Biometrics 69, 651–660
September 2013

DOI: 10.1111/biom.12053

Bayesian Quantile Regression for Censored Data

Brian J. Reich, Luke B. Smith

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
email: brian_reich@ncsu.edu

Summary. In this paper we propose a semiparametric quantile regression model for censored survival data. Quantile regression permits covariates to affect survival differently at different stages in the follow-up period, thus providing a comprehensive study of the survival distribution. We take a semiparametric approach, representing the quantile process as a linear combination of basis functions. The basis functions are chosen so that the prior for the quantile process is centered on a simple location-scale model, but flexible enough to accommodate a wide range of quantile processes. We show in a simulation study that this approach is competitive with existing methods. The method is illustrated using data from a drug treatment study, where we find that the Bayesian model often gives smaller measures of uncertainty than its competitors, and thus identifies more significant effects.

Key words: Accelerated failure time model; Markov chain Monte Carlo; Quantile regression; Survival data.

1. Introduction

Survival data analysis typically relies on a parametric assumption about the relationship between the covariates and the survival distribution, for example, the proportional hazards, proportional odds, or accelerated failure time models. While these methods have attractive features and rich histories, in this paper we pursue quantile regression. The linear quantile regression model assumes that each quantile of the survival (or log survival) distribution is a linear combination of the covariates. The covariates are allowed to have different effects on each quantile level, and thus varying effects at different stages of the follow-up period.

For example, we consider a drug treatment study where time until relapse is modeled in terms of treatment and other factors including compliance, IV drug use, age, and race. Quantile regression provides a comprehensive analysis, as we can study treatment effects early and late in the follow-up period. This could be used to identify subjects and times after treatment that would be aided by further intervention. Also, flexible model-based methods such as those proposed here provide straight-forward predictions of the relapse time of individual subjects, which could be used to identify the optimal treatment for individuals.

There are several frequentist approaches to quantile regression for censored data (e.g., Powell, 1984; Lindgren, 1997; Portnoy, 2003; Koenker, 2008; Peng and Huang, 2008). These model-free methods are geared towards estimating covariate effects at a single quantile level (e.g., the median). The algorithm is then applied in separate analyses to determine the effects at different quantile levels. There are also many Bayesian approaches to estimate effects at a single quantile level (Yu and Moyeed, 2001; Kottas and Gelfand, 2001; Hanson and Johnson, 2002; Kottas and Krnjajić, 2009; Reich, Bondell, and Wang, 2010). Focusing on survival data, Lin et al. (2012) develop a semi-parametric Bayesian median regression model. In particular, this approach models only the median survival time as a linear function of the covariates, and thus would not permit inference on differential covariate effects early versus late in follow-up period. For non-censored data it has been shown that modeling quantile levels simultaneously provides better estimates and more power for identifying significant effects than separate analyses (Bondell, Reich, and Wang, 2010; Reich, Fuentes, and Dunson, 2011). In this paper, we propose a Bayesian quantile regression model for censored data that jointly analyzes all quantile levels.

In a Bayesian setting, modeling quantiles simultaneously amounts to specifying a survival distribution with the desired quantiles. Quantile regression models have been proposed for non-censored data that allow for different covariate effects at different quantile levels (Dunson and Taylor, 2005; Hjort and Walker, 2009; Reich et al., 2011; Todkar and Kadane, 2011; Reich, 2012). The most similar to our approach is Reich (2012), who use a piecewise quantile model with prior centered on the normal distribution. They allow the quantile function to vary over space and time but without covariates. In this paper we generalize this to censored data with arbitrary centering distribution, and to include covariates.

The proposed model has several nice properties. Unlike Reich et al. (2011) and Todkar and Kadane (2011), the model permits a simple closed-form for the likelihood, which facilitates straight-forward MCMC sampling to explore the posterior. Given this likelihood, it is possible to handle any type of censoring, that is, left-censoring, right-censoring, or interval-censoring. Despite this computational simplicity, we show that the model is flexible enough to fit any valid quantile process at any finite set of quantile levels. As with many semi-parametric Bayesian models, we allow for a wide class of models while centering the quantile process on a parametric model, for example, the accelerated failure time model. We show via simulation that incorporating valid prior information can substantially improve estimation over model-free methods. Also, to deal with high-dimensional problems

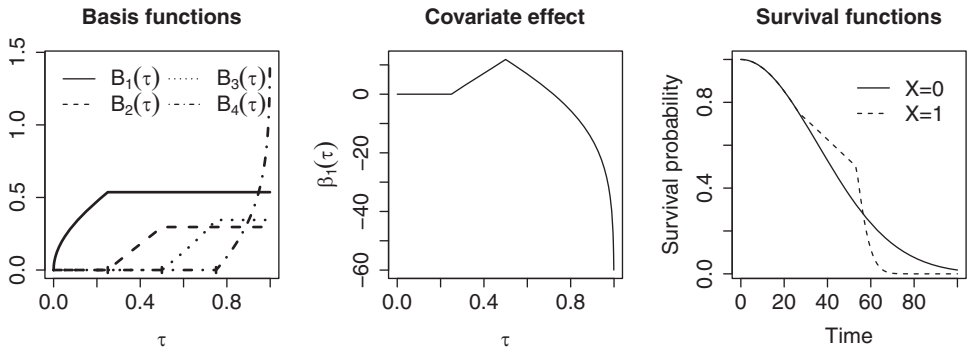

Figure 1. Illustration of the quantile model with $L = 4$ basis functions, $p = 1$ covariate, $(\alpha_{00}, \ldots, \alpha_{0L}) = (0, 50, 50, 50, 50)$, $(\alpha_{10}, \ldots, \alpha_{1L}) = (0, 0, 40, -40, -40)$, and β_0 equal to the Weibull quantile function with shape equal 2 and scale equal 1. Plotted are the basis functions $B_l(\tau)$ (left), the covariate effect $\beta_1(\tau)$ (middle), and the survival function of T_i for $X = 0$ and $X = 1$ (right).

we incorporate Bayesian variable selection techniques to eliminate unneeded covariates from the model.

2. Semiparametric Quantile Regression Model

Denote T_i and C_i as the survival and censoring times, respectively, for subject $i = 1, \ldots, n$. We observe the follow-up time $Y_i = \min\{T_i, C_i\}$, censoring indicator $\delta_i = I(T_i \leq C_i)$, and covariates $\mathbf{X}_i = (X_{i0}, X_{i1}, \ldots, X_{ip})$, where $x_{i0} = 1$ for the intercept. Our objective is to model the quantile function of T_i or $Z_i = \log(T_i)$ as a function of the covariate $\mathbf{X}_i$. We first describe the quantile function for Z_i in Section 2.1, and then discuss T_i in Section 2.2.

