Supplementary material for 'Mid-quantile regression for discrete responses'

Marco Geraci*
Sapienza University of Rome, Italy
University of South Carolina, USA
and
Alessio Farcomeni
University of Rome "Tor Vergata", Italy

Abstract

This report contains supporting materials for the paper entitled 'Mid-quantile regression for discrete responses', hereinafter referred to as the 'Manuscript'. Section A contains technical details on inference. Section B reports additional tables with results from the simulation study. Section C reports a brief tutorial to illustrate mid-quantile regression routines available in the `R` package `Qtools`.

Keywords: Conditional CDF; Healthcare; Kernel estimator; Maximum score estimation; NHANES

*Correspondence to Marco Geraci, MEMOTEF Department, School of Economics, Sapienza University of Rome, Via del Castro Laurenziano 9, Rome 00161, Italy. E-mail: marco.geraci@uniroma1.it

A Supplementary theoretical results

In this section, we prove Theorems 1 and 2 in the Manuscript. We begin by providing some auxiliary results. We assume, throughout, that $\hat{G}^c_{Y|X}(\cdot|x)$ is a linear interpolant. While the validity of the Theorems still holds for other types of interpolants (e.g., polynomial), analytical expressions are more tractable in the linear case.

A.1 Auxiliary results

Our objective function and estimator are given by

$$\psi_n(\beta;p) = \frac{1}{n}\sum_{i=1}^{n}\left\{p - \hat{G}^c_{Y|X}\left(\eta_i|x_i\right)\right\}^2 \tag{A.1}$$

and

$$\hat{\beta}(p) = \underset{\beta\in\mathbb{R}^q}{\arg\min}\,\psi_n(\beta;p), \tag{A.2}$$

respectively. The equation of the interpolating function can be written explicitly as

$$\hat{G}^c_{Y|X}\left(\eta_i|x_i\right) = b_{j_i}(\eta_i - z_{j_i}) + \hat{\pi}_{j_i} \qquad\qquad z_{j_i} \leq \eta_i \leq z_{j_i+1},$$

where $b_{j_i} = \dfrac{\hat{\pi}_{j_i+1} - \hat{\pi}_{j_i}}{z_{j_i+1} - z_{j_i}}$ and $\hat{\pi}_{j_i} = \hat{G}_{Y|x_i}(z_{j_i})$. The index $j_i = 1,\ldots,k-1$ identifies, for a given $i = 1,\ldots,n$, the value z_{j_i} among the z's such that $\hat{G}_{Y|x_i}(z_{j_i}) \leq p \leq \hat{G}_{Y|x_i}(z_{j_i+1})$.

Then, the derivative of ψ_n with respect to the hth element of β is given by

$$\frac{\partial\psi_n(\beta;p)}{\partial\beta_h} = \frac{1}{n}\sum_{i=1}^{n} 2\left\{p - \hat{G}^c_{Y|X}\left(h^{-1}(x_i^\top\beta)|x_i\right)\right\}\left\{-\frac{\partial\hat{G}^c_{Y|X}\left(h^{-1}(x_i^\top\beta)|x_i\right)}{\partial\beta_h}\right\},$$

where

$$\frac{\partial\hat{G}^c_{Y|X}\left(h^{-1}(x_i^\top\beta)|x_i\right)}{\partial\beta_h} = x_{ih}b_{j_i}\frac{\partial h^{-1}(\eta_i)}{\partial\eta_i},$$

the existence of which follows from the differentiability of h.

