Current Epidemiology Reports (2019) 6:191–199
https://doi.org/10.1007/s40471-019-00204-6

EPIDEMIOLOGIC METHODS (P HOWARDS SECTION EDITOR)

Applications for Quantile Regression in Epidemiology

Ying Wei[1] · Rebecca D. Kehm[2] · Mandy Goldberg[2] · Mary Beth Terry[2,3]

Published online: 20 May 2019

Abstract

Purpose of Review To illustrate the utility of quantile regression in epidemiology for outcomes that are continuous and when exposure effects may differ across the distribution of the outcome. Linear regression methods estimate only the effects at the mean level which may be an incomplete and biased summary of the effect of exposures for some continuous health outcomes.

Recent Findings There are several variations of the quantile regression method including classical linear quantile regression, nonparametric quantile regression for growth trajectories, and the modified quantile regression for case–control designs. Such methods offer several applications including (1) the use of quantile regression to test whether the effects of exposure are similar across quantiles, (2) the use of quantile regression for risk prediction, and (3) the use of quantile regression to examine the effects of growth trajectories over time.

Summary Quantile regression is an important tool for understanding continuous health outcomes, especially outcomes that are not normally distributed, as it offers insight into the relation of exposures with respect to the distribution of the outcome. Quantile regression methods have the potential to deepen and expand the existing quantitative evidence from more common mean-based analyses.

Keywords Quantile regression · Epidemiology · Statistical methods · Continuous outcomes · Growth trajectories

Introduction

Epidemiology research heavily relies on generalized linear models, which quantify the associations between exposures on the mean of outcomes. For example, the linear, logistic, Poisson, and relative risk regression models all estimate the effects of exposures on the mean of the continuous outcome, log odds, the log rate, and the log risk, respectively. However, associations between exposures and outcomes often exist for specific parts of the outcome distribution and therefore estimates of the mean of Y in a generalized linear model may be an incomplete picture of the association between X and Y. Exposures may also have an effect not just on the outcome but also on the variance of the outcome [1, 2]. For example, Yang et al. (2012) [1] found that a genetic variant in the FTO gene and its association with the outcome of body mass index (BMI) was different in size and variance by percentile of BMI.

Quantile regression [3] is one way to investigate exposure–outcome associations beyond the mean level. It models the conditional quantiles of an outcome of interest as a function of covariates (exposures) without assuming equal normally distributed errors or homoscedasticity. By estimating associations across different quantile levels, we can assess how the change of exposure affects the distribution of the outcome. As an illustration, we generated two data sets. The first data set is heteroscedastic, where the outcome Y follows a normal distribution, but both its mean and variance increase with X. In the second data set, the outcome Y is skewed. Figure 1 a displays

This article is part of the Topical Collection on *Epidemiologic Methods*

✉ Mary Beth Terry
mt146@columbia.edu

Ying Wei
yw2148@cumc.columbia.edu

Rebecca D. Kehm
rk2967@cumc.columbia.edu

Mandy Goldberg
mg3431@cumc.columbia.edu

[1] Department of Biostatistics, Columbia University Mailman School of Public Health, New York, NY, USA

[2] Department of Epidemiology, Columbia University Mailman School of Public Health, 722 West 168th St, New York, NY 10032, USA

[3] Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York, NY, USA

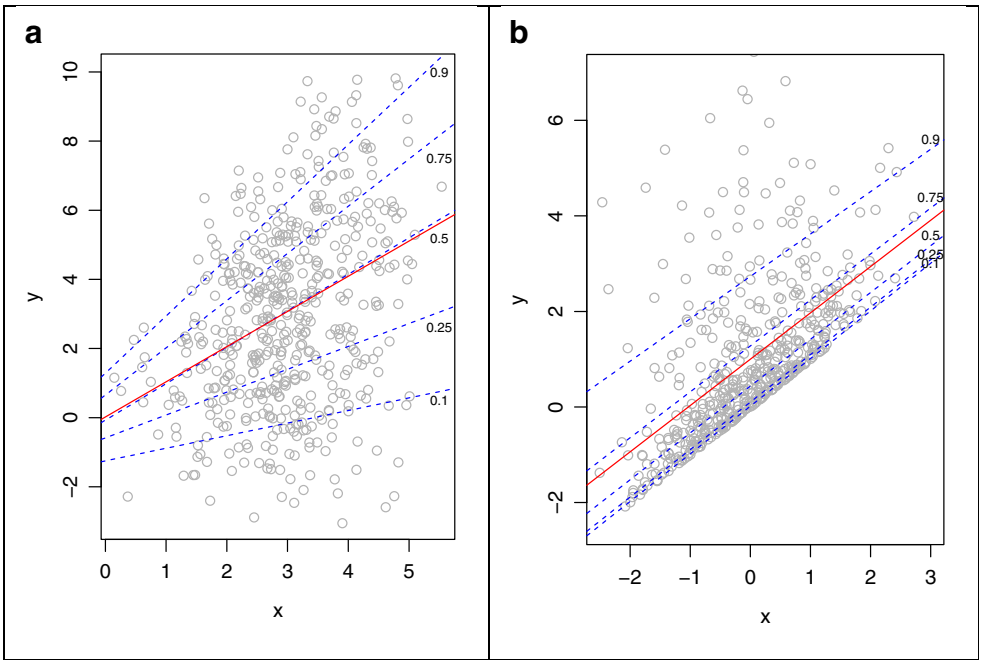

Fig. 1 Fitted regression quantiles over **a** heteroscedastic data and **b** skewed data. Panel A shows the estimated regression quantile functions (blue dashed lines) and mean regression (red line) from a simulated heteroscedastic data set. Panel B illustrates the same for a simulated homoscedastic data set with a skewed outcome