2.1. *Model Formulation*

The quantile function, denoted $q(\tau|\mathbf{X}_i)$, is defined as the function satisfying $\text{Prob}\,[Z_i < q(\tau|\mathbf{X}_i)] = \tau \in [0, 1]$. For example, with $\tau = 0.5$, $q(0.5|\mathbf{X}_i)$ is the median log survival time of a subject with covariate vector $\mathbf{X}_i$. Linear quantile regression assumes that the τ^{th} quantile is a linear combination of the covariates, $q(\tau|\mathbf{X}_i) = \sum_{j=0}^{p} X_{ij}\beta_j(\tau)$. In this model, the vector of regression coefficients $\boldsymbol{\beta}(\tau) = [\beta_0(\tau), \ldots, \beta_p(\tau)]^T$ is different for each quantile level τ, allowing for different covariate effects on different aspects of the survival distribution, for example, $\boldsymbol{\beta}(0.1)$ measures the effects early in the follow-up while $\boldsymbol{\beta}(0.5)$ determines median log survival. We center our Bayesian model on the heteroskedastic accelerated failure time model $Z_i = \mathbf{X}_i\boldsymbol{\alpha}_0 + (\mathbf{X}_i\boldsymbol{\alpha}_1)\varepsilon_i$, where $\boldsymbol{\alpha}_0 = (\alpha_{00}, \ldots, \alpha_{0p})^T$ and $\boldsymbol{\alpha}_1 = (\alpha_{10}, \ldots, \alpha_{1p})^T$ control the location and scale (with $\mathbf{X}_i\boldsymbol{\alpha}_1$ restricted to be positive), respectively, and ε_i are independent errors with quantile function $q_0(\tau)$. This model has quantile function

$$\begin{aligned} q(\tau|\mathbf{X}_i) &= \mathbf{X}_i\boldsymbol{\alpha}_0 + (\mathbf{X}_i\boldsymbol{\alpha}_1)q_0(\tau) \\ &= \sum_{j=0}^{p} X_{ij}\left[\alpha_{0j} + \alpha_{1j}q_0(\tau)\right] = \sum_{j=0}^{p} X_{ij}\beta_j(\tau), \end{aligned} \quad (1)$$

and thus the quantile function for covariate j, $\beta_j(\tau) = \alpha_{0j} + \alpha_{1j}q_0(\tau)$, varies with τ if $\alpha_{1j} \neq 0$.

We extend this to allow for a richer span of models for the quantile function. Models must satisfy the restriction that $q(\tau|\mathbf{X}_i)$ is continuous and monotonically increasing in τ for all $\mathbf{X}_i$. Generalizing (1), we model the derivative piece-wise over $L > 1$ intervals separated by breakpoints $0 = \kappa_0 < \kappa_1 < \cdots < \kappa_{L-1} < \kappa_L = 1$,

$$\frac{dq(\tau|\mathbf{X}_i)}{d\tau} = \sum_{l=1}^{L} \mathrm{I}\left(\kappa_{l-1} < \tau \leq \kappa_l\right)(\mathbf{X}_i\boldsymbol{\alpha}_l)\frac{dq_0(\tau)}{d\tau}. \quad (2)$$

The effect of the covariates on the derivative of the quantile function for $\tau \in [\kappa_{l-1}, \kappa_l]$ is determined by $\boldsymbol{\alpha}_l = (\alpha_{l0}, \ldots, \alpha_{lp})^T$. For the quantile function to be increasing, the derivative must be positive for all τ which is true if and only if $\mathbf{X}_i\boldsymbol{\alpha}_l > 0$ for all l and for all $\mathbf{X}_i$.

The continuous quantile function corresponding to (2) is $c_i + \sum_{l=1}^{L}(\mathbf{X}_i\boldsymbol{\alpha}_l)B_l(\tau)$, where c_i is a constant and B_l are known functions of q_0 (Figure 1),

$$B_1(\tau) = \begin{cases} q_0(\tau), & \tau \leq \kappa_1 \\ q_0(\kappa_1), & \tau > \kappa_1 \end{cases} \quad \text{and}$$

$$B_l(\tau) = \begin{cases} 0, & \tau \leq \kappa_{l-1} \\ q_0(\tau) - q_0(\kappa_{l-1}), & \kappa_{l-1} < \tau \leq \kappa_l \\ q_0(\kappa_l) - q_0(\kappa_{l-1}), & \tau > \kappa_l \end{cases} \quad (3)$$

for $l > 1$. To retain the connection with (1), we take the constant to be $c_i = \mathbf{X}_i\boldsymbol{\alpha}_0$, giving

$$\begin{aligned} q(\tau|\mathbf{X}_i) &= \mathbf{X}_i\boldsymbol{\alpha}_0 + \sum_{l=1}^{L}(\mathbf{X}_i\boldsymbol{\alpha}_l)B_l(\tau) \\ &= \sum_{j=0}^{p} X_{ij}\left[\alpha_{0j} + \sum_{l=1}^{L} B_l(\tau)\alpha_{lj}\right] = \sum_{j=0}^{p} X_{ij}\beta_j(\tau). \end{aligned} \quad (4)$$

With this choice of c_i, this model reduces to the parametric location-scale model (1) if $L = 1$.

The quantile function for covariate j is $\beta_j(\tau) = \alpha_{0j} + \sum_{l=1}^{L} B_l(\tau)\alpha_{lj}$, which is a linear combination of fixed basis functions B_l with coefficients α_{lj}. In Figure 1, $\beta_1(\tau)$ is

positive for $\tau \in (0.25, 0.75)$ and negative for $\tau > 0.75$. The corresponding survival curves have changepoints at probabilities $\{0.25, 0.50, 0.75\}$, with the subject with $X = 1$ having higher survival probabilities for times (30, 60) and lower survival probability for times over 60. Therefore, this flexible model can accommodate, among other things, crossing survival curves.

The derivative of the quantile process is positive if an only if $\mathbf{X}_i\boldsymbol{\alpha}_l > 0$ for all $\mathbf{X}_i$ and all $l = 1, \ldots, L$. To ensure these constraints are satisfied, we use a latent variable approach similar to Reich et al. (2011) and Reich (2012) for related models. We assume that the covariates are scaled so that $X_{ij} \in [-1, 1]$, that is, $\mathbf{X}_i \in \mathcal{S} = \{(X_0, \ldots, X_p) | X_0 = 1, X_1, \ldots, X_p \in [-1, 1]\}$. In this case, $\mathbf{X}_i\boldsymbol{\alpha}_l$ is minimized by the $\mathbf{X}_i$ with $X_{ij} = -1$ for covariates with $\alpha_{lj} > 0$ and $X_{ij} = 1$ for covariates with $\alpha_{lj} < 0$. In this worst case ("WC"), $\mathbf{X}_i\boldsymbol{\alpha}_l$ equals $WC(\boldsymbol{\alpha}) = \alpha_{l0} - \sum_{j=1}^{p} |\alpha_{lj}|$. To satisfy this criteria for all $\mathbf{X} \in \mathcal{S}$, we build the prior using latent unconstrained coefficients $\boldsymbol{\alpha}_j^* = (\alpha_{l0}^*, \ldots, \alpha_{lp}^*)$, and set

$$\alpha_{lj} = \begin{cases} \alpha_{lj}^*, & WC(\boldsymbol{\alpha}_l^*) > 0 \\ \epsilon I(j = 0), & WC(\boldsymbol{\alpha}_l^*) < \epsilon, \end{cases} \tag{5}$$

where $\epsilon > 0$ is a small constant. Although these restrictions on $\boldsymbol{\alpha}_l$ may seems prohibitive, this provide a very flexible model. To demonstrate the flexibility of this approach, we state and prove (in Web Appendix A) the following theorem.