Now, consider the second derivative of the objective function

$$\begin{aligned}\frac{\partial^2\psi_n(\beta;p)}{\partial\beta_h\partial\beta_u} = &-\frac{2}{n}\sum_{i=1}^{n}\left[p - \hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top\beta)|x_i\right\}\right]\frac{\partial^2\hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top\beta)|x_i\right\}}{\partial\beta_h\partial\beta_u}\\ &-\frac{\partial\hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top\beta)|x_i\right\}}{\partial\beta_h}\frac{\partial\hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top\beta)|x_i\right\}}{\partial\beta_u},\end{aligned}$$

where

$$\frac{\partial^2 \hat{G}^c_{Y|X}\left(h^{-1}(x_i^\top\beta)|x_i\right)}{\partial\beta_h\partial\beta_u} = x_{ih}x_{iu}b_{j_i}\frac{\partial^2 h^{-1}(\eta_i)}{\partial\eta_i}.$$

In summary, we obtain

$$\frac{\partial^2\psi_n(\beta;p)}{\partial\beta_h\partial\beta_u} = -\frac{2}{n}\sum_{i=1}^n x_{ih}x_{iu}b_{j_i}\left[p - \hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top\beta)|x_i\right\}\right]\frac{\partial^2 h^{-1}(\eta_i)}{\partial\eta_i} - x_{ih}x_{iu}\left\{b_{j_i}\frac{\partial h^{-1}(\eta_i)}{\partial\eta_i}\right\}^2.$$

Clearly, if h is the identity function, then

$$\frac{\partial^2\psi_n(\beta;p)}{\partial\beta_h\partial\beta_u} = \frac{2}{n}\sum_{i=1}^n x_{ih}x_{iu}b_{j_i}^2.$$

A.2 Proof of Theorem 1

Proof. Under the conditions stated (Li and Racine, 2008)

$$\max_z\left|\hat{G}_{Y|X}(z) - G_{Y|X}(z)\right| \to 0$$

as $n \to \infty$. We can also verify that

$$\sup_z\left|\hat{G}^c_{Y|X}(z|x) - G^c_{Y|X}(z|x)\right| \to 0.$$

Consequently,

$$\Pr\left(\lim_n \hat{G}^c_{Y|X}\left[h^{-1}\left\{x^\top\beta(p)\right\}|x\right] = G^c_{Y|X}\left[h^{-1}\left\{x^\top\beta(p)\right\}|x\right]\right) = 1.$$

Consider now $\gamma(p) \neq \beta^*(p)$. It is straightforward to verify that

$$\left(p - G^c_{Y|X}\left[h^{-1}\left\{x^\top\beta^*(p)\right\}|x\right]\right)^2 \leq \left(p - G^c_{Y|X}\left[h^{-1}\{x^\top\gamma(p)\}|x\right]\right)^2.$$

In fact, if $h^{-1}\left\{x^\top\beta^*(p)\right\} = y_j$ for some value of p and $y_j \in \mathcal{S}_Y$, then $(p - G^c_{Y|X}[h^{-1}\{x^\top\beta^*(p)\}|x])^2 = 0$; while all other values are obtained through interpolation. A consequence is that $\beta^*(p)$ is, eventually, a solution of the minimization problem in (A.2). Additionally, there is only one such solution, since, by assumption, $\Pr(Y = y|X) > 0$ for all $y \in \mathcal{S}_Y$, and $G^c(\eta(p)|x)$ is monotonic for $\pi_1 < p < \pi_k$, where π_1 and π_k are the mid-probabilities corresponding to, respectively, the smallest and largest discrete value (if $k = \infty$, then $\pi_1 < p < 1$). This implies consistency of $\hat{\beta}(p)$, the minimizer in (A.2). Consistency of the predicted mid-quantiles follows directly. □

A.3 Proof of Theorem 2

Proof. Since the differentiability of $\psi_n(\beta;p)$ follows from the assumptions, we can apply a first-order Taylor expansion to obtain

$$\nabla_\beta \psi_n(\hat{\beta};p) = \nabla_\beta \psi_n(\beta^*;p) + \nabla^2_\beta \psi_n(\beta^+;p)(\hat{\beta}-\beta^*), \tag{A.3}$$

where β^+ is a point in the interior of the hypercube delimited by $\hat{\beta}$ and β^*. Expressions for $\nabla_\beta \psi_n$ and $\nabla^2_\beta \psi_n$ are given in Section A.1. Note that $\nabla_\beta \psi_n(\hat{\beta};p) = 0$ since $\hat{\beta}$ is the minimizer in (A.2). The assumption on the design matrix guarantees that the Hessian $\nabla^2_\beta \psi_n(\beta^+;p)$ is positive definite. Hence, we can rewrite (A.3) as

$$\sqrt{n\prod_j \lambda_j}(\hat{\beta}-\beta^*) = -(\nabla^2_\beta \psi_n(\beta^+;p))^{-1}\sqrt{n\prod_j \lambda_j}\nabla_\beta \psi_n(\beta^*;p). \tag{A.4}$$