the estimated 0.1th, 0.25th, 0.5th, 0.75th, and 0.9th regression quantile functions (the dotted lines) from heteroscedastic data. We see that the upper quantiles of Y increase with X more rapidly (steeper slope) than the lower quantiles, and consequently capture heteroscedasticity of the data. Figure 1 b displays homoscedastic data with a skewed outcome. As a result, we observe parallel regression quantiles, but the spacing (differences in intercepts) between upper quantiles are much wider than those of lower quantiles. In both cases, a single mean regression (as shown by the red line) would fail to capture the full association that is pictured.

In real-world health applications, the exposure–outcome associations may be even more complex than these examples. It is very likely that exposure only changes part of the Y distribution. Quantile regression would be ideal to explore complex associations. In addition to its modeling flexibility, quantile regression is also robust against outliers. It could be a better fit even in estimating the conditional means in the presence of outliers or heavy tail outcomes [4].

Here, we present several quantile regression models that could be useful for epidemiology studies, including the classical linear quantile regression, nonparametric quantile regression for growth trajectories, and quantile regression models for a case–control design. We highlight several applications of quantile regression under each of the three quantile models, including (1) the use of quantile regression to assess and test quantile treatment effects of exposure across quantiles, (2) the use of quantile regression for risk prediction, and (3) the use of quantile regression to examine the effects of growth trajectories over time.

Linear Quantile Regression Model

Model Description Here, we briefly summarize the basic assumptions of the quantile regression model. Let Y be an outcome of interest, and let X represent the vector of the covariates of interest. Classical linear regression models the conditional mean of Y given X, that is, $E(Y \mid X) = X'\beta$. The regression coefficients β can be interpreted as the change of mean Y due to a unit increase of X. In quantile regression, we model the conditional quantile of Y as a linear function of X without assuming that the coefficients are constant across quantile levels. A linear conditional quantile model can be written as follows:

$$Q_\tau(Y|X) = X'\beta(\tau)$$

where $Q_\tau(Y \mid X)$ stands for the τth conditional quantile of Y given a covariate profile X, and τ is quantile level that ranges between 0 and 1. For example, τ would equate to 0.5 if modeling the median (50th percentile) of Y. The quantile coefficient $\beta(\tau)$ can be interpreted as the change of the τth quantile due to a unit increase of X. It is estimated separately at each quantile level, and hence the estimation of $\beta(\tau)$ is entirely data-driven and does not rely on any distributional assumption. Since quantile regression does not assume any parametric form for $\beta(\tau)$, it is a flexible modeling tool and can accommodate any continuous distribution of Y. Thus, Y does not need to be normally distributed which is a key advantage for health data.

Whereas a linear regression model is fit by minimizing the squared loss, defined as the difference between the observed and expected Y for each observation, a quantile regression

model uses the absolute loss rather than the squared loss. When $\tau = 0.5$, the model estimates the effect of X on the conditional median of Y, and the loss function is calculated by the absolute value of residuals. To estimate quantiles other than the median, we generalize the absolute loss function by weighting the absolute loss by the quantile level. One can obtain an unbiased estimate for the quantile coefficient $\beta(\tau)$ from a sample (Y_i, X_i)'s by minimizing the following objective function over all the possible βs

$$\hat{\beta}(\tau) = \arg\min_{\beta} \sum_i \rho_\tau(y_i - X_i\beta)$$

where $\rho_\tau(u) = \tau u I\{u \geq 0\} - (1-\tau)uI\{u < 0\}$ is the asymmetric absolute loss function. For example, to estimate the 90th quantile, the absolute loss function for observations above the 90th quantile would be the absolute loss multiplied by 0.9. For observations below the 90th quantile, the loss function would be the absolute loss multiplied by 0.10. As a result, there are $\tau \times 100\%$ data points above the fitted regression plane $X'\hat{\beta}(\tau)$, and $(1-\tau) \times 100\%$ observations below the regression plane.

