Censored Quantile Regression Redux

Roger Koenker
University of Illinois at Urbana-Champaign

Abstract

This vignette is a slightly modified version of Koenker (2008a). It was written in plain latex not Sweave, but all data and code for the examples described in the text are available from either the JSS website or from my webpages. Quantile regression for censored survival (duration) data offers a more flexible alternative to the Cox proportional hazard model for some applications. We describe three estimation methods for such applications that have been recently incorporated into the R package **quantreg**: the Powell (1986) estimator for fixed censoring, and two methods for random censoring, one introduced by Portnoy (2003), and the other by Peng and Huang (2008). The Portnoy and Peng-Huang estimators can be viewed, respectively, as generalizations to regression of the Kaplan-Meier and Nelson-Aalen estimators of univariate quantiles for censored observations. Some asymptotic and simulation comparisons are made to highlight advantages and disadvantages of the three methods.

Keywords: quantile regression, censored data.

1. Introduction

Powell (1984, 1986) initiated an "era of econometric perestroika" for the censored regression model, liberating it from the oppressive Gaussian specification that had prevailed since its introduction by Tobin (1958) in the midst of the cold war. Given the linear latent variable model,

$$T_i = x_i^\top \beta + u_i$$

with u_i assumed to be iid with distribution function F, Powell noted that if censoring values, C_i, are observed for all $i = 1, \cdots, n$ and we observe $Y_i = \max\{C_i, T_i\}$ then the conditional quantile functions,

$$Q_{Y_i|x_i}(\tau|x_i) = F^{-1}(\tau) + x_i^\top \beta$$

can be consistently estimated, setting $\rho_\tau(u) = u(\tau - I(u < 0)$, by,

$$\hat{\beta} = \text{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \max\{C_i, x_i^\top b\}),$$

provided that the design matrix, $X = (x_i)$, contains an intercept to absorb the τ dependent contribution $F^{-1}(\tau)$. This observation follows immediately from the monotonicity of the mapping $T_i \to Y_i$, and the fact that for any monotonically increasing function, h, and scalar random variable Z, $P(Z \leq z) = P(h(Z) \leq h(z))$. The result generalizes nicely to a variety of non-iid latent variable settings; in particular to other linear conditional quantile latent

variable models, permitting linear scale shift and other more general forms of heterogeneity in the covariate effects. Right censoring, as is more typical of duration modeling applications, is easily accommodated by replacing max by min above. Often, in econometric applications the C_i's take a constant value as in the original tobit model where $C_i = 0$, or in wage equation top-coding, but this is not essential. What *is* necessary – and we shall see that this is not without its unfortunate consequences – is that the C_i's are known for all observations. Following Powell, we will refer to this situation as fixed censoring.

Random censoring, in contrast, refers to situations in which censoring values, C_i, are only observed for the censored observations. In effect, we observe only the event times, Y_i and a censoring indicator, δ_i, taking the value one if the observation is uncensored and zero if the observation is censored. Random censoring has received much less attention in the econometric literature, and it is not difficult to conjecture why. Analysis of randomly censored data requires that censoring times are independent of event times, or, in regression settings, that they are independent conditional on covariates. This assumption is frequently implausible in econometric applications where censoring is due to endogenous influences. In biostatistics, where random censoring is more often considered, the dominant empirical strategy has been the Cox proportional hazard model. However, there has also been a recognition that the proportionality assumption underlying the Cox model is sometimes inappropriate, necessitating stratification of the baseline hazard or some other weakening of the proportional hazards condition. Much more flexible models can be constructed by modeling conditional quantiles of the event time distribution. For uncensored survival data this approach has been explored by Koenker and Geling (2001), but censoring poses some new challenges. Fitzenberger and Wilke (2006) provide a valuable survey of applications of censored quantile regression methods in econometric duration modeling.

An early alternative approach to Powell, suggested by Lindgren (1997), simply bins the data in covariate space and computes local Kaplin Meier estimates in each bin. The obvious difficulty with this approach is that the binning quickly becomes impractical as the number of covariates grows.

Portnoy (2003) proposed an ingenious method of recursively estimating linear conditional quantile functions from censored survival data and established consistency and $\sqrt{n}$-convergence of the proposed estimators. Portnoy's method can be regarded as a generalization to regression of the Kaplan Meier estimator. Recently, Peng and Huang (2008) have proposed a closely related method. Rather than building on the linkage to Kaplan-Meier, they instead develop an approach linked to the Nelson-Aalen estimator of the cumulative hazard function. The main advantage of the latter approach is that it enables them to employ counting process methods to establish a martingale property for their estimating equation from which a more complete asymptotic theory for the estimator flows.

