American Journal of Epidemiology

Vol. 180, No. 3

Research Letter

QUANTILE REGRESSION—OPPORTUNITIES AND CHALLENGES FROM A USER'S PERSPECTIVE

Quantile regression is a statistical technique used to model quantiles (i.e., percentiles) within a regression framework. Although median regression, a special case of quantile regression, dates back to as early as 1760 (1), quantile regression has been introduced to the statistical community mainly by the works of Roger Koenker during the last decade (2, 3). Although since then it has been of greater interest to statistical methodologists and is implemented in standard statistical packages, it appears to be quite underused in medical research.

Obviously, distributions may not differ only by their means, but also (or even only) by their lower or upper parts (Figure 1). Thus, modeling only the mean, as is done in linear regression, may miss important aspects of the association between the outcome and its predictors, especially if the outcome distribution is skewed, as is frequently the case in medical data. Quantile regression allows one to model any quantile of the outcome distribution, including the median (i.e., the 0.5 quantile). Although the computation of the regression coefficients is somewhat different compared with linear regression (because it is based on minimizing the sum of weighted absolute residuals instead of squared residuals), quantile regression can be applied in the same way, allowing adjustment for potential confounders and calculation of interaction terms and variable selection, while being more robust to statistical outliers and yielding much more information about the underlying associations. There is also established methodology covering, for example, nonlinear and longitudinal quantile regression, as well as applications in survival analysis and growth reference calculation (4–6). It might be argued that logistic regression could be used in addition to linear regression to assess associations with extreme values of the outcome variable. However, logistic regression answers a slightly different question (i.e., the risk of lying below or above a predefined cut-off) and requires—in contrast to quantile regression— categorization of the outcome variable, thus resulting in a substantial loss of information.

Indeed, quantile regression has successfully been applied in medical research. For example, large meta-analyses had indicated that breastfeeding was associated with a significant reduction of a child's risk of overweight later in life (7–9), but there was no difference found in mean body mass index (BMI) (weight (kg)/height $(m)^2$) values between breastfed and formula-fed children (10). These seemingly contradictory results fit well together when quantile regression analyses on a German data set showed that breastfeeding was associated with both a decrease in the upper BMI percentiles and an increase in the lower BMI percentiles at the ages of 5–6 years, and thus with no difference in mean BMI (11). Quantile regression was also helpful in showing that there may be different risk factors for low and high birth weight (12).

As these examples demonstrate, quantile regression appears to be useful if the associations of explanatory variables with the extreme values of an outcome distribution are of particular interest. It may be used either to assess associations with 1 specific percentile (e.g., the 90th BMI percentile in studies of overweight) or to examine whether associations are different for low, medium, and high percentiles. In the latter case, multiple testing issues should be considered and can, for example, be addressed by specific tests assessing trends in quantile regression coefficients across percentiles (2).

Moreover, median regression has been suggested as a way to obtain adjusted medians in clinical research (13), which might be a compelling alternative to the frequently used combination of nonparametric Mann-Whitney *U* tests and linear regression as a way to obtain unadjusted and adjusted estimates from nonnormally distributed data. This approach is quite doubtful from a statistical perspective, because nonparametric tests and linear regression are based on different assumptions and may therefore lead to considerably different results in the unadjusted case (in which linear regression simplifies to a 2-sample Student's *t* test). This is illustrated in a simple example in Figure 2. The values of samples 1 and 2 were drawn from normal distributions, and the distribution of the values from sample 3 shows heavy tails in its upper part. Using Mann-Whitney *U* tests and linear regression, we found relatively similar results for the comparison of samples 1 and 2 ($P = 0.64$ and $P = 0.49$ for Mann-Whitney *U* test and linear regression, respectively) but substantially different results for the comparisons of samples 1 and 3 ($P = 0.39$ and $P = 0.01$, respectively) and samples 2 and 3 ($P = 0.37$ and $P = 0.04$, respectively).

Thus, it appears rather surprising that there has been no greater use of quantile regression in epidemiologic and clinical studies so far. One reason might be that quantile regression is based on sample-specific quantiles, whereas predefined cutoffs or sex- and age-specific percentiles from external references are often the main focus of epidemiologic research. However, this problem may be solved by assessing the percentage of observations at or below the respective threshold (e.g., 86%) and then modeling the associated (i.e., the 0.86)

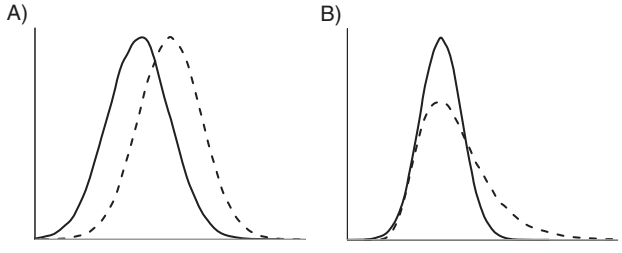

Figure 1. Two distributions may differ with respect to A) their mean only or B) specific quantiles.