Estimation of quantile coefficients $\hat{\beta}(\tau)$ depends on the local data sparsity around the target quantiles. In general, we have better power for estimating median and mid-range quantile levels, where data are relatively dense, while we have less power to estimate extreme quantiles at the two tails (e.g., 0.1 and 0.9) where data are sparse. Consequently, we need a large sample size to estimate the extreme quantiles well, while smaller sample sizes are sufficient to obtain a reliable estimation of the median. Since $\hat{\beta}(\tau)$ is estimated using an absolute loss function, we do not assume that the standard errors are identically distributed and thus standard Wald-type inferences (i.e., test statistics based on the ratio between the estimate and its standard error) are difficult to estimate, particularly at extreme quantile levels where there is sparse data. A few alternative inference tools have been proposed. The rank score test (inference) proposed by Gutenbrunner and colleagues [5] is often recommended due to its robust performance. Bootstrapping is another commonly used way to achieve quantile inference. Extensions to conventional bootstrapping approaches include the Markov chain marginal bootstrap (MCMB) [6] and Wild-bootstrap [7]. Using simulation methods, Kocherginsky and colleagues [8] recommended using the rank score method for small datasets (e.g., the sample size (n) is smaller than 1000, and the number of covariates (p) is smaller than 10) and using the MCMB for moderately large datasets (e. g. , $1 \times 10 \wedge 4 < np < 2 \times 10 \wedge 6$). In very large datasets, Wald-type inference tests can be used because sparse data is no longer a limiting factor.

If X_i is a binary treatment indicator (treatment or placebo, exposed or unexposed), then $\hat{\beta}(\tau) = Q_\tau(Y|\text{treatment}) - Q_\tau(Y|\text{placebo/control})$ is the quantile treatment effect, which measures the change of the τth quantile of Y due to a treatment or an exposure (compared with the control group). Figure 2 illustrates the advantage of evaluating the quantile treatment effect in comparison with the mean treatment effect (*i. e.* , $E(Y \mid \text{treatment}) - E(Y \mid \text{placebo/control})$) under three scenarios including two-condition location shift (i.e., $Y = \mu + X'\beta + e$), multiplicative scale model (i.e., $Y = \mu + (X'\gamma)e$), and location-scale models (i.e., $Y = \mu + X'\beta + (X'\gamma)e$). In the first column, the red and black curves represent the densities of Y under two conditions (e.g., exposed and unexposed), the second column is the corresponding distribution functions, and the third column is the resulting quantile effect. In panel A, the change of X only causes a location shift of the distribution/density of Y. In such case, the quantile effect $\beta(\tau) \equiv \beta$ is constant across quantile levels and is equivalent to the mean effect. In panel B, the exposure X only inflates the variance of Y but does not change the mean of Y. In such case, $\beta(\tau) \equiv \gamma F_e^{-1}(\tau)$, where $F_e^{-1}(\tau)$ is the τth quantile of the error e. As shown in the figure, we observe positive quantile effect at upper quantiles, while negative effect at lower quantiles. In panel C, the exposure changes not only the mean of Y, but also the variance. In this location-scale model, the quantile effects $\beta(\tau) \equiv \beta + \gamma F_e^{-1}(\tau)$ increase with quantile levels and are most evident at upper tails. By estimating the exposure effect across quantile levels, quantile regression provides richer information on how the exposure X impacts the outcome Y, especially for non-normal outcomes and heteroscedastic data.

Computer Syntax Given the advances in estimating the standard errors, most standard statistical packages now estimate quantile regression. Computation packages for quantile regression are readily available in R (the quantreg package), SAS (the quantreg procedure), and Stata (qreg). In the Appendix, we have included the syntax of quantile regression in both R and SAS.

Use of Linear Quantile Regression Model to Test Whether the Effects of an Exposure Are Similar Across Quantiles

In many epidemiology applications, hypotheses focus on not only the mean of the distribution but specifically how exposure might affect the tail of the distribution. Particularly with the onset of precision medicine, there is a growing recognition that exposure may operate differently depending on absolute risk of disease. We can also think of outcomes like BMI where we might hypothesize that exposure affects only one part of the distribution, but not the entire distribution as reflected by the mean of the outcome. For example, people with a high BMI are predisposed to diabetes, cancers, and many other disorders, [9] and therefore we may be most interested in

