

# Mid-quantile regression for discrete responses

Statistical Methods in Medical Research

2022, Vol. 31(5) 821–838

© The Author(s) 2022

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/09622802211060525

[journals.sagepub.com/home/smm](https://journals.sagepub.com/home/smm)Marco Geraci<sup>1,2</sup>  and Alessio Farcomeni<sup>3</sup> 

## Abstract

We develop quantile regression methods for discrete responses by extending Parzen's definition of marginal mid-quantiles. As opposed to existing approaches, which are based on either jittering or latent constructs, we use interpolation and define the conditional mid-quantile function as the inverse of the conditional mid-distribution function. We propose a two-step estimator whereby, in the first step, conditional mid-probabilities are obtained nonparametrically and, in the second step, regression coefficients are estimated by solving an implicit equation. When constraining the quantile index to a data-driven admissible range, the second-step estimating equation has a least-squares type, closed-form solution. The proposed estimator is shown to be strongly consistent and asymptotically normal. A simulation study shows that our estimator performs satisfactorily and has an advantage over a competing alternative based on jittering. Our methods can be applied to a large variety of discrete responses, including binary, ordinal, and count variables. We show an application using data on prescription drugs in the United States and discuss two key findings. First, our analysis suggests a possible differential medical treatment that worsens the gender inequality among the most fragile segment of the population. Second, obesity is a strong driver of the number of prescription drugs and is stronger for more frequent medications users. The proposed methods are implemented in the R package *Qtools*. Supplemental materials for this article, including a brief R tutorial, are available as an online supplement.

## Keywords

Conditional CDF, health care, kernel estimator, maximum score estimation, National Health and Nutrition Examination Survey

## 1 Introduction

In its classical formulation,<sup>1</sup> quantile regression (QR) provides a distribution-free approach to the modelling and estimation of quantile treatment effects<sup>2</sup> (QTEs) for *continuous* response variables. QR has become a successful analytic method in many fields of science because of its ability to draw inferences about individuals that rank below or above the population conditional mean. The ranking within the conditional distribution of the outcome can be considered as a natural index of individual latent characteristics which cause heterogeneity at the population level.<sup>3</sup> The value of estimating QTEs in medical and public health research has been illustrated in several studies.<sup>4–11</sup>

While most of the progress in QR methods has revolved around continuous responses, relatively less contributions have been made in the discrete case so far. Discrete response variables are ubiquitous in medical research and the literature is arguably dominated by generalized linear models.<sup>12</sup> The reasons are manifold and include convenient interpretation of the regression coefficients, e.g., as (log) odds ratios in logistic regression or (log) rate ratios in Poisson regression, universal

<sup>1</sup>MEMOTEF Department, Sapienza University of Rome, Rome Italy

<sup>2</sup>Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, USA

<sup>3</sup>Department of Economics and Finance, University of Rome "Tor Vergata" Rome, Italy

## Corresponding author:

Marco Geraci, MEMOTEF Department, School of Economics, Sapienza University of Rome, Via del Castro Laurenziano 9, Rome 00161, Italy. marco.geraci@uniroma1.it

availability in statistical software, and the benefits of a well-developed, unifying maximum likelihood theory. However, research has been increasingly directed toward the development of nonparametric (distribution-free) methods to overcome situations in which traditional approaches are unsatisfactory or, more in general, when the goal of the inference transcends the conditional mean of the response.

QR for discrete responses presents some hurdles. Major hindrances include lack of a general theory for handling different types of discreteness, practical estimation challenges, and the troublesome asymptotic behaviour of sample quantiles in the presence of ties. Thus, it is not surprising that existing approaches to discrete QR rely on some notion of continuity, either postulated or artificially induced. Early works in the former category date back to the 1950s.<sup>13</sup>

Prominent in the econometric literature, maximum score estimation deals with conditional median models of binary<sup>14–16</sup> and ordered discrete<sup>17</sup> response variables. More recently, Kordas<sup>18</sup> extended Horowitz's<sup>16</sup> estimator for binary outcomes to quantiles other than the median. In the maximum score estimation approach, the key assumption is the existence of a continuous latent variable, say  $Y^*$ , which undergoes the working of a threshold mechanism resulting in the observable binary outcome  $Y = I(Y^* > 0)$ , where  $I(\cdot)$  denotes the indicator function. The conditional quantiles of the observable outcome are obtained as transformed quantiles of the latent outcome. However, maximum score estimation is computationally expensive as it involves nonconvex loss functions. Jittering is another strategy used for quantile estimation with discrete responses. A (pseudo)continuous variable, say  $Z$ , is obtained by adding random noise, say  $U$ , to the observable discrete outcome, i.e.  $Z = Y + U$ . Estimation then proceeds by applying standard algorithms for convex quantile loss functions (e.g., linear programming) and, successively, by averaging the noise out. This approach, which has been adopted for modelling count<sup>19</sup> and ordinal<sup>20</sup> response variables, may lack generality as it requires that adjacent values in the support of  $Y$  are equally spaced. Another estimation method for quantiles of count data has been recently proposed.<sup>21</sup> An approximation to continuity is introduced in the sense that the response becomes  $Z_E = Y + E(U)$ , where  $E(U) = 0.5$ . As explained by the authors, the resulting estimator is asymptotically equivalent whether it is applied to  $Z_E$  or the jittered response  $Z$  as defined above. Efficiency gains of this approach derive from the parametric modelling of a sequence of quantiles, but they are offsetted by the risk of overfitting<sup>21</sup> and the inability to model a single quantile.

For the sake of completeness, we briefly mention possible alternatives. Recently, Chernozhukov et al<sup>22</sup> proposed inference methods to construct simultaneous confidence bands for quantile and quantile effect functions of possibly discrete random variables. However, their study does not provide any strategy for regression modelling or point estimation. Other proposed methods for estimating non-central summaries of discrete responses include M-QR.<sup>23,24</sup> Unfortunately, conditional M-quantiles suffer from a lack of interpretability as their relationship with conditional quantiles is obscure. They also depend on global properties of the parent distribution, despite they attempt to describe a local property of such distribution.<sup>25</sup> Another class of models that may be used to estimate conditional quantiles can be collectively referred to as distributional regression. Whether distributional regression is used within a non-, semi- or fully parametric framework, the common goal is to model the conditional distribution function (CDF) as flexibly as possible by means of covariate-dependent distributional parameters. For example, covariates may enter into the model via parameters related to the location, scale, and shape of the distribution.<sup>26</sup> While the conditional quantile function (QF) can be obtained by inverting the conditional CDF (though in general this is not guaranteed to yield an analytical expression), the former will typically depend on the regression parameters in a complicated (often implicit and non-linear) fashion. That is, distributional regression parameters may not give a simple and straightforward summary of the QTE, in contrast to the parameters of QR models.<sup>2</sup> Finally, in the hope of remedying a misconception that has crept into the literature of QR, we note that the method proposed by Bottai et al,<sup>27</sup> despite being called 'logistic quantile regression', is applied to *continuous* responses only and has nothing to do with conditional quantiles of binary responses.

In this paper, we build on mid-quantiles<sup>28</sup> to introduce an alternative estimation approach for conditional quantiles of discrete responses. Sample mid-quantiles, which are based on essentially the same idea of the mid- $p$ -value,<sup>29</sup> offer a unifying theory for quantile estimation with continuous or discrete variables and are well-behaved asymptotically.<sup>30</sup> In our approach, we develop a two-step estimator that can be applied to a large variety of discrete responses, including binary, ordinal, and count variables, and is shown to have good theoretical properties. In a simulation study, we gather empirical evidence that conditional mid-quantile estimation is more efficient than jittering. However, this evidence is contextual as is based on a single simulated scenario and may not be generalizable.

The rest of the paper is organized as follows. In the next section, we discuss modelling, estimation, and theoretical properties of conditional mid-quantile estimators, with technical details on inference given in the online Supplemental material. In Section 3, we report the results of a simulation study to assess bias and efficiency of the proposed estimator, as well as confidence interval coverage. We also illustrate an application to data on prescription drugs usage in the United States. We conclude with final remarks in Section 4.

## 2 Methods

### 2.1 Marginal mid-quantiles

Let  $Y$  be a discrete random variable with probability mass function  $m_Y(y) = \Pr(Y = y)$  and cumulative distribution function (CDF)  $F_Y(y) = \sum_{u \leq y} m_Y(u)$ . The  $p$ th quantile of  $Y$ , denoted by  $\xi_p$ , is defined as  $\xi_p \equiv \inf\{y \in \mathbb{R} : F_Y(y) \geq p\}$  for any  $0 < p < 1$ . We may define the QF of  $Y$ ,  $Q_Y(p)$ , as the generalized inverse of the CDF of  $Y$ , that is  $Q_Y(p) \equiv F_Y^{-1}(p)$ . In the discrete case, the CDF is not injective, thus a discrete QF is not the standard inverse of the CDF. Now let  $Y_1, Y_2, \dots, Y_n$  be an independent sample of size  $n$  from the population  $F_Y$ . The sample CDF is defined as  $\hat{F}_Y(y) = n^{-1} \sum_{i=1}^n I(Y_i \leq y), y \in \mathbb{R}$ , while the (ordinary) sample QF, defined as the inverse of the sample CDF (see for example Hyndman and Fan<sup>31</sup> for a detailed overview of alternative sample quantiles). The sample QF, too, is discrete.

In general, sample quantiles as defined above may not be consistent for the population quantiles when the underlying distribution is discrete.<sup>32</sup> Additionally, the sample median  $\hat{Q}_Y(0.5) = Y_{(\lfloor n/2 \rfloor)}$  lacks asymptotic normality if  $Y$  is discrete.<sup>33</sup> Throughout this article, we use  $\lceil x \rceil$  ( $\lfloor x \rfloor$ ) to denote the smallest (largest) integer that is larger (smaller) than or equal to  $x$ .

We now introduce the mid-CDF,<sup>28,34</sup> a modification of the standard CDF that plays an important role in discrete modelling and in samples with ties. For a random variable  $Y$  with CDF  $F_Y(y)$ , the function

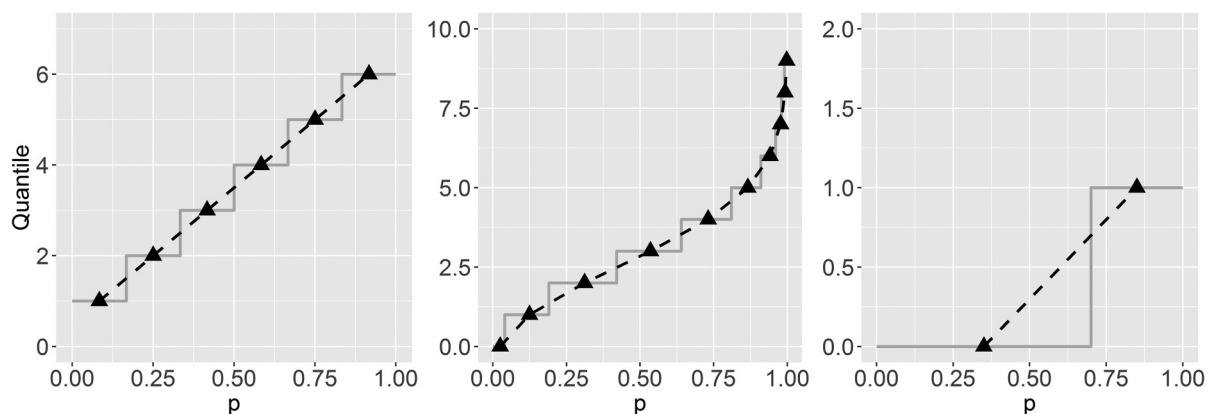
$$G_Y(y) \equiv \Pr(Y \leq y) - 0.5 \cdot \Pr(Y = y) \tag{2.1}$$

is called mid-distribution function (mid-CDF). Since  $Y$  is discrete,  $G_Y(y)$  is a step function (a downward-shifted version of  $F_Y(y)$ ). Note that, if  $Y$  were instead continuous, then  $G_Y(y)$  would reduce to  $F_Y(y)$  since, in that case,  $\Pr(Y = y) = 0$ .

