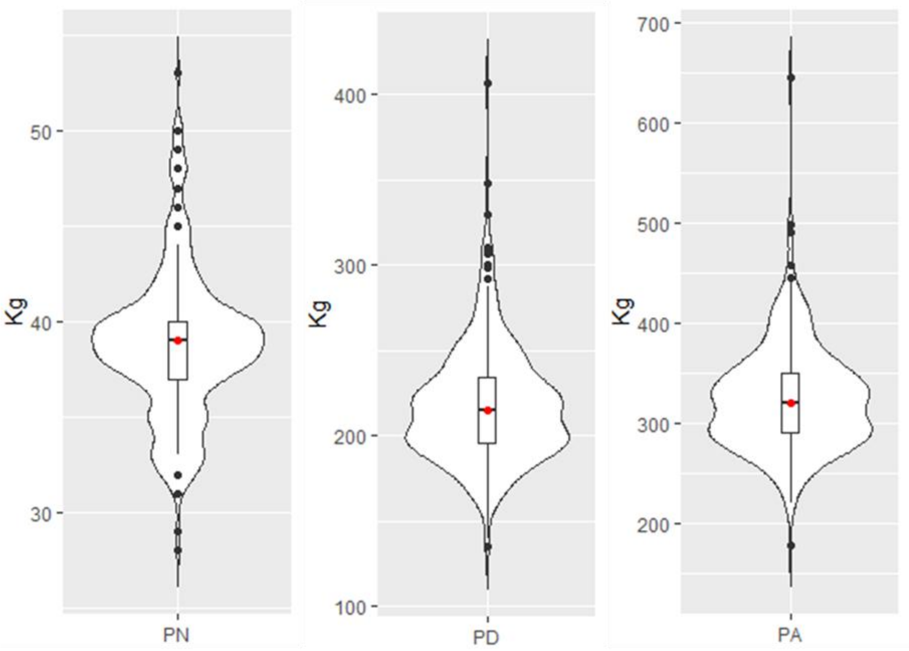
The phenotypic and pedigree information of the Braunvieh cattle population was obtained from the database of the Mexican Association of Breeders of Registered Swiss Cattle. Records of birth (BW), weaning (WW) and yearling (YW) weights were used for analysis. Phenotype editing was similar for BW, WW and YW, records of animals not genetically related to those genotyped or with missing information for herd, dam's age and management were discarded. Contemporary groups (CG) were defined by removing animals in CG of 2 individuals or with variance equal to zero. For BW, the CGs were defined by combining the effects of the herd (8 herds), year (1998 to 2016) and season of birth; the seasons of birth were defined considering the Julian calendar, from 80 to 171d, spring; from 172 to 264 d, summer; from 265 to 354 d, autumn; from 355 to 366 d and from 1 to 79 d, winter. After editing data, for BW, 330 records were obtained. For WW and YW, the CGs were defined by combining the effects of the herd (6 herds), year (from 1998 to 2016), season of birth (same as BW) and management. In the case of WW, the management groups were defined in three ways: calves fed their mother's milk; their mother's milk plus balanced feed; and milk from their mother and nurse plus a balanced diet. For YW, the management groups were defined in three ways: grazing animals; grazing animals with feed concentrate; and housed animals with a balanced diet. The edition of WW and YW data ended with 267 and 232 records for further analyses. Table 1 shows a summary of the number of animals genotyped, and phenotyped for BW, WW and YW. Figure 1 shows the violin plots for BW, WW and YW, the sample mean is represented by the red dot and the sample median by the horizontal line within the box, from the plot, it is clear that the response variables have an asymmetric distribution and the circles with solid filling in it suggest the presence of outliers.

Table 1: Number of animals genotyped and phenotyped for the analysis of birth, weaning and yearling weights of a Braunvieh cattle population

Group	**Birth weight**	**Weaning weight**	**Yearling weight**
Genotyped	300	300	300
Genotyped and phenotyped	232	218	191
Phenotyped in QRM and GBLUP	232	218	191
Phenotyped in QRH and ssGBLUP	330	267	232

QRM=Quantile regression using marker information, QRH=Quantile regression using marker and pedigree information, GBLUP=Genomic best linear unbiased predictor, ssGBLUP=Single-step genomic evaluation.

Figure 1: Violin plots of birth (PN=BW), weaning (PD=WW) and yearling (PA=YW) weights in a Braunvieh cattle population



The sample mean is represented by the red dot and the sample median by the horizontal line inside the box

Models

Quantile regression model with markers (QRM)

The model for quantile regression is:

$$y_i = \mu + \boldsymbol{x}_i^t \boldsymbol{\beta} + w_i,$$

where y_i is the value of the phenotype of the i-th animal; μ is an intercept; $\boldsymbol{x}_i^t = (x_{i1}, \dots, x_{ip})$ represents the i-th row of the marker matrix, $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)^t$ is the vector of regression coefficients associated with markers and w_i are independent random variables such that their quantile $\theta \in (0,1)$ is zero. The estimation of the regression coefficients for a fixed interest quantile θ is obtained by solving the following minimization problem:

$$min\left\{\sum_{i=1}^{n} \rho_\theta \left(y_i - \mu - \boldsymbol{x}_i^t \boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} \left|\beta_j\right|\right\},$$

where $\sum_{j=1}^{p} \left|\beta_j\right|$ is the sum of the absolute values of the regression coefficients; λ is the penalty parameter that controls the intensity of regularization; and $\rho_\theta(\cdot)$ is the function defined as[1]:

$$\rho_\theta(t_i) = \begin{cases} \tau \times t_i & \text{If } t_i \geq 0 \\ -(1-\tau) \times t_i & \text{If } t_i < 0, \end{cases}$$

where $t_i = y_i - \mu - \boldsymbol{x}_i^t \boldsymbol{\beta}$. After estimating the parameters of the model, the breeding values estimated by markers (GEBV) are obtained by the following expression:

$GEBV(\tau) = \hat{y}_i(\tau) = \sum_{j=1}^{p} x_{ij}\hat{\beta}_j(\tau)$,

where $\hat{\beta}_j(\tau)$ is the effect of the *j*-th marker, defined by the functional relationship obtained for the quantile of interest.

The QR model can be extended to include other terms, in particular for growth characteristics, the following model is used:

$y_i = \mu + \boldsymbol{s}_i^t\boldsymbol{\varsigma} + \boldsymbol{c}_i^t\boldsymbol{\varrho} + \boldsymbol{x}_i^t\boldsymbol{\beta} + \boldsymbol{w}_i$,

where y_i is the value of the phenotype of the analyzed characteristic (BW, WW or YW) of the *i*-th animal, μ is an intercept; $\boldsymbol{s}_i^t = (s_{i1}, \ldots, s_{if})$ the *i*-th row of the incidence matrix for fixed effects (sex, dam's age, management), $\boldsymbol{\varsigma} = (\varsigma_1, \ldots, \varsigma_f)^t$ the regression coefficients for fixed effects, $\boldsymbol{c}_i^t = (c_{i1}, \ldots, c_{it})$ the *i*-th row of the incidence matrix for random effects of contemporary group (54, 43 and 37 for BW, WW and YW), $\boldsymbol{\varrho} = (\varrho_1, \ldots, \varrho_t)^t$ random effects of contemporary group, the rest of the terms as described above.

GBLUP

The model is given by:

$y_i = \mu + \boldsymbol{s}_i^t\boldsymbol{\varsigma} + \boldsymbol{c}_i^t\boldsymbol{\varrho} + \boldsymbol{z}_i^t\boldsymbol{u} + e_i$,

where $\boldsymbol{z}_i^t = (z_{i1}, .., z_{in})$ is the *i*-th row of the matrix that connects phenotypes with genotypes, $\boldsymbol{u} = (u_1, \ldots, u_n)^t$ is the vector of random effects for animals. Additive, contemporary group and residual genetic variances are assumed $Var(\boldsymbol{u}) = \boldsymbol{G}\sigma_u^2$, $Var(\boldsymbol{c}) = \mathbf{I}\sigma_{cg}^2$, and $Var(\boldsymbol{e}) = \mathbf{I}\sigma_e^2$, respectively. The matrix of genomic relationships, $\boldsymbol{G}$, is calculated as described by Lopez-Cruz *et al*(14) and Pérez-Rodríguez *et al*(15); briefly, **G** = **WW**'/*p*, where **W** is the standardized and centered marker matrix (each marker centered by subtracting the mean allele frequency and standardized by dividing by the standard deviation of the sample of the allele frequency), *p* is the total number of markers, e_i normal and independent random variables with normal distribution with mean 0 and variance σ_e^2.

Single-step quantile regression (QRH) model

This method is considered an extension of the quantile model for a relationship matrix constructed using matrices of relationships for genotyped and non-genotyped animals and of which a pedigree is available. The resulting matrix is known in the literature as matrix **H**(16,17), this matrix is given by:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_a^{-1} - \mathbf{A}_{gg}^{-1} \end{bmatrix},$$

where, $\mathbf{A}_{gg}$ is a submatrix of **A** for genotyped animals, $\mathbf{G}_a = \beta\mathbf{G} + \alpha$; β and α are obtained by solving the system of equations:

$$\begin{cases} Avg(diag(\mathbf{G}))\beta + \alpha = Avg\big(diag(\mathbf{A}_{gg})\big) \\ Avg(\mathbf{G})\beta + \alpha = Avg(\mathbf{A}_{gg}) \end{cases}.$$

The QRH model is given by:
$y_i = \mu + \boldsymbol{s}_i^t \boldsymbol{\varsigma} + \boldsymbol{c}_i^t \boldsymbol{\varrho} + \boldsymbol{z}_i^t \boldsymbol{u} + w_i,$
where $Var(\boldsymbol{u}) = \sigma_H^2 \boldsymbol{H}$, the rest of the terms as described above.

Single-step GBLUP regression (ssGBLUP) model

The ssGBLUP model is equivalent to the GBLUP model described above with the difference that the genomic relationship matrix **G** is replaced with the extended genetic relationship matrix **H**, it is assumed that $Var(\boldsymbol{u}) = \boldsymbol{H}\sigma_H^2$.