The main objective of this paper is to describe an implementation of all the foregoing methods appearing in recent versions of my **quantreg** package for R. This package seeks to provide a comprehensive implementation of quantile regression methods for the R (R Development Core Team 2008) language. The package is available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=quantreg`. It incorporates both linear and nonlinear in parameters methods as well as non-parametric additive model fitting techniques. The new censored quantile regression methods are accessible through the new fitting function `crq`, which extends the functionality of the existing functions `rq`, `nlrq` and `rqss` that are used/rearrage/ for fitting linear, nonlinear, and nonparametric models respectively. After a

brief overview of the implementation, we will consider the three new methods in turn and provide some comparisons and offer some advice on their strengths and weaknesses.

2. Overview

Model fitting in R typically proceeds by specifying a formula describing the model, a data frame containing the data, and possibly some some further fitting options. For censored quantile regression these arguments are passed to the function `crq`. Formulae are specified for the two random censoring methods using the function `Surv` from the package **survival**, see Therneau and Lumley (2008)

The accelerated failure time model,

$$\log(Y_i) = x_i^\top \beta + u_i,$$

with u_i iid with distribution function F is a common model for survival data. When the data are uncensored the model can be simply estimated by least squares, or using quantile regression as in Koenker and Geling (2001). The latter approach offers some distinct advantages since it permits the researcher to focus attention on narrow slices of the conditional survival distribution. In Koenker and Geling (2001) where the interest is in mortality of medflies it was particularly valuable to focus attention on the upper tail of the lifetime distribution where it was found that there was a crossover in gender survival prospects at advanced ages. It is difficult, even impossible, to see such effects in some classical survival models where attention typically focuses on covariate effects on mean survival prospects. For further details on quantile regression methods and their implementation in R, see Koenker (2005) and the vignette available with the package **quantreg**, Koenker (2008b). For censored data, and parametric choice of F, the model can be easily estimated by maximum likelihood. Relaxing the parametric restriction and the iid error assumption leads naturally to the censored quantile regression model,

$$Q_{\log(Y_i)|x_i}(\tau|x_i) = x_i^\top \beta(\tau).$$

The choice of the log transformation, although traditional, is entirely arbitrary and may be replaced by any monotone transformation. In applications with random censoring such models can be estimated in R using `crq` using the formula,

```
Surv(log(y), delta) ~ x
```

where `delta` denotes the vector of censoring indicators. For fixed censoring of the type considered by Powell, formulae take the form,

```
Curv(log(y), c, type= "left") ~ x
```

Here, `Curv` is a slightly modified version of `Surv` designed to accommodate the provision of the censoring times instead of the censoring indicators to the fitting routine. The `type` argument indicates whether the censoring is from the left, as in the classical Tobit model, or from the right as in the case of top coding. Other arguments can be supplied to fitting function including: `taus` a list of quantiles to be estimated, `data` a data frame where the formula variables reside, etc. The argument `method` is used to specify one of three currently available methods: `"Powell"` for the Powell estimator, `"Portnoy"` for Portnoy' censored

quantile regression estimator, and `"PengHuang"` for Peng and Huang's version of the censored quantile regression estimator. Partial argument matching in R permits these strings to be abbreviated to the shortest distinguishable substrings: `"Pow"`, `"Por"` and `"Pen"`. Further arguments can be specified to the specific fitting routines, notably `start` to specify a initial value for the coefficients for the Powell method, and `grid` to specify the evaluation grid for the random censoring methods.

Given fixed censoring data it is always possible to fit random censoring models, and we will argue that this may often be advantageous, but since the Powell estimator requires censoring times for all observations, it can generally not be applied to randomly censored data. We will focus in the remainder of the paper on the case of right censoring but it should be understood that all of the methods discussed can be adapted to left censoring as well. Applications involving interval censoring are the subject of active current research and we hope to incorporate new methods when they become available.

3. The Powell Estimator

Given censoring times C_i and event times $Y_i \leq C_i$ with associated covariate vectors $x_i \in \mathbb{R}^p$, the Powell estimator minimizes,

$$R_\tau(b) = \sum \rho_\tau(Y_i - \min\{C_i, x_i^\top b\}).$$

The piecewise linear form of the response function poses some real computational challenges. Unlike the uncensored quantile regression problem, the objective function, $R_\tau(b)$ is no longer convex, so local optimization methods like steepest descent may terminate at a local minimum that is not the global minimum. Fitzenberger (1996) describes an algorithm that adapts the classical Barrodale and Roberts (1974) simplex algorithm for ℓ_1 regression to this end. In effect, Fitzenberger's algorithm is steepest descent: due to the piecewise linear form of the objective function solutions can be characterized by an exact fit to p observations, so careful computation of the directional derivatives at successive "basic" solutions in the directions obtained by deleting one of the p points from the "basis" ensures convergence to a local optimum.