Further, let  $\mathcal{S}_Y = \{y_1, \dots, y_s\}$ , with  $y_j < y_{j+1}$  for all  $j = 1, \dots, s - 1$ , be the set of  $s$  distinct values in the population that the discrete random variable  $Y$  can take on, with corresponding probabilities  $p_1, \dots, p_s$ . We also define the mid-probabilities  $\pi_1 = p_1/2$  and  $\pi_j = G(y_j) = \sum_{u=1}^{j-1} p_u + p_j/2$ , for  $j = 2, \dots, s$ . The following function

$$H_Y(p) = \begin{cases} y_1, & \text{if } p < \pi_1 \\ y_j, & \text{if } p = \pi_j, j = 1, \dots, s \\ (1 - \gamma)y_j + \gamma y_{j+1}, & \text{if } p = (1 - \gamma)\pi_j + \gamma\pi_{j+1} \\ & 0 < \gamma < 1, j = 1, \dots, s - 1 \\ y_s, & \text{if } p > \pi_s \end{cases} \tag{2.2}$$

is called mid-QF.<sup>30</sup> If  $s = \infty$ , then the last category is suppressed. Examples of  $H_Y(p)$  when  $Y$  is discrete uniform, count, or binary are given in Figure 1. The mid-QF is piecewise linear and connects the points  $(\pi_j, y_j)$  (dashed lines in Figure 1). One can verify that  $H_Y\{G_Y(y_j)\} = y_j, j = 1, \dots, s$ . In general, mid-quantiles cannot be obtained by inverting  $G_Y(y)$  at points  $y \notin \mathcal{S}_Y$ . However, we can define  $G_Y^c(y)$ , the continuous version of  $G_Y(y)$ ,<sup>34</sup> as the piecewise linear function that connects the values  $G_Y^c(y_j) = \pi_j, j = 1, \dots, s$ , and satisfies  $G_Y^c(y) \equiv H_Y^{-1}(y)$ , for all  $y \in \mathbb{R}$ . Related to this, we have the equivariance property  $h^{-1}\{H_h(Y)(\pi_j)\} = H_Y(\pi_j), j = 1, \dots, s$ , for a monotone transformation  $h$ . Equivariance of  $H_Y(p)$  no longer applies if  $H_Y(\pi_j) < p < H_Y(\pi_{j+1})$  unless  $h$  is linear.



**Figure 1.** True quantile function (grey solid line) and mid-quantiles (black filled triangles). Left: Discrete uniform on (1,6). Centre: Poisson with mean 3. Right: Bernoulli with probability 0.3.

The sample mid-CDF corresponding to (2.1) is  $\hat{G}_Y(y) = \hat{F}_Y(y) - 0.5 \cdot \hat{m}_Y(y)$ , where  $\hat{m}_Y(y) = n^{-1} \sum_{i=1}^n I(Y_i = y)$  is the sample relative frequency of  $y$ . To define the sample mid-quantiles, we need to introduce some more notation. Let  $z_j$ ,  $j = 1, \dots, k$ , be  $k$  distinct values that occur in the sample, with  $z_j < z_{j+1}$  for all  $j = 1, \dots, k-1$ , and let  $\hat{p}_j = \hat{m}_Y(z_j)$ ,  $j = 1, \dots, k$ , be the corresponding relative frequencies. Then  $\hat{H}_Y(p) = (1 - \gamma)z_j + \gamma z_{j+1}$ , where the index  $j = 1, \dots, k-1$  is such that  $p = (1 - \gamma)\hat{G}_Y(z_j) + \gamma\hat{G}_Y(z_{j+1})$ ,  $0 \leq \gamma \leq 1$ , and  $\hat{G}_Y(z_j) = \hat{F}_Y(z_j) - 0.5 \cdot \hat{p}_j$ . Moreover,  $\hat{H}_Y(p) = z_1$  if  $p < \hat{G}_Y(z_1)$  and  $\hat{H}_Y(p) = z_k$  if  $p > \hat{G}_Y(z_k)$ . A natural estimator of  $G_Y^c(y)$  is then  $\hat{G}_Y^c(y) = (1 - \gamma)\hat{\pi}_j + \gamma\hat{\pi}_{j+1}$  for  $z_j \leq y \leq z_{j+1}$ .

In samples with ties,  $\hat{H}_Y(p)$  is the piecewise linear function connecting the values  $\hat{H}_Y\{\hat{G}_Y(z_j)\} = \hat{H}_Y\{\hat{G}_Y^c(z_j)\} = z_j$ . It has been shown that if the underlying distribution  $F_Y$  is absolutely continuous, then the sample mid-quantiles have the same asymptotic properties as the ordinary sample quantiles.<sup>30</sup> More importantly, if  $F_Y$  is discrete, then the sample mid-quantiles are consistent estimators of the population mid-quantiles and their sampling distribution is normal.<sup>30</sup>

As far as interpretation goes, mid-quantiles can be viewed as fractional order statistics.<sup>35,33</sup> A mid-quantile is a quantile of  $Y$  when  $p = \pi_j$ ,  $j = 1, \dots, k$ . Otherwise, when  $p \neq \pi_j$ , its interpretation involves an underlying form of continuity that captures the smooth progression from one quantile to the next. It is in this spirit that Wang and Hutson<sup>36</sup> proposed smooth quantiles for discrete distributions based on fractional-order statistics. Compared with ordinary sample quantiles, it can be argued that mid-quantiles offer a sensible approach to quantifying the differences between discrete distributions. For example, consider two binary samples with different proportions of successes,  $\{0, 0, 0, 0, 1\}$  and  $\{0, 0, 0, 1, 1\}$ .<sup>30</sup> The sample median is 0 in both cases; the sample mid-median corresponds to the proportion of 1's, that is,  $1/5$  in the former sample, but  $2/5$  in the latter. In this example, the mid-median helps discriminate between the two samples in an intuitive way (as formally shown in Section 2.2, our proposed conditional mid-median inherits the same properties).

Another advantage of mid-quantiles is that they can be easily relabelled to obtain the (ordinary) quantiles. Suppose  $\hat{H}_Y(p) = z_j$  and the goal is to recover the sample quantile  $\hat{\xi}_p$ . First, the mid-probabilities are obtained from the inversion  $\hat{H}_Y^{-1}(z_j) = \hat{G}_Y(z_j)$ , while the sample CDF is calculated recursively from  $\hat{F}(z_j) = 2\hat{G}_Y(z_j) - \hat{F}(z_{j-1})$ ,  $j = 1, \dots, k$ , with the convention that  $\hat{F}(z_0) = 0$ . This leads to  $\hat{\xi}_p = z_j$  for  $\hat{F}(z_{j-1}) < p \leq \hat{F}(z_j)$ ,  $j = 1, \dots, k$ . Otherwise, if  $\hat{H}_Y(p) = z \neq z_j$ , then one takes  $\hat{H}_Y^{-1}(z_{j^*})$  where  $z_{j^*}$  is the largest of the  $z_j$ 's that satisfy  $z > z_j$ . This is an important property that sets mid-quantiles apart from so-called 'quantile-like' alternatives like expectiles. The latter have a non-trivial relationship with quantiles<sup>37</sup> and require non-trivial relabelling procedures that may lead to poor approximations on the tails.<sup>25</sup> M-quantiles, which represent a generalization of expectiles, suffer from a similar limitation.

## 2.2 Conditional mid-quantiles

Analogously to (2.1), we define the *conditional* mid-CDF as

$$G_{Y|X}(y|x) \equiv F_{Y|X}(y|x) - 0.5 \cdot m_{Y|X}(y|x) \quad (2.3)$$

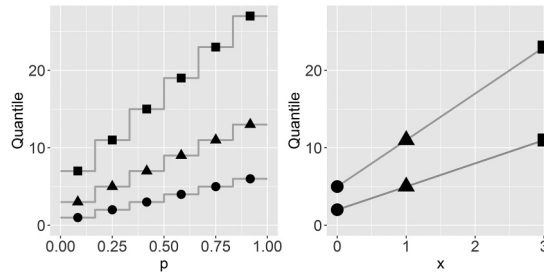
where  $Y$  is a random variable with image  $\mathcal{S}_Y \subset \mathbb{R}$  and  $X$  is a  $q$ -dimensional vector of covariates, which may include a constant equal to one for the intercept,  $F_{Y|X}(y|x) = \Pr(Y \leq y|X = x)$ , and  $m_{Y|X}(y|x) = \Pr(Y = y|X = x)$ . Although the definition of conditional mid-CDF applies to both continuous and discrete response variables (as in the case of the marginal mid-CDF (2.1)), we assume that  $\mathcal{S}_Y = \{y_1, \dots, y_s\}$  is a finite or countably infinite ( $s = \infty$ ) subset of  $\mathbb{R}$ . In particular,  $Y$  can be binary, ordinal, or count, with positive or negative values, not necessarily equally spaced. Values need not be integers either. However, we do exclude nominal variables with more than two categories from the application of (2.3) as they do not have a natural ordering.

We use the conditional mid-CDF (2.3) as the springboard for defining conditional mid-quantiles. Let  $\pi_j = G_{Y|X}(y_j|x)$  (for the sake of simplicity, we have suppressed the dependence on  $x$  from  $\pi_j$ 's notation). The *conditional* mid-QF  $H_{Y|X}(p)$  is defined as the piecewise linear connecting the values  $G_{Y|X}^{-1}(\pi_j|x)$ ,  $j = 1, \dots, s$ , for given  $x$ . We assume a quantile-specific model that is linear on the scale of  $h$ , i.e.,

$$H_{h(Y)|X}(p) = x^\top \beta(p) \quad (2.4)$$

where  $h$  is a known monotone and differentiable 'link' function, and  $\beta(p)$  is a vector of  $q$  unknown regression coefficients for a given  $p \in (0, 1)$ . In our approach,  $h$  may simply be the identity or a linear transformation, the logarithmic function – which is typically used in the modelling of counts,<sup>19</sup> the logistic function, or belong to a family of flexible transformation models.<sup>38–41</sup> These often involve the Box-Cox<sup>42</sup> or Aranda-Ordaz<sup>43</sup> families. As in the marginal case, conditional mid-quantiles cannot be obtained by inverting  $G_{Y|X}(y|x)$  at points  $y \notin \mathcal{S}_Y$ . Again, we define the piecewise linear function  $G_{Y|X}^c(y|x)$  that connects the values  $G_{Y|X}(y_j|x) = \pi_j$ ,  $j = 1, \dots, s$ , and satisfies  $G_{Y|X}^c(y|x) \equiv H_{Y|X}^{-1}(y|x)$ , for all  $y \in \mathbb{R}$ .

The QTE interpretation of the  $j$ th coefficient  $\beta_j(p)$  is immediate if  $x_j$  is discrete. Otherwise, it can be defined in terms of the partial derivative of  $H$ . For example, if  $h$  is linear and there are no additional terms that depend on  $x_j$  (e.g., interactions,



**Figure 2.** True quantile function (grey solid line) and mid-quantiles (black filled circles, triangles and squares). Left: Quantiles as a function of  $p$  by values of  $x \in \{0, 1, 3\}$ . Right: Quantiles as a function of  $x$  by values of  $p \in \{0.25, 0.75\}$ .

quadratic terms), then the QTE associated with  $x_j$  is given by

$$\frac{\partial H_{Y|X}(p)}{\partial x_j} = \beta_j(p)$$

In Figure 2, we provide a simple example using the heteroscedastic model  $Y = \lfloor x \rfloor + \lfloor x + 1 \rfloor \epsilon$ , for  $x \geq 0$ , where  $\epsilon$  is discrete uniform between 1 and 6 (die rolling). As a function of  $p$ , for fixed  $x$ , mid-quantiles follow the uniform model we already seen in Figure 1. As a function of  $x$ , for fixed  $p$ , we obtain our proposed mid-QF (2.4). In particular, the first and third mid-quantiles have equations  $H(0.25) = 2 + 3x$  and  $H(0.75) = 5 + 6x$ , respectively, which entail a QTE equal to 3 and 6, respectively.