To derive the asymptotic distribution of $\hat{\beta}$, it suffices to study the asymptotic distribution of the right-hand side of (A.4). First, let $J(b) = E\left\{\nabla^2_\beta \psi_n(\beta;p)\Big|_{\beta=b}\right\}$. By using the consistency results in Theorem 1 and the triangle inequality, it is immediate to show that $\nabla^2_\beta \psi_n(\beta^+;p)$ weakly converges element-wise to $J(\beta^*)$. Using the results in Section A.1, we then can write

$$\begin{aligned}\sqrt{n\prod_j \lambda_j}\nabla_\beta \psi_n(\beta^*;p) = &-2\sqrt{\frac{1}{n}\prod_j \lambda_j}\sum_{i=1}^n \nabla_\beta \hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top \beta^*)|x\right\} \\ &\times\left[p-\hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top \beta^*)|x\right\}\right].\end{aligned}$$

We need to demonstrate that the expression above converges in distribution, thus we expand the quantities on the right-hand side as follows:

$$\begin{aligned}\sqrt{n\prod_j \lambda_j}\nabla_\beta \psi_n(\beta^*;p) = &-\frac{2}{n}\sum_{i=1}^n x_i \dot{h}^{-1}(\eta_i)\frac{p}{z_{j_i+1}-z_{j_i}}\sqrt{n\prod_j \lambda_j}\hat{G}_{Y|x_i}(z_{j_i+1}) \\ &+\frac{2}{n}\sum_{i=1}^n x_i \dot{h}^{-1}(\eta_i)\frac{p}{z_{j_i+1}-z_{j_i}}\sqrt{n\prod_j \lambda_j}\hat{G}_{Y|x_i}(z_{j_i}) \\ &+\frac{2}{n}\sum_{i=1}^n x_i \dot{h}^{-1}(\eta_i)\frac{\hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top \beta^*)|x_i\right\}}{z_{j_i+1}-z_{j_i}}\sqrt{n\prod_j \lambda_j}\hat{G}_{Y|x_i}(z_{j_i+1}) \\ &-\frac{2}{n}\sum_{i=1}^n x_i \dot{h}^{-1}(\eta_i)\frac{\hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top \beta^*)|x_i\right\}}{z_{j_i+1}-z_{j_i}}\sqrt{n\prod_j \lambda_j}\hat{G}_{Y|x_i}(z_{j_i}),\end{aligned} \tag{A.5}$$

where $\dot{h}^{-1}(\eta_i) = \frac{\partial h^{-1}(\eta_i)}{\partial \eta_i}$. First of all, as shown in Li and Racine (2008), $\sqrt{n \prod_j \lambda_j} \hat{G}_{Y|x_i}(z_{j_i})$ converges in distribution to a Gaussian random variable for all i. Additionally, the assumptions on the bandwidths guarantee asymptotic independence of $\hat{G}_{Y|x_h}(z)$ and $\hat{G}_{Y|x_l}(z)$ for $x_l \neq x_h$ and all z. To see this, note that $K_\lambda(X_i, x) \to 0$ for all $X_i \neq x$. According to the dominated convergence theorem, the asymptotic covariance of $\hat{G}_{Y|x_h}(z)$ and $\hat{G}_{Y|x_l}(z)$ is zero. Asymptotic independence follows by the Cramer-Wold device. Furthermore, $\Pr(z_{j_i+1} - z_{j_i} \neq 0) = 1$ since Y is discrete. Finally, note that by our Theorem 1, $\hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top \beta^*)|x_i\right\}$ converges in probability to a constant value. By combining the results above with the assumptions on the design matrix (namely, that $1/n \sum_i x_i$ converges to a bounded vector), we obtain convergence in distribution of the right-hand side of (A.5) to a Gaussian random variable.