Fitzenberger and Winker (2007) investigate a modified version of this BRCENS algorithm that employs a threshold accepting outer loop somewhat like simulated annealing to improve the chances of converging to the global optimum. Ironically, it is far from obvious that this more diligent search for the global Powell solution is justified. Simulations by Fitzenberger and Winker, and supported by my own simulations, suggest that in many censored regression problems the global optimizer performs much worse than its more myopic counterparts. Starting the BRCENS iterations at $\beta = 0$ or some other plausible value and taking steepest descent steps acts as a shrinkage technique, thereby avoiding embarrassing globally optimal points further away. In the **quantreg** implementation the default starting value is the naive `rq` estimate ignoring the censoring; this has the dubious advantage that it retains the usual equivariance properties of the conventional quantile regression estimators.

In simulations, where exhaustive search for the R_τ minimizer is feasible, the global optimizer is prone to find, at least occasionally, solutions that are absurdly far from the parameters used to generate the data, and at least from a mean squared error perspective these realizations wreck havoc with performance. Asymptotic theory assures us that this is only an evanescent

"finite sample problem," but such assurances may not offer much consolation to the applied researcher who generally lacks the patience to let data accumulate in asymptopia. Fortunately, other methods may offer some rather unexpected advantages.

The function `crq` implements a new fortran version of the algorithm described in Fitzenberger (1996) for the method `"Powell"`. This version is considerably simpler than the original BRCENS version and more modular. I have also included an implementation of an exhaustive global search algorithm that pivots through *all* $\binom{n}{p}$ basic solutions and chooses the one that minimizes the Powell objective function. This option is selected by specifying the option `start = "global"`, but it should be recognized that for problem with even a moderately large sample size the resulting search becomes impractical. It would be quite easy to embed the current implementation into a global optimization method such as the `anneal` function of the R package **subselect**, see Cerdeira, Silva, Cadima, and Minhoto (2007), but we have not (yet) done this.

4. Random Censoring

In one-sample settings with random censoring the Kaplan-Meier product-limit estimator is known to be an efficient estimation technique and can be interpreted as a nonparametric maximum likelihood estimator, see e.g. Andersen, Borgan, Gill, and Keiding (1991). In the simplest case, without tied event times, the Kaplan-Meier estimator of the survival function, $S(t)$ can be written as,

$$\hat{S}(t) = \prod_{i:y_{(i)} \le t} (1 - 1/(n - i + 1))^{\delta_{(i)}},$$

where $y_{(i)}$'s denote the ordered event times, and the $\delta_{(i)}$'s denote the associated censoring indicators. Efron (1967) interpreted $\hat{S}$ as shifting mass of the censored observations to the right, distributing it in accordance with the subsequent uncensored event times.

4.1. Kaplan-Meier Quantiles as Argmins

Portnoy (2003) observed that quantiles of the Kaplan-Meier distribution function, $\hat{F}(t) = 1 - \hat{S}(t)$ could be expressed as solutions to a weighted quantile optimization problem in which weight associated with censored observations was split into two pieces. A part of the mass associated with each censored observation is left in its initial position at the censoring time, and the remainder is shifted to right, in effect to $+\infty$.

To see this, recall that in one-sample settings without censoring the ordinary sample quantiles can be expressed as,

$$\hat{\xi}(\tau) = \text{argmin}_\xi \sum_{i=1}^{n} \rho_\tau(Y_i - \xi)$$

to obtain the step function,

$$\hat{\xi}(\tau) = y_{(i)} \qquad \text{for } \tau \in ((i-1)/n, i/n].$$

It is helpful to view this as parametric in τ: as τ increases from 0, $y_{(1)}$ is the solution until we reach, $\tau = 1/n$, at which point $y_{(2)}$ is also a minimizer, and so on.

When there are censored observations we can proceed in a similar fashion, except that when we encounter a τ_i such that $\hat{\xi}(\tau_i) = y_{(i)}$ and $\delta_{(i)} = 0$, we split the censored observation into two pieces: one piece remains at its original position, $y_{(i)}$, and receives weight

$$w_i(\tau) = \frac{\tau - \tau_i}{1 - \tau_i}$$

at all subsequent τ, and the other piece is shifted to $y_\infty = +\infty$ and gets weight $1 - w_i(\tau)$. This reweighting assures that $\hat{\xi}(t)$ is constant in an open neighborhood of any such τ_i, and the remaining mass, the $1 - w_i$ part of each censored observation, gets distributed appropriately. The crucial insight is simply that the quantiles only depend on how much mass is below and how much is above – shifting part of the censored mass to $+\infty$ ensures that all the subsequent uncensored observations receive their fair share of the "credit" for each of the censored points.

