Conditional concordance of the signs of the residuals of quantiles regressions of multivariate outcomes

Concordanza condizionata del segno dei residui di regressione quantilica per variabili dipendenti multivariate.

Silvia Columbu and Matteo Bottai

Abstract The extension of the concept of quantile of a univariate response variable to a multivariate response is not straightforward because of the absence of a natural definition of multivariate quantile. This paper describes an approach to studying the dependence of a multivariate response variable that is based on the comparison of signs of the residuals of univariate quantile regressions. The method permits to capture important aspects of the association structure of dependent variables given a set of covariates.
Abstract L'estensione del concetto di quantile al caso di variabili risposta multivariate è resa complicata dall'assenza di una definizione naturale e univoca di quantile multivariato. In questo lavoro presentiamo un metodo sviluppato per lo studio della struttura di dipendenza per una variabile risposta multivariata basato sul confronto dei segni dei residui di regressione quantilica univariata . Il metodo permette di catturare aspetti importanti legati alla struttura di associazione delle variabili dipendenti dato un insieme di covariate.

Key words: quantile regression, conditional quantiles, multivariate regression, sign of residuals

1 Introduction

When multiple characteristics are observed on a sample of subjects or experimental units, it is often the case that more than one of these represent important outcomes of scientific interest. For example, the effect of a treatment on hypertensive patients

Silvia Columbu
Dipartimento di Matematica e Informatica, Università di Cagliari, e-mail: silvia.columbu@unica.it

Matteo Bottai
Unit of Biostatistics, Institute of Environmental Health, Karolinska Institutet, Stockholm e-mail: matteo.bottai@ki.se

may have a beneficial effect on systolic blood pressure and a detrimental effect on hepatic function. Evaluating the treatment effect on blood pressure and hepatic function jointly, rather than separately, may therefore offer important scientific insight.

Unlike inference on conditional quantiles of univariate outcomes given covariates, which has gained popularity in many fields of applied research, multivariate quantile methods are not widely utilized. This is caused by lack of a natural definition of multivariate quantile, contrarily to the multivariate mean. In the literature various multivariate quantile methods have been explored, but none has shown a clear advantage over the others [4].

In the present paper we propose a nonparametric method based on the signs of residuals of univariate quantile regression, which permits to study the dependence structure of multivariate outcomes conditional on a possible set of covariates. We define a statistic based on a comparison of the signs of the regression residuals. The statistic is specific to the quantile considered and can depend on a set of covariates.

In the following sections we describe the proposed method and apply it to studying the multivariate structure in a dataset on the study of obstructive pulmonary diseas in the north of Italy.

2 Quantile regression residuals's sign concordance

In this section we discuss the case of a bivariate response variable. The results can be generalized to the multivariate case. Consider a 2-dimensional response variable of n observations $\boldsymbol{y}^i = \{y_1^i, \ldots, y_n^i\}$, $i = 1,2$, and $\boldsymbol{X}$ a $n \times q$ matrix of covariates. The quantile regression model for each of the two outcomes is

$$\boldsymbol{y}^i = \boldsymbol{X}\beta_\tau^i + \boldsymbol{\varepsilon}^i \quad i = 1,2$$

where $\beta_\tau^i = \{\beta_{\tau,0}^i, \beta_{\tau,1}^i, \ldots, \beta_{\tau,q}^i\}$ indicate the unknown regression parameters. For any given $\tau \in (0,1)$ we assume that $Q_\varepsilon(\tau|\boldsymbol{X} = x) = 0$. The τ-th quantile of the conditional distribution of $\boldsymbol{y}^i$ given $\boldsymbol{X}$ is defined as

$$Q_{\boldsymbol{y}^i}(\tau) = \boldsymbol{X}\beta_\tau^i.$$

We define $\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$ a random vector with

$$Z_j = \begin{cases} 1 & \text{if } (y_j^1 \leq X_j\beta_{\tau,j}^1 \wedge y_j^2 \leq X_j\beta_{\tau,j}^2) \vee (y_j^1 > X_j\beta_{\tau,j}^1 \wedge y_j^2 > X_j\beta_{\tau,j}^2) \\ 0 & \text{otherwise} \end{cases}$$

The variable Z_j is an indicator of concordance of residuals of quantile regression. The following statistic

$$\hat{\sigma} = \frac{1}{n}\sum_{j=1}^{n} Z_j$$

represents the proportion of observations with concordant residuals signs of quantile regression for the two outcomes. The marginal distribution of sign of the quantile regression residuals is known according [1, Theorem 3.4]. From marginal distributions three limit values of $\hat{\sigma}$ can be calculated under different hypothesis of dependence association.