Cross-validation

The predictive capacity of the models was evaluated by cross-validation, which was performed as follows. The dataset was divided into five groups of the same size $\{S_1, S_2, \dots, S_5\}$, 80 % of the data was used for training of the model, the remaining 20 % for validation. For example, $\{S_2\}$ is used as a validation group and the set $\{S_1, S_3, \dots, S_5\}$ for training of the model. The models were fitted using the training set, and the fitted model was used to obtain predictions for the validation set. This procedure was repeated five times and predictions were obtained for each group. Correlations between observed and predicted phenotypes were calculated and averaged for the test sets(18). Note that because these are actual values, the true breeding values are not known, but only the observed phenotypes are available, the fitted model provides predictions for breeding values and predictions of other fixed and environmental effects, with which a prediction of the phenotype is obtained, which is contrasted with the true value of the phenotype.

Simulation

In order to evaluate the predictive power of the QR model against GBLUP, an asymmetric data simulation with the presence of outliers was also carried out; the simulation of the present work is analogous to that used by Pérez-Rodríguez *et al*(10). The main idea is to highlight that the quantile regression model works adequately in the presence of atypical observations, inhomogeneous variances and response variables with responses with asymmetric distribution. In the context of selection, for example, it is not unusual to have asymmetric distributions for phenotypes due to the process itself, since, if one selects for some characteristic Y, and if there is in addition to this another characteristic of interest *O*, then the conditional distribution of *Y |O>o*(19) is asymmetric. In the context of genomic selection, it is also common to find subsets of observations that differ significantly from the rest and these observations could be considered atypical. Montesinos-López *et al*(20) proposed a model with Laplace errors and showed that it predicts adequately even in the presence of outliers, the proposed model is a special case of the quantile regression model that is studied in the present work. The 9,628 SNPs resulting from the quality control described above for 300 animals were considered, the simulation of the data was carried out considering the linear model:

$y_i = \mu + \sum_{j=1}^{9,628} \boldsymbol{x}_{ij}\beta_j + e_i$,

where $i = 1, \dots, 300$, with $\mu = 39$ for BW, it was assumed that the errors come from a biased normal distribution (SN_c) with mean 0, variance σ^2 (scale parameter σ) and asymmetry index γ_1, that is $e_i \sim SN_C(0, \sigma, \gamma_1)$, with $\sigma = \sqrt{1 - h^2}$, h^2 with a value of 0.35, $\gamma_1 = \sqrt{\frac{2}{\pi}}\rho^3\left(\frac{4}{\pi} - 1\right)\left(1 - \frac{2\rho^2}{\pi}\right)^{-3/2}$, $\rho \in \{0.950, 0.975, 0.999\}$ were considered, leading to different degrees of positive bias. Only positive values of γ_1 were considered since the negative bias is obtained simply by changing the sign of the $e_i's$ and therefore the conclusions obtained for the case of positive bias will also be valid for the negative case[21,22]. Fifty markers with non-zero effect were fixed, simulating them from a normal distribution with mean 0 and variance $\sqrt{1 - h^2}/50$, the rest of the markers were set at 0; the positions of the sampled markers were taken at random. To introduce outliers in the phenotypes, a certain proportion of the residues of $e_i \sim SN_C(0{,}3, \gamma_1)$ were randomly generated, two proportions were considered, 5 and 10 %, so samples from a mixture of two components of biased normal distributions were taken. Six datasets were generated, three different asymmetry coefficients 0.950, 0.975, 0.999 with their two alternatives of outlier proportion 5 % and 10 %. The asymmetric normal distribution has been used in genomic prediction[22] and its use in channeled selection has also been suggested[23]. Once the data were generated, the QRM model was fitted with $\theta = \{0.25, 0.50, 0.75\}$ to compare it with GBLUP. The selection of quantiles was made according to Nascimento *et al*[11], who consider these three possibilities when the distribution of phenotypes is asymmetric $\theta \in \{0.25{,}0.75\}$ or when the distribution is symmetric θ 0.50, since our fundamental interest in this work focused on the modeling of possibly asymmetric data and with the presence of outliers. The selection of the parameters was also made for computational convenience since the fitting of the model is done by using intensive computational techniques based on Markov chain Monte Carlo, as mentioned in the section on software and fitting of the models. For each analysis, the correlation between true and estimated $\boldsymbol{\beta}'s$, the correlation between true $\boldsymbol{X\beta}$ and estimated $\boldsymbol{X\hat{\beta}}$ signals and the component of variance associated with the residuals for each model, which is a way to evaluate the goodness of fit of the models, were calculated. The Deviation Information Criterion (DIC) was also considered, which can be used to select candidate models; models with lower DIC are preferred to models with higher DIC[24].

Software and model fitting

The quantile regression models were fitted using a computational strategy similar to that described by Pérez-Rodríguez *et al*[10]. Adaptations of algorithms to include fixed and random effects do not present great computational difficulty. The codes for the fitting of the models were developed in the programming languages R[25] and C. The codes for the fitting of the models were organized in such a way that they can be easily run from the statistical software R and are available by requesting them to the first author of the present study. In all cases, three quantiles were selected, $\theta = \{0.25, 0.50, 0.75\}$. The GBLUP and ssGBLUP models were fitted with the BGLR library of R[26].