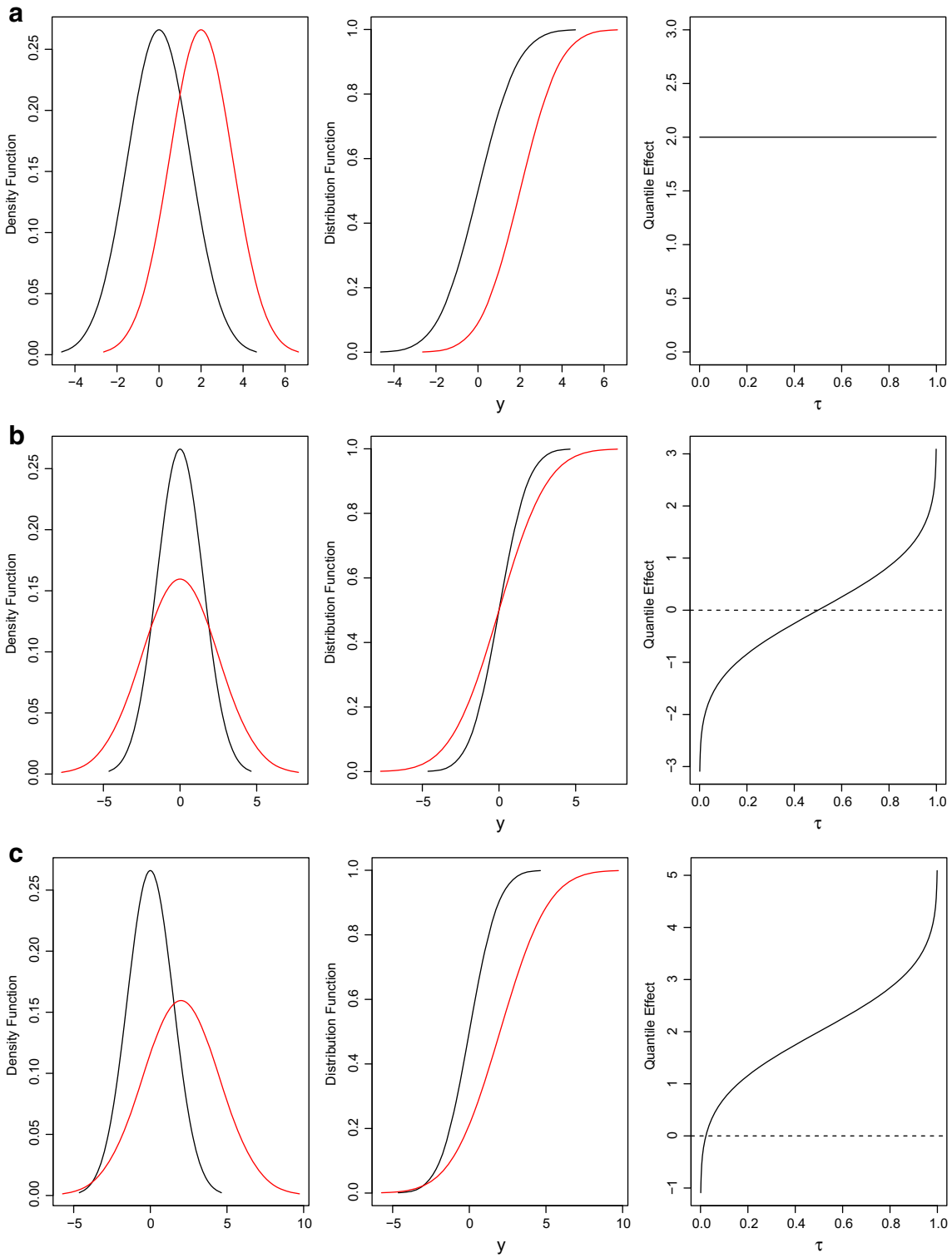

Fig. 2 The quantile effect under location, scale, and location-scale models for a binary exposure. In the first column, the red and black curves represent the densities of the outcome, *Y*, under two conditions, exposed and unexposed. The second column is the corresponding distribution functions, and the third column is the resulting quantile effect. In panel A, the exposure causes a shift in the mean of *Y*, but does not change the variance. In panel B, the exposure inflates the variances of *Y* but does not change the mean. In panel C, the exposures affect both the mean and variance of *Y*

modeling factors that drive the higher levels of BMI and not just the mean of BMI. We previously investigated pre- and postnatal influences on adult BMI [10]. Here, we take a portion of the dataset from this example to illustrate how to interpret quantile regression.

To illustrate quantile regression methods, Table 1 displays the estimated quantile coefficients across different quantile levels for three risk factors labeled X1, X2, and X3. For this example, the covariates were specifically related to maternal factors and early-life growth [10]. The coefficients from the mean model are listed in the last column of Table 1. These coefficients suggest that factors X1, X2, and X3 are all positively associated with the outcome *Y* (in this example, BMI). However, the coefficients differ across the quantile levels, and the pattern of these differences is risk factor-specific. To some extent, one can view the quantile level as a partition of a population: the coefficients at lower quantile levels represent the impact of risk factors for the subpopulation with lower BMI, while those at upper quantiles represents their effect for the group with higher BMI. For example, we observed that risk factor X1 makes little difference in the lower quantiles of adult BMI (e.g., a unit change in X1 was associated with a 0.002 decrease in 10th percentile of BMI and a 0.04 increase in the median BMI), but has a large positive association with BMI in the upper quantile (e.g., a unit change in X1 was associated with a 0.24 increase in the 90th percentile of BMI), which suggests that X1 has a substantial impact in increasing the risk of obesity, although it does not change the distribution below the median. Although it is also significant in the mean model, the population-average coefficient overestimates the association of X1 and BMI in the lower quantiles and underestimates the association in the higher quantiles, leading to an overall underestimation of the potential role of X1 on the outcome (in this case, adult BMI). In contrast, risk factor X2 has a significant impact only on the lower quantiles and not the highest quantile, a distinction that is missed in the mean model. In contrast, for risk factor X3, all quantiles are statistically significant, but even with this example the advantage of quantile regression over the mean is that the association with X3 and *Y* is monotonic across the quantiles and we get an estimate of the range with the estimate for the 90th percentile double that of the 10th percentile. The linear regression estimate for X3 still leads to similar inferences but the differences in the magnitude of estimates across quantiles are not revealed.