Our definition of conditional mid-quantiles is general as it applies to any type of discrete response variable that can be ordered. An interesting special case is when  $Y$  is binary, thus  $\mathcal{S}_Y = \{0, 1\}$ . According to (2.3), the mid-CDF is then given by  $G_{Y|X}(y|x) = (y + 1 - \mu(x))/2$ , where  $\mu(x) = \Pr(Y = 1|X = x)$ . Therefore,  $\pi_1 = G_{Y|X}(0|x) = 0.5 - 0.5\mu(x)$  and  $\pi_2 = G_{Y|X}(1|x) = 1 - 0.5\mu(x)$ . By definition, mid-quantiles are equal to 0 if  $p < \pi_1$  and to 1 if  $p > \pi_2$ . Otherwise,  $H_{Y|X}(p) = (1 - \gamma)y_1 + \gamma y_2 = \gamma$  for  $p = (1 - \gamma)\pi_1 + \gamma\pi_2$ . From the latter expression we get  $\gamma = (p - \pi_1)/(\pi_2 - \pi_1) = 2p - 1 + \mu(x)$ . In summary, we obtain the following conditional mid-QF

$$H_{Y|X}(p) = \begin{cases} 0 & \text{if } p < \pi_1 \\ 2p - 1 + \mu(x) & \text{if } \pi_1 \leq p \leq \pi_2 \\ 1 & \text{if } p > \pi_2 \end{cases}$$

Therefore the conditional mid-median is exactly equal to  $\mu(x)$ , while all the other conditional mid-quantiles are shifted by  $2p - 1$ .

### 2.3 Estimation

Consider a sample  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ , with corresponding observations  $(x_i, y_i)$ . Also, let  $z_j$ ,  $j = 1, \dots, k$ , be the  $j$ th distinct observation of  $Y$  that occurs in the sample, with  $z_j < z_{j+1}$  for all  $j = 1, \dots, k - 1$ .

Estimation of model (2.4) proceeds in two steps. In the first step, we estimate the mid-CDF. Let

$$\hat{G}_{Y|X}(y|x) \equiv \hat{F}_{Y|X}(y|x) - 0.5 \cdot \hat{m}_{Y|X}(y|x) \tag{2.5}$$

denote the sample equivalent of (2.3). The estimation of  $F_{Y|X}$  plays a key role in our approach. In our formulation, we require an estimator that can be applied to discrete responses and that admits continuous and discrete covariates (or a mix thereof). In line with the nonparametric flavour of our modelling strategy, we considered the conditional CDF estimator proposed by Li and Racine.<sup>44</sup> This takes the form

$$\hat{F}_{Y|X}(y|x) = \frac{n^{-1} \sum_{i=1}^n I(Y_i \leq y) K_\lambda(X_i, x)}{\hat{\delta}_X(x)} \tag{2.6}$$

where  $K_\lambda$  is the (product) kernel with bandwidth vector  $\lambda$  and  $\hat{\delta}_X(x)$  is the kernel estimator of the marginal density of  $X$ . The implementation of (2.6) involves choosing kernel types, as well as determining a number of tuning and estimation parameters. In our simulation study (Section 3), we adopted the default settings of the `np` package,<sup>45</sup> which include, for



example, least-squares cross-validation for bandwidth selection.<sup>46</sup> Since we obtained satisfactory empirical results using those settings, we will not dwell on this matter but instead refer the reader to the relevant literature for technical and implementation details.<sup>44–48</sup>

By applying (2.6) to the sample observations, we obtain  $\hat{F}_{Y|X}(z_j|x)$ ,  $j = 1, \dots, k$ , and  $\hat{m}_{Y|X}(z_j|x) = \hat{F}_{Y|X}(z_j|x) - \hat{F}_{Y|X}(z_{j-1}|x)$ , which we plug into (2.5) to obtain  $\hat{G}_{Y|X}(z_j|x)$ . Here, we set  $\hat{m}_{Y|X}(z_1|x) = \hat{F}_{Y|X}(z_1|x)$ , hence  $\hat{G}_{Y|X}(z_1|x) = 0.5 \cdot \hat{F}_{Y|X}(z_1|x)$ .

We note that, unfortunately, nonparametric estimation of  $F_{Y|X}$  entails a loss of performance when the dimension of  $X$  is large, the design is sparse, or both. In these cases, a semiparametric approach may be preferred. A natural choice is to obtain  $\hat{F}_{Y|X}(z_j|x)$  as the estimate of the binomial probability  $\Pr\{Y \leq z_j|x\}$ , for  $j = 1, \dots, k$ . This idea was indeed considered by some authors<sup>49,50</sup> to address the curse of dimensionality of nonparametric estimators and, while originally a logit estimator was proposed, in principle any other link function can be employed. Since applying the binomial estimator is tantamount to fitting  $k$  separate binomial regressions, one for each value  $z_j$ , one must verify whether the estimates  $\hat{F}_{Y|X}(z_j|x)$  are monotone. Monotonicity can be imposed a priori<sup>50</sup> or a posteriori.<sup>51</sup>

Now we move to the second step of model estimation. For a given  $x$ , define  $\hat{G}_{Y|X}^c(y|x)$  as the function interpolating the points  $(z_j, \hat{G}_{Y|X}(z_j|x))$ , where the ordinates have been obtained in the first step. The function  $\hat{G}_{Y|X}^c$  estimates  $G_{Y|X}^c$  and is the conditional extension of  $\hat{G}_Y^c$  introduced in Section 2.1. The goal is to estimate  $\beta(p)$  in (2.4) by solving the implicit equation  $p = \hat{G}_{Y|X}^c(\eta(p)|x)$ , where  $\eta(p) = h^{-1}\{x^\top \beta(p)\}$ . Our objective function and estimator are thus given by

$$\psi_n(\beta; p) = n^{-1} \sum_{i=1}^n \left\{ p - \hat{G}_{Y|X}^c(\eta_i|x_i) \right\}^2 \quad (2.7)$$

where  $\eta_i = h^{-1}\{x_i^\top \beta\}$ , and

$$\hat{\beta}(p) = \underset{\beta \in \mathbb{R}^q}{\operatorname{argmin}} \psi_n(\beta; p) \quad (2.8)$$

respectively. This estimation approach, which is an extension of the ‘inverse-CDF’ technique for marginal quantiles, has been recently proposed by De Backer et al<sup>52</sup> for fitting censored QR models. They, too, approach nonparametrically the estimation of the CDF (though in the continuous domain) via a double-kernel estimator akin to that of Li and Racine.<sup>44</sup> In summary, we echo De Backer et al’s<sup>52</sup> advocacy of the squared distance for its computational advantages and of the double-kernel estimation for its asymptotic and finite sample efficiency gains.

We can make (2.8) explicit by using the following linear interpolating function:

$$\hat{G}_{Y|X}^c(\eta_i|x_i) = b_{j_i}(\eta_i - z_{j_i}) + \hat{\pi}_{j_i} \quad z_{j_i} \leq \eta_i \leq z_{j_i+1}$$

where  $b_{j_i} = (\hat{\pi}_{j_i+1} - \hat{\pi}_{j_i})/(z_{j_i+1} - z_{j_i})$  and  $\hat{\pi}_{j_i} = \hat{G}_{Y|X}(z_{j_i}|x_i)$ . The index  $j_i = 1, \dots, k-1$  identifies, for a given  $i = 1, \dots, n$ , the value  $z_{j_i}$  among the  $z$ ’s such that  $\hat{G}_{Y|X}(z_{j_i}|x_i) \leq p \leq \hat{G}_{Y|X}(z_{j_i+1}|x_i)$ . If we restrict  $p \in \mathcal{I}$ , where  $\mathcal{I} = [\max_i \hat{G}_{Y|X}(z_1|x_i), \min_i \hat{G}_{Y|X}(z_k|x_i)]$ , then we find that our estimator, conditionally on  $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_k, \dots, \hat{\pi}_n, \dots, \hat{\pi}_n)^\top$ , has the form

$$\hat{\beta}(p; \hat{\pi}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top u \quad (2.9)$$

where  $\mathbf{X}$  is a  $n \times q$  matrix with  $i$ th row  $x_i$  and  $u$  is a  $n \times 1$  vector with  $i$ th element  $h\{[(p - \hat{\pi}_{j_i})/b_{j_i}] + z_{j_i}\}$ . It is straightforward to verify that (2.9) is a minimizer by plugging it into (2.7). The closed-form of (2.9) is, clearly, computationally convenient. Of course, one can still obtain an estimate of  $\beta(p)$  using (2.9), regardless of whether  $p$  is within the interval  $\mathcal{I}$ . However, when  $p \notin \mathcal{I}$ , all the elements of  $u$  such that  $\hat{\pi}_{j_i} > p$  or  $\hat{\pi}_{k_i} < p$  are ‘censored’, that is, constrained to be  $h(z_1)$  or  $h(z_k)$ . As a result, the linear predictor of the mid-quantile model will flatten out if  $p$  approaches 0 or 1, with ‘slopes’ tending to 0 and the intercept tending to the sample minimum or maximum, at a speed that depends on the censoring rate. This is a reasonable behaviour since it reflects all we can say about the relationship between  $Y$  and  $X$  when we have little or no information. Note also that, in principle, one can consider other general minimizers of (2.7) should the optimization problem become more complex (e.g., because of the addition of nonlinear constraints or penalties). This is a topic for future research.