THEOREM 1. Let $\tilde{\boldsymbol{\beta}}(\tau) = [\tilde{\beta}_0(\tau), \ldots, \tilde{\beta}_p(\tau)]$ be any valid set of quantile functions so that $\tilde{q}(\tau|\mathbf{X}) = \mathbf{X}\tilde{\boldsymbol{\beta}}(\tau)$ is monotonically increasing in τ for all $\mathbf{X} \in \mathcal{S}$, and let $q_0(\tau)$ be any monotonically-increasing base quantile function. Then there exist values of ϵ and $\{\alpha_{lj}\}$ satisfying $WC(\boldsymbol{\alpha}_l) > \epsilon$ for all $l = 1, \ldots, L$ so that $\boldsymbol{\beta}(\tau) = \tilde{\boldsymbol{\beta}}(\tau)$ at the interior breakpoints $\tau \in \{\kappa_1, \ldots, \kappa_{L-1}\}$.

Therefore, if interest is restricted to a finite set of quantile levels $\{\kappa_1, \ldots, \kappa_{L-1}\}$, then this semiparametric model with any choice of base quantile function q_0 spans the entire class of valid quantile functions at these quantile levels. Also, this result suggests that for large L the semiparametric model can approximate a wide class of quantile functions.

2.2. *Prior Selection*

As with many semi-parametric methods that make use of a basis expansion, a crucial step in applying this method is to select the form of the basis functions (q_0) as well as the number of basis functions (L). By construction, $\sum_{l=1}^{L} B_l(\tau) = q_0(\tau)$, and so if $\boldsymbol{\alpha}_1 = \cdots = \boldsymbol{\alpha}_L$, then the quantile function reduces to the heteroskedastic model (1). Therefore, q_0 determines the shape of the residual distribution in this parametric special case, and if prior information from parametric modeling exists this can be used to select q_0. In the absence of prior information, exploratory analysis using classical estimates (e.g., Portnoy, 2003; Peng and Huang, 2008) computed separately for several quantile levels and plotted against quantile level may suggest some reasonable choices for q_0. In general, we recommend fitting a few combinations of q_0 and L and comparing fits using goodness-of-fit criteria, as illustrated in Section 4.

For identification purposes, q_0 should have location fixed at zero and scale fixed at one. Examples of symmetric quantile functions include the standard normal $q_0(\tau) = \Phi^{-1}(\tau)$, the standard logistic $q_0(\tau) = \log[\tau/(1-\tau)]$, and the standard t quantile function with $\phi > 0$ degrees of freedom. For asymmetry, we also consider the asymmetric Laplace quantile function Kotz, Kozubowski, and Podgorski, 2001 with shape parameter $\phi \in (0, 1)$. These final two quantile functions have a shape parameter, ϕ, which we treat as an unknown parameter to estimated.

To complete the Bayesian formulation, we must specify the priors for the coefficients that define the likelihood, $\{\alpha_{lj}^*\}$. The basis coefficients (excluding the location effect α_{0j}) for covariate j, $\boldsymbol{\alpha}_j^* = (\alpha_{1j}^*, \ldots, \alpha_{Lj}^*)^T$, have multivariate normal priors with $\mathrm{E}(\alpha_{lj}^*) = \mu_j$ and autoregressive covariance $\mathrm{Cov}(\alpha_{lj}^*, \alpha_{kj}^*) = \sigma_j^2 \rho_j^{|k-l|}$. Therefore, if the prior variances σ_j^2 are near zero, the model reduces to (1) with scale $\mathbf{X}_i\boldsymbol{\mu}$, where $\boldsymbol{\mu} = (\mu_0, \ldots, \mu_p)^T$. The prior for the remaining parameters in the model for log survival are taken to be $\alpha_{0j}, \mu_j \overset{iid}{\sim} \mathrm{N}(0, c^2)$, $\sigma_j^{-2} \overset{iid}{\sim} \mathrm{Gamma}(a, b)$, $\phi \sim f(\phi)$, and $\rho_j \sim \mathrm{Unif}(0,1)$.

Within this semiparametric framework, it is also straightforward to model survival rather than log survival. To model survival directly, we must select a prior so that $q(0|\mathbf{X}_i) = 0$. If we simply fix $\boldsymbol{\alpha}_0 = 0$ and select q_0 to be the quantile function of a density with lower bound zero. Then $q_0(0) = 0$, $B_l(0) = 0$ for all l, and $q(0|\mathbf{X}_i) = 0$ for all $\mathbf{X}_i$. A consequence of this is that $\beta_j(0) = 0$ for all j, which is reasonable because the lower bound of survival for all subjects is assumed to be zero regardless of $\mathbf{X}_i$. Possible base quantile functions q_0 for survival are the gamma quantile function with shape ϕ and scale one, and the Weibull quantile function with shape ϕ and scale one.

We also note that in this framework it is not necessary to allow all covariates to affect all quantile levels. For example, it may be of interest to intensively study the effect of one covariate (e.g., treatment) over quantile levels, while simply accounting for the effects of other covariates in the location component of the model. This is accomplished by setting $\alpha_{1j} = \cdots = \alpha_{Lj} = 0$ for location-only covariates.

A final consideration when specifying the prior for this model for the quantile function of each covariate in terms of the base quantile function q_0, is that if the base quantile function is chosen to have no upper bound, that is, $q_0(1) = \infty$, then $|\beta_j(1)| = \infty$ with probability one. If this is a concern, a remedy is to fix the final coefficient $\alpha_{Lj}^* = 0$, so that $\beta_j(\tau) = \beta_j(\kappa_{L-1})$ for all $\tau > \kappa_{L-1}$. In our analysis we do not fix the final term to zero because our focus is not on the extreme tail and we wish to maximize flexibility for estimating the quantiles of interest.

2.3. *Censored Likelihood and Computing Details*

The density of Z_i corresponding to (4) has the relatively simple form

$$f(z|\mathbf{X}_i, \boldsymbol{\alpha}) = \sum_{l=1}^{L} \frac{I\left[q(\kappa_{l-1}|\mathbf{X}_i) < z < q(\kappa_l|\mathbf{X}_i)\right]}{\mathbf{X}_i\boldsymbol{\alpha}_l} f_0 \times \left[\frac{z - \mathbf{X}_i\boldsymbol{\alpha}_0 - I(l > 1)\mathbf{X}_i\boldsymbol{\alpha}_l q_0(\kappa_l)}{\mathbf{X}_i\boldsymbol{\alpha}_l}\right], \tag{6}$$

where f_0 is the density corresponding to q_0 and $\boldsymbol{\alpha} = \{\alpha_{lj}\}$. Therefore, while the quantile functions $\beta_j(\tau)$ are continuous functions of τ, the density function (6) has discontinuities at the interior breakpoints $q(\kappa_l|\mathbf{X}_i)$. Discontinuous densities are common in Bayesian nonparametrics (e.g., Ferguson, 1973, 1974; Lavine, 1992, 1994). In this case, the breakpoints are random functions of the unknown regression coefficients $\boldsymbol{\alpha}$, and thus the posterior mean of the density averaging over uncertainty in these coefficients is almost surely continuous.