Table 1 Example of quantile regression for a continuous outcome

Risk factor	Quantile			
	10th	50th	90th	Mean
X1	− 0.002	0.04	0.24**	0.10**
X2	1.54**	1.85*	0.62	1.47*
X3	0.03**	0.03**	0.06*	0.05**

*p ≤ 0.05; **p ≤ 0.01

As shown in the example, quantile regression allows for heterogeneous effects across different quantile levels, which can provide useful insights into biological pathways and identify high-risk groups that may benefit most from health promotion interventions. It has a key advantage in that it can identify factors that increase the risk at both tails of the distribution; this U-shape association is likely to be missed in a linear or logistic model. The mean model assumes that the effect of a risk factor is to shift the entire distribution of the continuous outcome, as was the case for risk factor X3. However, in making this assumption, the mean model may miss or underestimate the effect of risk factors that exert their influence on the tail of distribution, such as risk factor X1. We recommend using quantile regression across multiple quantile levels to investigate complex associations. Typical choices of quantile levels include 0.1 and 0.9 plus the three quartiles (0.25, 0.5, 0.75), or 5–10 evenly spaced quantile levels. One could consider the quantile process, a data-driven approach to determine the quantile levels [11, 12].

Why Not Just Use Logistic Regression? Logistic regression models can also be used to examine high-risk groups by dichotomizing a continuous outcome. For example, for BMI, one could dichotomize the outcome based on cut points such as 25 kg/m^2 or 30 kg/m^2 to indicate overweight or obese status. One could also model multiple categories using a polytomous or multinomial regression model. However, both binary and multinomial logistic regression models will be sensitive to the cutoffs used to define high-risk.

Quantile Regression for Prediction

When there are heterogeneous effects across different quantile levels, quantile regression could serve as a better tool for risk prediction than a standard linear model. From the quantile regression model, one could estimate the conditional quantile function of the outcome given different risk/covariate profiles. For example, to construct a 90% prediction interval of *Y*, one can naturally construct the interval by the 5th and 95th conditional quantile of *Y* given *X*. Such prediction is free from distributional assumptions and is hence more flexible to fit the underlying true distributions. If there are clinically defined diagnosis cutoffs, we can easily calculate the probability of risk from the conditional quantiles and the predicted risk of multiple cutoffs can be jointly calculated. For example, Fig. 3 displays the estimated quantile functions of adult BMI given different combination categories of early-life risk factors, which we define as scenarios A, B, and C. The gray lines of 25 and 30 are clinical cutoffs for overweight and obesity. The portion exceeding these lines demonstrates the risk of

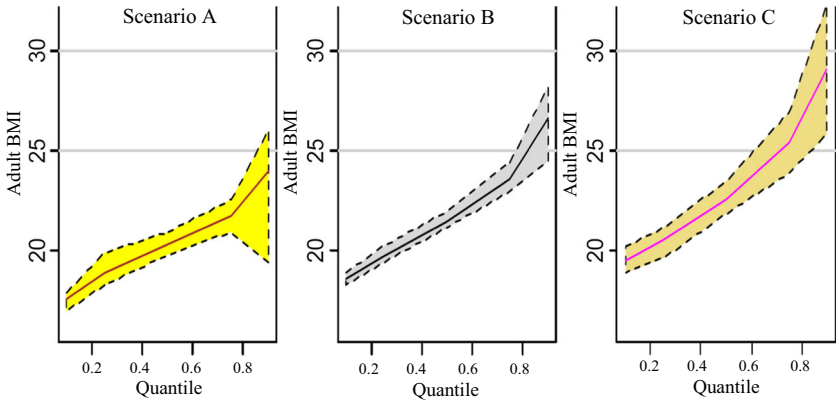

Fig. 3 Estimated quantile functions of adult BMI given three scenarios of different combination categories of early-life risk factors

overweight and obesity. As illustrated, there is a clear impact of early-life risk factors on the probability of being overweight in adulthood. Individuals with risk factors in scenario A have a very low predicted risk of being overweight in adulthood, which is illustrated by the fact that the conditional quantiles Q01-Q09 all fall below 25 kg/m^2. Conversely, individuals with early-life risk factors in scenario C have a much higher probability (approximately 30% chance) of being overweight in adulthood given that the quantile function crosses the 25 kg/m^2 mark at about Q07.

Nonparametric Quantile Regression Model for Growth Trajectories

In addition to using quantile regression to examine how exposures are associated with a continuous outcome at one point in time, quantile regression methods can also be used to examine repeated measures of continuous outcomes and investigate whether quantile-specific exposure effects vary across time. Nonparametric quantile regression has shown to be a powerful tool to model growth trajectories and construct pediatric growth charts [13, 14].

Using BMI trajectories as an example, we denote Y_t as the BMI measured at age t, and then expand the linear quantile model to

$$Q_\tau(Y_t|X) = g_\tau(t) + X'\beta_\tau(t)$$

where $Q_\tau(Y_t \mid X)$ is τth quantile of BMI at age t and covariate profile X.