- $\sigma_{indep} = 1 + 2\tau^2 - 2\tau$; independence
- $\sigma_{dep+} = 1$; largest possible positive dependence
- $\sigma_{dep-} = |2\tau - 1|$; largest possible negative dependence

The limit values of $\hat{\sigma}$ are ordered as $0 \leq \sigma_{dep-} < \sigma_{indep} \leq \sigma_{dep+}$. Quantiles of residuals are for construction independent from explicative variables, $Q_\varepsilon(\tau|\boldsymbol{X} = x) = 0$, but the joint distribution of residuals can still depend on them. The probability of concordance of residuals is modeled by logistic regression

$$\text{logit}(\sigma_x) = \log\left(\frac{\sigma_x}{1-\sigma_x}\right) = \boldsymbol{X}\gamma_\tau + \eta$$

with $\sigma_x = P(\boldsymbol{Z} = 1|\boldsymbol{X} = x)$. The dependence of the concordance on the covariates $\boldsymbol{X}$ for different quantiles can be summarized by odds ratios and showed through graphical representations.

The dependent variable in the logistic model is based on the residuals based on the estimated quantile regression. The variance of quantile regression coefficients, β_τ^1 and β_τ^2, needs to be considered when estimating the variance of the coefficient vector γ_τ. We utilized the bootstrap with 200 replications to obtain robust estimates of errors in the application showed in the next section. Alternatively, the method of moments can provide large sample approximations.

3 Obstructive pulmonary disease in northern Italy

We apply the proposed method to analyze the multivariate structure in 2000 individuals randomly selected from the Po River Delta epidemiological study, a prospective study conducted on the general population of a rural area in northern Italy (near Venice) [3]. A baseline survey was carried out during $1980-1982$ on $3,284$ subjects of whom $2,136$ (65%) were followed in $1988-1991$, to whom 705 new subjects were added.

We consider a set of four variables, two outcomes and two explicative variables. The dependent variables considered are two lung function measures, namely Forced Vital Capacity (FVC) and Forced Expiratory Volume in one second (FEV1). A computerized pneumotachograph (Fleisch No.3) ($47804/s$ Pulmonary System; Hewlett-Packard, Waltham, MA) was used for measuring flow and volume of air that can be inhaled and exhaled. The covariates considered are height and sex. Height (in centimeters) was measured in standing position without shoes in subjects wearing clothes.

We estimate a quantile regression model for each of the two outcomes considering for 9 quantiles $\tau = \{0.1, 0.2, \ldots, 0.9\}$

$$FVC = \beta^1_{\tau,0} + \beta^1_{\tau,1} sex + \beta^1_{\tau,2} height$$

$$FEV_1 = \beta^2_{\tau,0} + \beta^2_{\tau,1} sex + \beta^2_{\tau,2} height$$

The *rq* function of *quantreg* package in R [2] is used to estimate regression parameters. We extract the residuals of estimate in both models and use them to define the indicator variable of concordance Z as described in the previous section. The limit values of the σ statistic for the selected quantiles are reported in Table 1.

Table 1 σ statistic values under hypothesis of perfect negative dependence, perfect independence and perfect positive dependence for 9 quantiles

	$\tau = 0.10$	$\tau = 0.20$	$\tau = 0.30$	$\tau = 0.40$	$\tau = 0.50$	$\tau = 0.60$	$\tau = 0.70$	$\tau = 0.80$	$\tau = 0.90$
σ_{dep-}	0.8	0.6	0.4	0.2	0.0	0.2	0.4	0.6	0.8
σ_{indep}	0.82	0.68	0.58	0.52	0.50	0.52	0.58	0.68	0.82
σ_{dep+}	1	1	1	1	1	1	1	1	1

Coefficients estimates of the logistic regression model for the probability of concordance are reported in Table 2. When $\tau = \{0.30, 0.40, 0.50\}$ the estimates of the coefficient associated with height are statistically significant showing a decreasing effect of concordance between the residuals of FVC and FEV1 along with height. The probability of concordance of residuals' signs changes with height for 3 of the 9 quantiles considered. The effect of sex on concordance was statistically significant for $\tau = \{0.40, 0.50, 0.60\}$.

Table 2 Estimates of parameters of logistic regression model. 200 Bootstrap replications are used to estimate standard errors.