Similarly, the distribution function is

$$F(z|\mathbf{X}_i, \boldsymbol{\alpha}) = \begin{cases} F_o\left[\frac{z-\mathbf{X}\boldsymbol{\alpha}_0}{\mathbf{X}_i\boldsymbol{\alpha}_1}\right], & z < q(\kappa_1|\mathbf{X}_i) \\ F_o\left[\frac{z-q(\kappa_{l-1}|\mathbf{X}_i)+(\mathbf{X}_i\boldsymbol{\alpha}_l)q_0(\kappa_l)}{\mathbf{X}_i\boldsymbol{\alpha}_l}\right], & z \in [q(\kappa_{l-1}|\mathbf{X}_i), \\ & q(\kappa_l|\mathbf{X}_i)] \text{ for } l>1 \end{cases} \tag{7}$$

where F_0 is the distribution function corresponding to q_0. Combining these results, the censored likelihood for right-censored data $[\log(Y_1), \delta_1], \ldots, [\log(Y_n), \delta_n]$ is

$$\prod_{i=1}^{n} f[\log(Y_i)|\mathbf{X}_i, \boldsymbol{\alpha}]^{\delta_i} \left\{1 - F[\log(Y_i)|\mathbf{X}_i, \boldsymbol{\alpha}]\right\}^{1-\delta_i}. \tag{8}$$

Other types of censoring (interval censoring) are also easily accommodated within this likelihood-based approach. With this closed-form expression of the likelihood, MCMC sampling proceeds using the standard Metropolis within Gibbs algorithm (Chib and Greenberg, 1995) as described in Web Appendix B. We draw 25,000 samples and discard the first 5000 as burn-in. Convergence is monitored using trace plots of several representative parameters. The methods introduced in this paper are implemented in the R package `BSquare`.

2.4. *Variable Selection*

We use Bayesian variable selection methods to determine the subset of covariates to be included in the model. We assume that the number of predictors is not so large that it is infeasible to include all p parameters in the location component of the model, $\mathbf{X}_i\boldsymbol{\alpha}_0$. Even in this moderate case, including a large number of covariates in shape/scale component $\sum_{l=1}^{L} \mathbf{X}_i\boldsymbol{\alpha}_l q_0(\tau)$ is problematic because there are pL parameters, which is cumbersome for large L, and the constraint on $WC(\boldsymbol{\alpha}_l)$ becomes more restrictive when there are many parameters. Therefore, we focus our attention on determining the subset of variables to include in this component of the model.

We introduce binary indicators θ_j to index the model for covariate j, so that if $\theta_j = 1$ then covariate j affects the shape and scale of the survival distribution, and if $\theta_j = 0$ then covariate j is only a location-shift parameter. In Bayesian variable selection (e.g., O'Hara and Sillanpaa, 2009), we seek the posterior of $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_p)$. If all the shape/scale parameters for variable j are zero, that is, $\alpha_{1j} = \cdots = \alpha_{Lj} = 0$, then $\beta_j(\tau) = \alpha_{0j}$ for all τ, and thus variable j only affects the survival distribution via a location shift. This suggests a way to compute the posterior distribution of $\boldsymbol{\theta}$ by treating it as unknown parameter in the Bayesian model and analyzing its posterior samples from the MCMC algorithm. The model becomes,

$$\alpha_{lj} = \begin{cases} \theta_j\alpha^*_{lj}, & WC(\boldsymbol{\theta}\boldsymbol{\alpha}^*_l) > 0 \\ \epsilon I(j=0), & WC(\boldsymbol{\theta}\boldsymbol{\alpha}^*_l) < \epsilon, \end{cases} \tag{9}$$

where $\boldsymbol{\theta}\boldsymbol{\alpha}^*_l$ denotes $(\theta_0\alpha^*_{l0}, \ldots, \theta_p\alpha^*_{lp})$ and all other aspect of the model remain the unchanged. In this formulation, when $\theta_j = 0$ then α_{lj} is forced to be zero for all l. Therefore, we report the posterior mean of the parameters θ_j as the posterior probability that covariate j is included in shape and scale, and thuse has a non-constant effect across quantile levels. We fix $\theta_0 = 1$ to include the intercept, and use priors $\theta_j \sim$ Bernoulli(0.5) for $j = 1, \ldots, p$ to reflect the prior information that all subsets of the covariates are equally likely.

3. Simulation Study

We compare the proposed method with the frequentist procedures of Portnoy (2003) and Peng and Huang (2008) implemented in `quantreg` package (Koenker, 2010) in `R` (R Development Core Team, 2010), as well as the parametric heterodskedastic logistic model (1) with $L = 1$ and $q_0(\tau) = \log[\tau/(1-\tau)]$. We fit three versions of the semiparametric model by varying the base quantile function and the number of basis function: logistic q_0 with $L = 4$ and $L = 8$, and the asymmetric Laplace q_0 with $L = 4$. For priors we select $c = 10$, $a = b = 0.1$, and $\rho_j \overset{iid}{\sim}$ Unif(0, 1) for all models, $\phi \sim$ Unif(0, 1) for the asymmetric Laplace, and because there is only a single covariate we fix $\theta_1 = 1$ so it is included with probability 1. We consider three simulation designs:

(1) $\beta_0(\tau) = \log[\tau/(1-\tau)]$; $\beta_1(\tau) = 2$
(2) $\beta_0(\tau) = \text{sign}(0.5-\tau)\log(1-2|0.5-\tau|)$; $\beta_1(\tau) = 2\tau$
(3) $\beta_0(\tau) = \Phi^{-1}(\tau)$; $\beta_1(\tau) = 2\min\{\tau - 0.5, 0\}$

where β_0 for the second simulation is the double exponential quantile function. Data are generated by sampling $X_{1i} \overset{iid}{\sim}$ Unif(-1,1), $U_i \overset{iid}{\sim}$ Unif(0,1), and setting $Z_i = \beta_0(U_i) + X_{1i}\beta_1(U_i)$. Censoring times are generated as $C_i \sim$ Unif(-2,7) for all designs, giving 20–30% censoring for three scenarios. Sample datasets from each design are plotted in Figure 2. For each design we generated $S = 200$ data sets with $n = 250$. Models are compared in terms of estimating the covariate's $\beta_1(\tau)$ using root mean squared error $RMSE(\tau) = \sqrt{\frac{1}{S}\sum_{s=1}^{S}[\beta_1(\tau) - \hat{\beta}_1^{(s)}(\tau)]^2}$, where $\hat{\beta}_1^{(s)}(\tau)$ is the estimate (posterior mean for Bayesian methods) of $\beta_1(\tau)$ for dataset s. We also compute the coverage of 95% intervals.

