Silvia Columbu*, Paolo Frumento and Matteo Bottai

Modeling sign concordance of quantile regression residuals with multiple outcomes

https://doi.org/10.1515/ijb-2022-0020
Received December 17, 2020; accepted June 15, 2022; published online July 11, 2022

Abstract: Quantile regression permits describing how quantiles of a scalar response variable depend on a set of predictors. Because a unique definition of multivariate quantiles is lacking, extending quantile regression to multivariate responses is somewhat complicated. In this paper, we describe a simple approach based on a two-step procedure: in the first step, quantile regression is applied to each response separately; in the second step, the joint distribution of the signs of the residuals is modeled through multinomial regression. The described approach does not require a multidimensional definition of quantiles, and can be used to capture important features of a multivariate response and assess the effects of covariates on the correlation structure. We apply the proposed method to analyze two different datasets.

Keywords: conditional correlation; multinomial model; multiple quantiles; multivariate regression; sign-concordance.

1 Introduction

The study of association between multiple outcomes is common in the medical literature. For instance, clinical trials often have multiple primary endpoints, and limiting the analysis to one single endpoint may be considered undesirable. This problem is particularly important when the trialed treatment is suspected to have potentially different effects on different endpoints. For example, an anticoagulant treatment after a primary stroke may reduce the risk of a second stroke and at the same time increase the risk of additional bleeding [1].

When there is an interest in exploring the effect of covariates locally on specific quantiles of a univariate response, quantile regression (e.g., [2]) can be used. However, if multivariate outcomes need to be considered extending quantile regression is not straightforward, as no natural definition of multivariate quantile is available.

The analysis of multivariate quantiles has been discussed in the existing literature. An excellent review is provided in [3]. Most of the proposed methods use geometrical definitions of multidimensional quantiles that are mainly based on vector-valued ranks, and use the orientation information to identify directional quantiles in a multidimensional data cloud [4–13]. Other scholars proposed joint modeling of quantiles in a likelihood framework, avoiding a mathematical definition of multivariate quantile [14–16]. Bayesian methods have also been described in the literature [17–19].

In this work, we propose a method that does not require a definition of multivariate quantile. The joint distribution, $F_{Y^{(1)}Y^{(2)},\dots,Y^{(d)}}(y^{(1)}, y^{(2)}, \dots, y^{(d)})$, of a d-dimensional response, is not modeled parametrically and is estimated locally at values that correspond to univariate conditional quantiles. The method is implemented in two steps: first, quantile regression is applied to each response separately; then, a suitable regression model is used to investigate the joint distribution of the signs of the residuals and their conditional association. The

***Corresponding author: Silvia Columbu**, University of Cagliari, Cagliari, Italy, E-mail: silvia.columbu@unica.it
Paolo Frumento, University of Pisa, Pisa, Italy
Matteo Bottai, Karolinska Institute, Solna, Stockholm, Sweden

method aims to explore the local association between multiple outcomes by investigating the joint behavior of their conditional quantiles. This approach bears some similarities with that described in [20], in which copulas are used to model the bivariate quantile-specific conditional distribution.

Our proposal requires formulating a multivariate binary response model in which the binary outcomes are defined by the sign of quantile regression residuals. In the literature, the problem of analyzing correlated binary outcomes has been tackled in different ways, which may be broadly grouped into three categories. The first category comprises methods based on generalized estimating equations that do not require specifying the joint distribution of the multivariate response [21–24]. The second category consists of generalized linear mixed models [25–28]. The third and most recent category includes methods that use copulas to model the multivariate association [29–32].

In this paper, we suggest modeling the full joint distribution of the binary responses by fitting a multinomial logistic regression model. This approach does not require strong simplifying assumptions, but may become unfeasible if the response vector is high-dimensional. In many real-data settings, however, the number of responses is two or three.

The proposed method is illustrated through two different applications. The first one considers a dataset on lung function capacity with correlated spirometrics outcome measures. The second application refers to the National Merit Twins Study, and allows for a comparison with the quantile association method proposed by [20].

The paper is structured as follows. In Section 2 we describe the method and discuss how to measure the correlation between the signs of quantile regression residuals. In Section 3 we present two different simulation studies to illustrate the finite-sample performance of the proposed method. Sections 4 and 5 describe the applications.

2 The proposed model

2.1 Sign-concordance of quantile regression residuals

We denote by $\boldsymbol{x}$ a q-dimensional vector of observed covariates, and by $(Y^{(1)}, Y^{(2)})$ a pair of continuous response variables. Following standard quantile regression notation [33], we assume the following univariate, quantile-specific linear model to hold for each response:

$$Y_i^{(j)} = \mathbf{x}_i^T \boldsymbol{\beta}_\tau^{(j)} + \boldsymbol{\varepsilon}_i^{(j)} \quad j = \{1, 2\}, \quad i = \{1, \dots, n\} \tag{1}$$

with $P\left(\varepsilon_i^{(j)} \leq 0 \mid \mathbf{x}_i\right) = \tau$. In this model, $Q_{y^{(j)}}\left(\tau \mid \mathbf{x}_i^T\right) = \mathbf{x}_i^T \boldsymbol{\beta}_\tau^{(j)}$ represents the τth conditional quantile of the jth response, and $\boldsymbol{\beta}_\tau^{(j)}$ is a vector of model coefficients, $\tau \in (0, 1)$.

An estimate of the unknown quantile regression coefficients, $\boldsymbol{\beta}_\tau^{(1)}$ and $\boldsymbol{\beta}_\tau^{(2)}$, is obtained by minimizing

$$\sum_{i=1}^{n} \rho_\tau(y_i^{(j)} - \mathbf{x}_j^T \boldsymbol{\beta}_\tau^{(j)}), \quad j = \{1, 2\}, \tag{2}$$

where $y_i^{(j)}$ is a realization from $Y_i^{(j)}$, and $\rho_\tau(u) = (I(u \leq 0) - \tau)u$ is a loss function. We denote by $\hat{\boldsymbol{\beta}}_\tau^{(j)}$ the estimated regression coefficients, and by $\hat{\varepsilon}_i^{(j)} = y_i^{(j)} - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\tau^{(j)}$ the corresponding quantile-specific regression residuals.

We define two binary indicators, namely $\omega_i^{(1)} = I(Y_i^{(1)} \leq \mathbf{x}_i^T \boldsymbol{\beta}_\tau^{(1)})$ and $\omega_i^{(2)} = I(Y_i^{(2)} \leq \mathbf{x}_i^T \boldsymbol{\beta}_\tau^{(2)})$, such that 0 and 1 indicate positive and negative residuals, respectively. We introduce the following random variable:

$$Z_i = \begin{cases} 1 & \text{if } \omega_i^{(1)} = 0 \text{ and } \omega_i^{(2)} = 0, \\ 2 & \text{if } \omega_i^{(1)} = 1 \text{ and } \omega_i^{(2)} = 1, \\ 3 & \text{if } \omega_i^{(1)} = 0 \text{ and } \omega_i^{(2)} = 1, \\ 4 & \text{if } \omega_i^{(1)} = 1 \text{ and } \omega_i^{(2)} = 0. \end{cases} \tag{3}$$

For a more intuitive notation, in the rest of the manuscript we will associate the labels {"00", "11", "01", "10"} to the values $\{1, 2, 3, 4\}$ that form the support of Z. The variable Z_i carries information on the concordance of the signs of the residuals from the two regression equations defined in Eq. (1). Discordance between signs indicates negative dependence between $Y^{(1)}$ and $Y^{(2)}$, given $\boldsymbol{x}$. Similarly, concordance between signs suggests a positive correlation.