Results

Real data

Tables 2, 3, and 4 show the results of the experiment conducted with BW, WW, and YW data from a Braunvieh cattle population, evaluated under two scenarios 1) with marker information only, and 2) marker and pedigree information. In general, the highest correlations between observed and predicted values were obtained with QR, except for BW, where the correlations of GBLUP and ssGBLUP were higher than those obtained with QRM and QRH, however, the correlations of QRM $\theta = 0.75$ and QRH $\theta = 0.75$ were close to those obtained with GBLUP and ssGBLUP (0.7902 *vs* 0.7924), (0.6981 *vs* 0.7055), respectively. The lowest MSE values were obtained with QRM $\theta = 0.75$ and QRH $\theta = 0.75$ in the WW characteristic, while in the BW and YW characteristics, the lowest values were obtained with GBLUP and ssGBLUP. The variance components associated with the error obtained with QRM and QRH were lower than those obtained with GBLUP and ssGBLUP. In general, the lowest DIC values were obtained with QRM $\theta = 0.75$ and QRH $\theta = 0.75$, except for BW with the markers-only scenario, where the lowest DIC was obtained with QRM $\theta = 0.25$.

Table 2: Averages of Pearson correlation and standard deviation (in parentheses) between observed phenotypic values ($\boldsymbol{y}$) and predicted phenotypic values ($\hat{\boldsymbol{y}}$), mean squared error, variance components associated with the error (σ_e^2, σ_w^2) and deviation information criterion (DIC) for birth weight

Model	**Cor($\boldsymbol{y},\hat{\boldsymbol{y}}$)**	**MSE**	$\boldsymbol{\sigma_e^2}$ **or** $\boldsymbol{\sigma_w^2}$	**DIC**
QRM $\boldsymbol{\theta = 0.25}$	0.7521	3.9973	2.7260	**513.5014**
	(0.0753)	(1.6108)	(1.9762)	(531.5701)
QRM $\boldsymbol{\theta = 0.50}$	0.5619	7.3249	8.6297	970.7680
	(0.1501)	(0.4561)	(0.2660)	(6.9791)
QRM $\boldsymbol{\theta = 0.75}$	0.7902	3.6535	**2.4268**	716.4237
	(0.0766)	(0.0943)	(0.4829)	(35.7161)
GBLUP	**0.7924**	**2.3269**	3.0035	803.0675
	(0.0874)	(0.2063)	(0.5578)	(31.9814)
QRH $\boldsymbol{\theta = 0.25}$	0.6713	3.5026	**2.3645**	872.3949
	(0.1329)	(1.2848)	(1.9670)	(432.0737)
QRH $\boldsymbol{\theta = 0.50}$	0.6816	2.9988	2.7372	**659.1450**
	(0.1253)	(0.7769)	(1.8239)	(1079.8674)
QRH $\boldsymbol{\theta = 0.75}$	0.6981	4.1405	2.8610	1077.2027
	(0.1140)	(0.6187)	(0.8666)	(60.6781)
ssGBLUP	**0.7055**	**2.4463**	3.2641	1189.4282
	(0.1191)	(0.2204)	(0.4244)	(26.5023)

Cor($\boldsymbol{\beta},\hat{\boldsymbol{\beta}}$)=correlation between observed and predicted phenotypes, MSE=mean squared error, σ_e^2 or σ_w^2=components of variance associated with the error, DIC=deviation information criterion.

Table 3: Averages of Pearson correlation and standard deviation (in parentheses) between observed phenotypic values ($\boldsymbol{y}$) and predicted phenotypic values ($\hat{\boldsymbol{y}}$), mean squared error, variance components associated with the error (σ_e^2, σ_w^2) and deviation information criterion (DIC) for weaning weight

Model	**Cor($\boldsymbol{y},\hat{\boldsymbol{y}}$)**	**MSE**	$\boldsymbol{\sigma_e^2}$ **or** $\boldsymbol{\sigma_w^2}$	**DIC**
QRM $\boldsymbol{\theta = 0.25}$	0.5661	476.5293	419.4138	1550.5339
	(0.2212)	(17.4612)	(23.3216)	(13.9644)
QRM $\boldsymbol{\theta = 0.50}$	**0.5695**	357.7328	396.8138	1576.8871
	(0.2307)	(8.9681)	(47.7433)	(21.5826)
QRM $\boldsymbol{\theta = 0.75}$	0.5493	**175.1298**	**67.9660**	**737.2216**
	(0.2196)	(47.6181)	(82.0807)	(1150.7340)
GBLUP	0.5677	294.5807	376.7794	1583.2355
	(0.2377)	(36.6279)	(24.1379)	(16.2187)
QRH $\boldsymbol{\theta = 0.25}$	**0.4816**	644.1278	551.5150	1962.1296
	(0.0672)	(50.8464)	(64.8091)	(20.9916)
QRH $\boldsymbol{\theta = 0.50}$	0.4797	366.5940	356.9005	1537.7760
	(0.0274)	(56.8604)	(238.5303)	(903.3492)
QRH $\boldsymbol{\theta = 0.75}$	0.3918	**216.1753**	**5.9471**	**-706.1573**

	(0.0544)	(53.2417)	(11.7834)	(2034.7757)
ssGBLUP	0.4712	303.0404	421.8316	1982.3314
	(0.0502)	(37.6933)	(55.2774)	(21.9229)

Cor($\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}$)=correlation between observed and predicted phenotypes, MSE=mean squared error, σ_e^2 or σ_w^2=components of variance associated with the error, DIC=deviation information criterion.