Thus, denoting the index set of the censored observations encountered up to τ by $K(\tau)$, the quantiles of the Kaplan-Meier distribution, $\hat{F}$ can be expressed as a solution to the problem:

$$\min \sum_{i \notin K(\tau)} \rho_\tau(Y_i - \xi) + \sum_{i \in K(\tau)} [w_i(\tau)\rho_\tau(Y_i - \xi) + (1 - w_i(\tau))\rho_\tau(y_\infty - \xi)].$$

The advantage of this formulation is that it generalizes nicely to the regression setting where the scalar ξ is replaced by the inner product $x_i^\top \beta$.

4.2. Portnoy's Censored Quantile Regression Estimator

Portnoy (2003) describes in detail an algorithm for the regression analogue of this problem. There are several complications in the regression setting that do not arise in the one-sample context; the most important of these is the possibility that censored observations that are "crossed" by estimated quantile regression process and thus have negative residuals, may return to the optimal basis and have zero residuals for some subsequent τ. This cannot happen in the one-sample setting by the monotonicity of the Kaplan-Meier estimator, but may occur using the reweighting due to the weaker nature of the monotonicity condition in the p-dimensional regression setting. Portnoy describes an effective way to deal with these pivoting anomalies as well as discussing complications due to an excess of censored observations in the upper tail that limit range $\tau \in [0, 1]$ for which the model is inestimable. The latter is a familiar problem even in the one-sample setting where censored observations above the largest uncensored observation imply a "defective" Kaplan-Meier survival function.

Portnoy provided a Fortran implementation of his estimator based on a "pivoting" method that is similar to that described in Koenker and D'Orey (1987), but adapted to the "recrossing" problems alluded to above. Starting τ near zero, at each step it is possible to evaluate the length of the interval of τ's for which the current solution to the weighted quantile regression problem:

$$\min \sum_{i \notin K(\tau)} \rho_\tau(Y_i - x_i^\top \beta) + \sum_{i \in K(\tau)} [w_i(\tau)\rho_\tau(Y_i - x_i^\top \beta) + (1 - w_i(\tau))\rho_\tau(y_\infty - x_i^\top \beta)].$$

remains optimal, the problem is then updated and resolved at the upper bound of the interval, and iteration proceeds until $\tau = 1$ is reached or the process is halted because there are only non-reweighted censored observations with positive residuals remaining. Portnoy has also

suggested an alternative approach in which the process is evaluated on a grid of $\tau \in [0, 1]$. In large samples the latter approach is generally preferred since the inherent accuracy of the estimated $\hat{\beta}(\tau)$ process is $O_p(1/\sqrt{n})$ making the evaluation of the process at $O_p(n \log n)$ points using the pivoting method rather excessive. The algorithm written by Steve Portnoy was originally made available in the R package **crq**, prepared in collaboration with Tereza Neocleous and myself. The functionality of this package has now been folded into the quantreg package.

To illustrate this technique we estimate the model appearing in (Portnoy 2003, Section 6.3), adapted from Hosmer and Lemeshow (1999), using the R code fragment:

```
R> require("quantreg")
R> data("uis")
R> fit <- crq(Surv(log(TIME), CENSOR) ~  ND1 + ND2 + IV3 +
      TREAT + FRAC + RACE + AGE * SITE, data = uis, method = "Por")
R> Sfit <- summary(fit, 1:19/20)
R> PHit <- coxph(Surv(TIME, CENSOR) ~  ND1 + ND2 + IV3 +
      TREAT + FRAC + RACE + AGE * SITE, data = uis)
R> plot(Sfit, CoxPHit = PHit)
```