Given the estimates $\hat{\boldsymbol{\beta}}_\tau^{(j)}$ and $\hat{\varepsilon}_i^{(j)}$ we denote by $\hat{\omega}_i^{(j)} = I(y_i^{(j)} \leq \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\tau^{(j)}) = I(\hat{\varepsilon}_i^{(j)} \leq 0)$ the estimated binary indicators of negative residuals. The observed values of z_i are defined by the four possible combinations of $\hat{\omega}_i^{(1)}$, which reflects the sign of $\hat{\varepsilon}_i^{(1)}$, and $\hat{\omega}_i^{(2)}$, which reflects the sign of $\hat{\varepsilon}_i^{(2)}$. The sign concordance can be summarized by a measure of correlation, i.e., the sample counterpart of the following population parameter:

$$\phi = \operatorname{cor}\left(\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}\right) = \frac{E\left[\boldsymbol{\omega}^{(1)}\boldsymbol{\omega}^{(2)}\right] - E\left[\boldsymbol{\omega}^{(1)}\right] E\left[\boldsymbol{\omega}^{(2)}\right]}{\sqrt{E\left[\left(\boldsymbol{\omega}^{(1)} - E\left[\boldsymbol{\omega}^{(1)}\right]\right)^2\right] E\left[\left(\boldsymbol{\omega}^{(2)} - E\left[\boldsymbol{\omega}^{(2)}\right]\right)^2\right]}}. \tag{4}$$

The equation in formula (Eq. (4)) can be rewritten as

$$\phi = \frac{F_{Y^{(1)}Y^{(2)}}(Q_{y^{(1)}}(\tau \mid \mathbf{x}), Q_{y^{(2)}}(\tau \mid \mathbf{x})) - \tau^2}{\tau(1-\tau)}, \tag{5}$$

where $F_{Y^{(1)}Y^{(2)}}(Q_{y^{(1)}}(\tau \mid \mathbf{x}), Q_{y^{(2)}}(\tau \mid \mathbf{x}))$ is the joint, unconditional distribution function of the responses, evaluated at the conditional quantiles. Note that, by definition, $E[\boldsymbol{\omega}^{(j)}] = P(\boldsymbol{\varepsilon}^{(j)} \leq 0 \mid \mathbf{x}) = \tau$. This holds approximately true when the sample counterparts of $\boldsymbol{\omega}^{(j)}$ are used (33, Theorem 3.4). The joint distribution of the quantile regression residuals signs is illustrated in Table 1, where we used the notation $p_z = P(Z = z)$, $z \in \{$"00", "11", "01", "10"$\}$.

Following Table 1, the value of the ϕ statistic in Eq. (4) can be also expressed as:

$$\phi = \frac{p_{11}p_{00} - p_{01}p_{10}}{\sqrt{p_{0*} \times p_{1*} \times p_{*0} \times p_{*1}}} \tag{6}$$

with $p_{0^*} = p_{*0} = 1 - \tau$ and $p_{1^*} = p_{*1} = \tau$.

There is no direct link between ϕ and the global correlation structure: our method is essentially nonparametric and can be seen as an approach to assess correlation locally, showing how ϕ varies across quantiles and how it depends on covariates. The bounds of the ϕ-coefficient can be obtained by calculating its value in

Table 1: Contingency table showing the joint distribution of the signs of quantile regressions residuals. By definition, the margins are given by $P(Y^{(j)} > \mathbf{x}^T \boldsymbol{\beta}_\tau^{(j)}) = 1 - \tau$ and $P(Y^{(j)} \leq \mathbf{x}^T \boldsymbol{\beta}_\tau^{(j)}) = \tau$, $j = \{1, 2\}$.

		$\operatorname{sign}\left(\boldsymbol{Y^{(2)}} - \mathbf{x}^T \boldsymbol{\beta}_\tau^{(2)}\right)$		
		+	−	
$\operatorname{sign}\left(Y^{(1)} - \mathbf{x}^T \boldsymbol{\beta}_\tau^{(1)}\right)$	+	p_{00}	p_{01}	$p_{0^*} = 1 - \tau$
	−	p_{10}	p_{11}	$p_{1^*} = \tau$
		$p_{*0} = 1 - \tau$	$p_{*1} = \tau$	

Table 2: Joint distribution (relative frequencies) of the signs of quantile regression residuals in case of independence, perfect positive dependence (Max) and perfect negative dependence (Min) between the two outcomes.

			$\text{sign}\left(Y^{(2)} - \mathbf{x}^T\beta_\tau^{(2)}\right)$		
			+	−	
$\text{sign}\left(Y^{(1)} - \mathbf{x}^T\beta_\tau^{(1)}\right)$	+	Independence	$(1-\tau)^2$	$\tau - \tau^2$	$1-\tau$
		Max	$1-\tau$	0	
		Min ($\tau \le 0.5$)	$1-2\tau$	τ	
		Min ($\tau \ge 0.5$)	0	$1-\tau$	
	−	Independence	$\tau - \tau^2$	τ^2	τ
		Max	0	τ	
		Min ($\tau \le 0.5$)	τ	0	
		Min ($\tau \ge 0.5$)	$1-\tau$	$2\tau - 1$	
			$1-\tau$	τ	

the three limiting situations summarized in Table 2. In the case of independence, the joint distribution of the residuals corresponds to the product of the two marginal probabilities. Under perfect positive dependence, the joint distribution is derived assuming that there are no observations generating discordant residuals. The perfect negative dependence requires some additional reasoning. Under this scenario, one may expect to have no observations generating concordant residuals. However, as a consequence of the asymmetric structure of quantiles, the cells on the principal diagonal of Table 1 can never be simultaneously empty, unless $\tau = 0.5$. We must therefore separate the cases above and below the median and allow for a small proportion of observations in the cell corresponding to negative concordance for $\tau > 0.50$, and in the cell of positive concordance when $\tau < 0.50$.

By applying Eq. (6) to Table 2 we have:

- $\phi_{\text{Indep}} = 0$, independence;
- $\phi_{\text{Max}} = 1$, largest possible positive dependence;
- $\phi_{\text{Min}} = \left\{ \begin{array}{ll} -\tau/(1-\tau) & \tau \le 0.50 \\ -(1-\tau)/\tau & \tau \ge 0.50 \end{array} \right\}$, largest possible negative dependence.

The above theoretical bounds for the ϕ statistic depend only on the quantile being estimated, and not on the data. The largest possible value of ϕ is the same as that of the Pearson's correlation coefficient. Instead, the lower limit of the coefficient is greater than −1, unless $\tau = 0.5$.

2.2 Modeling the conditional correlation

The correlation coefficient ϕ, that describes the association between the signs of the residuals of univariate quantile regression, is usually a function of the predictors. In particular, while the *margins* of Table 1 are always equal to τ and $1-\tau$, the *joint* distribution of $\omega_i^{(1)} = I(Y_i^{(1)} \le \mathbf{x}_i^T\boldsymbol{\beta}_\tau^{(1)})$ and $\omega_i^{(2)} = I(Y_i^{(2)} \le \mathbf{x}_i^T\boldsymbol{\beta}_\tau^{(2)})$ can depend on covariates. For example, the correlation between outcomes could be larger in smokers than in non-smokers. Note that the factors influencing this association may not coincide with those used in the univariate quantile regressions.