Therefore, $\sqrt{n \prod_j \lambda_j} \nabla_\beta \psi_n(\beta^*; p)$ is asymptotically normal with variance

$$D(\beta^*) = \mathrm{Var}\left(\frac{2\sqrt{\prod_j \lambda_j}}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\beta \hat{G}^c_{Y|X}\left\{h^{-1}(x_i^\top \beta^*)|x_i\right\} \left[p - \hat{G}^c_{Y|X}\left\{h^{-1}\left(x_i^\top \beta^*\right)|x_i\right\}\right] \right). \tag{A.6}$$

By letting

$$V(\beta^*) = J(\beta^*)^{-1} D(\beta^*) J(\beta^*)^{-1}, \tag{A.7}$$

we obtain

$$V(\beta^*)^{-1/2} \sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, I_q).$$

$\square$

A consistent estimator of $V(\beta^*)$ could be found by calculating sample averages of the quantities involved in $J(\beta^*)$, and computing $D(\beta^*)$ via resampling. However, using expression (2.10) in the Manuscript leads to an analytical calculation of the variance of $\hat{\beta}$ with clear computational advantages.

B Supplementary simulation results

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

	$n = 100$		$n = 500$		$n = 1000$		
p	Bias	RMSE	Bias	RMSE	Bias	RMSE	$\bar{H}$
0.2	−0.046	0.803	−0.037	0.528	−0.036	0.453	8.995
0.3	0.071	0.827	0.016	0.535	0.000	0.456	9.995
0.4	0.122	0.849	0.034	0.537	0.014	0.455	10.995
0.5	0.156	0.854	0.046	0.532	0.022	0.451	11.995
0.6	0.197	0.851	0.055	0.521	0.031	0.439	12.995
0.7	0.245	0.837	0.067	0.507	0.041	0.425	13.995
0.8	0.346	0.839	0.111	0.491	0.069	0.412	14.995

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

	$n = 100$		$n = 500$		$n = 1000$		
p	Bias	RMSE	Bias	RMSE	Bias	RMSE	$\bar{H}$
0.2	−0.463	1.838	−0.324	1.227	−0.344	1.114	13.988
0.3	−0.545	2.167	−0.394	1.462	−0.390	1.343	16.986
0.4	−0.562	2.431	−0.457	1.719	−0.431	1.591	19.983
0.5	−0.501	2.662	−0.507	1.972	−0.463	1.848	22.981
0.6	−0.228	2.857	−0.474	2.211	−0.461	2.104	25.978
0.7	0.175	3.060	−0.275	2.455	−0.300	2.353	28.976
0.8	0.843	3.376	0.196	2.749	0.108	2.659	31.973

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

	$n = 100$		$n = 500$		$n = 1000$		
p	Bias	RMSE	Bias	RMSE	Bias	RMSE	$\bar{H}$
0.2	−18.167	36.048	−13.598	27.366	−12.075	24.612	216.351
0.3	−9.786	23.088	−8.372	19.552	−7.651	17.930	220.421
0.4	−3.072	14.097	−4.141	13.343	−3.996	12.624	223.926
0.5	3.416	11.066	0.076	7.978	−0.371	7.560	227.223
0.6	9.946	14.939	4.761	7.626	3.551	6.309	230.542
0.7	16.268	21.987	9.473	12.590	7.860	10.549	234.117
0.8	26.405	35.420	15.210	20.174	12.691	16.860	238.331

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

	$n = 100$		$n = 500$		$n = 1000$		
p	Bias	RMSE	Bias	RMSE	Bias	RMSE	$\bar{H}$
0.5	−0.000	0.067	0.000	0.029	0.000	0.021	0.577