We begin by loading the **quantreg** package, if it is not already loaded, and then loading the Hosmer and Lemeshow data. The model formula in the call to `crq` specifies that the logarithm of the "time to relapse" of subjects in a drug treatment program depends on the number of prior treatments, `ND1` and `ND2`; the treatment indicator, `TREAT` taking the value 1 for subjects taking the "long" course, and 0 for subjects taking the "short" course; an indicator for prior intravenous drug use, `IV3`; a compliance variable, `FRAC`; subject's race; and the main and interaction effects of subjects age and site of treatment. The object `fit` produced by the call to `crq` evaluates, by default, the Portnoy estimator on an equally spaced grid with increments of about 0.006, for this sample of size 575. The function `summary` computes bootstrapped standard errors for the quantile regression estimates. In this example this step generates several warning messages indicating that estimation of the bootstrapped samples result in a "premature stop." This is quite common and occurs whenever excessive censoring prevents estimation of the upper conditional quantiles. In the usual terminology of survival analysis this results in a "defective" estimate of the survival distribution. To compare with the Cox proportional hazard model, we estimate the same model with the **survival** package's function `coxph`. This enables us to compare the fitted models in the coefficient plots appearing in Figure 1.

The solid blue line in these plots is the point estimate of the respective quantile regression fits, and the lighter blue region indicates a 95% confidence region. The solid (horizontal) black line in some of the plots indicates a null effect. The red line in each of the plots indicates the estimated conditional quantile "effects" implied by the estimated Cox model, see Koenker and Geling (2001) and Portnoy (2003) for further details on how this is done. A feature of the Cox model is that all of the red lines are proportional to one another; they are forced to all have the same shape determined by the estimate of the baseline hazard function. This shape is quite consistent with the quantile regression estimates for some of the covariate effects, but for the treatment and compliance effects the estimates are quite disparate.

Because the baseline hazard function is non-negative, another feature of the Cox estimates is that they must lie entirely above the horizontal "effect equals zero" axis, or entirely below

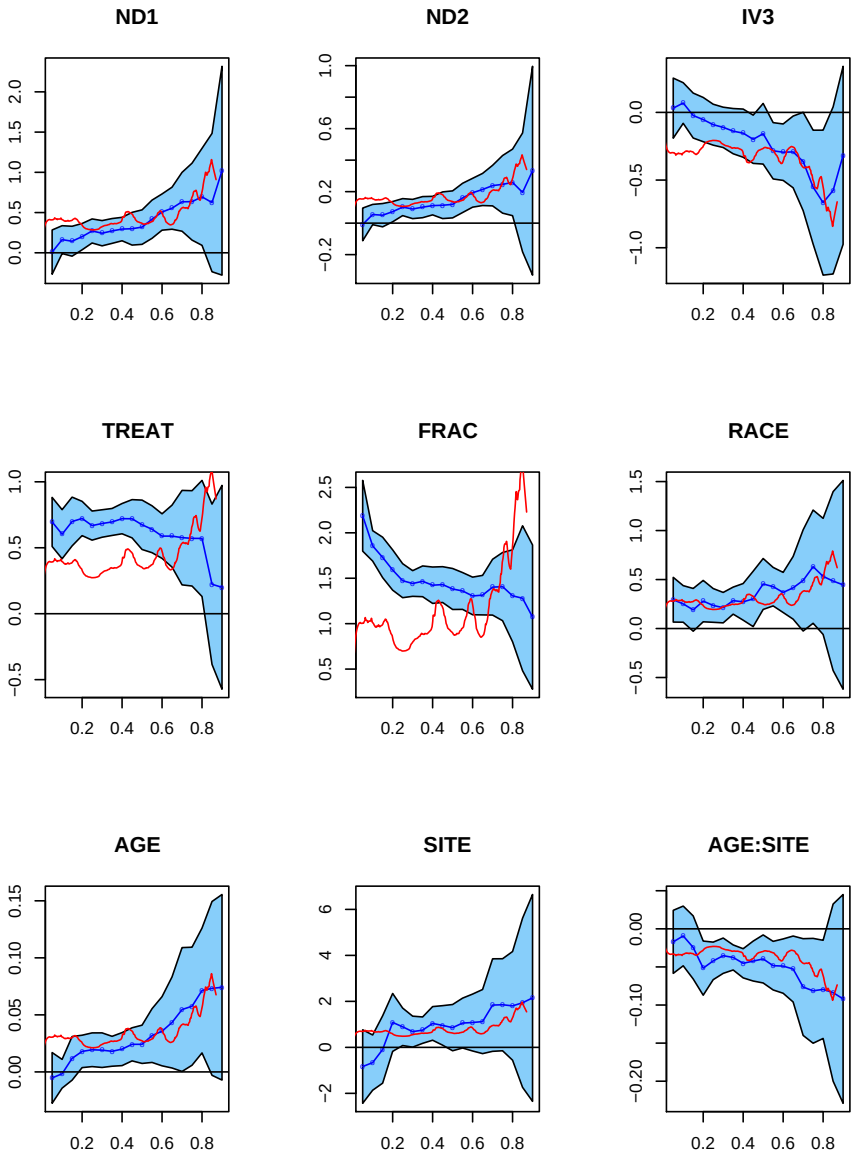

Figure 1: Censored Quantile Regression Coefficients Plots for the Hosmer-Lemeshow Data: The solid blue line indicates the quantile regression point estimates, the lighter blue region is a pointwise 95% confidence band, and the red curve in each plot illustrates the estimated conditional quantile "effect" estimated for the Cox proportional hazard model.