In this paper we suggest using a multinomial logistic regression to model the distribution of Z:

$$\log\left(\frac{P\left(Z_i = z|\mathbf{x}_i^T\right)}{P\left(Z_i = \text{“00”}|\mathbf{x}_i^T\right)}\right) = \mathbf{x}_i^T\boldsymbol{\gamma}_{z:\tau}, \quad i = \{1,\ldots,n\}, \quad z = \{\text{“11”},\text{“01”},\text{“10”}\}. \tag{7}$$

where $\boldsymbol{\gamma}_{z:\tau}$ is the vector of parameters of the regressive model, which will depend both on the quantile considered in the quantile regression models at first step and on the z.

The predicted probabilities, $\{p_{00}(\mathbf{x}), p_{11}(\mathbf{x}), p_{01}(\mathbf{x}), p_{10}(\mathbf{x})\}$, are then combined (Eq. (6)) to compute an estimate $\hat{\phi}(\mathbf{x})$ of the conditional correlation coefficient. By plotting $\mathbf{x}$ versus $\hat{\phi}(\mathbf{x})$, it is possible to show how the correlation structure depends on covariates. The estimates can be compared with the three limit values shown in Section 2.1. Note, however, that such limits may be surpassed at some values of $\mathbf{x}$, due to a combination of finite-sample variability and model misspecification.

The proposed estimator is implemented in two steps, where the first step requires estimating a quantile regression model on each response separately, and the second step consists of a multinomial regression model applied to the joint distribution of the signs of the residuals. Both estimators are supported by most standard software. The variability associated with the first-step estimation of quantile regression coefficients, $\hat{\boldsymbol{\beta}}_\tau^{(1)}$ and $\hat{\boldsymbol{\beta}}_\tau^{(2)}$, must be taken into account to evaluate correctly the variance of the second-step estimator $\hat{\boldsymbol{\gamma}}_{z:\tau}$. In principle, one could use well-known results on two-step estimators [34, 35]. However, bootstrap is commonly used for inference on quantile regression, and represents a very convenient approach in the current framework.

3 Simulation study

We conducted a simulation study to illustrate the finite-sample performance of the proposed method. We considered a simplified problem in which the second-step response was the binary indicator $Z_i = \omega_i^{(1)} \wedge \omega_i^{(2)} = I(Y_i^{(1)} \leq \mathbf{x}_i^T \boldsymbol{\beta}_\tau^{(1)} \wedge Y_i^{(2)} \leq \mathbf{x}_i^T \boldsymbol{\beta}_\tau^{(2)})$, and we applied logistic regression to evaluate $P(Z = 1|\mathbf{x})$:

$$\log\left(\frac{P(Z_i = 1|\mathbf{x})}{1 - P(Z_i = 1|\mathbf{x})}\right) = \mathbf{x}_i^T \boldsymbol{\gamma}_\tau, \quad i = \{1, \ldots, n\}. \tag{8}$$

Two different scenarios were considered. In scenario 1, we directly controlled the dependence structure of the responses, and generated $Y^{(1)}$ and $Y^{(2)}$ from a bivariate normal heteroskedastic model, $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in which parameters were linear functions of a continuous covariate $x \sim \mathrm{U}(0, 1)$ such that $\boldsymbol{\mu} = \{\mu_1,\ \mu_2\} = \{(3 + 2x), (2 - 3x)\}$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 + x & 2(x - 0.5) \\ 2(x - 0.5) & 1 - 0.5x \end{pmatrix}.$$

In this scenario, the second-step logistic model is not the true model, i.e., the log-odds ratio is not a linear function of x. This generates a small bias that can be eliminated by using a sufficiently flexible parametric model, i.e., by introducing x in the regression equation by means of polynomials or splines. In Table 3, we consider three different values of the predictor, $x = \{0.1, 0.5, 0.9\}$ and, for three quantiles $\tau = \{0.25, 0.5, 0.75\}$, we report: the true probability of negative concordance, $P(Z = 1|x)$; its average estimate from a logistic model in which x was included linearly; and the estimates obtained when x was included in the regression equation by means of natural cubic splines with 2 internal knots at the empirical tertiles. For each estimator, we also report the empirical standard errors.

In scenario 2, we directly generated the binary indicator Z as follows:

$$P(Z = 1|x) = \frac{\mathrm{e}^{bx}}{1 + \mathrm{e}^{bx}} \tag{9}$$

where $x \sim \mathrm{U}(0.5, 1.5)$. We fixed the quantile regression equations, $Q_{y^{(1)}}\left(\tau \mid \mathbf{x}_i^T\right) = 1 + 2x$ and $Q_{y^{(2)}}\left(\tau \mid \mathbf{x}_i^T\right) = 3 + 4x$. To simulate data, we first generated two random variables $E_i^{(1)} \sim \mathrm{Exp}(2)$ and $E_i^{(2)} \sim \mathrm{Exp}(1/2)$, $i = 1, \ldots, n$. If $Z_i = 1$, which implies $\omega_i^{(1)} = \omega_i^{(2)} = 1$, we defined $Y_i^{(j)} = Q_{y^{(j)}}\left(\tau \mid \mathbf{x}_i^T\right) - E^{(j)}$. When $Z_i = 0$, we used the probabilities reported in Table 4 to assign the values of $\omega_i^{(1)}$ and $\omega_i^{(2)}$. In case of positive concordance, that is $\omega_i^{(1)} = \omega_i^{(2)} = 0$, we defined $Y_i^{(j)} = Q_{y^{(j)}}\left(\tau \mid \mathbf{x}_i^T\right) + E_i^{(j)}$; in case of discordance, if $\omega_i^{(1)} = 1$ and $\omega_i^{(2)} = 0$ then $Y^{(1)} = Q_{y^{(1)}}\left(\tau \mid \mathbf{x}_i^T\right) - E_i^{(1)}$ and $Y_i^{(2)} = Q_{y^{(2)}}\left(\tau \mid \mathbf{x}_i^T\right) + E_i^{(2)}$; otherwise, $Y_i^{(1)} = Q_{y^{(1)}}\left(\tau \mid \mathbf{x}_i^T\right) + E_i^{(1)}$ and $Y_i^{(2)} = Q_{y^{(2)}}\left(\tau \mid \mathbf{x}_i^T\right) - E_i^{(2)}$. To generate plausible data, a different value of b (Eq. (9)) was used for each value of τ.

Table 3: Results of simulation 1. Data were generated from a bivariate normal heteroskedastic model. At different quantiles and different values of x, we report the true probabilities of negative concordance, $P(Z = 1|x)$, and their average estimates and standard errors under two different specifications of the second-step logistic regression model: either a linear specification, in which x was included linearly, or a spline-based model, in which x was included by means of natural cubic splines with 2 internal knots at the empirical tertiles.

