Quantile regression modelling of some Brazilian anthropometric data

Luna Hidalgo Carneiro*
IBGE, Rio de Janeiro, Brasil - lunahidalgo@gmail.com

Pedro Luis do Nascimento Silva
IBGE – Escola Nacional de Ciências Estatísticas, Rio de Janeiro, Brasil
pedro-luis.silva@ibge.gov.br

Abstract

In Brazil, the main recent official source of basic anthropometric data (height and weight) is the Consumer Expenditure Survey (POF) conducted by the IBGE in 2002/03 and 2008/09. Estimates of the median height (length for babies) and weight (body mass) are released by sex and age group, with federation units as the lowest level of geographic disaggregation. The age groups used vary from 0 to 19 in 1 year intervals, 20 to 34 in 5 year intervals, 35 to 74 in 10 year intervals. 75 and above is the last interval. As a consequence, only small samples are available to estimate these medians in many of the federation units and sex × age groups. Nevertheless, there is increasing demand for such estimates, and for estimates for even lower levels of geographic disaggregation. In this paper, linear quantile regression models were used to obtain estimates for selected quantiles (centiles 10, 50 and 90) of height and weight by sex and age in each federation unit, using data from the POF 2008/09. The Quantile Regression estimates yielded noticeable improvements in comparison to direct estimates for the same parameters.

Keywords: anthropometry; sample survey; quantile regression; small domains.

1. Introduction

In Brazil, the main recent official source of basic anthropometric data (height and weight) is the Consumer Expenditure Survey (POF) conducted by the Brazilian Institute of Geography and Statistics (IBGE) in 2002/03 and 2008/09. The 2008-2009 edition of the POF obtained measurements of height and weight of all the residents in surveyed households.

In accordance with the sampling design applied, estimates of the median height (length for babies) and weight (body mass) are released by sex and age group, with federation units as the lowest level of geographic disaggregation. The age groups used vary from 0 to 19 in one year intervals, 20 to 34 in five year intervals, 35 to 74 in ten year intervals. 75 years old and above is the last interval.

Consequently, only small samples are available to estimate these medians in many of the federation units× sex × age groups. Nevertheless, there is increasing demand for such estimates and also for estimates for even lower levels of geographic disaggregation.

This type of situation calls for the application of small domain estimation (SDE) methods which can provide more reliable estimates for domains where the sample sizes are small (PFEFFERMANN, 2002; RAO, 2003). The application of SDE methods to enable more detailed estimates for the anthropometric measures is the main goal of this paper.

The sampling design adopted by POF 2008-2009 was a stratified two-stage cluster sampling design, with census enumeration areas (CEAs) as the primary sampling units (PSUs) and households (HHs) as the secondary sampling units (SSUs). Within sampled households, all residents were included in the survey, though anthropometric measurements might be missing for some residents who were not at home during the entire interviewing period. The sampling weights were based on post-stratification (IBGE, 2010).

In line with the sampling design used in POF 2008-2009, the basic sampling weights were obtained considering the two-stage stratified cluster sampling design. These weights were subsequently adjusted by a post-stratification procedure to compensate for household and person level non-response.

Direct estimates of the median height and weight per federation unit× sex × age group were obtained here using standard weighted estimators of the median, such as described in Särndal, Swensson & Wretman (1992), using the same approach used by IBGE to obtain the published estimates. The sampling design and weights were incorporated into the inference using the «survey» package (LUMLEY, 2012) from the R software (R Core Team, 2013).

2. Quantile Regression

Koenker and Bassett (1978) proposed the linear quantile regression model (LQRM) as a robust alternative to the linear regression model (LRM). While the LRM models the relationship between the conditional mean of y and covariates x, E[y|x], LQRM describes the behaviour of the quantiles of y given x, namely $Q_t(y|x)$, where t designates the desired quantile (0< t <1).

Equation (1) describes the general LQRM considered in this paper:

$$Q_t(y|\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta}_t + \varepsilon_t \tag{1}$$

Where $Q_t(y|\boldsymbol{x})$ is the quantile t (0 < t < 1) of the response variable y conditional on the covariates $\boldsymbol{x}$;

$\boldsymbol{x} = (1, x_1, \ldots, x_p)'$;

x_j denotes the jth explanatory variable ;

$\beta_{t,0}$ is the intercept of the quantile t,

$\beta_{t,j}$ is the coefficient associated to variable $\boldsymbol{X}_j$ for quantile t, $j = 1, 2, \ldots, p$,

$\boldsymbol{\beta}_t = (\beta_{t,0}, \ldots, \beta_{t,p})'$; and

ε_t is the error term of quantile t.