In the first design, the effect of the covariate is constant across quantile levels, and the model-based Bayesian methods are far more effective than the frequentist approaches (Figure 3) because they share information across quantile levels. For the Bayesian methods, the results are not very sensitive to the base quantile function or the number of knots. For these methods, the true quantile curve is obtained by setting $\alpha_{01} = 2$ in the location and $\alpha_{11} = \cdots = \alpha_{L1} = 0$ in the shape/scale component of the model. Therefore, one might expect the $L = 1$ model to be optimal because there are fewer unnecessary parameters. However, we find this model actually performs

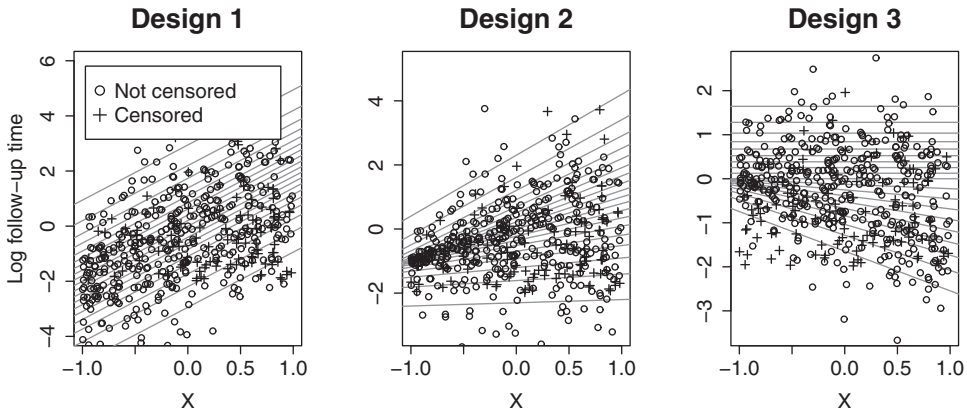

Figure 2. A sample from each simulation design. The gray lines represent the true quantiles for $\tau = 0.05, 0.10, \ldots, 0.95$.

the worst in the tails, perhaps because of inflexibility in the shape of the quantile function.

For designs two and three, the true quantile curves cannot be fit exactly with a finite number of basis functions for the Bayesian methods. However, the semiparametric approach remains competitive with the non-parametric frequentist methods in these cases. For the second design with smooth (linear) $\beta_1(\tau)$ the model with logistic base quantile function and $L = 4$ terms has smaller RMSE than the frequentist methods for all quantile levels, with the largest difference in the tails. The true quantile function for the third design has a point of non-differentiability at $\tau = 0.5$, and the Bayesian methods

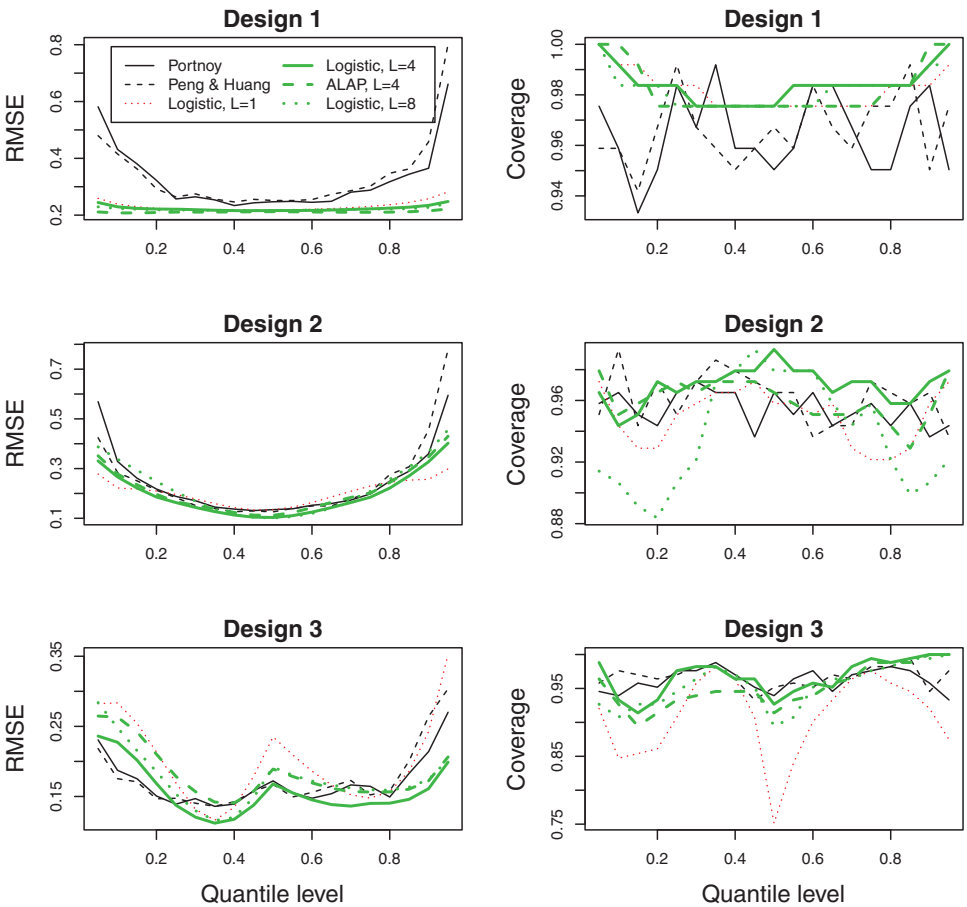

Figure 3. Root mean squared error and coverage of 95% intervals for $\beta_1(\tau)$ for the simulation study at quantile levels $\tau = 0.05, 0.10, \ldots, 0.95$. This figure appears in color in the electronic version of this article.