If, for example, X is a binary indicator of a particular exposure, then $g_\tau(t)$ is the $\tau \times 100$ percentile curve of BMI for subjects without the exposure while $\beta_\tau(t)$ is the change of the $\tau \times 100$ percentile curve of BMI for those with the exposure. Due to the nature of growth trajectories, we assume that $g_\tau(t)$ and $\beta_\tau(t)$ are nonparametric functions with certain smoothness to avoid pre-specifying the shape of trajectories. To estimate $g_\tau(t)$ and $\beta_\tau(t)$, one can use spline approximation [15•]. In theory, any smooth functions can be well approximated by a linear combination of B-splines basis functions.

As an example application of nonparametric quantile regression modeling, we used this method to examine the influence of the early-life risk factors on growth trajectories from birth to 7 years in the National Collaborative Perinatal Project (NCPP) [15•]. This was an expansion of previous NCPP analyses that used linear quantile regression to examine pre- and postnatal influences on body size at specific ages [10, 16, 17•]. We also point the reader to work by Briollais and Durrieu (2017) looking at the effect of the genetic variant in the TCF4 gene on childhood BMI in the Western Australian Pregnancy Cohort (RAINE) as another example of nonparametric quantile regression modeling of growth trajectories [18•]. In this example, Briollais and Durrieu [18•] investigated the effect of the genetic variant in the TCF4 gene on the childhood growth trajectories. In the Appendix, we have provided R codes for nonparametric quantile splines.

Quantile Regression for Intermediate Markers Within Studies that Use Case–Control Studies

Epidemiology studies often employ case–control designs, where researchers sample cases and controls from a disease population and a disease-free population, respectively. It is a cost-effective way to identify disease-associated risk factors, especially for rare diseases. In addition to the disease status, case–control studies also collect secondary outcomes/disease biomarkers. For example, in a case–control study for diabetes, glucose levels and BMI are often included. Here, we discuss the use of quantile regression for these secondary analyses of intermediate markers such as glucose levels or BMI. These intermediate markers are often continuous but for proper inference, we need to consider the original case–control sampling methods. Analyses of intermediate markers within cohort studies do not face this same empirical challenge if the

sampling is based on outcome as with a case–control study. With analyses of continuous intermediate markers within a case–control study, we can conduct quantile-based analyses using inverse probability weighting (IPW) and likelihood-based approaches. The main idea of IPW is to reweight individual observations by the inverse of their selection probabilities and conduct weighted regressions. IPW is simple and can be easily applied to quantile regression. However, it requires information on the case–control sampling scheme to construct the weights and is often biased when there are unobserved confounders. Case–control studies may also be used to jointly model the primary binary disease status and secondary continuous outcomes. The semiparametric maximum likelihood (SPML) proposed by Lin and Zeng [19] is the most widely recognized approach in this group as it largely improves the efficiency over IPW. Since quantile regression does not assume any parametric likelihood, likelihood-based approaches cannot be applied to quantile regression. Based on the concept of counterfactual outcomes, Wei and colleagues [20] proposed weighted estimating equations (WEE) to estimate the conditional quantile of secondary outcomes in a case–control study. Suppose $S_\tau(x, y, \beta)$ is the original quantile regression score function, we can then construct the unbiased estimating equations by

$$\sum_i S\_\tau(x_i, y_i, \beta)p(d_i|x_i) + S\_\tau(x_i, \tilde{y}_i, \beta)p(1-d_i|x_i) = 0$$

where y_i (e.g., BMI) is observed the secondary outcome of the ith subject in the sample, d_i is the binary indicator for case/control status, x_i is the exposure level, and $\tilde{y}i$ is counterfactual secondary outcomes $\tilde{y}i$ under the alternative disease status but the same exposure x_i .The weight $p(d_i \mid x_i)$ is the probability of being the observed disease status given exposure x_i, and $p(1 - d_i \mid x_i)$ is the probability of being the counterfactual disease status.

The counterfactual secondary outcomes are not observed but can be simulated from stratified quantile process modeling. The WEE is computationally simple and straightforward. The computation packages for the WEE approach using linear regression, logistic regression, and quantile regression are available on github (https://github.com/songxiaoyu/secondary-analysis-in-case-control-studies) together with sample codes. As an example, this approach was applied to an asthma case–control GWAS from the New York University Bellevue Asthma Registry [21] to identify the association between the TSLP gene and IgE. This analysis led to the discovery of new SNPs that would have been missed if only mean-based analysis had been used [22]. Some SNPs only affect the upper tails of IgE. Consequently, the distributions of IgE between two genotypes coincided with one another except at the upper tail of the distribution. The mean effect essentially average the differences between the two distribution/quantile functions and consequently underestimated and overlooked the tail differences [22].