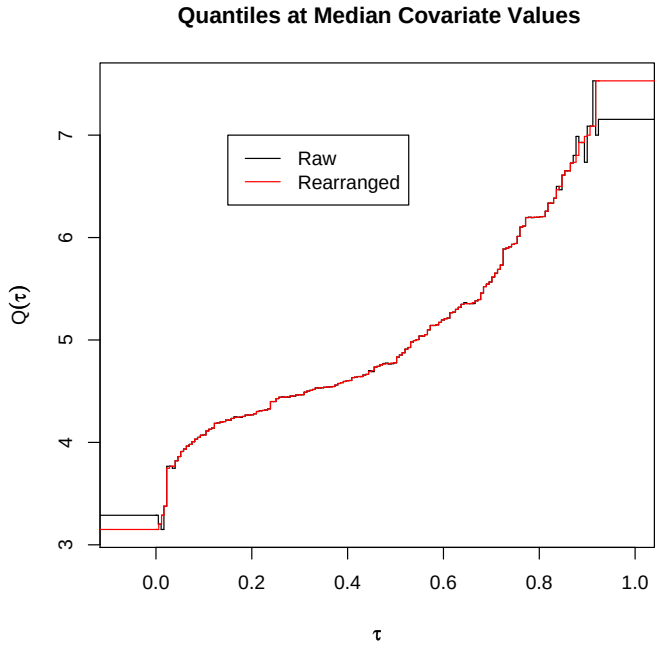

Figure 2: Predicted Conditional Quantile Function Plots for the Hosmer-Lemeshow Data: The solid black line indicates the predicted quantile function based on the censored quantile regression estimator of Portnoy, evaluated at median values of the each of the covariates. The monotonized red line is the "rearranged" version of the black line.

it. Thus, covariates must either increase hazard over the whole time scale, or decrease it; the model forbids the possibility that treatments may increase hazard for a time and then decrease them. Such crossovers are, however, sometimes quite plausible, and an advantage of the quantile regression approach is that they are more easily revealed. An interesting example of this phenomenon is the cross-over in gender mortality rates discussed in Koenker and Geling (2001).

Given the fitted `crq` object the conditional quantile function can be estimated at any setting of the covariates and plotted using something similar to the following code:

```
R> formula <- ~ ND1 + ND2 + IV3 + TREAT + FRAC + RACE + AGE * SITE - 1
R> X <- data.frame(model.matrix(formula, data=uis))
R> newd <- as.list(apply(X, 2, median))
R> pred <- predict(fit, newdata=newd, type = "stepfun")
R> plot(pred, xlab = expression(tau), ylab = expression(Q(tau)),
                do.points = FALSE, main = "Quantiles at Median Covariate Values")
R> plot(rearrange(pred), add=TRUE, do.points=FALSE,
                col.vert ="red", col.hor="red")
R> legend(.15, 7, c("Raw","Rearranged"), lty = 1:2,
                col=c("black","red"))
```

We first construct a data frame representing the variables of the model formula and then

compute medians of these variables to represent the setting of the covariates at which we wish to predict. The function `predict` takes the fitted object and the new data `newd` and returns a step function representing the predicted quantile function. If the covariate setting is chosen to be the means of the covariates, $\bar{x}$, then the predicted quantile function is guaranteed to be monotone increasing, (Koenker 2005, Theorem 2.5) but at other settings there can be violations of monotonicity. This eventuality appears in the present example in the extremes of the plotted function in Figure 1 where the estimated function is least precisely estimated, and in some nearly invisible smaller violations occurring in the central region of the plot. A simple and theoretically attractive way of dealing with these violations has been recently introduced by Chernozhukov, Fernández-Val, and Galichon (2006). Their procedure has been embodied in the **quantreg** function `rearrange` as used in the plotting command above.