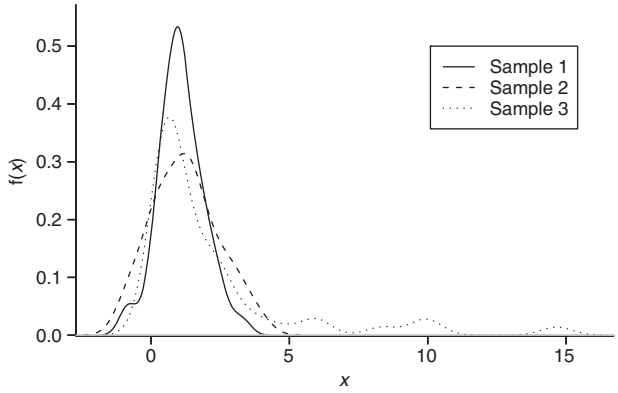

Figure 2. Density plots of 3 samples ($n = 50$ each) drawn from a normal distribution with a mean and a standard deviation of 1 (sample 1), a normal distribution with a mean of 1.3 and a standard deviation of 1 (sample 2), and a log normal distribution with a mean of 1.3 and a standard deviation of 1.5 (sample 3).

quantile. One major reason why quantile regression is still not widely used in medical research is probably that its interpretation seems rather unintuitive. A quantile regression coefficient quantifies how much a specific quantile of the outcome distribution is shifted by a 1-unit increase in the predictor variable. However, this interpretation is basically very similar to that of linear regression, in which the regression coefficient tells the reader how much the mean of the outcome changes in relation to the respective predictor variable. The only actual difference is that we can speak of the latter as an "average difference," whereas we have no appropriate terms in our common language to easily describe results from quantile regression. Furthermore, the interpretation of a single measure, such as that obtained from linear regression, may appear to be more straightforward than the interpretation of a number of quantile regression coefficients, which may not combine to form a simple picture. However, sometimes only the pattern of regression coefficients over the whole range of quantiles can reveal the true underlying associations.

Simplicity in interpretation is certainly an important criterion for the choice of a statistical method. However, quantile regression is not considerably inferior to linear regression in this respect, and it offers much more information and is less sensitive with respect to the distribution of the outcome variable.

Acknowledgments

Conflict of interest: none declared.

References

1. Stigler SM. Studies in the history of probability and statistics XL Boscovich, Simpson and a 1760 manuscript note on fitting a linear relation. *Biometrika*. 1984;71(3):615–620.
2. Koenker R. *Quantile Regression*. 1st ed. New York, NY: Cambridge University Press; 2005.
3. Koenker R, Hallock KF. Quantile regression. *J Econ Perspect*. 2001;15(4):143–156.
4. Fenske N, Fahrmeir L, Hothorn T, et al. Boosting structured additive quantile regression for longitudinal childhood obesity data. *Int J Biostat*. 2013;9(1):1–18.
5. Peng L, Huang Y. Survival analysis with quantile regression models. *J Am Stat Assoc*. 2008;103(482):637–649.
6. Wei Y, Pere A, Koenker R, et al. Quantile regression methods for reference growth charts. *Stat Med*. 2006;25(8):1369–1382.
7. Arenz S, Rückerl R, Koletzko B, et al. Breast-feeding and childhood obesity—a systematic review. *Int J Obes Relat Metab Disord*. 2004;28(10):1247–1256.
8. Harder T, Bergmann R, Kallischnigg G, et al. Duration of breastfeeding and risk of overweight: a meta-analysis. *Am J Epidemiol*. 2005;162(5):397–403.
9. Owen CG, Martin RM, Whincup PH, et al. Effect of infant feeding on the risk of obesity across the life course: a quantitative review of published evidence. *Pediatrics*. 2005;115(5):1367–1377.
10. Owen CG, Martin RM, Whincup PH, et al. The effect of breastfeeding on mean body mass index throughout life: a quantitative review of published and unpublished observational evidence. *Am J Clin Nutr*. 2005;82(6):1298–1307.
11. Beyerlein A, Toschke AM, von Kries R. Breastfeeding and childhood obesity: Shift of the entire BMI distribution or only the upper parts? *Obesity (Silver Spring)*. 2008;16(12):2730–2733.
12. Wehby GL, Murray JC, Castilla EE, et al. Prenatal care effectiveness and utilization in Brazil. *Health Policy Plan*. 2009;24(3):175–188.
13. McGreevy KM, Lipsitz SR, Linder JA, et al. Using median regression to obtain adjusted estimates of central tendency for skewed laboratory and epidemiologic data. *Clin Chem*. 2009;55(1):165–169.

Andreas Beyerlein[1,2]
(e-mail: andreas.beyerlein@helmholtz-muenchen.de)
[1] *Institute of Diabetes Research, Helmholtz Zentrum München, Munich, Germany*
[2] *Forschergruppe Diabetes der Technischen Universität München, Munich, Germany*

DOI: 10.1093/aje/kwu178; Advance Access publication: July 2, 2014