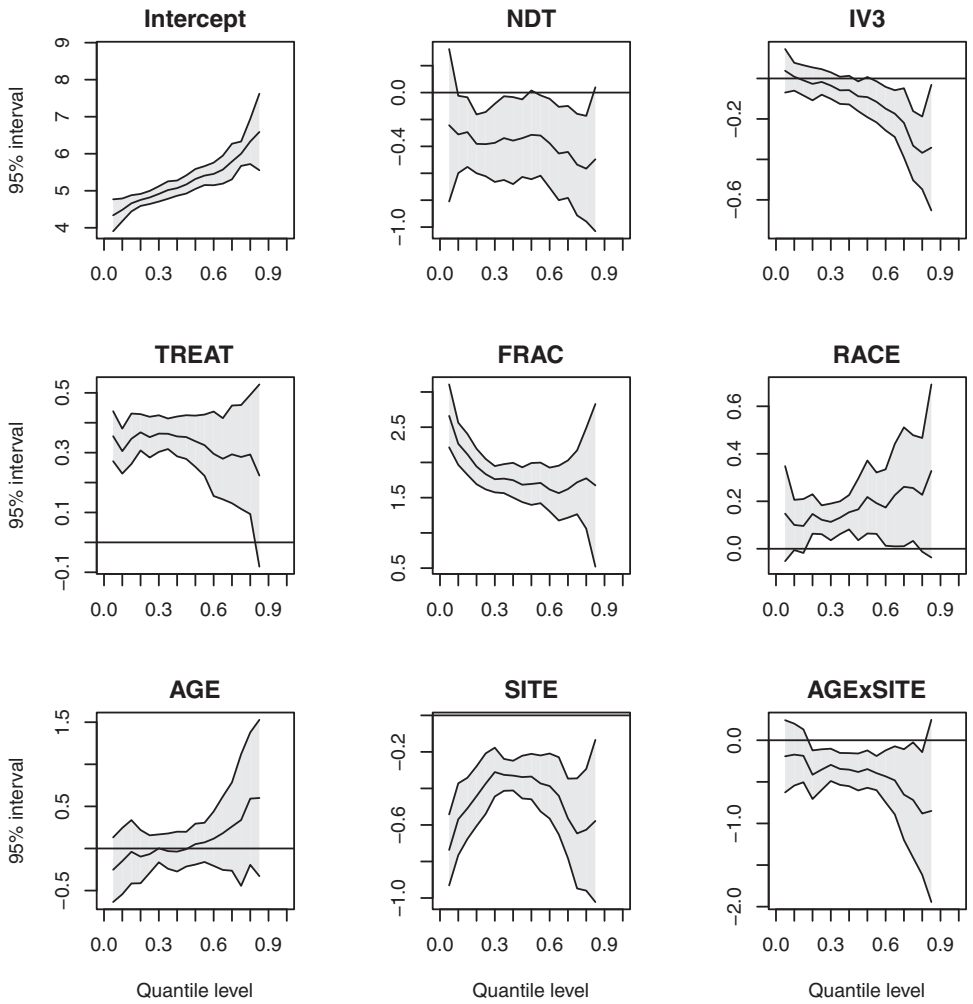

Figure 4. Posterior 95% intervals for the quantile function $\beta_j(\tau)$ for the Portnoy method (no value was returned for $\tau \geq 0.9$).

have a peak in RMSE at this value. However, even in this difficult setting, the Bayesian methods with $L > 1$ have the appropriate coverage and the smallest RMSE for most quantile levels.

In addition, we consider a five-predictor design with $\beta_0(\tau) = 2\Phi^{-1}(\tau)$, $\beta_1(\tau) = 2\min\{\tau - 0.5, 0\}$, $\beta_2(\tau) = 2\tau$, $\beta_3(\tau) = 2$, $\beta_4(\tau) = 1$, and $\beta_5(\tau) = 0$. The covariates are drawn $X_{ji} \overset{iid}{\sim}$ Unif(-1,1); all other settings, models, and priors the same as above. For the Portnoy method (Peng and Huang is similar), the RMSE averaged over quantile levels (plots of RMSE and coverage by quatile level can be found in Web Appendix C) are 0.313 (β_1), 0.330 (β_2), and 0.331 (average of $\beta_3 - \beta_5$). For the Bayesian model with logistic q_0, the corresponding RMSEs are 0.290, 0.296, and 0.264 for $L = 1$, 0.282, 0.306, and 0.239 for $L = 4$, and 0.286, 0.327, and 0.238 for $L = 10$. For the asymmetric Laplace model with $L = 4$, the corresponding RMSEs are 0.295, 0.316, and 0.248. Therefore, as in the single predictor settings, the Bayesian models are competitive with other approaches for complex quantile functions (β_1 and β_2), and provide substantial improvements for simple quantile functions ($\beta_3 - \beta_5$).

4. Analysis of the UIS Data

To illustrate the Bayesian model and compare with previous approaches, we use the UIS drug treatment study data available in the `quantreg` package in `R`, from Hosmer and Lemeshow (1998) and analyzed using quantile regression in Portnoy (2003) and Koenker (2008). The response is time until relapse, and there are $n = 575$ observations with complete data. We use $p = 8$ predictors: number of previous drug treatments (NDT), IV drug use (IV; "Yes" $= 1$, "No" $= -1$), treatment (TRT; "long" $= 1$, "short" $= -1$), compliance fraction ($FRAC$), race ($RACE$; "white" $= 1$, "Non-white" $= -1$), age (AGE), site ($SITE$; "A" $= 1$, "B" $= -1$), and the interaction between age and site. All variables are scaled to lie in $[-1, 1]$ via the transformation $2[X - min(X)]/[max(X) - min(X)] - 1$.

We first select the base quantile function, q_0, and the number of basis functions, L, using test set validation. The data are split into $m_1 = 0.8n$ training observations, $\mathbf{y}_1$, and $m_2 = 0.2n$ testing observations, $\mathbf{y}_2$. Since the usual summaries such as mean squared error are inappropriate for censored survival data, we compare models using the log pseudo maximum likelihood (LPML) statistic (Ibrahim, Chen, and

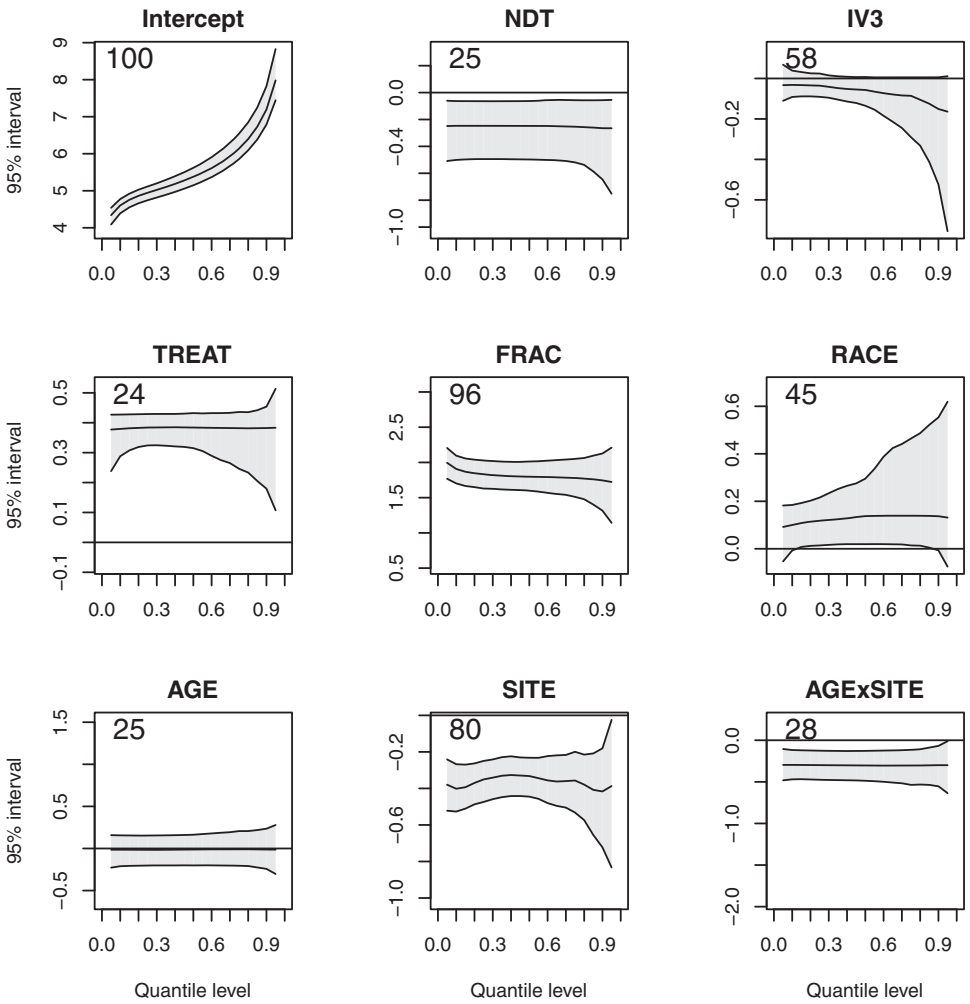

Figure 5. Posterior 95% intervals for the quantile function $\beta_j(\tau)$ for the Bayesian method with asymmetric Laplace base quantile function and $L = 8$ basis functions. The value in the upper left corner of each plot is the posterior probability (multiplied by 100) that the variable has non-constant effect across quantile levels (that is, the posterior probability that $\theta_j = 1$).