We can derive the variance-covariance of  $\hat{\beta}(p)$  via the total variance law<sup>53</sup> as follows:

$$\operatorname{var}(\hat{\beta}(p)) = E_{\hat{\pi}} \left\{ \operatorname{var}_{\hat{\beta}}(\hat{\beta}|\hat{\pi}) \right\} + \operatorname{var}_{\hat{\pi}} \left\{ E_{\hat{\beta}}(\hat{\beta}|\hat{\pi}) \right\} \quad (2.10)$$

We estimate the first term in the right-hand side of (2.10) using a Huber-White variance-covariance estimator, which is given by  $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{D} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}$ , where  $\mathbf{D} = \operatorname{diag}(\hat{e}_1^2, \dots, \hat{e}_n^2)$  and  $\hat{e}_i = h(y_i) - x_i^\top \hat{\beta}(p; \hat{\pi})$ ,  $i = 1, \dots, n$ . To obtain an

estimate of the second term, we note that, by the delta method,

$$\text{var}_{\hat{\pi}} \left\{ E_{\hat{\beta}}(\hat{\beta} | \hat{\pi}) \right\} \approx \nabla_{\hat{\pi}} \hat{\beta}(p; \hat{\pi})^\top \text{var}(\hat{\pi}) \nabla_{\hat{\pi}} \hat{\beta}(p; \hat{\pi}) \quad (2.11)$$

The validity of (2.11) relies on regularity conditions on moments,<sup>54</sup> which in our case are guaranteed by the asymptotic normality established by Theorem 2. The expression for  $\text{var}(\hat{\pi})$  depends on the variance of the estimator  $\hat{F}_{Y|X}$ , which we discuss further below. We omit the tedious algebra for the Jacobian  $\nabla_{\hat{\pi}} \hat{\beta}(p; \hat{\pi})$ , which can be easily obtained via numerical differentiation. Also, note that the latter is carried out efficiently since the Jacobian is sparse, with sparsity no less than  $1 - [2n/(nk)^2]$ . This follows from the fact that the  $i$ th partial derivatives of  $\hat{\beta}(p; \hat{\pi})$  with respect to elements of  $\hat{\pi}$  with indices other than  $j_i$  and  $j_{i+1}$  are null (hence, there are at most  $2n$  non-zero partial derivatives). The diagonal elements of the  $nk \times nk$  matrix  $\text{var}(\hat{\pi})$  are given by

$$\begin{aligned} \text{var}(\hat{\pi}_{ij}) &= \frac{1}{4} \text{var}\{\hat{F}_{Y|X}(z_{j-1}|x_i)\} + \frac{1}{4} \text{var}\{\hat{F}_{Y|X}(z_j|x_i)\} \\ &\quad + \frac{1}{2} \text{cov}\{\hat{F}_{Y|X}(z_{j-1}|x_i), \hat{F}_{Y|X}(z_j|x_i)\}, \end{aligned}$$

$i = 1, \dots, n, j = 2, \dots, k$ , and  $\text{var}(\hat{\pi}_{i1}) = \frac{1}{4} \text{var}\{\hat{F}_{Y|X}(z_1|x_i)\}$ . In the expression above, we can neglect the covariance between  $\hat{F}_{Y|X}(z_j|x_i)$  and  $\hat{F}_{Y|X}(z_{j'}|x_i)$ ,  $j \neq j'$ , as this is asymptotically zero as shown in the proof of Theorem 2 (Section 2.5). This means that the off-diagonal elements of  $\text{var}(\hat{\pi})$  are also asymptotically negligible. Finally, the expression for the variance of  $\hat{F}_{Y|X}$  is given elsewhere.<sup>44</sup>

We conclude this section by noting that we would arrive at the estimator (2.9) also by starting from the conditional extension of the marginal mid-quantiles in (2.2), say

$$H_{Y|X=x_i}(p) = \begin{cases} y_{j_i}, & \text{if } p < \pi_{j_i} \\ y_{j_i}, & \text{if } p = \pi_{j_i}, j_i = 1, \dots, k \\ (1 - \gamma_i)y_{j_i} + \gamma_i y_{j_i+1}, & \text{if } p = (1 - \gamma_i)\pi_{j_i} + \gamma_i \pi_{j_i+1} \\ y_{j_k}, & \text{if } p > \pi_{j_k} \end{cases}$$

The dependence of  $\pi_{j_i}$ ,  $y_{j_i}$  and  $\gamma_i$  on  $i$  is the consequence of *conditioning* on  $x_i$ . Clearly, one would need to calculate all the mid-probabilities  $\pi_{j_i}$ ,  $j = 1, \dots, k$ , because the value of  $H_{Y|X=x_i}(p)$ ,  $i = 1, \dots, n$ , depends on where  $p$  lies with respect to the  $\pi_{j_i}$ 's. However, if one starts from the implicit equation  $p = \hat{G}_{Y|X}^c(\eta(p)|x)$  as we did, it becomes easier to work out the asymptotic properties of the estimator using the objective function (2.7).

## 2.4 Prediction of conditional mid-quantiles and recovery of ordinary quantiles

So far, we have focused our attention on the  $\hat{\beta}(p)$ , which is and remains the primary goal of inference in our modelling approach. However, it is immediate to obtain a prediction of the conditional mid-quantiles of  $Y$  conditional on  $X = x$  with  $\hat{H}_{Y|X}(p) = h^{-1}(x^\top \hat{\beta}(p))$ . The approximate sampling distribution of  $\hat{H}_{Y|X}(p)$  is given in Corollary 2.1.

It might also be of some interest to know that we are able to recover ordinary conditional quantiles from the conditional mid-quantiles. As this result represents a by-product of our models, we do not place it in direct competition with alternative approaches,<sup>22</sup> which may well be preferable to ours since ours is not specifically designed for such a purpose. The goal is to relabel  $\hat{H}_{Y|X}(p)$  to obtain an estimate of the (ordinary) conditional quantiles  $Q_{Y|X}(p) \equiv \inf\{y \in \mathbb{R} : F_{Y|X}(y) \geq p\}$ . Let us denote such an estimate by  $\hat{Q}_{Y|X}(p)$ . This can be achieved in few simple steps. Since the distribution function  $F_{Y|X}$  has been already estimated in order to fit the mid-QR model, it is sufficient to plug  $x$  into  $\hat{F}_{Y|X}(y|x)$ . The latter is used to obtain  $\hat{G}_{Y|X}(y|x)$  as described in the previous section. Successively, we identify the index  $j$  such that  $\hat{G}_{Y|X}(z_j|x) \leq p < \hat{G}_{Y|X}(z_{j+1}|x)$ . If  $Y$  is a variable whose values are irregularly spaced, then one would naturally take  $\hat{Q}_{Y|X}(p) = z_{j+1}$  if  $p > \hat{F}_{Y|X}(z_j|x)$  or  $\hat{Q}_{Y|X}(p) = z_j$  otherwise, with the understanding that  $\hat{Q}_{Y|X}(p) = z_1$  if  $p < \hat{F}_{Y|X}(z_1|x)$ . If, on the other hand,  $Y$  is a variable with equally spaced values (e.g., a count), then we can take  $\hat{Q}_{Y|X}(p) = \lceil \hat{H}_{Y|X}(p) \rceil$  if  $p > \hat{F}_{Y|X}(z_j|x)$  or  $\hat{Q}_{Y|X}(p) = \lfloor \hat{H}_{Y|X}(p) \rfloor$  otherwise, with the understanding that  $\hat{Q}_{Y|X}(p) = \lceil \hat{H}_{Y|X}(p) \rceil$  if  $p < \hat{F}_{Y|X}(z_1|x)$ . A demonstration is given in Figure 7. We do not pursue the calculation of standard errors for  $\hat{Q}_{Y|X}(p)$ , which can be derived from the distribution of  $\hat{H}_{Y|X}(p)$  (Corollary 2.1), as it is not within our main interests.

## 2.5 Theoretical results

We first show consistency of  $\hat{\beta}(p)$  under general assumptions. We then provide its asymptotic distribution. Here, we assume that the conditional CDF estimator of Li and Racine<sup>44</sup> is used to obtain  $\hat{F}_{Y|X}(y|x)$ . The identity matrix of order  $m$  will be denoted by  $I_m$ . The proofs of the theorems in this section are given in the Supplemental material.

**Theorem 1.** *Generate  $n$  independent conditional responses from a discrete distribution with parameters satisfying (2.4). Assume that the marginal density of the continuous covariates in  $X$  is strictly positive and that  $0 < F_{Y|X}(y|x) < 1$ . Assume also that the kernel  $K_\lambda(X, x)$  in (2.6) is symmetric, bounded, and compactly supported, and that  $n \prod_j \lambda_j \rightarrow \infty$ , while  $\lambda_j \rightarrow 0$  for all  $j = 1, \dots, q$ . For a fixed  $p \in (0, 1)$ , let  $\hat{\beta}(p)$  denote the solution in (2.8) and let  $\beta^*(p)$  be its population counterpart.*

*Then, as  $n \rightarrow \infty$ ,  $\sup_z |\hat{G}_{Y|X}^c(z|x) - G_{Y|X}^c(z|x)| \rightarrow 0$ . Additionally,  $\|\hat{\beta}(p) - \beta^*(p)\| \rightarrow 0$  and  $|\hat{H}_{h(Y)|X}(p) - H_{h(Y)|X}(p)| \rightarrow 0$ , for all  $p \in (0, 1)$ .*

**Theorem 2.** *In addition to the assumptions in Theorem 1, assume that*

*$\hat{G}_{Y|X}^c(h^{-1}(x^\top \beta)|x)$  is differentiable with respect to  $\beta$ . Assume also that the design matrix is full rank, that  $\lim_n n^{-1} X X^\top$  exists and is a positive-definite matrix. Finally, assume that  $\sqrt{n \prod_j \lambda_j} (\sum_j \lambda_j)^2 = O(1)$ . Then,*

$$V(\beta^*(p))^{-1/2} \sqrt{n \prod_j \lambda_j} (\hat{\beta}(p) - \beta^*(p)) \rightarrow N(0, I_q) \quad \text{in distribution,}$$

where  $V(\beta^*(p)) = J(\beta^*(p))^{-1} D(\beta^*(p)) J(\beta^*(p))^{-1}$ ,

$$J(\beta^*) = E\{\nabla_{\beta}^2 \psi_n(\beta; p)|_{\beta=\beta^*}\}, \text{ and } D(\beta^*) = \text{Var}\left\{\sqrt{n \prod_j \lambda_j} \nabla_{\beta} \psi_n(\beta^*; p)|_{\beta=\beta^*}\right\}.$$

**Corollary 2.1.** *The approximate sampling distribution of  $\hat{H}_{Y|X}(p) = h^{-1}(x^\top \hat{\beta}(p))$  follows from the results in Theorems 1 and 2. Let  $\Sigma_p = V(\beta^*(p))(n \prod_j \lambda_j)^{-1}$  and let  $\delta(b) = \nabla_{\beta} h^{-1}(x^\top \beta)|_{\beta=b}$ . Then for large  $n$ , the distribution of  $\hat{H}_{Y|X}(p)$  is approximately normal with mean  $h^{-1}(x^\top \beta^*(p))$  and variance  $\delta(\beta^*(p))^\top \Sigma_p \delta(\beta^*(p))$ .*

## 3 Results

In this section, we illustrate the performance of mid-QR on both simulated and real data. The application concerns prescription drugs usage in the United States (US).

### 3.1 Simulation study

Data were generated according to six distinct models. The first four had the homoscedastic discrete uniform, heteroscedastic discrete uniform, Poisson, and Bernoulli errors, respectively. Each of these four models was considered with either one discrete covariate (1a, 2a, 3a, and 4a), or with two continuous covariates (1b, 2b, 3b, and 4b). The fifth model was defined as the ratio of two Poisson random variables. Lastly, the response variable for the sixth model was randomly sampled (with replacement) from the empirical (marginal) distribution of prescription drugs analysed in Section 3.2 and depicted in Figure 5. While the responses generated with the first four models are equally spaced integers from standard, *regular* distributions, the last two provide instances of non-standard features like irregularly spaced values, fractional values, zero-excess, and outliers. In symbols, data were generated as follows:

- (1a)  $Y = \lfloor 1 + 2w \rfloor + \epsilon$ , where  $w \sim \text{DU}(0, 5)$  and  $\epsilon \sim \text{DU}(1, 10)$ ;
- (1b)  $Y = \lfloor 1 + 2w_1 + w_2 \rfloor + \epsilon$ , where  $w_1 \sim \text{U}(0, 5)$ ,  $w_2 \sim 1/3\chi_3^2$ , and  $\epsilon \sim \text{DU}(1, 10)$ ;
- (2a)  $Y = \lfloor 1 + 2w \rfloor + \lfloor w + 1 \rfloor \epsilon$ , where  $w \sim \text{DU}(0, 5)$  and  $\epsilon \sim \text{DU}(1, 10)$ ;
- (2b)  $Y = \lfloor 1 + 2w_1 + w_2 \rfloor + \lfloor w_1 + 1 \rfloor \epsilon$ , where  $w_1 \sim \text{U}(0, 5)$ ,  $w_2 \sim 1/3\chi_3^2$ , and  $\epsilon \sim \text{DU}(1, 10)$ ;
- (3a)  $Y = \epsilon$ , where  $\epsilon \sim \text{Poisson}(\mu)$ ,  $\mu = \exp(0.5 + 2w)$ , and  $w \sim \text{DU}(1, 3)$ ;
- (3b) as in scenario (3a) with  $\mu = \exp(0.5 + 2w_1 + 0.3w_2)$ ,  $w_1 \sim \text{U}(1, 3)$ , and  $w_2 \sim 1/3\chi_3^2$ ;
- (4a)  $Y = \epsilon$ , where  $\epsilon \sim \text{Bernoulli}(\mu)$ ,  $\mu = 1/[1 + \exp\{- (3 + w)\}]$ , and  $w \sim \text{DU}(0, 5)$ ;
- (4b) as in scenario (4a) with  $\mu = 1/[1 + \exp\{- (3 + w_1 + w_2)\}]$ ,  $w_1 \sim \text{U}(0, 5)$ , and  $w_2 \sim 1/3\chi_3^2$ ;
- (5)  $Y = \epsilon_1/(\epsilon_2 + 1)$ , where  $\epsilon_h \sim \text{Poisson}(\mu)$ ,  $h = 1, 2$ ,  $\mu = \exp(0.5 + 1w)$ , and  $w \sim \text{DU}(1, 3)$ ;