Table 4: Averages of Pearson correlation and standard deviation (in parentheses) between observed phenotypic values ($\boldsymbol{y}$) and predicted phenotypic values ($\widehat{\boldsymbol{y}}$), mean squared error, variance components associated with the error (σ_e^2, σ_w^2) and deviation information criterion (DIC) for yearling weight

Model	**Cor($\boldsymbol{y}, \widehat{\boldsymbol{y}}$)**	**MSE**	**$\boldsymbol{\sigma_e^2}$ or $\boldsymbol{\sigma_w^2}$**	**DIC**
QRM $\boldsymbol{\theta = 0.25}$	**0.5421**	1037.6529	953.6807	1487.1104
	(0.1350)	(175.2648)	(261.8652)	(35.8873)
QRM $\boldsymbol{\theta = 0.50}$	0.5341	868.3651	964.4477	1524.0511
	(0.1355)	(34.0429)	(113.1832)	(12.4648)
QRM $\boldsymbol{\theta = 0.75}$	0.5115	938.8244	**700.7849**	**1284.0829**
	(0.1290)	(241.2205)	(465.2109)	(402.9787)
GBLUP	0.5330	**725.7579**	924.8388	1526.7596
	(0.1389)	(71.3999)	(90.0089)	(11.6346)
QRH $\boldsymbol{\theta = 0.25}$	**0.5306**	1277.9493	1172.2877	1850.7122
	(0.1411)	(44.0948)	(108.7991)	(17.2025)
QRH $\boldsymbol{\theta = 0.50}$	0.5098	894.4148	1061.3157	1883.6773
	(0.1700)	(35.3996)	(129.4702)	(15.4422)
QRH $\boldsymbol{\theta = 0.75}$	0.5027	915.1871	**666.8830**	**1706.4933**
	(0.1748)	(162.7629)	(413.5046)	(209.8455)
ssGBLUP	0.4712	**778.6416**	1071.3096	1891.9029
	(0.0502)	(84.9871)	(128.2878)	(17.5592)

Cor($\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}$)=correlation between observed and predicted phenotypes, MSE=mean squared error, σ_e^2 or σ_w^2=variance components associated with the error, DIC=deviation information criterion.

Simulated data

The results of the simulation exercise where QR is compared with GBLUP under different degrees of asymmetry and proportions of outliers are shown in Table 5. Column 2 records the correlations between the "true" marker effects and the estimated marker effects, the correlations obtained with QR were higher than those obtained with GBLUP. Column 3 shows the correlations between the "true signals" and the estimated ones, the highest correlations were obtained with QR. Column 4 records the estimation of the variance components associated with the error and column 5 the DIC values, the lowest values in both columns were obtained with QR $\theta = 0.75$.

Table 5: Averages of Pearson correlation and standard deviation (in parentheses) between “true” and estimated marker effects, “true” and estimated signals, variance components associated with the error and DIC values for simulated data with different degrees of asymmetry and proportion of outliers

Model	**Cor($\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}$)**	**Cor($\boldsymbol{X\beta}, \boldsymbol{X}\widehat{\boldsymbol{\beta}}$)**	$\boldsymbol{\sigma_e^2}$ **or** $\boldsymbol{\sigma_w^2}$	**DIC**
$\boldsymbol{\rho = 0.95}$, 5% outliers				
QR $\boldsymbol{\theta = 0.25}$	**0.0784**	**0.4963**	0.6821	620.5455
	(0.0034)	(0.0336)	(0.1806)	(49.3305)
QR $\boldsymbol{\theta = 0.50}$	0.0766	0.4643	0.6644	665.8219
	(0.0042)	(0.0493)	(0.0703)	(16.3032)
QR $\boldsymbol{\theta = 0.75}$	0.0606	0.4269	**0.1438**	**290.6870**
	(0.0132)	(0.0386)	(0.1421)	(148.9695)
GBLUP	0.0722	0.4910	0.7375	691.6503
	(0.0064)	(0.0398)	(0.0723)	(19.9391)
$\boldsymbol{\rho = 0.95}$, 10% outliers				
QR $\boldsymbol{\theta = 0.25}$	0.0614	0.4369	0.4683	407.6496
	(0.0183)	(0.0329)	(0.4030)	(330.6304)
QR $\boldsymbol{\theta = 0.50}$	**0.0728**	**0.4579**	0.7947	706.7931
	(0.0045)	(0.0420)	(0.1063)	(20.5797)
QR $\boldsymbol{\theta = 0.75}$	0.0574	0.4061	**0.4482**	**381.4644**
	(0.0092)	(0.0399)	(0.3225)	(474.7138)
GBLUP	0.0654	0.4556	0.8717	731.9104
	(0.0057)	(0.0314)	(0.0890)	(21.8563)
$\boldsymbol{\rho = 0.975}$, 5% outliers				
QR $\boldsymbol{\theta = 0.25}$	**0.0773**	**0.4835**	0.5578	582.4254
	(0.0087)	(0.0562)	(0.2523)	(83.0548)
QR $\boldsymbol{\theta = 0.50}$	0.0771	0.4689	0.6369	662.0337
	(0.0074)	(0.0515)	(0.0868)	(23.8018)
QR $\boldsymbol{\theta = 0.75}$	0.0598	0.4169	**0.2398**	**219.1691**
	(0.0128)	(0.0450)	(0.2033)	(444.5060)
GBLUP	0.0703	0.4804	0.7316	692.6392
	(0.0056)	(0.0333)	(0.0831)	(24.0645)
$\boldsymbol{\rho = 0.975}$, 10% outliers				
QR $\boldsymbol{\theta = 0.25}$	0.0731	0.4386	0.8739	677.0858
	(0.0081)	(0.0789)	(0.1077)	(23.5472)
QR $\boldsymbol{\theta = 0.50}$	**0.0734**	**0.4529**	0.8154	711.2935
	(0.0078)	(0.0615)	(0.0845)	(14.9809)
QR $\boldsymbol{\theta = 0.75}$	0.0541	0.3945	**0.3628**	**385.6030**
	(0.0056)	(0.0583)	(0.2572)	(393.1935)
GBLUP	0.0640	0.4491	0.8913	736.7880
	(0.0077)	(0.0517)	(0.0654)	(14.8343)
$\boldsymbol{\rho = 0.999}$, 5% outliers				
QR $\boldsymbol{\theta = 0.25}$	0.0615	0.5286	0.1535	205.6973