Sinha, 2001). The LPML statistic is the log density of $\mathbf{y}_2$ given $\mathbf{y}_1$, that is, $\log f(\mathbf{y}_2|\mathbf{y}_1) = \int \log f(\mathbf{y}_2|\mathbf{y}_1, \boldsymbol{\alpha}) f(\boldsymbol{\alpha}|\mathbf{y}_1) d\boldsymbol{\alpha} = \int \log f(\mathbf{y}_2|\boldsymbol{\alpha}) f(\boldsymbol{\alpha}|\mathbf{y}_1) d\boldsymbol{\alpha}$. This is approximated using MCMC output as

$$\text{LPML} \approx \frac{1}{S} \sum_{s=1}^{S} \sum_{i=1}^{m_2} \delta_i^* \log\{f[\log(Y_i^*)|\mathbf{X}_i^*, \boldsymbol{\alpha}^{(s)}]\} + (1 - \delta_i^*) \log\left\{1 - F[\log(Y_i^*)|\mathbf{X}_i^*, \boldsymbol{\alpha}^{(s)}]\right\},$$

where S is the number of MCMC samples, $\boldsymbol{\alpha}^{(s)}$ is the draw of $\boldsymbol{\alpha}$ for sample s $(Y_1^*, \mathbf{X}_1^*, \delta_1^*), \ldots, (Y_{m_2}^*, \mathbf{X}_{m_2}^*, \delta_{m_2}^*)$ are the test set data, and f and F are defined in (6) and (7), respectively. An advantage of this criterion is that it evaluates the entire response density, that is, all quantile levels. Models with larger LPML are preferred.

We fit the model using logistic, asymmetric Laplace and t quantile functions for the base distribution q_0. For each base distribution, we fit the model with $L = 1$, 4, 8, and 12 using same uninformative priors as in Section 3 except that $\theta_j \sim$ Bernoulli(0.5) for $j = 1, \ldots, p$ to perform variable selection. Models with $L = 1$ refer to the parametric location-scale model in (1) with residual distribution determined by q_0. For each base quantile function $L = 8$ maximized LPML, therefore for each base distribution a semiparametric fit with $L > 1$ is preferred to the parametric model with $L = 1$. For all L, the asymmetric Laplace base quantile function maximized LPML. Therefore, we present the results assuming the asymmetric Laplace base distribution and $L = 8$.

Figures 4 and 5 plot the results for the Portnoy (2003) and Bayesian methods, respectively. Using either method, long treatment, high compliance fraction, and white race have positive associations with survival, while the number of previous drug treatments and site A have negative associations.

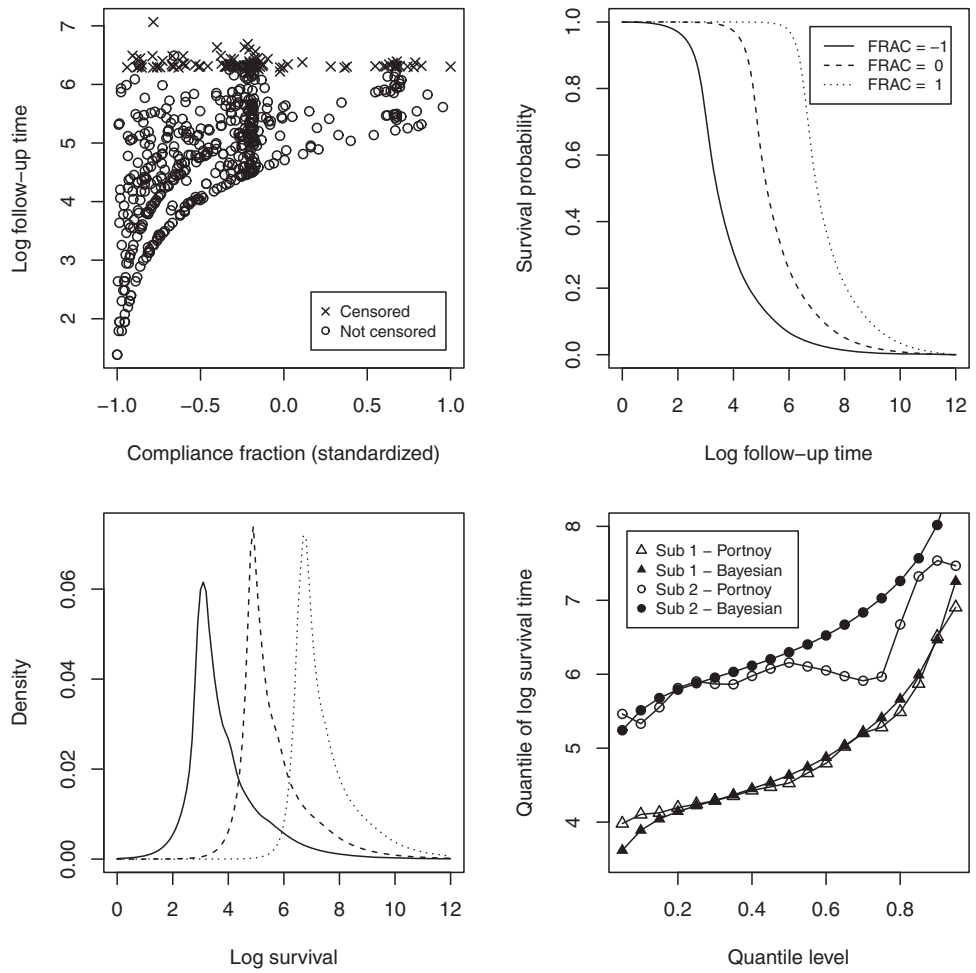

Figure 6. Data (top left) and posterior mean survival function (top right) and density (bottom left) for three levels of compliance fraction (transformed to the interval [-1,1]) with all other covariates fixed at their median (continuous covariates) or mode (binary covariates). The bottom right panel plots the predicted quantile values at $\tau \in \{0.05, \ldots, 0.95\}$ for two subjects.

The most striking difference between the fits is that the Bayesian estimates are far smoother across quantile levels. This borrowing of information across quantile levels leads to narrower intervals than the frequentist approach. As a result, the intervals for the number of previous drug treatments, treatment, and interaction between age and site exclude zero for the Bayesian model for all quantile levels.