			Linear			Splines		
			$x = 0.1$	$x = 0.5$	$x = 0.9$	$x = 0.1$	$x = 0.5$	$x = 0.9$
$\tau = 0.25$		True	0.0021	0.0625	0.1656	0.0021	0.0625	0.1656
	$n = 300$	Estimated	0.0140	0.0509	0.1808	0.0044	0.0722	0.1653
		se	0.0071	0.0120	0.0236	0.0055	0.0292	0.0281
	$n = 600$	Estimated	0.0136	0.0508	0.1787	0.0040	0.0702	0.1653
		se	0.0049	0.0087	0.0167	0.0040	0.0186	0.0195
$\tau = 0.5$		True	0.1069	0.2500	0.3930	0.1069	0.2500	0.3930
	$n = 300$	Estimated	0.1217	0.2352	0.4076	0.1002	0.2557	0.3975
		se	0.0179	0.0152	0.0223	0.0329	0.0345	0.0355
	$n = 600$	Estimated	0.1209	0.2353	0.4088	0.1000	0.2555	0.3986
		se	0.0121	0.0103	0.0160	0.0234	0.0253	0.0255
$\tau = 0.75$		True	0.5021	0.56254	0.6656	0.5021	0.5625	0.6656
	$n = 300$	Estimated	0.4967	0.5767	0.6528	0.5128	0.5577	0.6722
		se	0.013	0.0101	0.0184	0.0404	0.0364	0.0344
	$n = 600$	Estimated	0.4936	0.5764	0.6551	0.5103	0.5573	0.6733
		se	0.0097	0.0069	0.0129	0.0301	0.0271	0.0240

Table 4: Contingency table reporting the joint distribution of the signs of quantile regressions residuals for a fixed probability of negative concordance $P(Z = 1 \mid x)$ with respect to the marginal distribution in each response.

		$\text{sign}\left(\mathbf{Y^{(2)}} - \mathbf{x}^T \boldsymbol{\beta}_\tau^{(2)}\right)$		
		+	−	
$\text{sign}\left(Y^{(1)} - \mathbf{x}^T \boldsymbol{\beta}_\tau^{(1)}\right)$	+	$1 - 2\tau + P(Z = 1\|x)$	$\tau - P(Z = 1\|x)$	$1 - \tau$
	−	$\tau - P(Z = 1\|x)$	$P(Z = 1\|x)$	τ
		$1 - \tau$	τ	

In this scenario, both the step-1 and the step-2 model are correct. Results are summarized in Table 5, which reports the same information as Table 3 for three different quantiles, $\tau = \{0.25, 0.5, 0.75\}$ and three different values of the predictor, $x = \{0.5, 1, 1.5\}$.

The bias was always very small, even in simulation 1 where the logit-linear model was misspecified. Moreover, in both simulations showed the estimator had a reliable finite-sample performance with relatively small standard errors.

4 Dependence between lung function measures

Spirometric indexes are used to assess lung function impairment [36], and the diagnosis of many pulmonary diseases is based on comparing observed measures with the tails of the distribution in the healthy population [37]. Analyzing and interpreting percentiles of spirometric indexes allows identifying risk factors of respiratory impairment, diagnosing pulmonary diseases, and selecting appropriate treatments [38].

Table 5: Results of simulation 2 (see text for details). At different quantiles and different values of x, we report the true probabilities of negative concordance, $P(Z = 1|x)$, and their average estimates and standard errors.

				$x = 0.75$	$x = 1$	$x = 1.25$
$b = -2.3$	$\tau = 0.25$		True	0.0534	0.0308	0.0175
		$n = 300$	Estimated	0.0570	0.0365	0.0236
			se	0.0143	0.0135	0.0118
		$n = 600$	Estimated	0.0556	0.0344	0.0213
			se	0.0102	0.0094	0.0078
$b = -1.15$	$\tau = 0.5$	$n = 300$	True	0.2968	0.2405	0.1919
			Estimated	0.2930	0.2443	0.2018
			se	0.0183	0.0148	0.0193
		$n = 600$	Estimated	0.2939	0.2430	0.1986
			se	0.0136	0.0110	0.0141
$b = 0.4$	$\tau = 0.75$		True	0.5744	0.5987	0.6225
		$n = 300$	Estimated	0.5780	0.5976	0.6168
			se	0.0152	0.0118	0.0162
		$n = 600$	Estimated	0.5756	0.5971	0.6182
			se	0.0105	0.0082	0.0113

We investigated the effect of a variety of predictors on two important spirometric indexes: forced vital capacity (FVC, in liters), and forced expiratory volume in 1 s (FEV1, also in liters). FVC is a measure of the volume change in the lung between a full inspiration to total lung capacity, and a maximal expiration to residual volume. FEV1 represents the volume exhaled during the first second of a forced expiratory maneuver started from the level of total lung capacity. We used data from 945 individuals from the Po river delta study [39], a prospective study conducted to investigate obstructive pulmonary diseases in the general population of a rural area in northern Italy. The patients' age ranged 18 to 64. We only analyzed males, which represented about forty-nine percent (466 subjects) of the entire dataset. We considered four covariates: height (cm, centered at its sample mean), age (years, centered at its sample mean), an indicator of comorbidities such as asthma, cough or wheeze, and an indicator of smoking (0 = never smoker, 1 = ever smoker). All considered predictors are known relevant determinants of lung function [36, 37].

Exploratory analyses were performed to assess the relationship between the responses considered. Figure 1 suggested a strong positive association between FVC and FEV1 (Spearman's correlation = 0.878).

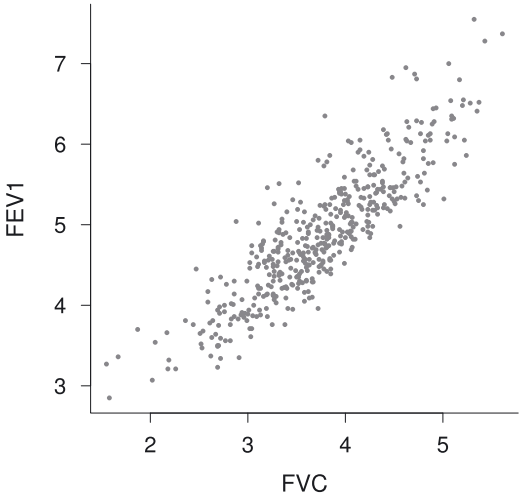

Figure 1: Scatterplot of the sample values of FEV1 (liters) against FVC (liters).

We estimated the following quantile regression models for $\tau = \{0.1, 0.5, 0.9\}$:

$$Q_{\text{FVC}}(\tau) = \beta^{(1)}_{\tau,0} + \beta^{(1)}_{\tau,1}(\text{age} - 37) + \beta^{(1)}_{\tau,2}(\text{height} - 172) + \beta^{(1)}_{\tau,3}\text{comorb.} + \beta^{(1)}_{\tau,4}\text{smoke},$$

$$Q_{\text{FEV1}}(\tau) = \beta^{(2)}_{\tau,0} + \beta^{(2)}_{\tau,1}(\text{age} - 37) + \beta^{(2)}_{\tau,2}(\text{height} - 172) + \beta^{(2)}_{\tau,3}\text{comorb.} + \beta^{(2)}_{\tau,4}\text{smoke}.$$

The estimated quantile regression coefficients are reported in Table 6. Standard errors were obtained from 1000 tilted bootstrap replicates [40, 41]. Results suggested that age and height were two important predictors of lung function.