(6)  $Y$  is a discrete variable with values in  $S_Y = \{0, 1, 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 361, 400\}$  and corresponding probabilities  $\{5.0 \times 10^{-1}, 1.8 \times 10^{-1}, 9.6 \times 10^{-2}, 5.4 \times 10^{-2}, 4.9 \times 10^{-2}, 3.4 \times 10^{-2}, 2.4 \times 10^{-2}, 2.0 \times 10^{-2}, 1.1 \times 10^{-2}, 8.9 \times 10^{-3}, 5.1 \times 10^{-3}, 2.8 \times 10^{-3}, 3.3 \times 10^{-3}, 2.3 \times 10^{-3}, 2.3 \times 10^{-3}, 9.3 \times 10^{-4}, 9.3 \times 10^{-4}, 1.9 \times 10^{-3}, 4.7 \times 10^{-4}, 4.7 \times 10^{-4}\}$ , and  $w \sim \text{DU}(1, 3)$ ,

where  $\text{DU}(a, b)$  and  $\text{U}(a, b)$  denote random variables with, respectively, discrete and continuous uniform distribution on  $(a, b)$ . Samples  $(y_i, w_i)$  of size  $n \in \{100, 500, 1000\}$  were independently drawn from each model for  $R = 1000$  replications. We then fitted the linear mid-quantile model  $H_{Y|X}(p) = x^T \beta(p)$  with data generated under models 1, 2, 5, and 6 (note, however, that the sampling probabilities under model 6 do not depend on  $w$ ); the log-linear mid-quantile model  $H_{Y|X}(p) = \exp \{x^T \beta(p)\}$  with data generated under model 3; and the logistic mid-quantile model  $H_{Y|X}(p) = [1 + \exp \{-x^T \beta(p)\}]^{-1}$  with data generated under model 4. All models were estimated for seven deciles,  $p \in \{0.2, 0.3, \dots, 0.8\}$ , except the logistic model, which was estimated for the median only.

Let  $H_i(p_k) \equiv H_{Y_i|X_i}(p_k)$  denote the true mid-quantile at level  $p_k$  for a given  $x_i = (1, w_i)^T$  under any of the data-generating models defined above and  $\hat{H}_i^{(r)}(p_k)$  be the corresponding estimate for replication  $r$ . We assessed the performance of the proposed methods in terms of average bias and root mean squared error (RMSE) of the mid-QF, i.e.

$$\frac{1}{R} \sum_{r=1}^R \left\{ n^{-1} \sum_{i=1}^n \hat{H}_i^{(r)}(p_k) - H_i(p_k) \right\}$$

and

$$\left[ \frac{1}{R} \sum_{r=1}^R \left\{ n^{-1} \sum_{i=1}^n \left( \hat{H}_i^{(r)}(p_k) - H_i(p_k) \right)^2 \right\} \right]^{1/2}$$

respectively. We also report the average true mid-quantiles at  $n = 1000$

$$\bar{H}(p_k) = n^{-1} \sum_{i=1}^n H_i(p_k)$$

as a term of comparison for assessing the relative magnitude of the bias. Finally, we calculated 95% confidence intervals to assess coverage of the slope parameter in mid-quantile models for  $p \in \{.3, .5, .7\}$  when data were generated under scenarios 1a, 2a, and 3a. The corresponding standard errors were computed based on expression (2.10).

Estimated bias and RMSE of the proposed estimator are shown in Tables 1–6 for scenarios 1a, 2a, 3a, 4a, 5, and 6, and in Supplemental Tables 1–4 for scenarios 1b, 2b, 3b, and 4b. The bias was, in general, small, never exceeding 2.3% of the average mid-quantile for the homoscedastic discrete uniform model (1a and 1b), 3.3% for the heteroscedastic discrete uniform model (2a and 2b), 0.6% for the Poisson model with a discrete covariate (3a), and 2.5% for the Poisson ratio model (5). The estimated bias and RMSE of the proposed estimator for the Bernoulli model (4a and 4b) were extremely small at all sample sizes. In contrast, the bias was relatively higher (up to 11% of the average mid-quantile) in the Poisson scenario with continuous covariates (3b), although this issue was limited to the tail quantiles at smaller sample sizes. The bias and RMSE for model (6) were notable at smaller sample sizes. In particular, the bias was larger at  $p = .6$ . This can be explained by examining the mid-CDF in Figure 5. The function goes from 0.25 at  $y = 0$  to just shy of 0.6 at  $y = 1$ .

**Table 1.** Bias and root mean squared error (RMSE) of predicted quantiles for data generated using the homoscedastic discrete uniform model (1a).

$p$	$n = 100$		$n = 500$		$n = 1000$		$\bar{H}$
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
0.2	0.015	0.508	-0.023	0.221	-0.006	0.153	8.494
0.3	0.097	0.548	0.004	0.246	0.011	0.172	9.494
0.4	0.113	0.574	0.008	0.263	0.016	0.182	10.494
0.5	0.121	0.587	0.010	0.271	0.016	0.186	11.494
0.6	0.121	0.579	0.012	0.269	0.013	0.183	12.494
0.7	0.135	0.550	0.014	0.251	0.012	0.167	13.494
0.8	0.211	0.532	0.045	0.223	0.028	0.146	14.494

**Table 2.** Bias and root mean squared error (RMSE) of predicted quantiles for data generated using the heteroscedastic discrete uniform model (2a).

$p$	$n = 100$		$n = 500$		$n = 1000$		$\bar{H}$
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
0.2	-0.417	1.689	-0.281	1.056	-0.071	0.791	14.737
0.3	-0.482	1.940	-0.389	1.108	-0.229	0.856	18.234
0.4	-0.397	2.038	-0.181	1.071	0.048	0.817	21.731
0.5	-0.444	2.100	-0.418	1.068	-0.297	0.810	25.228
0.6	-0.308	2.134	-0.322	1.138	-0.290	0.887	28.725
0.7	-0.073	1.933	-0.252	0.967	-0.168	0.724	32.222
0.8	0.484	1.864	0.273	0.817	0.259	0.594	35.719

**Table 3.** Bias and root mean squared error (RMSE) of predicted quantiles for data generated using the Poisson model (3a).

$p$	$n = 100$		$n = 500$		$n = 1000$		$\bar{H}$
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
0.2	0.232	3.369	0.190	1.848	0.226	1.552	243.938
0.3	0.711	2.921	0.482	1.476	0.492	1.199	247.933
0.4	0.554	2.794	0.340	1.358	0.377	1.037	251.596
0.5	0.862	2.892	0.635	1.434	0.642	1.144	254.593
0.6	0.891	2.983	0.605	1.454	0.610	1.122	257.921
0.7	1.265	3.462	0.856	1.842	0.844	1.504	261.251
0.8	1.580	4.288	0.888	2.388	0.827	2.050	265.580

**Table 4.** Bias and root mean squared error (RMSE) of predicted quantiles for data generated using the Bernoulli model (4a).

$p$	$n = 100$		$n = 500$		$n = 1000$		$\bar{H}$
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
0.5	0.001	0.047	0.001	0.022	-0.000	0.015	0.424

**Table 5.** Bias and root mean squared error (RMSE) of predicted quantiles for data generated using the Poisson ratio model (5).

$p$	$n = 100$		$n = 500$		$n = 1000$		$\bar{H}$
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
0.2	-0.005	0.063	-0.010	0.031	-0.008	0.024	0.643
0.3	-0.008	0.065	-0.012	0.032	-0.011	0.024	0.736
0.4	-0.007	0.069	-0.011	0.033	-0.010	0.025	0.821
0.5	-0.000	0.076	-0.006	0.033	-0.006	0.023	0.905
0.6	-0.009	0.087	-0.008	0.041	-0.005	0.031	1.015
0.7	-0.018	0.107	-0.028	0.063	-0.029	0.055	1.144
0.8	-0.017	0.134	-0.022	0.064	-0.021	0.050	1.304

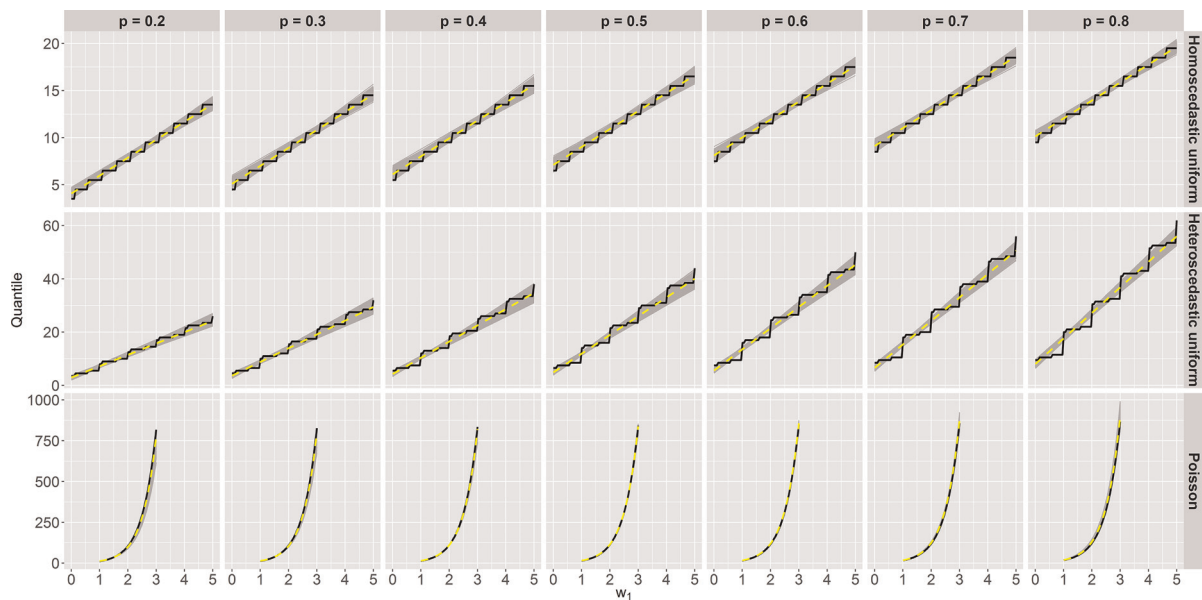
Conditioning on  $w$  skews the distribution of  $\hat{\beta}(0.6)$  since  $w$  is unrelated to the response (further investigation revealed that, at  $p = .6$ , the estimator is median-unbiased). In general, both bias and RMSE decreased with  $n$  at approximately the expected rate for all six models.