	(0.0144)	(0.0271)	(0.1657)	277.5807
QR $\boldsymbol{\theta = 0.50}$	**0.0741**	**0.5514**	0.4860	614.2282
	(0.0037)	(0.0167)	(0.0663)	15.7647
QR $\boldsymbol{\theta = 0.75}$	0.0467	0.4855	**0.0166**	**-271.4761**
	(0.0112)	(0.0150)	(0.0192)	288.4509
GBLUP	0.0737	0.5428	0.5305	625.9703
	(0.0030)	(0.0121)	(0.0353)	11.3632
$\boldsymbol{\rho = 0.999}$, 10% outliers				
QR $\boldsymbol{\theta = 0.25}$	**0.0768**	**0.4807**	0.7817	650.8593
	(0.0080)	(0.0687)	(0.0888)	22.8417
QR $\boldsymbol{\theta = 0.50}$	0.0696	0.4630	0.6154	511.5287
	(0.0148)	(0.0600)	(0.3369)	412.6645
QR $\boldsymbol{\theta = 0.75}$	0.0507	0.3967	**0.0204**	**-160.1660**
	(0.0031)	(0.0505)	(0.0127)	213.0462
GBLUP	0.0659	0.4649	0.7876	709.7240
	(0.0065)	(0.0418)	(0.0528)	14.8566

Cor($\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}$)=correlation between "true" and estimated marker effects, Cor($\boldsymbol{X\beta}, \boldsymbol{X}\hat{\boldsymbol{\beta}}$)=correlation between "true" and estimated signals, $\boldsymbol{\sigma_e^2}$ or $\boldsymbol{\sigma_w^2}$=variance components associated with the error, DIC=deviation information criterion.

Discussion

In this study, QR analysis methodologies were compared with GBLUP and ssGBLUP. This comparison was made with simulated phenotypes with different degrees of asymmetry and proportions of outliers and actual data for birth, weaning and yearling weights.

Real data

The observed and predicted phenotype correlations obtained from cross-validation with actual data were higher when using QRM and QRH in the WW and YW characteristics. For BW, the highest correlations were obtained with GBLUP and ssGBLUP; however, in this study, only three quantiles 0.25, 0.50 and 0.75 were tested, there is evidence in other studies where QR is better than GBLUP, as in the case of the work of Nascimento *et al*[(4)], who compared QR with models such as BLASSO, BayesB and RR-BLUP. These authors found a 15.15 % gain in the predictive capacity of QR compared to RR-BLUP, it should be noted that, mathematically, RR-is equivalent to GBLUP, in addition to the fact that the datasets used in this experiment presented asymmetry.

The values of the mean squared error (MSE) measure the average of the squared error, that is, the difference between the estimator and what is estimated, so low values are preferred; the MSE averages of QRM and QRH were lower than those obtained with GBLUP and ssGBLUP only for WW. The residual variance estimator is an indication that how well or poorly the model fits the observed data, low values are preferred; the smallest variance components of the error were obtained with QRM and QRH for the three characteristics analyzed. Finally, the DIC value is used to select candidate models and, like MSE and error variance components, low values are preferred. The lowest DIC values were obtained with QRM $\theta = 0.75$ and QRH $\theta = 0.75$, except in the marker-only scenario and BW, where the lowest DIC was obtained with QRM $\theta = 0.25$. The mean squared error, the residual variance and the DIC are values that help to choose the best fit model. When examining these values together, it is observed that QRM and QRH are better in some of them, while in others they are not, that is, QR has a performance equal to or greater than GBLUP and ssGBLUP; although it should be noted that only three quantiles were tested and that QR has advantages when used in asymmetric data and outliers, for this case there were only outliers and the distributions did not present asymmetry. Mendes *et al*[27] compared QR with the Bayesian method of LASSO (BLASSO), these authors reported a 6.7 % and 20.0 % increase in accuracy and considered quantiles 0.15 and 0.45 in the evaluation of carcass yield and bacon thickness, respectively, however, the characteristics evaluated in their study were asymmetric.