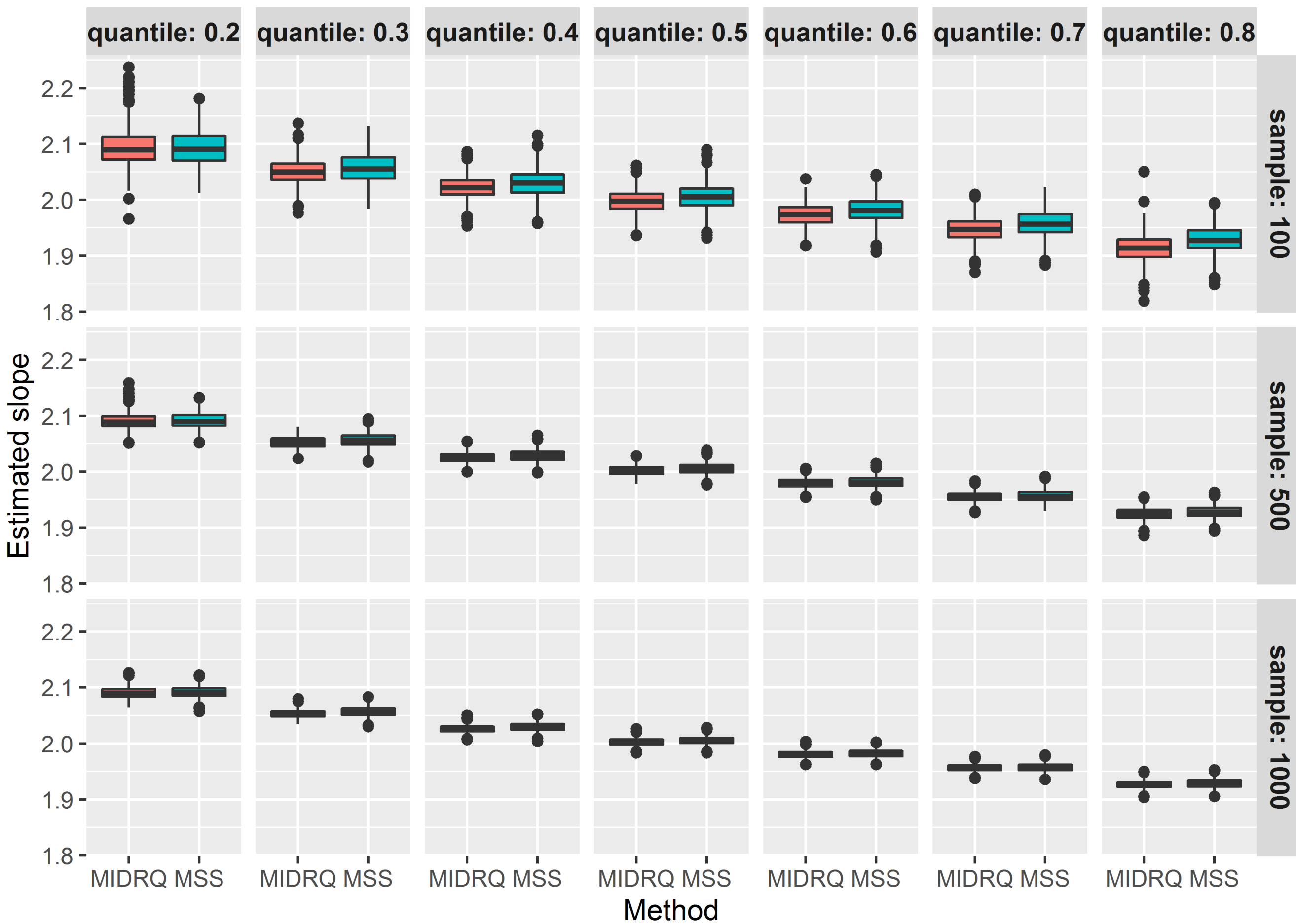

Figure 1: Boxplots of the estimates of the slope parameter from Machado and Santos Silva's (2005) estimator (MSS) and mid-quantile regression (MIDQR) for $p \in \{0.2, 0.4, 0.6, 0.8\}$ and $n \in \{100, 500, 1000\}$ when data are generated using the Poisson model (3a).