To appreciate how these results translate into model fitting, the estimated conditional mid-quantiles from all replications and the average estimated conditional mid-quantiles ( $n = 1000$ ) are shown in Figure 3 for scenarios 1b, 2b, and 3b, and in Figure 4 for scenario 4b. All mid-quantiles are plotted as functions of  $w_1$ , with  $w_2$  set equal to the median of  $1/3\chi_3^2$ .

**Table 6.** Bias and RMSE of predicted quantiles for data generated using the NHANES empirical distribution model (6).

$p$	$n = 100$		$n = 500$		$n = 1000$		$\bar{H}$
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
0.2	0.003	0.004	0.000	0.000	0.000	0.000	0.000
0.3	0.018	0.078	0.002	0.033	0.001	0.024	0.143
0.4	0.023	0.098	0.003	0.039	0.002	0.029	0.436
0.5	0.060	0.152	0.004	0.047	0.002	0.034	0.729
0.6	0.417	0.654	0.111	0.287	0.070	0.217	1.170
0.7	0.492	1.242	0.054	0.400	0.008	0.273	3.348
0.8	0.979	3.151	0.286	1.292	0.129	0.908	8.651

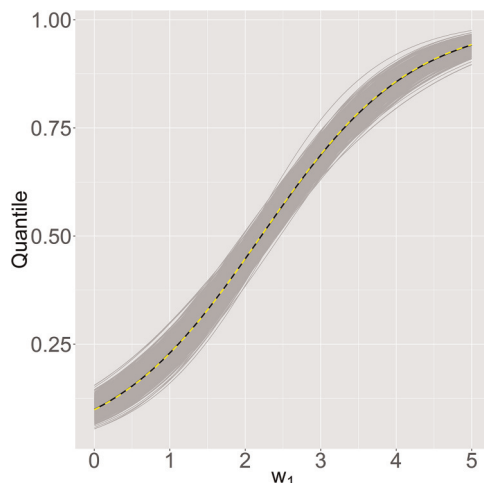
NHANES, National Health and Nutrition Examination Survey; RMSE, root mean squared error.



**Figure 3.** True quantile function (black line), estimated conditional mid-quantile functions for all replications (grey lines), and average estimated conditional mid-quantile function (dashed yellow) for three simulated scenarios with continuous covariates and  $n = 1000$ .

The observed coverage at the nominal 95% confidence level for the slope in selected scenarios is given in Table 7. The results are in general accurate, although frequencies are occasionally slightly away from the nominal level. This is not surprising since the sample estimator of (2.10) relies on the Huber-White estimator and on several approximations.

It would be remiss of us not to make a contrast between our proposed estimator and existing alternatives. The estimator developed by Machado and Santos Silva<sup>19</sup> (hereinafter referred to as MSS) is a natural candidate. However, the comparison is inevitably restricted: first of all, neither the ‘true’ coefficients nor the population quantiles underlying mid-quantile and jittering-based estimation are necessarily the same quantities. Indeed, the quantiles modelled by MSS are defined as  $Q_{Z|X}(p) = Q_{Y|X}(p) + \frac{p - F_{Y|X}\{Q_{Y|X}(p) - 1\}}{m_{Y|X}(Q_{Y|X}(p))}$ , where  $Z = Y + U$  and  $U$  is uniformly distributed on  $[0, 1)$ . The jittered quantiles  $Q_{Z|X}(p)$  dominate the true quantiles since  $F_{Y|X}\{Q_{Y|X}(p) - 1\} \leq p$ , uniformly over  $p$ . In contrast, mid-quantiles interpolate the true quantiles. In addition, the comparison must be restricted to when the response is a count or ordinal variable, as these are required by jittering-based estimation. For these reasons, we considered only Poisson data as in scenario 3a, in which case we expected the two estimators to target the same functional relationship (i.e., slope). The mean and variance of the estimates using our estimator (MIDQR), as well as the ratio of means and variances comparing the MSS estimator relative to MIDQR are given in Table 8. The two estimators gave similar estimates of the slope with MSS:MIDQR ratios of the means close to 1 across quantiles and sample sizes. Compared to the MSS estimator, our estimator had lower variance (MSS:MIDQR ratios of the variances  $> 1$ ) consistently for  $.2 < p < .8$ , but higher variance (MSS:MIDQR ratios of the variances lower than 1) at  $p = .2$  and  $p = .8$ . The boxplot of the estimates shown in Supplemental Figure 1 provides some



**Figure 4.** True quantile function (black line), estimated conditional mid-quantile functions for all replications (grey lines), and average estimated conditional mid-quantile function (dashed yellow) for the Bernoulli simulated scenario with continuous covariates and  $n = 1000$ .

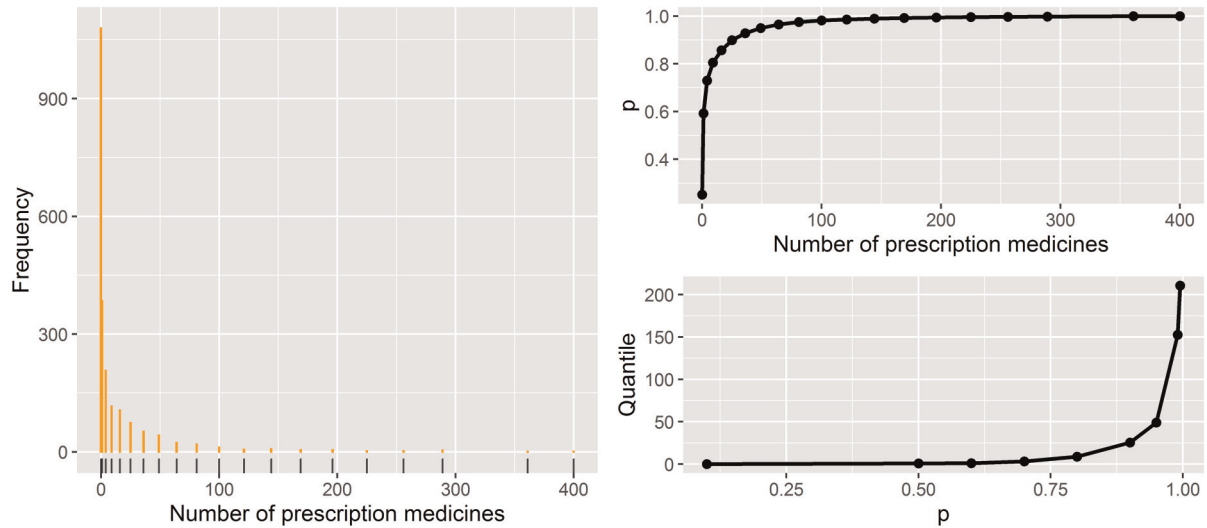
**Table 7.** Observed coverage at the nominal 95% confidence level for the slope in mid-quantiles models for data generated using the homoscedastic discrete uniform (1a), heteroscedastic discrete uniform (2a), and Poisson (3a) models.

$p$	1a			2a			3a		
	100	500	1000	100	500	1000	100	500	1000
0.3	97.70	97.70	97.90	96.60	95.90	93.30	93.70	94.59	95.09
0.5	95.90	94.90	96.10	94.60	94.60	92.50	93.90	95.19	95.09
0.7	98.30	96.70	98.50	96.00	96.00	96.80	97.60	97.49	96.99

**Table 8.** Mean and variance ( $\times 1000$ ) of the slope's estimates using the proposed approach (MIDQR), and ratio of mean and variance of the estimates using Machado and Santos Silva (MSS)'s<sup>19</sup> estimator compared to MIDQR for data generated using the Poisson model (3a). A MSS:MIDQR ratio greater than 1 indicates a larger MSS value.

$p$	$n = 100$		$n = 500$				$n = 1000$					
	MIDQR		MSS:MIDQR		MIDQR		MSS:MIDQR		MIDQR		MSS:MIDQR	
	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.	Mean	Var.
0.2	2.094	0.997	1.000	0.878	2.091	0.189	1.000	0.895	2.091	0.089	1.001	0.978
0.3	2.051	0.489	1.003	1.316	2.053	0.097	1.002	1.312	2.053	0.055	1.002	1.214
0.4	2.022	0.409	1.004	1.384	2.025	0.081	1.002	1.307	2.026	0.043	1.002	1.364
0.5	1.997	0.386	1.004	1.442	2.002	0.073	1.002	1.257	2.003	0.039	1.001	1.308
0.6	1.974	0.372	1.004	1.357	1.979	0.072	1.001	1.332	1.980	0.037	1.001	1.399
0.7	1.947	0.423	1.005	1.250	1.955	0.081	1.001	1.220	1.956	0.040	1.000	1.387
0.8	1.913	0.578	1.008	0.939	1.924	0.114	1.002	1.025	1.926	0.058	1.001	1.072

insight on why the performance of MIDQR was less satisfactory on the tails. For example, when  $n = 100$  and  $p = 0.2$ , the empirical distributions of the two estimators were nearly identical between the first and third quartiles. However, because of a few extreme MIDQR estimates, the variance of MSS was about 12% lower. The advantage of the MSS estimator on the tails withered away as the sample size increased. At  $p = 0.2$ , the ratio of the variances went from 0.878 ( $n = 100$ ) to 0.978 ( $n = 1000$ ), while at  $p = 0.8$ , it went from 0.939 ( $n = 100$ ) to 1.072 ( $n = 1000$ ). That is, for  $n = 1000$ , the two estimators performed similarly on the tails. We conclude this section by remarking that this comparison is limited to a single scenario



**Figure 5.** Number of prescription medicines using the National Health and Nutrition Examination Survey data, 2015–2016. Left: Frequency bar plot with rug plot. Right: Estimated mid-cumulative distribution function  $\hat{G}^c(y)$  with filled circles marking points  $y \in \{0, 1, 4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289, 361, 400\}$  (top) and estimated mid-QF  $\hat{H}(p)$  with filled circles marking points  $p \in \{0.1, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.995\}$  (bottom).

and results are clearly not generalizable. However, if a lesson can be drawn, it is that the limited loss of efficiency of our estimator at lower sample sizes and outer regions of the response distribution is more than compensated by the larger efficiency gains for most of the quantiles and by the flexibility of dealing with different types of discrete responses.

### 3.2 Prescription drugs

In this section, we illustrate an application of mid-QR using data on prescription medications from the National Health and Nutrition Examination Survey (NHANES).<sup>55</sup> The US is the worldwide leader in per capita prescription drug spending,<sup>56</sup> and its pharmaceutical market represents a major economic sector worth hundreds of billions of dollars. In a recent QR analysis of NHANES data, Hong et al<sup>57</sup> found a higher opioid use (morphine milligram equivalent) in adults with long-standing physical disability and those with inflammatory conditions as compared to individuals with other conditions. Differences were markedly larger at the 75th and 95th percentiles than those at lower percentiles. In the context of medication use, a higher percentile can be interpreted as an index of diminished health, lower quality of life, and higher financial burden. A QR analysis of prescription medications use is therefore of both public health and health economics interest.