In the analysis of real data, a limitation of the present study is the sample size, which can impact the variability of the parameters estimated with the models and consequently the variability of the predictions, however, all the models were fitted using the same information and therefore the comparison of the predictive capacity of the models is considered reasonable, the ideal would be to have large sample sizes, but, due to economic limitations, this is not always possible. On the other hand, it is currently very common to use prediction models in which the number of phenotypic records is smaller than the number of predictors (SNPS), that is $n \ll p$, even in this context, numerous studies have shown that Bayesian methods provide sophisticated tools that allow deriving reasonable predictions as long as the regularization parameters are selected properly, for example using cross-validation methods[28–30].

Simulated data

In the simulated data experiment, the correlations between “true” marker effects and estimated effects as well as correlations of “true” and estimated signals were higher when QR was used compared to GBLUP. These results are similar to those obtained by other researchers[10], who simulated data with three different coefficients of asymmetry 0.75, 0.95, 0.999 with 5 % and 10 % of outliers and found that the correlations obtained with QR were higher than those obtained with Bayesian ridge regression (BRR), in addition, this pattern was more evident with a greater asymmetry and proportion of outliers. In this study, simulations with asymmetry coefficients of 0.950, 0.975, 0.999 were carried out and the quantiles with which higher correlations were obtained varied between 0.25 and 0.50; the advantage of QR is that different quantiles can be tested, obtaining better results depending on the quantile used, this advantage in the ability to predict the effects of markers and signals has been taken advantage of by other researchers[4] , who found no trait association using the traditional GWAS model of single SNP, but, when using QR with extreme quantiles such as 0.1, the model was able to find up to 7 SNPs associated with the characteristics studied.

The coefficients of variance of the error indicate how well the proposed model fits the studied data, the smaller this value, the better the fit, the DIC is another value that is used to compare candidate models. Models with a smaller DIC are preferred to models with a larger DIC[24]. The lowest residual variance estimators and DIC values were obtained with QR $\theta = 0.75$, perhaps this is because high asymmetry coefficients 0.950, 0.975, 0.999 were used in the simulation, so therefore a quantile that fits best is expected to be the highest, in this case 0.75. QR performed equally or better than GBLUP and ssGBLUP to predict growth characteristics BW, WW and YW, the advantages of this method are more noticeable when the data are more biased and present a higher proportion of outliers, as in the case of the simulation experiment.

Conclusions and implications

The predictive performance of QR both with marker information alone and with information of markers plus pedigree, with the actual dataset, was better than the GBLUP and ssGBLUP methodologies for WW and YW. For BW, GBLUP and ssGBLUP were better; however, only quantiles 0.25, 0.50 and 0.75 were used, and the BW distribution was not asymmetric. In the simulated data experiment, correlations between “true” marker effects and estimated effects, as well as correlations of “true” and estimated signals were higher when QR was used compared to GBLUP. The advantages of QR were more noticeable with asymmetric distribution of phenotypes and with a higher proportion of outliers, as was the case with the simulated dataset.

Acknowledgements

To the National Council of Science and Technology, Mexico, for the financial support for the first author during his doctoral studies. The authors also thank the Mexican Association of Breeders of Registered Swiss Cattle for allowing the use of their databases, and the cooperating breeders for their kind cooperation in this study.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Literature cited:

1. Koenker R, Bassett G. Regression quantiles. Econometrica 1978;46(1):33. https://doi.org/10.2307/1913643.
2. Briollais L, Durrieu G. Application of quantile regression to recent genetic and -omic studies. Hum Genet 2014;133(8):951–966. https://doi.org/10.1007/s00439-014-1440-6.
3. Wang L, Wu Y, Li R. Quantile regression for analyzing heterogeneity in ultra-high dimension. J Am Stat Assoc 2012;107(497):214-222. https://doi.org/10.1080/01621459.2012.656014.
4. Nascimento M, Nascimento ACC, Silva FF, Barili LD, Do Vale NM, Carneiro JE, Cruz CD, Carneiro PCS, Serão NVL. Quantile regression for genome-wide association study of flowering time-related traits in common bean. PLoS One 2018;13(1):1-14. https://doi.org/10.1371/journal.pone.0190303.

5. Fisher E, Schweiger R, Rosset S. Efficient construction of test inversion confidence intervals using quantile regression. J Comput Graph Stat 2016;29:140-148, http://arxiv.org/abs/1612.02300.

6. Logan JAR, Petrill SA, Hart SA, Schatschneider C, Thompson LA, Deater-Deckard K, de Thorne LS, Bartlett C. Heritability across the distribution: An application of quantile regression. Behav Genet 2012;42(2):256–267. https://doi.org/10.1007/s10519-011-9497-7.

7. Vinciotti V, Yu K. M-quantile regression analysis of temporal gene expression data. Stat Appl Genet Mol Biol 2009;8(1). https://doi.org/10.2202/1544-6115.1452.

8. Gianola D, Cecchinato A, Naya H, Schön CC. Prediction of complex traits: Robust alternatives to best linear unbiased prediction. Front Genet 2018;9:195. https://doi.org/10.3389/fgene.2018.00195.