	Intercept	Sex	Height
$\tau = 0.10$	1.445 (-0.769, 3.659)	-0.192 (-0.571, 0.188)	0.006 (-0.008, 0.021)
$\tau = 0.20$	3.267* (1.212, 5.321)	0.110 (-0.198, 0.417)	-0.010 (-0.023, 0.003)
$\tau = 0.30$	3.617* (1.704, 5.529)	0.126 (-0.180, 0.433)	−0.013* (-0.026, -0.001)
$\tau = 0.40$	3.659* (1.531, 5.788)	0.391* (0.094, 0.687)	−0.015* (-0.028, -0.001)
$\tau = 0.50$	3.977* (1.732, 6.221)	0.388* (0.069, 0.706)	−0.018* (-0.032, -0.003)
$\tau = 0.60$	2.990* (0.791, 5.190)	0.303* (0.016, 0.591)	-0.011 (-0.025, 0.003)
$\tau = 0.70$	3.437 (1.244, 5.631)	0.141 (-0.142, 0.424)	-0.012 (-0.026, 0.001)
$\tau = 0.80$	2.161 (-0.102, 4.425)	-0.007 (-0.268, 0.254)	-0.003 (-0.018, 0.011)
$\tau = 0.90$	3.935 (1.113, 6.758)	-0.032 (-0.377, 0.314)	-0.010 (-0.027, 0.008)

* Significantly different from zero at the 5 per cent level.

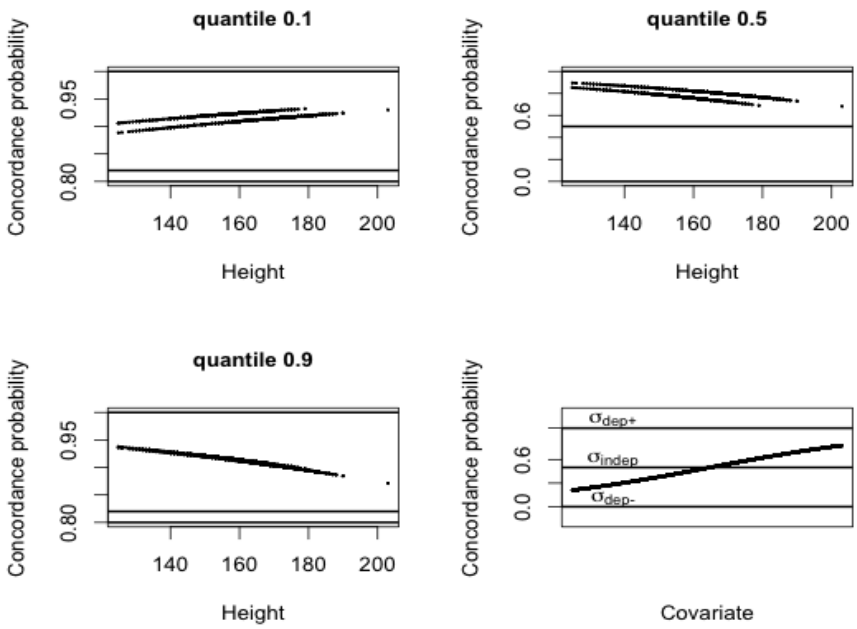

Fig. 1 Plots represent the logistic regression model for the probability of concordance relative to the 3 quantiles $\{0.10, 0.50, 0.90\}$ for men and women and horizontal lines corresponding to specific values of the σ statistic. The fourth graphic is reported as a reference.

Graphical representations may help understand the associations of the residuals' signs. Figure 1 shows the concordance probability for the 10th, 50th, and 90th percentile. The plots suggest that the concordance is positive and above independence. Considering the two lung function measures separately would not capture this dependence. In addition, the dependence structure varies at different values of height. For low values of height the predicted concordance probability for the median is close to prefect positive dependence, while for higher values of height it is closer to independence.

4 Summary

The proposed method can be used as a practical tool to study the association of the signs of residuals of quantile regression for multivariate outcomes. The easy mathematical formulation allows the method to be readily used and understood by non-statisical researchers. The derivation of the asymptotic properties of the statistic $\hat{\sigma}$ and of $\hat{\gamma}_\tau$ are work in progress.

References

1. Koenker, R., Bassett, G. : Regression Quantiles. Econometrica **46**, 33–50 (1978)
2. Koenker, R.: Quantile Regression, Econometric Society Monograph Series, Cambridge University Press (2005)
3. Pistelli, F., Bottai, M., Viegi, G., et al.: Smooth reference equations for slow vital capacity and flow-volume curve indexes. Am. J. Respir. Crit. Care Med. **161**, 899–905 (2000)
4. Serfling, R.: Quantile functionsfor multivariate analysis: approaches and applications. Statistica Neerlandica **56**, 214–232 (2002)