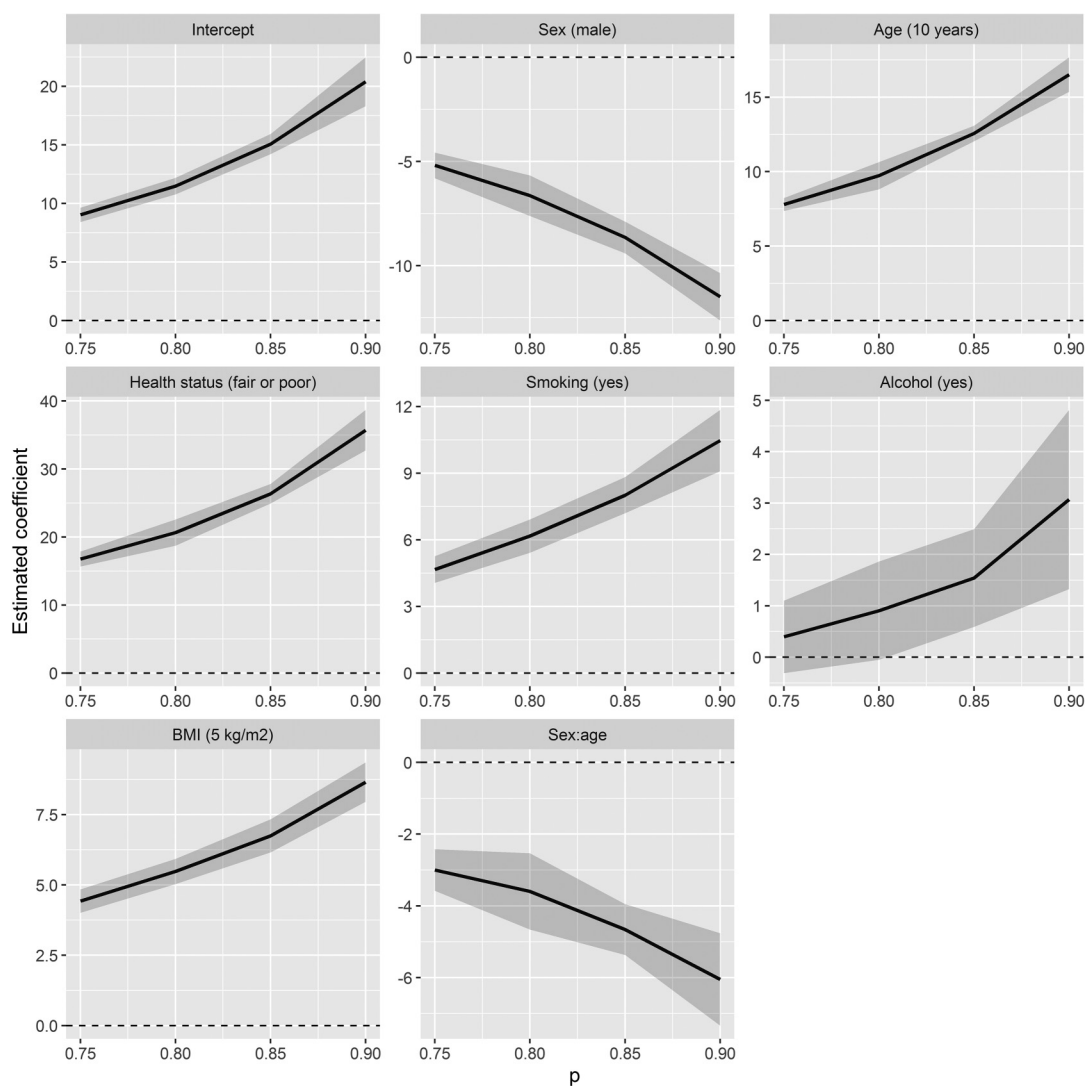
We abstracted data ( $n = 9971$ ) on the number of prescription medicines taken from 2015 to 2016 Dietary Supplement and Prescription Medication section of the Sample Person Questionnaire. We also obtained information on sex, age (years), perceived health status, smoking status (smoked at least 100 cigarettes in life), alcohol use (had at least 12 alcohol drinks in 1 year), weight (kg), height (m), and race. Before carrying out the analysis, we removed the effect of NHANES oversampling by first restricting the data set to all observations for White persons (about 30.7% of the overall sample), and subsequently adding observations for persons of other races that we subsampled with probabilities proportional to their NHANES weights. This resulted in a sample ( $n = 5058$ ) composed of about 60.6% of White persons and 49.3% females. We restricted the data set to adults aged 18–65 years and removed 378 incomplete observations. The final sample size for analysis was  $n = 2146$ . Figure 5 shows the marginal distribution and mid-QF of number of prescription medicines, while variables used for analysis are summarized in Table 9.

We investigated a linear model with sex (baseline: female), age (centred at 40 and scaled by 10), health status (baseline: good or excellent), smoking status (baseline: no), alcohol use (baseline: no), body mass index (BMI; centred at 30 and scaled by 5), and the interaction between sex and age. The admissible range  $\mathcal{I}$  for the application of (2.9) was  $[0.46, 0.92]$ , thus resulting in a large value of the lower bound due to the high proportion of zeros in the response. In Figure 6, we report the estimates of the coefficients obtained via (2.9) and their 95% confidence intervals based on (2.10) for  $p \in \{.75, .8, .85, .9\}$ . Older males tend to use less medications as compared to their female peers, and the inequality is more marked among high users (i.e., larger values of  $p$ ), with the largest difference of 27 prescriptions

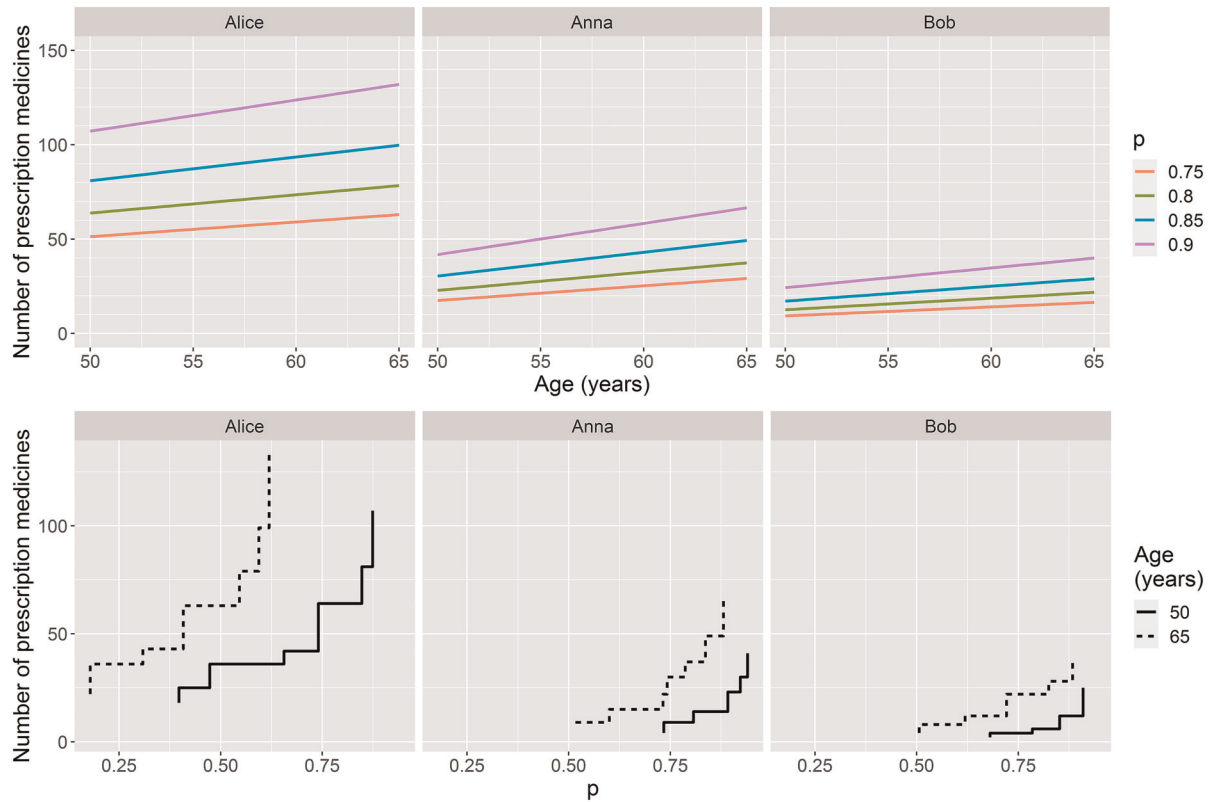


**Table 9.** Descriptive summary for the National Health and Nutrition Examination Survey data, 2015–2016.

Variable	Minimum	First quartile	Median	Third quartile	Maximum
Prescription medicines					
Overall	0.0	0.0	2.0	8.1	400.0
Female	0.0	0.2	2.5	9.5	400.0
Male	0.0	0.0	2.0	6.7	289.0
Age (years)	18.0	31.0	40.3	49.1	65.0
Body mass index (BMI) (kg/m <sup>2</sup> )	14.3	24.6	28.0	31.5	57.2
Variable	Frequency	Proportion (%)			
Sex (female)	1061	49.4			
Health status (good or excellent)	1723	80.3			
Smoking status (no)	1219	56.8			
Alcohol use (no)	547	25.5			



**Figure 6.** Estimated mid-QR coefficients ( $p \in \{0.75, 0.8, 0.85, 0.9\}$ ) and point-wise 95% confidence bands for the number of prescription medicines in the United States using the National Health and Nutrition Examination Survey data, 2015–2016.



**Figure 7.** Estimated conditional mid-quantiles (top row) and ordinary quantile functions (bottom row) of the number of prescription medicines for three hypothetical subpopulation profiles using the National Health and Nutrition Examination Survey data, 2015–2016.

between two 65-year-olds of opposite sex ranking in the top 10% of their respective distributions. However, due to a negative sex-age interaction, the inequality is actually reversed at younger ages, though differences are very small. Put differently, prescription medications count increases with age (as expected) and the rate of increase is sex-specific (steeper for females) and quantile-dependent (steeper at higher  $p$ ). A fair or poor perceived health status is associated with an estimated 17 ( $p = .75$ ) to 36 ( $p = .9$ ) more prescriptions than what is expected with a good or excellent health status. Individuals classified as smokers have an estimated 5 ( $p = 0.75$ ) to 10 ( $p = 0.9$ ) more prescriptions than non-smokers. However, one should be cautious with the interpretation of this result since the smoking variable used here does not capture smoking habits, past or recent. Alcohol, too, increases the prescription medications count and is quantile-dependent, though the magnitudes of the estimated effects are smaller than those associated with other factors. Finally, higher BMI is associated with a considerably larger number of prescription medications. For every 5 kg/m<sup>2</sup> increase in BMI, it is estimated that the 75th mid-quantile increases by about four prescriptions, while the 90th mid-quantile increases by about nine prescriptions.

To exemplify the results, we defined three hypothetical subpopulation profiles: Alice, a woman with fair or poor health status, non-smoker, non-alcohol-user, with a BMI equal to 50; Anna, a woman with good or excellent health status, smoker, alcohol user, with a BMI equal to 25; and Bob, a man with same attributes as Anna's. In Figure 7, we show the estimated mid-quantiles as a function of age (50–65 years) for the three subpopulations. Our model estimates that women like Alice who are in the top 25% of their subpopulation's distribution use as much as 51 prescription medications by the age of 50. The count raises to about 107 for those who are in the top 10%, with an additional 25 prescriptions when reaching the age of 65. Women like Anna, who enjoy a better health status and a lower BMI, are relatively better off than women like Alice, with substantially lower prescription drugs at all quantiles and all ages. Finally, as we noted above, estimates for (older) men start from lower values and increase more slowly with increasing age as compared to women. More specifically, a 65-year-old Bob in the top 10% of his subpopulation's distribution has approximately the same number of prescriptions as a 50-year-old Anna who ranks similarly in her subpopulation's distribution, all else being equal. In Figure 7, we also show the estimated ordinary QFs at age 50 and 65 that were recovered from the estimated conditional mid-quantiles following the procedure discussed in Section 2.4.

## 4 Discussion

We developed an approach to conditional quantile estimation with discrete responses. We established the theoretical properties of our conditional mid-quantile estimator under general conditions and showed its good performance in a simulation study with data generated from different discrete response models. Our two-step estimator is easy to implement. When constraining the quantile index to a data-driven admissible range, the second-step estimating equation has a least-squares type, closed-form solution, which is computationally efficient.