Conclusions

Quantile regression is now a readily accessible regression method that we believe has many different applications within epidemiology. In particular, when modeling continuous outcome data, either within a cohort, or as an intermediate marker within case–control designs, quantile regression does not require that the outcome data are normally distributed. Quantile regression has the flexibility of testing first if a given exposure differs by percentile of the outcome. If not, then it is a great launching point to support the use of linear regression method. However, when one starts with linear regression first for continuous outcomes without using quantile regression methods, it should be more broadly recognized that select associations with specific quantiles of Y may be missed. Thus, quantile regression can inform whether or not general linear regression can be used to correctly model exposure–outcome associations. Given these strengths, we recommend epidemiologists and population health scientists consider the use of quantile regression in future analyses of continuous outcomes, at least as a first step to inform future modeling decisions.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no potential conflicts of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

Appendix

Table 2. Computational packages and basic syntax of quantile regression

Software	packages
R	R package -- *quantreg*
SAS	SAS/STAT PROC QUANTREG
STATA	https://www.stata.com/features/overview/quantile-regression/
Matlab	Codes available at http://www.econ.uiuc.edu/roger/research/rq/rq.m

Basic Quantile Regression Syntax in R

> install.packages("quantreg").

> library(quantreg).

> fit = rq(y~×1 + ×2, tau = .5, data = data).

Note: tau is the quantile level(s) of interest. It could a single value for a fixed quantile level, or a vector of quantile levels,

tau = c(0.25, 0.5, 0.75). The function rq() will return regression quantiles from multiple quantiles. If tau is smaller than 0 or larger than 1, the function will return the entire quantile process.

Basic Quantile Regression Syntax in SAS

PROC QUANTREG.
DATA = sas-data-set;
CLASS X1;
MODEL Y = X1 X2 / QUANTILE = 0.25 0.5 0.75;
RUN;

Note: if the option QUANTILE = ALL, it returns the entire quantile process. Same as in R, the default value is 0.5, corresponding to the median.

Statistical Inference of Quantile Regression in R and SAS

To obtain statistical inference of quantile regression in R, we need to use the function *summary.rq*(object, se = "nid", ...), where object is the returned object from the function *rq*(), and the parameter *se* specify the inference methods. In SAS, the inference options are specified at the PROC QUANTREG Statement following the syntax "PROC QUANTREG CI= <NONE|RANK|...> ALPHA = value ;" where ALPHA is the significance level, and CI specifies the choice of inference. The table below lists the available methods in R and SAS.

		Options	
Inference method	Subcategories	R	SAS
Direct	i.i.d. model	se =" iid"	CI = SPARCITY/IID
	n.i.d. model	se = "nid"	CI = SPARCITY
Rank Score		se =" rank"	CI = RANK
resampling	Pairwise	se =" boot", bsmethod = "xy"	Not available
	Parzen, Wei and Ying	se =" boot", bsmethod = "pxy"	Not available
	MCMB	se =" boot", bsmethod = "mcmb"	CI = RESAMPLING
	Wild	se =" boot", bsmethod = "wild"	Not available

R script for the nonparametric quantile regression for growth trajectories with B-spline approximation

```
#The functions to fit quantile b spline

# ---------- R function to fit quantile bsplines---------#
#   *input*                                                  #
#      y: response variable                                  #
#      x: independent variable                               #
#    tau: a vector of quantiles, default is the median.      #
#   knots: internal knots                                    #
#    ord: order of spline, default is 3.                     #
# ------------------------------------------------------------#
qb=function(y,x,tau=0.5,knots,ord=3){

   require(splines)
   require(quantreg)

   boundry.knots=range(x,na.rm=T)
   knots=c(rep(boundry.knots[1],ord),knots,rep(boundry.knots[2],ord))
   xmodel=x[!is.na(x) & !is.na(y)]
   ymodel=y[!is.na(x) & !is.na(y)]

   Xm=splineDesign(knots,xmodel,ord)
   k=length(tau)
   n=length(knots)-ord
   coef=matrix(,nrow=k,ncol=n)

   for(i in 1:k)
   {
      fit=rq(ymodel~Xm-1,tau=tau[i],method="fn")
      coef[i,]=fit$coefficients
   }
   return(list(coef=coef,knots=knots,ord=ord,tau=tau,x=x,y=y))
}



# ------ R function to plot the fitted quantile bsplines ---------#
#   *input*                                                  #
#      fit: object from qb()                                 #
# ------------------------------------------------------------ - ------#

plot.qb=function(fit,lty=1,col=1,xlab=NULL,ylab=NULL){

    xlim=range(fit$x,na.rm=T); xlim[2]=xlim[2]+diff(xlim)/20
    plot(fit$x,fit$y,col="gray",xlab=xlab,ylab=ylab,xlim=xlim)
    x.pred=seq(min(fit$x,na.rm=T),max(fit$x,na.rm=T),,1000)
    xm.pred=splineDesign(fit$knots,x.pred,ord=fit$ord)
    y.pred=xm.pred%*%t(fit$coef)
    matlines(x.pred,y.pred,lty=lty,col=col)
    text(max(fit$x,na.rm=T)+diff(xlim)/40,y.pred[1000,],fit$tau,cex=0.6)

}



# --------- R function to estimate the quantile bsplines ----------------------#
#   *input*                                                                    #
#      fit: object from qb()                                                   #
#      new.x: estimate quantiles at new.x                                      #
# ---------------------------------------------------------------------------------#


pred.qb=function(fit,new.x){

   require(splines)
   Xm=splineDesign(fit$knots,new.x,fit$ord)
   pred.y=Xm%*%t(fit$coef)  ###each column represent predicted y at a specific tau
   return(list(pred=pred.y,new.x=new.x))

}


Example
>library(splines)
>library(quantreg)
> x = runif(500, 0, 3)
> y = log(x+5)/(3*x+1)
> knots = c(0.5,1,1.5,2);
> tau=c(0.03,0.1,0.25,0.5,0.75,0.9,0.97)
> fit = qb(y,x,tau=tau,knots=knots,ord=4)
>plot(fit)
>  x.pred = seq(0,2,,1000)
>  y.pred.un =pred.qb(fit,x.pred)$pred
```