4.3. Nelson-Aalen Quantiles as Argmins

Peng and Huang (2008) have recently suggested an alternative approach to censored quantile regression for censored survival data based on the well-known Nelson-Aalen estimator of the cumulative hazard function. To motivate the Peng and Huang estimator it is useful to briefly review the standard counting process development of the Nelson-Aalen estimator. As above, let $Y_i = \min\{T_i, C_i\}$ denote observed event times, and $\delta_i = I(T_i < C_i)$ the censoring indicators. The random variables T_i and C_i are assumed to be independent with distribution functions F and G, respectively. The distribution function, F, is assumed to be absolutely continuous with density f with respect to Lebesgue measure. Define the counting processes

$$\begin{aligned} N_i(t) &= I(\{T_i \le t\} \text{ and } \{\delta_i = 1\}) \\ R_i(t) &= I(\{T_i \ge t\}) \end{aligned}$$

and the corresponding aggregated processes $R(t) = \sum R_i(t)$ and $N(t) = \sum N_i(t)$. The cumulative hazard function,

$$\Lambda(t) \equiv \int_0^t \lambda(s)ds \equiv \int_0^t \frac{f(s)}{1 - F(s)}ds = -\log(1 - F(s))$$

has increments $\Lambda(s + h) = \Lambda(s) \approx \lambda(s)h$, so it is natural to estimate this quantity by the number of uncensored events occurring in the interval $[s, s + h]$ divided by the number of subjects at risk at time s, that is by $(N(s + h) - N(s))/R(s)$. Summing over all of $[0, t]$, we then have,

$$\hat{\Lambda}(t) = \int_0^t \frac{dN(s)}{R(s)}.$$

In principle, $dN(s)$ could accommodate both discrete and continuous components, but here we need only concern ourselves with the discrete component, $\Delta N(s) = N(s) - N(s-)$, which denotes the number of uncensored events occurring precisely at time s. Thus, we can express the Nelson-Aalen estimator in somewhat more concrete notation as

$$\hat{\Lambda}(t) = \sum_{\{i:y_i \le t\}} \frac{\Delta N(y_i)}{R(y_i)}.$$

Given the estimator, $\hat{\Lambda}(t)$, a natural estimator of the survival function would seem to be $\exp(-\hat{\Lambda}(t))$, but further reflection suggests that this is only really appropriate if $\hat{\Lambda}$ were

absolutely continuous. Alternatively, noting that

$$d\Lambda(s) = \frac{dF(s)}{1 - F(s-)}$$

we can write,

$$F(t) = \int_0^t dF(s) = \int_0^t (1 - F(s-))d\Lambda(s).$$

Then following Fleming and Harrington (1991), we can define recursively the estimator,

$$\hat{S}(t) = 1 - \int_0^t S(s-)d\hat{\Lambda}(s).$$

But since $\hat{S}(t-) - \hat{S}(t) = -\Delta\hat{S}(t) = \hat{S}(t-)\frac{\Delta N(t)}{R(t)}$, we have

$$\begin{aligned} \hat{S}(t) &= \hat{S}(t-)\left[1 - \frac{\Delta N(t)}{R(t)}\right] \\ &= \prod_{s \le t}\left[1 - \frac{\Delta N(s)}{R(s)}\right], \end{aligned}$$

which is recognizable as the Kaplan-Meier estimator.

The close relationship between the Nelson-Aalen and Kaplan-Meier estimators is not surprising; indeed both have some claim to the status of nonparametric maximum likelihood estimators, see e.g. (Andersen *et al.* 1991, Section IV.1.5). The martingale structure of the Nelson-Aalen estimator motivates the Peng and Huang approach to censored quantile regression, which we now briefly sketch.

4.4. Peng and Huang's Censored Quantile Regression Estimator

As above, let $Y_i = T_i \wedge C_i$ be a random event time and $\delta_i = I(T_i < C_i)$ be the associated censoring indicator. Denote, $F_i(t|x) = \mathsf{P}(T_i \le t|x_i)$, $\Lambda_i(t|x) = -\log(1 - F_i(t|x_i))$, and $N_i(t) = I(\{T_i \le t\}, \{\delta_i = 1\})$, then denoting $\min\{a, b\} = a \wedge b$,

$$M_i(t) = N_i(t) - \Lambda_i(t \wedge Y_i|x_i),$$

is a martingale process for $t \ge 0$. Adopting the accelerated failure time version of the quantile regression model,

$$\mathsf{P}(\log T_i \le x_i^\top \beta(\tau)) = \tau,$$

the martingale property, $\mathsf{E}M_i(t) = 0$ implies that,

$$\mathsf{E}[n^{-1/2}\sum x_i[N_i(\exp(x_i^\top\beta(\tau))) - \Lambda_i(\exp(x_i^\top\beta(\tau)) \wedge Y_i|x_i))] = 0.$$