C R code

In this section, we provide an example on how to do inference on mid-quantile regression models using the R package `Qtools` (Geraci, 2016). The latter is available on CRAN and can be installed as follows:

```
install.packages("Qtools")
```

We consider the dataset `esterase`, which is available in the Qtools package. The dataset contains data from an essay for the concentration of an enzyme esterase. The observed concentration of esterase was recorded (`esterase`), and then in a binding experiment the number of bindings were counted (`Count`). After loading the package, the following code shows how to attach the dataset and access the R documentation describing the variables:

```
library(Qtools)
data(esterase)
?esterase
```

```
> head(esterase)
  Esterase Count
1      3.1    28
2      5.6   166
3      6.1    52
4      6.4    84
5      6.5    85
6      6.7    86
```

We estimate the marginal mid-quantiles of the discrete variable `Count` using the function `midquantile`.

```
fit <- midquantile(esterase$Count, probs = 1:3/4)
```

```
> str(fit)
List of 5
```

```
 $ call: language midquantile(x = esterase$Count, probs = 1:3/4)
 $ x   : num [1:3] 0.25 0.5 0.75
 $ y   : num [1:3] 147 269 419
 $ fn  :function (v)
 $ data: int [1:113] 28 166 52 84 85 86 127 104 107 96 ...
 - attr(*, "class")= chr "midquantile"
```

The output is a list that contains the estimated mid-quantiles (`y`) at the specified probabilities (`x`). It also contains the interpolating mid-quantile function (`fn`) which can be plotted using the associated `plot.midquantile` function. Confidence intervals for mid-quantile estimates can be obtained using `confint.midquantile`.

Suppose we want to fit the linear model $H(p) = \beta_0 + \beta_1(p)x$ to estimate the 0.25 and 0.75 conditional mid-quantiles of `Count` as a function of `esterase`. We use the main command `midrq` where the argument `tau` specifies the level of the quantiles of interest.

```
fit <- midrq(Count ~ Esterase, tau = c(0.25, 0.75), data = esterase,
type = 3, control = midrqControl(method = "Nelder-Mead", ecdf_est = "npc"))

> fit
call:
midrq(formula = Count ~ Esterase, data = esterase, tau = c(0.25,
    0.75), type = 3, control = midrqControl(method = "Nelder-Mead",
    ecdf_est = "npc"))



Coefficients linear predictor:
                 0.25     0.75
(Intercept) -48.97063 16.02915
Esterase     15.61743 19.12168


Degrees of freedom: 113 total; 111 residual
```

There are three estimators available in `midrq` and these can be selected via the argument `type`. Using `type = 1`, the minimization of the objective function (2.7) in the Manuscript is carried out using a general purpose optimizer (by default, this is Nelder-Mead, although it can be changed via `midrqControl`). When `type = 2`, optimization is based on a CUSUM process (which is not discussed in the present work and should be considered experimental). Finally, `type = 3` gives the least-squares-type estimator in equation (2.9) of the Manuscript. On the other hand, the argument `ecdf_est` in `midrqControl` controls the conditional mid-CDF estimator (for example, `ecdf_est = "npc"` gives the kernel estimator by Hayfield and Racine (2008)).

The package provides several S3 methods for fitted `midrq` objects including: `summary`, which gives standard errors, p-values, and confidence intervals; `coef` to extract estimates of the regression coefficients; `vcov` to extract the variance-covariance matrix of the estimator $\hat{\beta}(p)$ defined in Section 2.3 of the Manuscript; and `predict` and `residuals`, whose names are self-explanatory. The function `midq2q` gives an estimate of ordinary quantiles using the procedure described in Section 2.4 of the Manuscript. Finally, we draw attention on the availability in the `Qtools` package of the functions `midecdf` and `cmidecdf` for estimating marginal and conditional mid-cumulative probabilities, respectively.

References

Geraci, M. (2016). Qtools: A collection of models and other tools for quantile inference. *R Journal 8*(2), 117–138.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*(5).

Li, Q. and J. S. Racine (2008). Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics 26*(4), 423–434.

Machado, J. A. F. and J. M. C. Santos Silva (2005). Quantiles for counts. *Journal of the American Statistical Association 100*(472), 1226–1237.