The model described in equation (1) can be estimated by:

$$\widehat{Q_t}(y|\boldsymbol{x}) = \boldsymbol{x}\widehat{\boldsymbol{\beta}_t} \tag{2}$$

Where $\widehat{Q_t}(y|\boldsymbol{x})$ is the estimated quantile t of the response variable y conditional on the covariates $\boldsymbol{x}$;

$\hat{\boldsymbol{\beta}}_t = (\hat{\beta}_{t,0}, \ldots, \hat{\beta}_{t,p})'$, where $\hat{\beta}_{t,j}$ is the estimated coefficient of quantile t associated to variable $\boldsymbol{x}_j$, and $\hat{\beta}_{t,0}$ is the estimated intercept.

In this paper, linear quantile regression models were fitted using the POF 2008-2009 data, taking y as one of the anthropometric variables, namely height (cm) or weight (kg). Only the centiles t=10%, 50% and 90% were considered here, but the approach is applicable more generally for any t. The covariates $\boldsymbol{x}$ contain indicator variables defined by federative unit of residence (26 categories), sex (1 category), and age group (27 categories). Therefore, a total of 54 predictor variables were considered. The quantile regression models fitted contained only the main effects. The interactions between the explanatory variables were not considered due to the reduced sample sizes in each combination of levels of the explanatory variables.

The model was fitted using package «quantreg» (KOENKER, 2012) from the R software (R Core Team, 2013). In order to incorporate the sampling design to estimation process of regression coefficients, the post-stratified weights were considered using the option *weights* of the *rq* function in «quantreg» package.

At the estimation of quantile regressions presented above, some coefficients had p-values higher than $\alpha = 0.05$. However, in order to have the same explanatory variables / category sets across all federation units, the non-significant coefficients were retained and the corresponding categories were not grouped with the baseline ones. This is justified by the use of the quantile regression model only as predictive tool, and not in an attempt to explain variation of the quantiles by the predictors for interpretation purposes. Also, the grouping of levels of sex and agre group is not acceptable for the publication of the results by the survey organization.

In figures 1 to 3 we compared the Direct and Quantile Regression (QR) estimates for the centiles 10, 50 and 90 of height for Roraima state, which has the smallest overall state level sample size in the POF 2008/09. They show that the QR method produces ‘smoother’ estimates and narrower confidence bands across all three centiles.

Similar results were observed for all other regions and for males. For height, the standard errors were generally larger for women than for men. For weight, the behavior is reversed. A small bias was observed in the QR estimates, but with different signals for the male and female populations.

Figure 1: Estimates and confidence intervals (CI) for centile 10 of height (in cm), for females in Roraima state, by age band, obtained by Direct and Quantile Regression (QR) methods

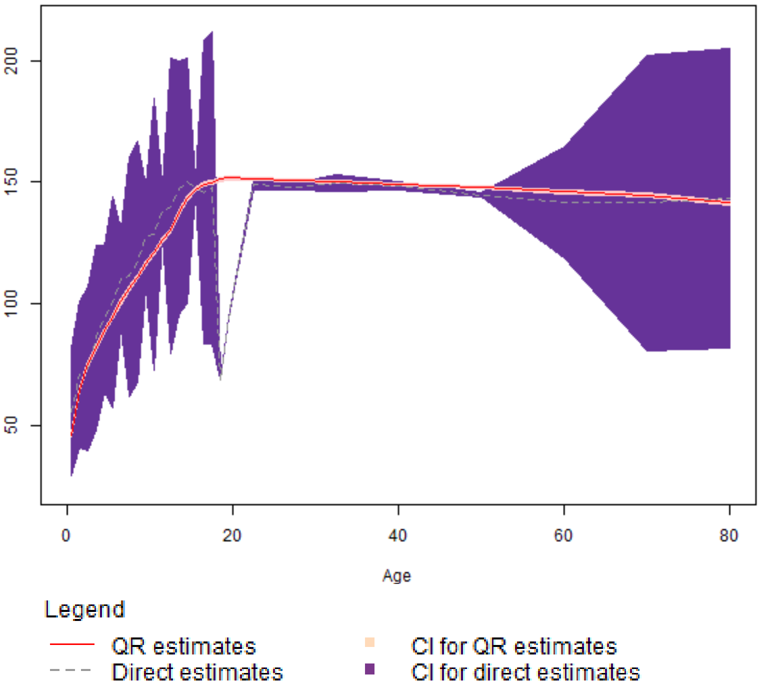


Figure 2: Estimates and confidence intervals (CI) for centile 50 of height (in cm), for females in Roraima state, by age band, obtained by Direct and Quantile Regression (QR) methods