Using the estimated quantile regression residuals, we calculated the concordance indicator z_i defined in Eq. (3), and modeled its conditional distribution using a multinomial regression:

$$\log\left(\frac{P(Z = z)}{P(Z = \text{“00”})}\right) = \gamma_{z:\tau,0} + \gamma_{z:\tau,1}s(\text{age}) + \gamma_{z:\tau,2}s(\text{height}) + \gamma_{z:\tau,3}\text{comorb.} + \gamma_{z:\tau,4}\text{smoke}, \tag{10}$$

$z = \{\text{“11”}, \text{“01”}, \text{“10”}\}$. In the regression equation, $s(x)$ denotes the basis of a natural cubic splines with two internal knots at the empirical tertiles. Using splines allows achieving any desired flexibility, but makes it difficult to interpret the γ model coefficients. Note, however, that the second-step multinomial regression is only used for prediction purposes. We remark that, in general, it is possible to use different sets of covariates in the first- and second-step model.

The conditional correlation $\hat{\phi}(\mathbf{x})$ was computed by applying Eq. (6) to the fitted probabilities from model (Eq. (10)). In Figures 2–4, we represent how correlation depends on covariates at different values of τ. To compute confidence intervals, we obtained bootstrap standard errors of logit $\left(\hat{\phi}(\mathbf{x})\right)$. In doing so the univariate quantile models of first step have been re-evaluated on each bootstrap replicate. The values corresponding to ϕ_{Indep}, ϕ_{Min}, and ϕ_{Max} are shown as horizontal lines.

Results showed a consistently positive correlation between FVC and FEV1. At $\tau = 0.1$, the estimated ϕ coefficients was considerable (close to 0.5), suggesting that patients are more often below or above the lower limit of normality with respect to both spirometric measurements. Interestingly, the correlation was slightly smaller in presence of comorbidities. This could be explained by the fact that some comorbidities may only affect one of the two response variables of interest, breaking the existing correlation. At $\tau = 0.5$, correlations were generally large, and unaffected by predictors. At $\tau = 0.9$, the estimated ϕ coefficient was typically around 0.5, but approached zero at young ages (<25 years), and in tall patients (>180 cm). This

Table 6: Estimated quantile regression coefficients with response FVC (top table) and FEV1 (bottom table).

	$\tau = 0.10$		$\tau = 0.50$		$\tau = 0.90$	
	Coefficient	**SE**	**Coefficient**	**SE**	**Coefficient**	**SE**
FVC						
Intercept	4.227	0.153*	4.977	0.118*	5.816	0.14*
Age−37	−0.019	0.004*	−0.024	0.003*	−0.022	0.004
Height−172	0.058	0.008*	0.056	0.005*	0.065	0.008*
Comorbidity	0.055	0.111	0.112	0.088	0.162	0.129
Ever smoker	0.037	0.164	−0.105	0.129	−0.259	0.149
FEV1						
Intercept	3.168	0.123*	3.961	0.080*	4.413	0.073*
Age−37	−0.027	0.003*	−0.027	0.003*	−0.029	0.003*
Height−172	0.045	0.006*	0.043	0.004*	0.043	0.006*
Comorbidity	−0.188	0.103	−0.029	0.073	0.044	0.106
Ever smoker	0.059	0.128	−0.183	0.088*	−0.113	0.085

The asterisk (*) indicates p-values less than 0.05.

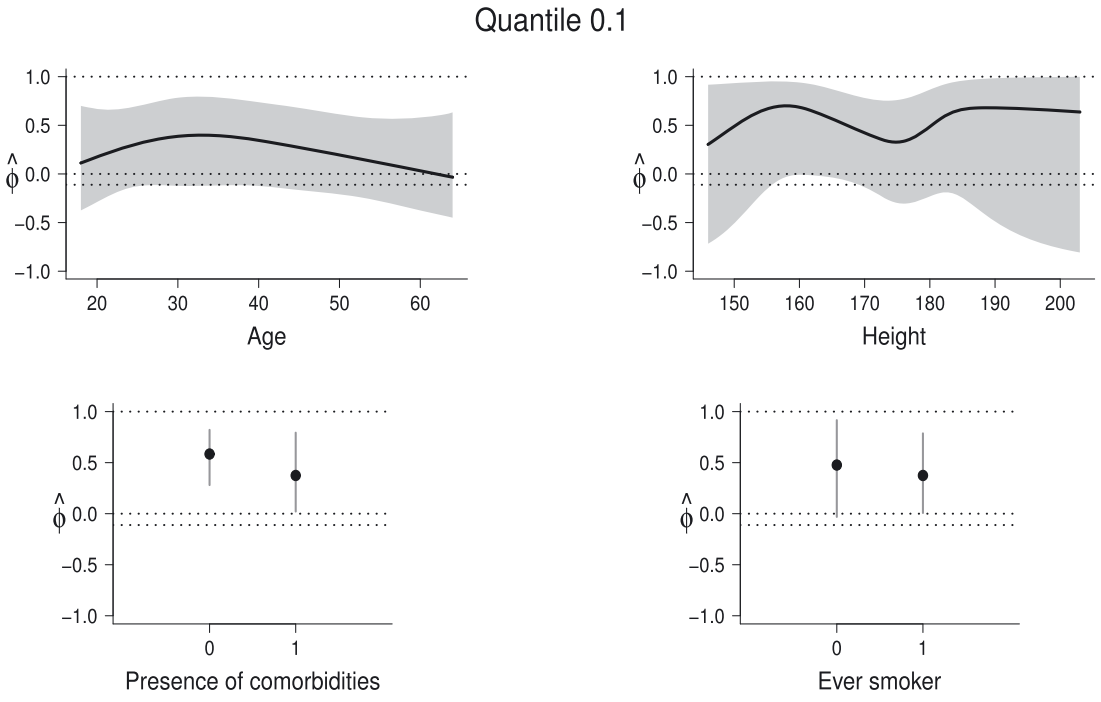

Figure 2: Predicted correlation at the 10th percentile, expressed as a function of the predictors. The horizontal lines indicate, from top to bottom, perfect positive correlation (ϕ_{Max}), independence (ϕ_{Indep}), and perfect negative correlation (ϕ_{Min}), respectively.

could be explained by the fact that particularly large values of both FVC and FEV1 indicate good health, without underlying pathological conditions that may induce correlation.

5 National Merit Twin Study

In this section we consider an application to the National Merit Twin Study, that was previously analyzed, among others, by [42] and [20]. Extensive questionnaires were administered to 839 adolescent twins identified among the roughly 600,000 US high school juniors who took the national merit scholarship qualifying test (NMSQT) in 1962. The dataset is available in the `mdhglm` R package [43], and includes 768 pairs of same-gender twins. The twins were classified as identical or fraternal based on a mail-in questionnaire. The NMSQT consists of five subtests, covering the domains of English, mathematics, social science, natural science, and vocabulary. A total score is calculated as the sum of the scores obtained from the five subtests. In our analysis, the bivariate outcome $(Y^{(1)}, Y^{(2)})$ is given by the total NMSQT scores of the twin pair. We considered the following binary covariates: Sex (0 = male, 1 = female), Income (an indicator of family income level being above 10,000 US dollars), Education (an indicator of whether at least one of the parents had education beyond high school), and Zygosity (1 for identical twins, and zero otherwise). The aim of the analysis was to study the association between twins in terms of academic abilities, conditional on the observed factors.