References

Papers of particular interest, published recently, have been highlighted as:
• Of importance

1. Yang J, Loos RJ, Powell JE, Medland SE, Speliotes EK, Chasman DI, et al. FTO genotype is associated with phenotypic variability of body mass index. Nature. 2012;490(7419):267–72. https://doi.org/10.1038/nature11401.
2. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science. 2007;316(5826):889–94. https://doi.org/10.1126/science.1141634.
3. Koenker R, Bassett G Jr. Regression quantiles. Econometrica. 1978;46:33–50.
4. Zhao Z, Xiao Z. Efficient regressions via optimally combining quantile information. Economic Theory. 2014;30(6):1272–314. https://doi.org/10.1017/s0266466614000176.
5. Gutenbrunner C, Jurečková J, Koenker R, Portnoy S. Tests of linear hypotheses based on regression rank scores. J Nonparametr Stat. 1993;2(4):307–31.
6. He X, Hu F. Markov chain marginal bootstrap. J Am Stat Assoc. 2002;97(459):783–95.
7. Feng X, He X, Hu J. Wild bootstrap for quantile regression. Biometrika. 2011;98(4):995–9.
8. Kocherginsky M, He X, Mu Y. Practical confidence intervals for regression quantiles. J Comput Graph Stat. 2005;14(1):41–55.
9. Hjartåker A, Langseth H, Weiderpass E. Obesity and diabetes epidemics. In: Innovative Endocrinology of Cancer: Springer; 2008. p. 72–93.
10. Terry MB, Wei Y, Esserman D. Maternal, birth, and early-life influences on adult body size in women. Am J Epidemiol. 2007;166(1):5–13. https://doi.org/10.1093/aje/kwm094.
11. Koenker RW, D'Orey V, Algorithm AS. 229: computing regression quantiles. J R Stat Soc: Ser C: Appl Stat. 1987;36(3):383–93. https://doi.org/10.2307/2347802.
12. Koenker R, d'Orey V, Remark AS. R92: a remark on algorithm AS 229: computing dual regression quantiles and regression rank scores. J R Stat Soc: Ser C: Appl Stat. 1994;43(2):410–4.
13. Wei Y, Pere A, Koenker R, He X. Quantile regression methods for reference growth charts. Stat Med. 2006;25(8):1369–82.
14. Wei Y, He X. Conditional growth charts. Ann Stat. 2006;34(5):2069–97.
15.• Wei Y, Ma X, Liu X, Terry MB. Using time-varying quantile regression approaches to model the influence of prenatal and infant exposures on childhood growth. Biostat Epidemiol. 2017;1(1):133–47. https://doi.org/10.1080/24709360.2017.1358137. **This is a paper that shows how to do repeated measures analysis with quantile regression.**
16. Terry MB, Wei Y, Esserman D, McKeague IW, Susser E. Pre- and postnatal determinants of childhood body size: cohort and sibling analyses. J Dev Orig Health Dis. 2011;2(2):99–111. https://doi.org/10.1017/s2040174411000067.
17.• Ester WA, Houghton LC, Lumey LH, Michels KB, Hoek HW, Wei Y, et al. Maternal and early childhood determinants of women's body size in midlife: overall cohort and sibling analyses. Am J Epidemiol. 2017;185(5):385–94. https://doi.org/10.1093/aje/kww222. **This analysis updated quantile-specific results from 2007 showing the association between maternal BMI and gestational weight gain and offspring BMI persists through midlife.**
18.• Briollais L, Durrieu G. Quantile regression for genetic and genomic applications. In: Handbook of quantile regression: Chapman and Hall/CRC; 2017. p. 409–27. **This paper is an example of applying quantile regression to genetic data.**
19. Lin D, Zeng D. Proper analysis of secondary phenotype data in case-control association studies. Genet Epidemiol. 2009;33(3):256–65.
20. Wei Y, Song X, Liu M, Ionita-Laza I, Reibman J. Quantile regression in the secondary analysis of case–control data. J Am Stat Asoc. 2016;111(513):344–54.
21. Liu M, Rogers L, Cheng Q, Shao Y, Fernandez-Beros ME, Hirschhorn JN, et al. Genetic variants of TSLP and asthma in an admixed urban population. PLoS One. 2011;6(9):e25099.
22. Song X, Ionita-Laza I, Liu M, Reibman J, We Y. A general and robust framework for secondary traits analysis. Genetics. 2016;202:1329–43.