9. Oliveira GF, Nascimento ACC, Nascimento M, Sant'Anna IdeC, Romero JV, Azevedo CF, Bhering LL, Moura ETC. Quantile regression in genomic selection for oligogenic traits in autogamous plants: A simulation study. PLoS One 2021;16(1):1-12. https://doi.org/10.1371/journal.pone.0243666.

10. Pérez-Rodríguez P, Montesinos-López OA, Montesinos-López A, Crossa J. Bayesian regularized quantile regression: A robust alternative for genome-based prediction of skewed data. Crop J 2020;8(5):713-722. https://doi.org/10.1016/j.cj.2020.04.009.

11. Nascimento M, e Silva FF, de Resende MDV, Cruz CD, Nascimento ACC, Viana JMS, Azebedo CF, Barroso LMA. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. Genet Mol Res 2017;16(1). https://doi.org/10.4238/GMR16019538.

12. Barroso LMA, Nascimento M, Nascimento ACC, Silva FF, Serão NVL, Cruz, CD, *et al*. Regularized quantile regression for SNP marker estimation of pig growth curves, J. Anim Sci Biotechnol 2017;8:59. https://doi.org/10.1186/s40104-017-0187-z.

13. Nascimento AC, Nascimento M, Azevedo C, Silva F, Barili L, Vale N, *et al*. Quantile regression applied to genome-enabled prediction of traits related to flowering time in the common bean. Agronomy 2019;9(12):1-11. https://doi.org/10.3390/agronomy9120796.

14. López-Cruz M, Crossa J, Bonnett D, Dreisigacker S, Poland J, Jannink JL, Singh RP, Autrique E, de los Campos G. Increased prediction accuracy in wheat breeding trials using a Marker × Environment interaction genomic selection model. G3 Genes Genom Genet 2015;5(4):569–582. https://doi.org/10.1534/g3.114.016097.

15. Pérez-Rodríguez P, Crossa J, Rutkoski J, Poland J, Singh R, Legarra A, *et al*. Single-step genomic and Pedigree Genotype × Environment interaction models for predicting wheat lines in international environments. Plant Genome 2017;10(2). https://doi.org/10.3835/plantgenome2016.09.0089.

16. Christensen O, Lund M. Genomic relationship matrix when some animals are not genotyped. Genet Sel Evol 2010;42(2):1-8. http://www.gsejournal.org/content/42/1/2.

17. Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J Dairy Sci 2010;93(2):743-752. https://doi.org/10.3168/jds.2009-2730.

18. Crossa J, Pérez P, de los Campos G, Mahuku G, Dreisigacker S, Magorokosho C. Genomic selection and prediction in plant breeding. J Crop Improv 2011;25(3):239-261. https://doi.org/10.1080/15427528.2011.558767.

19. Arnold BC, Beaver RJ. Hidden truncation models. Shankhya. Indian J Stat 2000;62(1):23–35. http://www.jstor.org/stable/25051286. Accessed Jul 6, 2022.

20. Montesinos-López A, Montesinos-López OA, Villa-Diharce ER, Gianola D, Crossa J. A robust Bayesian genome-based median regression model. Theor Appl Genet 2019;132(5):1587-1606. https://doi.org/10.1007/s00122-019-03303-6.

21. Pérez-Rodríguez P, Villaseñor-Alva JA. On testing the skew normal hypothesis. J Stat Plan Inference 2010;140(11):3148-3159. https://doi.org/10.1016/j.jspi.2010.04.013.

22. Pérez-Rodríguez P, Acosta-Pech R, Pérez-Elizalde S, Cruz CV, Espinosa JS, Crossa J. A Bayesian genomic regression model with skew normal random errors. G3 Genes|Genom|Genet 2018;8(5):1771–1785. https://doi.org/10.1534/g3.117.300406.

23. Domínguez-Viveros J. Parámetros genéticos en la varianza residual de variables de comportamiento en toros de lidia. Arch Zoot 2020;69(267):354–358. https://doi.org/10.21071/az.v69i267.5354.

24. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc: Series B Stat Methodol 2002;64(4):583–639. https://doi.org/10.1111/1467-9868.00353.

25. R Core Team. R: A language and environment for statistical computing. R Foundation for statistical computing 2021. Vienna, Austria. https://www.R-project.org/.

26. Pérez P, de los Campos G. Genome-wide regression and prediction with the BGLR statistical package. Genetics 2014;198(2):483-495. https://doi.org/10.1534/genetics.114.164442.

27. Mendes dos Santos P, Nascimento ACC, Nascimento M, Fonseca e Silva F, Azevedo CF, Mota RR, *et al*. Use of regularized quantile regression to predict the genetic merit of pigs for asymmetric carcass traits. Pesqui Agropecu Bras 2018;53(9):1011–1017. https://doi.org/10.1590/S0100-204X2018000900004.

28. Gianola D. Priors in whole-genome regression: The Bayesian alphabet returns. Genetics 2013;194(3):573-596. https://doi.org/10.1534/genetics.113.151753.

29. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 2013;193(2):327-345. https://doi.org/10.1534/genetics.112.143313.

30. Ferragina A, de los Campos G, Vazquez AI, Cecchinato A, Bittante G. Bayesian regression models outperform partial least squares methods for predicting milk components and technological properties using infrared spectral data. J Dairy Sci 2015;98(11):8133-8151. https://doi.org/10.3168/jds.2014-9143.