We first estimated two univariate quantile regression models ($\tau = 0.01, 0.02, \ldots, 0.99$) to describe the effect of covariates on the NMSQT scores $(Y^{(1)}, Y^{(2)})$. The regression equation also included an interaction term Zigosity × Income. Results showed that male students coming from families with high income and high parental education tend to score higher. The interaction was generally significant, suggesting that, in whealthier families, identical twins tend to perform better than heterozygous twins.

We then calculated the concordance indicator z_i defined in Eq. (3), and estimated the multinomial model of concordance. Following [20], we restricted the analysis to quantiles in the range [0.2, 0.8]. Because NMSQT scores in each twins' pair can be considered exchangeable (as the twins cannot be ordered), the discordant

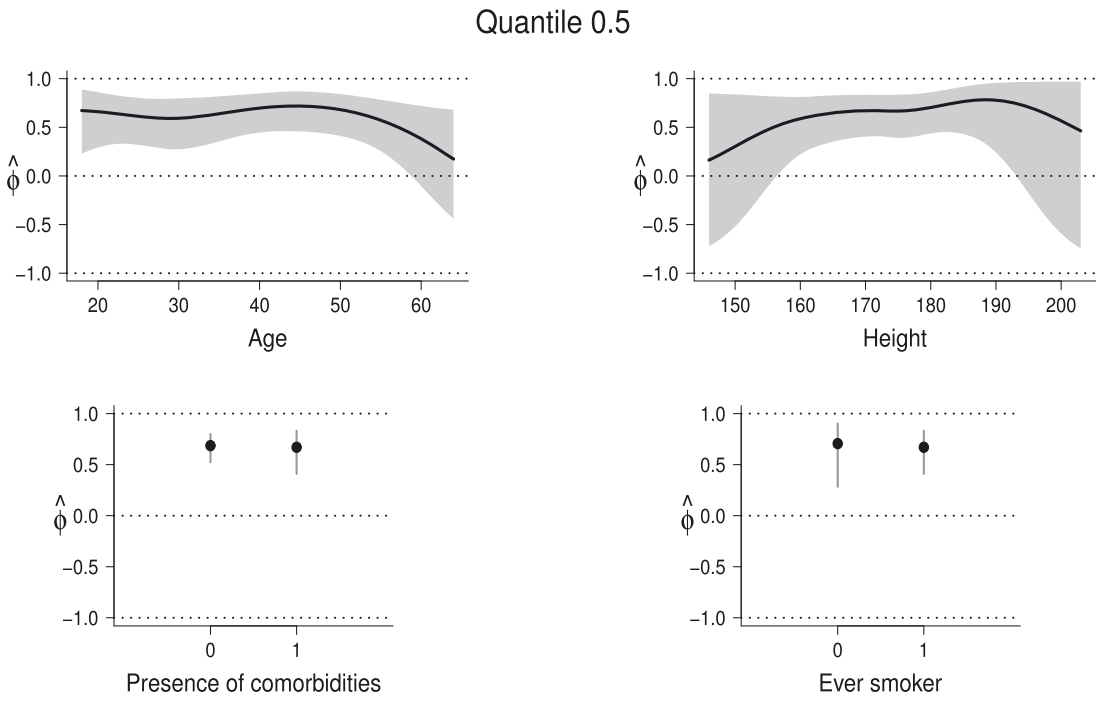

Figure 3: Predicted correlation at the median.

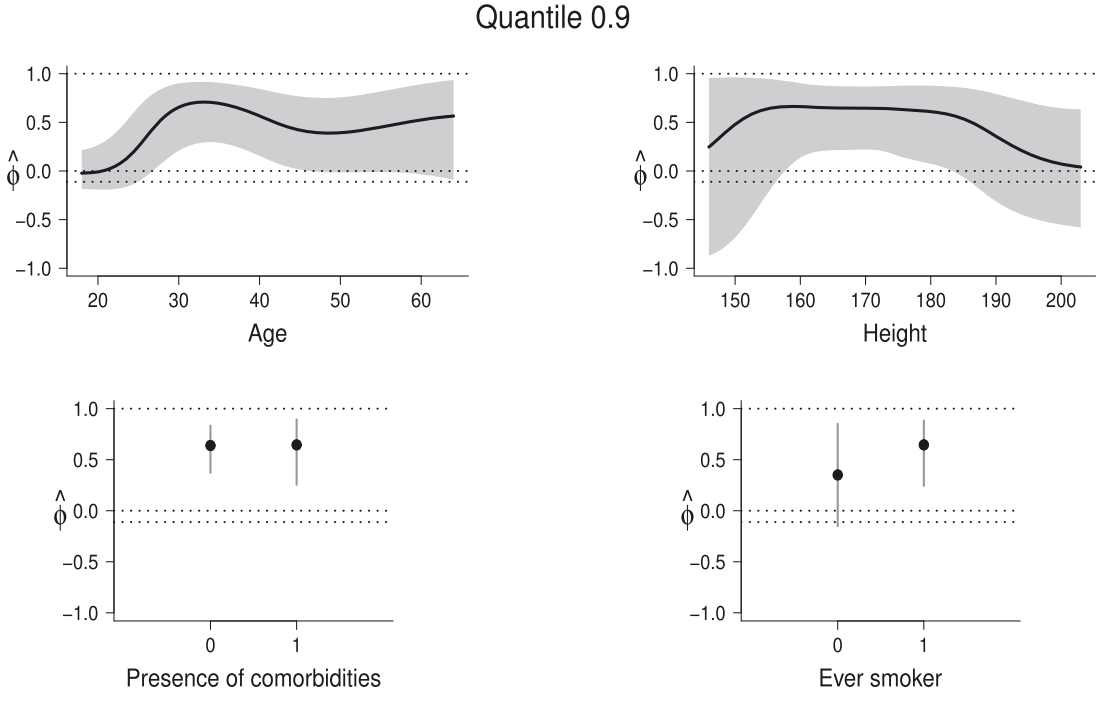

Figure 4: Predicted correlation at the 90th percentile.

residuals can be considered as a single category, simplifying the multinomial regression model. We initially included the same predictors used in the univariate quantile regression models; however, our final second-step model only included zigosity, that was the only significant predictor.