We note that the non-standard rate of convergence given in Theorem 2 may dampen the enthusiasm for the closed-form solution of estimator (2.9). However, faster rates of alternative estimators may come at a price. The  $\sqrt{n}$ -convergence of MSS's estimator,<sup>19</sup> for example, is restricted to finding an  $n$ -sequence of real numbers that depends on the model's transformation to linearity (see their assumption A6). This means that implementation of jittering requires ad hoc asymptotic calculations and software programming, neither of which are typically the source of excitement for the applied users. In general, while our proposed estimator does not achieve the much-coveted convergence rate of classical estimators, we can take solace in the fact that its bias and loss of efficiency can be ignored already at moderate sample sizes. As compared to jittering-based QR, mid-QR allows for the modelling of a wider range of discrete outcomes and it shows efficiency gains already at small sample sizes. Last but not least, the availability of mid-QR in the R package `Qtools`,<sup>58</sup> of which a brief tutorial is given in the Supplemental Material, may be particularly attractive in applied research studies.

In a real data analysis, conditional mid-quantiles revealed interesting aspects of prescription drug use in the US. In general, our results on gender and age inequalities are consistent with the literature.<sup>59,60</sup> While differences in medication use between men and women are expected due to differences in the incidence of disease (e.g., genitourinary infections, migraines, impaired thyroid function) or biological differences or differential preventive health care use, gender inequalities may be due also to unequal treatment.<sup>60</sup> Our analysis provides additional insight as it shows that differences between men and women (especially among older individuals) are heterogeneous, with a substantially higher gender gap among the top 10% of the distribution, after adjusting for perceived health status and other covariates. This raises the alarm about a possible differential medical treatment that worsens the inequality among the most fragile segment of the population. On the other hand, the heterogeneous association between the number of prescription medicines and BMI may be easier to explain on medical grounds. Higher prescribing levels are expected to be associated with higher BMI because of medical conditions that are known to be more prevalent in obese individuals.<sup>61</sup> If a higher percentile of the medications count distribution is interpreted as an (objective) index of poor health, then it is natural to expect that a given increase in BMI is more detrimental for someone ranking at, say, the 90th percentile than it is for the 'average' prescription drugs user. Owing to the fact that obesity increases drug prescribing in the most expensive prescribing categories,<sup>61</sup> it becomes clear that intervening on this modifiable risk factor has important connotations also for controlling health care expenditure.

We believe that mid-QR is amenable to several possible extensions, including estimation in the presence of censoring (survival analysis) and clustering (e.g., longitudinal analysis). Further research is needed to develop computationally efficient methods for high-dimensional data. Useful hints on how to tackle issues regarding censoring and higher-dimensional covariates may be gathered from the work by De Backer et al.<sup>52</sup>

### Acknowledgements

The authors are grateful to three anonymous referees for their helpful comments.

### Declaration of conflicting interests


The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

### ORCID iDs

Marco Geraci  <https://orcid.org/0000-0002-6311-8685>

Alessio Farcomeni  <https://orcid.org/0000-0002-7104-5826>

### Supplemental material

Supplemental material for this article is available online.

## References

1. Koenker R and Bassett G. Regression quantiles. *Econometrica* 1978; **46**: 33–50.
2. Koenker R. *Quantile regression*. New York, NY: Cambridge University Press, 2005.
3. Koenker R and Geling O. Reappraising medfly longevity. *J Am Stat Assoc* 2001; **96**: 458–468.
4. Austin PC, Tu JV, Daly PA, et al. The use of quantile regression in health care research: A case study examining gender differences in the timeliness of thrombolytic therapy. *Stat Med* 2005; **24**: 791–816.
5. Beyerlein A. Quantile regression—opportunities and challenges from a user’s perspective. *Am J Epidemiol* 2014; **180**: 330–331.
6. Ding R, McCarthy ML, Desmond JS, et al. Characterizing waiting room time, treatment time, and boarding time in the emergency department using quantile regression. *Acad Emerg Med* 2010; **17**: 813–823.
7. Mayfield CA, Geraci M, Dulin M, et al. Social and demographic characteristics of frequent or high-charge emergency department users: A quantile regression application. *J Eval Clin Pract* 2021; **27**: 1271–1280. DOI: doi:10.1111/jep.13537.
8. Rehkopf DH. Quantile regression for hypothesis testing and hypothesis screening at the dawn of big data. *Epidemiology* 2012; **23**: 665–667.
9. Wei Y and Terry MB. R: “Quantile regression—opportunities and challenges from a user’s perspective”. *Am J Epidemiol* 2015; **181**: 152–153.
10. Winkelmann R. Reforming health care: Evidence from quantile regressions for counts. *J Health Econ* 2006; **25**: 131–145.
11. Geraci M, Boghossian NS, Farcomeni A, et al. 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**: 1769–1786.
12. McCullagh P and Nelder JA. *Generalized linear models*. 2nd ed. New York: Chapman & Hall/CRC, 1989.
13. Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev* 1958; **65**: 386–408.
14. Manski CF. Maximum score estimation of the stochastic utility model of choice. *J Econom* 1975; **3**: 205–228.
15. Manski CF. Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *J Econom* 1985; **27**: 313–333.
16. Horowitz JL. A smoothed maximum score estimator for the binary response model. *Econometrica* 1992; **60**: 505–531.
17. Lee MJ. Median regression for ordered discrete response. *J Econom* 1992; **51**: 59–77.
18. Kordas G. Smoothed binary regression quantiles. *J Appl Econ* 2006; **21**: 387–407.
19. Machado JAF and Santos Silva JMC. Quantiles for counts. *J Am Stat Assoc* 2005; **100**: 1226–1237.
20. Hong HG and He X. Prediction of functional status for the elderly based on a new ordinal regression model. *J Am Stat Assoc* 2010; **105**: 930–941.
21. Frumento P and Salvati N. Parametric modeling of quantile regression coefficient functions with count data. *Stat Methods Appl* 2021; **30**: 1237–1258. DOI: 10.1007/s10260-021-00557-7.
22. Chernozhukov V, Fernández-Val I, Melly B, et al. Generic inference on quantile and quantile effect functions for discrete outcomes. *J Am Stat Assoc* 2019; **115**: 123–137.
23. Chambers R, Dreassi E and Salvati N. Disease mapping via negative binomial regression M-quantiles. *Stat Med* 2014; **33**: 4805–4824.
24. Chambers R, Salvati N and Tzavidis N. Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK. *J R Stat Soc A* 2016; **179**: 453–479.
25. Koenker R. Discussion: Living beyond our means. *Stat Modelling* 2013; **13**: 323–333.
26. Stasinopoulos MD, Rigby RA and De Bastiani F. GAMLSS: A distributional regression approach. *Stat Modelling* 2018; **18**: 248–273.
27. Bottai M, Cai B and McKeown RE. Logistic quantile regression for bounded outcomes. *Stat Med* 2009; **29**: 309–317.
28. Parzen E. Change PP plot and continuous sample quantile function. *Commun Stat - Theory Methods* 1993; **22**: 3287–3304.
29. Lancaster HO. Significance tests in discrete distributions. *J Am Stat Assoc* 1961; **56**: 223–234.
30. Ma Y, Genton MG and Parzen E. Asymptotic properties of sample quantiles of discrete distributions. *Ann Inst Stat Math* 2011; **63**: 227–243.
31. Hyndman RJ and Fan Y. Sample quantiles in statistical packages. *Am Stat* 1996; **50**: 361–365.
32. Jentsch C and Leucht A. Bootstrapping sample quantiles of discrete data. *Ann Inst Stat Math* 2016; **68**: 491–539.
33. Genton MG, Ma Y and Parzen E. Discussion of “sur une limitation très générale de la dispersion de la médiane” by M. Fréchet. *J Soc Franç Stat* 2006; **147**: 51–60.
34. Parzen E. Quantile probability and statistical data modeling. *Stat Sci* 2004; **19**: 652–662.
35. Stigler SM. Fractional order statistics, with applications. *J Am Stat Assoc* 1977; **72**: 544–550.
36. Wang D and Hutson AD. A fractional order statistic towards defining a smooth quantile function for discrete data. *J Stat Plan Inference* 2011; **141**: 3142–3150.
37. Jones MC. Expectiles and M-quantiles are quantiles. *Stat Probab Lett* 1994; **20**: 149–153.
38. Chamberlain G. Quantile regression, censoring, and the structure of wages. In: Sims C (ed.) *Advances in econometrics: Sixth world congress*. vol. 1. Cambridge, UK: Cambridge University Press; 1994.
39. Mu YM and He XM. Power transformation toward a linear regression quantile. *J Am Stat Assoc* 2007; **102**: 269–279.
40. Yin GS, Zeng DL and Li H. Power-transformed linear quantile regression with censored data. *J Am Stat Assoc* 2008; **103**: 1214–1224.

41. Geraci M and Jones MC. Improved transformation-based quantile regression. *Can J Stat* 2015; **43**: 118–132.
42. Box GEP and Cox DR. An analysis of transformations. *J R Stat Soc B* 1964; **26**: 211–252.
43. Aranda-Ordaz FJ. On two families of transformations to additivity for binary response data. *Biometrika* 1981; **68**: 357–363.
44. Li Q and Racine JS. Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *J Bus Econ Stat* 2008; **26**: 423–434.
45. Hayfield T and Racine JS. Nonparametric econometrics: The np package. *J Stat Softw* 2008; **27**: 1–32.
46. Li Q, Lin J and Racine JS. Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *J Bus Econ Stat* 2013; **31**: 57–65.
47. Wang MC and van Ryzin J. A class of smooth estimators for discrete distributions. *Biometrika* 1981; **68**: 301–309.
48. Li Q and Racine JS. *Nonparametric econometrics: Theory and practice*. Princeton, NJ: Princeton University Press, 2007.
49. Foresi S and Peracchi F. The conditional distribution of excess returns: An empirical analysis. *J Am Stat Assoc* 1995; **90**: 451–466.
50. Peracchi F. On estimating conditional quantiles and distribution functions. *Comput Stat Data Anal* 2002; **38**: 433–447.
51. Chernozhukov V, Fernandez-Val I and Galichon A. Quantile and probability curves without crossing. *Econometrica* 2010; **78**: 1093–1125.
52. De Backer M, El Ghouch A and Van Keilegom I. Linear censored quantile regression: A novel minimum-distance approach. *Scand J Stat* 2020; **47**: 1275–1306.
53. Mood AMF, Graybill FA and Boes DC. *Introduction to the theory of statistics*. New York: McGraw-Hill, 1974.
54. Oehlert GW. A note on the delta method. *Am Stat* 1992; **46**: 27–29.
55. National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. <https://www.cdc.gov/nchs/nhanes> (2019, accessed 4 November 2019).
56. The Kaiser Family Foundation. Follow the pill: Understanding the U.S. commercial pharmaceutical supply chain. <https://www.kff.org/other/report/follow-the-pill-understanding-the-u-s> (2015, accessed 4 December 2019).
57. Hong Y, Geraci M, Turk MA, et al. Opioid prescription patterns for adults with longstanding disability and inflammatory conditions compared to other users, using a nationally representative sample. *Arch Phys Med Rehabil* 2019; **100**: 86–94.e2.
58. Geraci M. Qtools: A collection of models and other tools for quantile inference. *R Journal* 2016; **8**: 117–138.
59. Roe CM, McNamara AM and Motheral BR. Gender- and age-related prescription drug use patterns. *Ann Pharmacother* 2002; **36**: 30–39.
60. Loikas D, Wettermark B, von Euler M, et al. Differences in drug utilisation between men and women: A cross-sectional analysis of all dispensed drugs in Sweden. *BMJ Open* 2013; **3**: e002378.
61. Counterweight Project Team. The impact of obesity on drug prescribing in primary care. *Br J Gen Pract* 2005; **55**: 743–749.