With high probability, most of the covariates are included in the location but not in the shape/scale (Figure 5, upper left corner of each plot). Three variables have posterior probability (i.e., the posterior mean of θ_j) of at least 0.5 of a non-constant quantile function: IV drug use, compliance fraction, and site. IV drug use has little effect early in the follow-up, but a negative effect on upper quantiles. Compliance fraction has a stronger positive effect early in the follow-up, while the quantile function for site is concave, with stronger negative effects in both tails than in the center of the distribution.

To illustrate the predictive model, Figure 6 plots the fitted survival curves and density for three values of compliance fraction. For each level of compliance the density is right skewed, resembling the asymmetric Laplace base quantile function (the posterior 95% interval of the shape parameter ϕ is $(0.20, 0.40)$, giving right-skewness). The most prominent effect of compliance fraction is the shift the density; the median log survival time is 3.5, 5.2, and 7.1 for the three increasing values of compliance fraction in the top right panel of Figure 6. However, compliance fraction also affects the shape of the survival distribution.

We conduct five-fold cross-validation to compare the predictive performance of the Bayesian model with the classical approach. Evaluating predictions is challenging because standard approaches such as the prediction mean squared error cannot be used do to censoring, and other measures cannot be used because the classical quantile method produces only linear quantile estimates and not a full predictive distribution. Therefore, we use the recently-proposed criteria of Saha-Chaudhuri and Heagerty (2013), described in Web Appendix D. The results in Web Appendix D show that the Bayesian method gives an improvement compared to the classical method at low quantile levels, and both methods give similar results for the median. The Bayesian model also provides more stable estimates. For example, denote

$R_i = \text{SE}[q_P(\tau|\mathbf{X}_i)]/\text{SE}[q_B(\tau|\mathbf{X}_i)]$ as the ratio of standard errors for the Portnoy (standard errors based on 1000 bootstrap samples) and Bayesian (posterior standard deviation) estimates of the τ quantile for subject i. The mean (90% interval) of R_i across subject is 1.55 (1.11, 2.24) for $\tau = 0.05$ and 1.32 (1.10, 1.57) for $\tau = 0.50$.

Figure 6 compares the predictive model for two subjects using the Bayesian and classical approaches. The bottom right panel plots the estimated (from the training set) quantiles $\hat{q}(\tau|\mathbf{X}_i) = \sum_{j=0}^{p} X_{ij}\hat{\beta}_j(\tau)$ for several τ. For the first subject, the quantiles are increasing in τ for both methods, and both methods produce similar estimates. For the second subject, the classical estimates are decreasing for 9 of the 19 quantile levels (the worst case in this dataset). In fact, the estimated median is larger than the estimated 0.75 quantile. These decreasing quantiles clearly prohibit predictive densities and survival probabilities as given in Figure 6 for the Bayesian model. Therefore, proposed approach not only provides an arbitrarily flexible model for the quantile function, it also provides more precise estimates of the quantile function and permits straight-forward predictions for individual subjects.

5. Discussion

In this paper, we propose a model for quantile regression for censored data. Unlike parametric models such as the accelerated failure time model, the proposed semi-parametric model is flexible enough to accommodate any valid quantile process at a finite number of quantile levels providing robustness to model misspecification. The simulation study shows that when data are generated from the location-scale model on which the prior is centered, the new method provides a large improvement over nonparametric frequentist methods. In other cases, it remains competitive with previous approaches. In the real data example, we find that the Bayesian method identifies similar broad scale features as the frequentist approach, but often has smaller uncertainty estimates and thus identifies more significant effects.

A drawback of the proposed method is computation time. The analysis of the UIS data took around 3.5 hours compared to a few seconds for the method of Portnoy (2003). Also, while there are clearly advantages to modeling all quantile levels simultaneously, it is also possible that this may over-smooth in some situations. A possible remedy to over-smoothing is to replace the autoregressive priors for the basis coefficients with independent priors.

Supplementary Materials

Web Appendices and Figures referenced in Sections 2.1, 3, and 4 are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

We thank the editor, associate editor, and reviewers for thoughtful and constructive comments, and Paramita Saha Chaudhuri (Duke) for providing `R` code for the methods in Saha-Chaudhuri and Heagerty (2013). This work was supported by NIH grant R01ES014843, R03DE021762.

References

Bondell, H. D., Reich, B. J., and Wang, H. (2010). Non-crossing quantile regression curve estimation. *Biometrika* **97**, 825–838.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* **49**, 327–335.

Dunson, D. B. and Taylor, J. A. (2005). Approximate Bayesian inference for quantiles. *Journal of Nonparametric Statistics* **17**, 385–400.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1**, 209–230.

Ferguson, T. S. (1974). Prior distribution on spaces of probability measures. *The Annals of Statistics* **2**, 615–629.

Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of Polya trees. *Journal of the American Statistical Association* **97**, 1020–1033.

Hjort, N. L. and Walker, S. G. (2009). Quantile pyramids for Bayesian nonparametrics. *The Annals of Statistics* **37**, 105–131.

Hosmer, D. W. and Lemeshow, S. (1998). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: John Wiley and Sons Inc.

Ibrahim, J. G., Chen, M., and Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer.

Koenker, R. (2008). Censored quantile regression redux. *Journal of Statistical Software* **27**, 1–25.

Koenker, R. (2010). *quantreg: Quantile Regression*. R package version 4.53.

Kottas, A. and Gelfand, A. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association* **96**, 1458–1468.

Kottas, A. and Krnjajić, M. (2009). Bayesian nonparametric modeling in quantile regression. *Scandinavian Journal of Statistics* **36**, 297–319.

Kotz, S., Kozubowski, T. J., and Podgorski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Birkhauser, Boston.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics* **20**, 1222–1235.

Lavine, M. (1994). More aspects of Polya tree distributions for statistical modelling. *The Annals of Statistics* **22**, 1161–1176.

Lin, J., Sinha, D., Lipsitz, S., and Polpo, A. (2012). Semiparametric Bayesian survival analysis using models with log-linear median. *Biometrics* **68**, 1136–1145.

Lindgren, A. (1997). Quantile regression with censored data using generalized L1 minimization. *Computational Statistics and Data Analysis* **23**, 509–524.

O'Hara, R. B. and Sillanpaa, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analaysis* **4**, 85–118.

Peng, L. and Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of American Statistical Association* **103**, 637–649.

Portnoy, S. (2003). Censored quantile regression. *Journal of American Statistical Association* **98**, 1001–1012.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics* **25**, 303–325.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reich, B. J. (2012). Spatiotemporal quantile regression for detecting distributional changes in environmental processes. *Journal of the Royal Statistical Society: Series C* **64**, 535–553.

Reich, B., Bondell, H., and Wang, H. (2010). Flexible Bayesian quantile regression for independent and clustered data. *Biostatistics* **11**, 337–352.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association* **106**, 6–20.

Saha-Chaudhuri, P. and Heagerty, P. (2013). Non-parametric estimation of time-dependent predictive accuracy curve. *Biostatistics* **14**, 42–59.

Todkar, S. T. and Kadane, J. B. (2011). Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Analysis* **6**, 1–12.

Yu, K. and Moyeed, R. (2001). Bayesian quantile regression. *Statistics and Probability Letters* **54**, 437–447.

Received August 2012. *Revised April* 2013. *Accepted April* 2013.