Figure 5(a) and (b) illustrate the parameters' estimates ($\hat{\gamma}_{z;\tau}$) together with 95% bootstrap confidence intervals. The figures compare the negative concordance (Z = "11") and the discordance (Z = "01" + "10") to

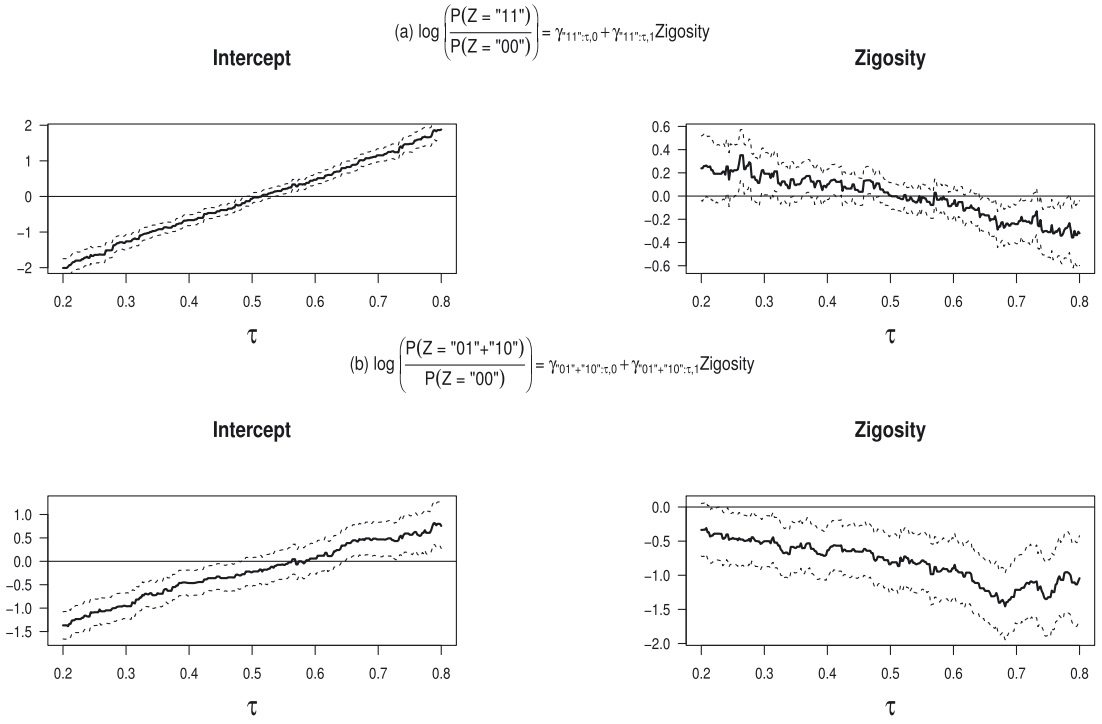

Figure 5: Estimated coefficients ($\hat{\gamma}_{z:\tau}$) of the multinomial logistic model, $\tau \in [0.2, 0.8]$.

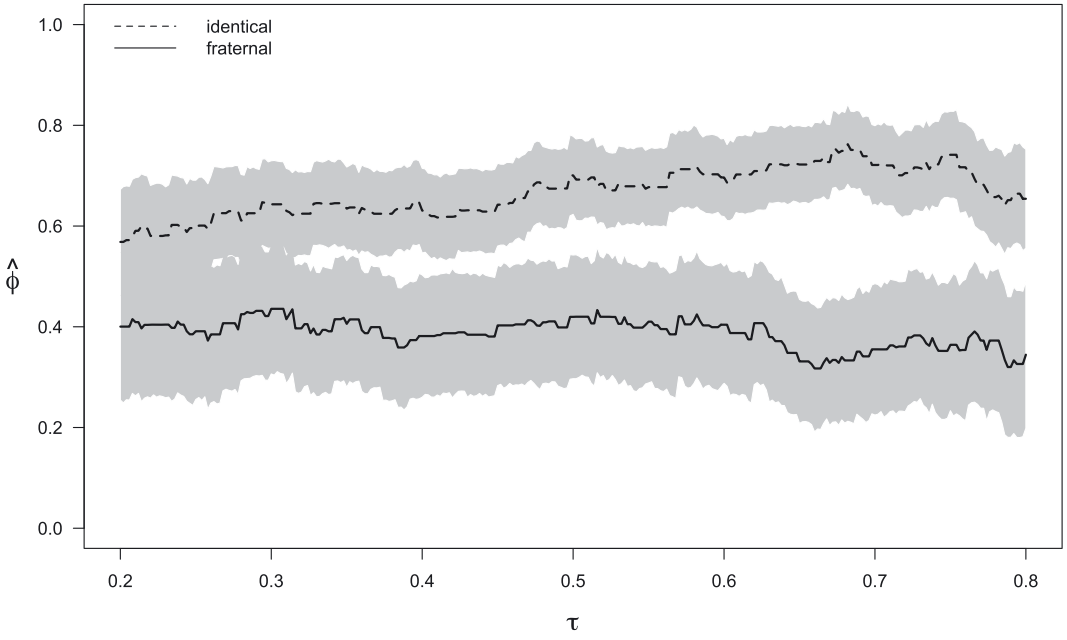

Figure 6: Predicted correlation $\hat{\phi}(\mathbf{x})$ for fraternal twins and identical twins, $\tau \in [0.2, 0.8]$.

the situation of positive concordance (Z = "00"). The model coefficients show that identical twins have a lower chance of being discordant and a higher chance of being positively concordant. The estimated coefficients are decreasing function of the quantile, showing a stronger difference between twins with higher performances in NMSQT test.

In Figure 6 we report the estimates of the conditional coefficient of correlation $\hat{\phi}(\mathbf{x})$. In the calculation of $\hat{\phi}(\mathbf{x})$, the predicted probability of discordant residuals in Eq. (6) was equally split in the two terms p_{01} and

p_{10}. The correlation was always positive for both identical and fraternal twins, and was higher for identical ones. These results are in line with those obtained by [20], and strongly support the presence of a genetic component in the students' performances.

6 Final remarks

Dependence between multiple response variables of interest is very common, and occurs in cross-sectional, case-control, and longitudinal studies. The method presented in this paper allows to investigate the conditional correlation structure of a multivariate response, by analyzing the signs of the residuals from univariate quantile regression models. In the paper, we assumed that the same quantile, τ, was estimated for all outcomes. However, depending on the purpose of the analysis, one could consider a different quantile, say $\tau^{(j)}$, for each outcome of interest, $j = 1, \dots, d$.

In principle, other families of regression models could be used in place of quantile regression: for example, one could apply the same method to the residuals of a linear (mean) regression. However, the residuals' signs are a natural outcome of quantile regression, in which by definition a proportion τ of the observations lie below the estimated regression line. Also, in Eq. (1), the conditional quantiles of $Y^{(j)}$ are assumed to be linear in the parameters. Although this parametrization is very popular and computationally convenient, the method presented in this paper can be applied to any nonlinear quantile function.

In our work, we mainly considered a bivariate response. When the response vector has more than two elements, the proposed approach can be modified by introducing higher-order sign-concordance probabilities, that can be combined into more complicated summary statistics. For example, the simple correlation coefficient ϕ defined in Eq. (6) could be replaced by some multivariate measure of correlation (e.g. [44]). Alternatively, one can calculate a standard measure of correlation for each pair of response variables.

The used multinomial logistic model is completely unstructured, and has a number of parameters which is proportional to 2^d where d is the number of outcomes being considered. Some simplifying assumptions may be used to investigate the joint distribution of multiple binary responses. For example, in the Twin Study presented in Section 5, we categorized the combinations of two responses into three groups only. Another possibility for the case $d = 2$ is to directly model a binary variable indicating concordance ($Z_i = \{00, 11\}$) using standard logistic regression. When d is large, identification problems may occur. For example, if $d = 10$ and all outcomes are independent, the relative frequency of the cell in which all responses are above the median is $0.5^{10} = 1/1024$. While an empty cell is simply ignored, a cell with very few observations will cause the multinomial regression model to be poorly identified and to have large standard errors. This could be solved by merging multiple cells, for example by defining a binary indicator of "at least 5 responses above the median".