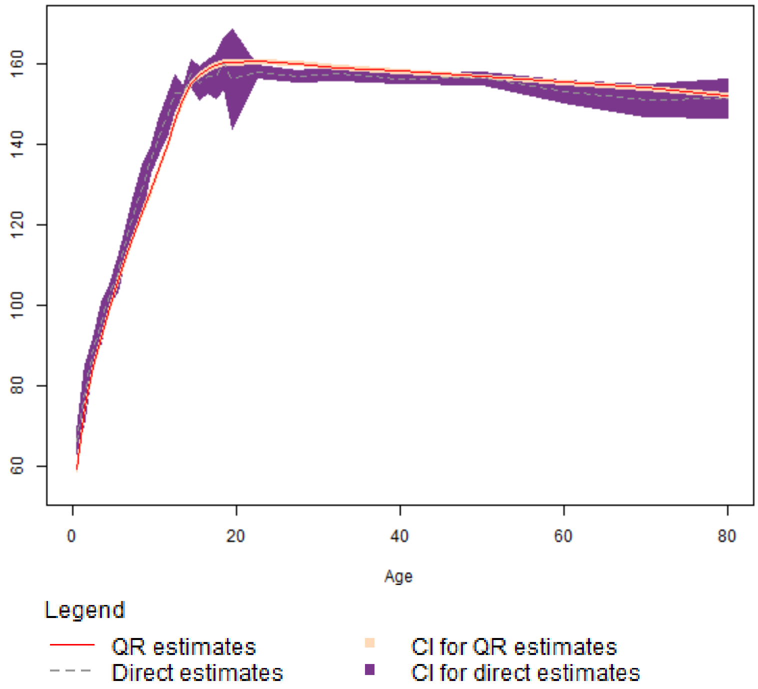
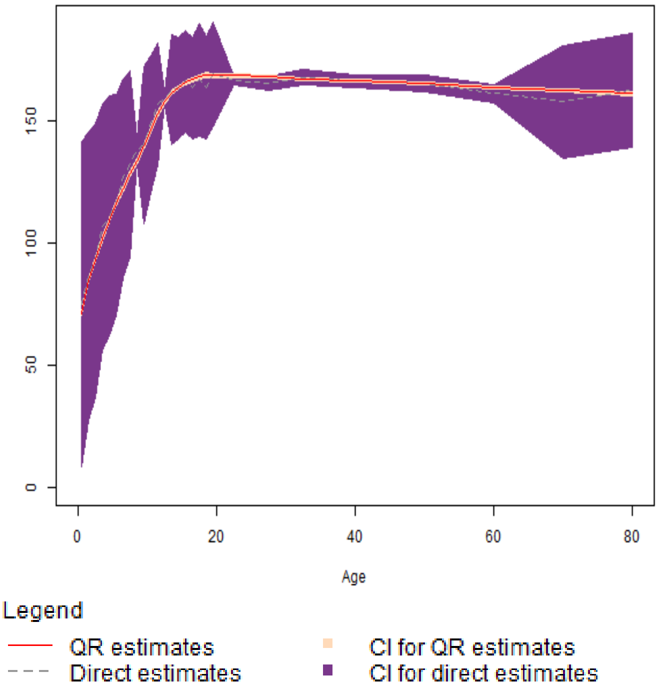


Figure 3: Estimates and confidence intervals (CI) for centile 90 of height (in cm), for females in Roraima state, by age band, obtained by Direct and Quantile Regression (QR) methods



3. Final Remarks

The best estimates of the selected centiles of height and weight were obtained through quantile regression, since there was a noticeable gain of accuracy with respect to coefficients of variation (CVs) of the point estimates, as well as regarding the smoothness of the curves of anthropometric estimates, closer to the expected smooth pattern of age variation for the centiles of these variables.

This study suggests that there is potential for adopting quantile regression methods for estimating the centiles of the anthropometric measures using data from the POF to enable producing more precise and smoother estimates for the required domains of interest. The approach considered here also enables publication of confidence limits for such estimates.

References

Hidalgo-Carneiro, L. (2013). Dados antropométricos da POF 2008/2009: uma estimação usado métodos de quantis para pequenos domínios. Rio de Janeiro: ENCE, Dissertação de Mestrado, 184 p..

IBGE (2010). Pesquisa de Orçamentos Familiares 2008 - 2009: antropometria e estado nutricional de crianças, adolescentes e adultos no Brasil. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística.

Koenker, R. (2012). *quantreg: Quantile Regression*. R package version 4.81. URL http://CRAN.R-project.org/package=quantreg.

Koenker, R. & Bassett, G. (1978). Regression quantiles. Econometrica, 46, pp. 33-50.

Lumley, T. (2012). *survey: analysis of complex survey samples*. R package version 3.28-2. URL http://CRAN.R-project.org/package=survey.

Pfeffermann, D. (2002). Small Area Estimation - New Developments and Directions. International Statistical Review, 70, pp. 125-143.R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rao, J.N.K. (2003). *Small Area Estimation*. New Jersey: Wiley-Interscience.

Särndal, C-E, Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Acknowledgements

This research was partly supported by a grant from the Brazilian Federal Agency for the Support and Evaluation of Graduate Education (CAPES).