Finally, the use of multinomial logistic regression is just a possible solution. Any other method that allows to predict probabilities could be used, including a multinomial model with a different link function (e.g., a probit model) or a probabilistic classifier.

All methods described in this paper are implemented in standard software. The R code used to analyze the two datasets presented in Sections 4 and 5 is available on GitHub repository at the link https://github.com/silviacolumbu/Sign-concordance-of-QR-residuals.

Acknowledgments: We thank Dr. Giovanni Viegi for allowing use of a subset of the data from the Po river delta epidemiological study.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: Silvia Columbu gratefully acknowledges Regione Autonoma della Sardegna for the financial support provided under the Operational Programme P.O.R. Sardegna F.S.E. (European Social Fund 2014-2020 - Axis III Education and Formation, Objective10.5, Line of Activity 10.5.12).

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

1. Paciaroni M, Agnelli G, Falocci N, Caso V, Becattini C, Marcheselli S, et al. Early recurrence and cerebral bleeding in patients with acute ischemic stroke and atrial fibrillation: effect of anticoagulation and its timing: the raf study. Stroke 2015;46:2175–82.
2. Koenker R. Quantile regression. Cambridge: Cambridge University Press; 2005.
3. Serfling R. Quantile functions for multivariate analysis: approaches and applications. Stat Neerl 2002;56:214–32.
4. Cai Y. Multivariate quantile function models. Stat Sin 2010.
5. Chakraborty B. On multivariate quantile regression. J Stat Plann Inference 2003;110:109–32.
6. Chauduri P. On a geometric notion of quantiles for multivariate data. J Am Stat Assoc 1996;91:862–72.
7. Chavas JP. On multivariate quantile regression analysis. Stat Methods Appl 2017:365–84. https://doi.org/10.1007/s10260-017-0407-x.
8. Dudley RM, Koltchinskii VI. The spatial quantiles. Unpublished Manuscript 1992.
9. Geraci M, Boghossian N, Farcomeni A, Horbar J. Quantile contours and allometric modelling for risk classification of abnormal ratios with an application to asymmetric growth-restriction in preterm infants. Stat Methods Med Res 2020;29:1769–86.
10. Hallin M, Paindaveine D, Siman M. Multivariate quantiles and multiple-output regression quantiles: from l1 optimization to halfspace depth. Ann Stat 2010;110:109–32.
11. Kong L, Mizera I. Quantile tomography: using quantiles with multivariate data. Statistics Sinica 2010;22:1589–610.
12. Liu X, Zuo Y. Computing halfspace depth and regression depth. Commun Stat Simulat Comput 2014;43:969–85.
13. Struyf AJ, Rousseuw PJ. Halfspace depth and regression depth characterize the empirical distribution. J Multivariate Anal 1999;69:135–53.
14. Alfo M, Marino F, Ranalli M, Salvati N, Tzavidis N. M-quantile regression for multivariate longitudinal data with an application to the millennium cohort study. J Roy Stat Soc: Series C (Appl Stat) 2020;70:9122–46.
15. Kulkarni H, Biswas J, Das K. A joint quantile regression model for multiple longitudinal outcomes. AStA Adv Stat Anal 2019. https://doi.org/10.1007/s10182-018-00339-9.
16. Petrella L, Raponi V. Joint estimation of conditional quantiles in multivariate linear regression models with an application to financial distress. J Multivariate Anal 2019. https://doi.org/10.1016/j.jmva.2019.02.008.
17. Drovandi C, Pettitt A. Likelihood-free Bayesian estimation of multivariate quantile distributions. Comput Stat Data Anal 2011. https://doi.org/10.1016/j.csda.2011.03.019.
18. Guggisberg MA. A Bayesian approach to multiple-output quantile regression. J Am Stat Assoc 2022. https://doi.org/10.1080/01621459.2022.2075369.
19. Waldmann E, Kneib T. Bayesian bivariate quantile regression. Stat Model Int J 2015. https://doi.org/10.1177/1471082x14551247.
20. Li R, Cheng Y, Fine JP. Quantile association regression models. J Am Stat Assoc 2014;109:230–42.
21. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.
22. Lipsitz SR, Laird NM, Harrington DP. Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. Biometrika 1991;78:153–60.
23. Lu M, Yang W. Multivariate logistic regression analysis of complex survey data with application to brfss data. J Data Sci 2012;10:157–73.
24. Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics 1988;44:1033–48.
25. Breslow NE, Calyton DG. Approximate inference in generalized linear mixed model. J Am Stat Assoc 1993;88:9–25.
26. Das A, Poole WK, Bada HS. A repeated measure approach for simultaneous modeling of multiple neurobehavioral outcomes in newborn exposed to cocaine in utero. Am J Epidemiol 2004;159:891–9.
27. Molenberghs G, Verbeke G. Models for discrete longitudinal data. New York: Springer; 2005.
28. Stiratelli R, Laird NM, Ware JH. Random effects models for serial observations with binary response. Biometrics 1984;40:961–71.
29. Gauvreau K, Pagano M. The analysis of correlated binary outcomes using multivariate logistic regression. Biom J 1997;39:309–25.
30. Genest C, Nikoloulopoulos AK, Rivest L, Fortin M. Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. Brazilian J Probab Stat 2013;27:265–84.
31. Meester SG, MacKay R. A parametric model for cluster correlated categorical data. Biometrics 1994;50:954–63.
32. Nikoloulopoulos AK, Karlis D. Multivariate logit copula model with an application to dental data. Stat Med 2008;27:6393–406.
33. Koenker R, Bassett G. Regression quantiles. Econometrica 1978:33–50. https://doi.org/10.2307/1913643.
34. Hardin JW. The robust variance estimator for two-stage models. Stata J 2002;2:253–66.

35. Murphy KN, Topel RH. Estimation and inference in two-step econometric models. Brazilian J Bus Econ Stat 1978;20:88–97.
36. Quanjer PH, Stanojevic S, Cole TJ, Baur X, Hall GL, Culver BH, et al. Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations. Eur Respir J 2012;40:1324–43.
37. Stanojevic S, Wade A, Stocks J. Reference values for lungfunction: past, present and future. Eur Respir J 2010;36:12–9.
38. Bottai M, Pistelli F, Pede FD, Baldacci S, Simoni M, Maio S, et al. Percentiles of inspiratory capacity in healthy nonsmokers: a pilot study. Respiration 2011;82:254–62.
39. Carrozzi L, Giuliano G, Viegi G, Paoletti P, Pede FD, Mammini U, et al. The po river delta epidemiological study of obstructive lung disease: sampling methods, environmental and population characteristics. Eur J Epidemiol 1990;6:191–200.
40. Ciccio TJD, Romano JP. A review of bootstrap confidence intervals. J Roy Stat Soc B 1988;50:338–54.
41. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. Stat Sci 1986;1:54–77.
42. Loehlin J, Nichols R. Heredity, environment, & personality: a study of 850 sets of twins. TX, Austin: University of Texas Press; 1976.
43. Lee Y, Molas M, Noh M. mdhglm: multivariate double hierarchical generalized linear models. In: R package version 1.8; 2018.
44. Wang J, Zheng N. Measures of correlation for multiple variables. arXiv:1401.4827v6.