Rewriting the Λ_i term as,

$$\Lambda_i(\exp(x_i^\top\beta(\tau)) \wedge Y_i|x_i) = H(\tau) \wedge H(F_i(Y_i|x_i)) = \int_0^\tau I(Y_i \ge \exp(x_i^\top\beta(u)))dH(u),$$

where $H(u) = -\log(1-u)$ for $u \in [0,1)$, yields the estimating equation,

$$\mathsf{E}[n^{-1/2}\sum x_i[N_i(\exp(x_i^\top\beta(\tau))) - \int_0^\tau I(Y_i \geq \exp(x_i^\top\beta(u)))dH(u)] = 0.$$

The integral can now be approximated on a grid, $0 = \tau_0 < \tau_1 < \cdots < \tau_J < 1$, as,

$$\alpha_i(\tau_j) = \sum_{k=0}^{j-1} I(Y_i \geq \exp(x_i^\top\hat\beta(\tau_k)))(H(\tau_{k+1}) - H(\tau_k)),$$

yielding Peng and Huang's final estimating equation,

$$n^{-1/2}\sum x_i[N_i(\exp(x_i^\top\beta(\tau))) - \alpha_i(\tau)] = 0.$$

Since the left hand side is not continuous an exact root may not exist. Peng and Huang consider "generalized solutions" citing Fygenson and Ritov (1994), who define a generalized estimating equation, $W(\beta)$, as a *monotone nondecreasing field*, if for any β and ξ in $\mathbb{R}^p$, $\xi^\top W(\beta + x\xi)$ is monotone nondecreasing in the scalar x. But this is precisely the condition that W be the subgradient of a convex function. Setting $r_i(b) = \log(Y_i) - x_i^\top b$, this convex function for the Peng and Huang problem takes the form

$$(Q) \qquad R(b,\tau_j) = \sum_{i=1}^n r_i(b)(\alpha_i(\tau_j) - I(r_i(b) < 0)\delta_i) = \min!$$

Theorem 1. *Fix τ, and define the n-vectors $\alpha = (\alpha_i(\tau))$, $\delta = (\delta_i)$ and $z = (\log(Y_i))$. The problem (Q) is equivalent to the linear programming problem:*

$$(P) \qquad \min\{\alpha^\top u + (\delta - \alpha)^\top v | z = Xb + u - v,\ u \geq 0,\ v \geq 0\}.$$

and its dual,

$$(D) \qquad \max\{z_1^\top a_1 | X_1^\top a_1 = X^\top(\delta - \alpha), a_1 \in [0,1]^m\}$$

where X_1 denotes the submatrix of X with m rows corresponding to uncensored observations, and z_1 denotes the associated subvector of z.

Proof: Note that $N_i(\exp(x_i^\top b)) = I(r_i(b) \leq 0)\delta_i$, and consequently splitting $r_i(b)$ into positive, u_i, and negative, v_i, parts yields (P). The formal dual is then,

$$\max_{d\in\mathbb{R}^n}\{z^\top d | X^\top d = 0, \alpha - d \geq 0, \delta - \alpha + d \geq 0\}$$

or equivalently, setting $a = \alpha - d$, 

$$\max_{a\in\mathbb{R}^n}\{z^\top a | X^\top a = X^\top(\delta - \alpha), a \in \Pi_{i=1}^n[0,\delta_i]\}$$

But the latter formulation implies that $a_i = 0$ for all i such that $\delta_i = 0$, so the dual problem can be reduced to focus only on the dual variables associated with the uncensored observations, which yields (D), after partitioning. ■

Remark: The dual formulation shows that solutions to the Peng and Huang problem must interpolate p *uncensored* observations. See the discussion in (Koenker 2005, Section 6.2). This

contrasts with both the Powell and Portnoy methods for which solutions also correspond to p-element subset solutions, but solutions may include censored as well as uncensored observations. ■

Implementation of the Peng and Huang estimator in the **quantreg** package requires that the process be evaluated on a prespecified grid. (There is no known "pivoting" form of the algorithm.) At each τ of the grid, the problem (D) is solved using a Fortran implementation of the Frisch-Newton algorithm described in Portnoy and Koenker (1997). This requires only a rather minor modification of the standard quantile regression procedure, replacing the usual right hand side of the dual equality constraints by the expression $X^\top(\delta - \alpha)$. In the case that $\delta_i \equiv 1$ so there is no censoring, this new right hand side reduces to approximately its original form $(1-\tau)X^\top 1_n$. This reduction is exact in the one-sample setting. Repeating the model fitting and prediction exercises described above using `method = "PengHuang"` rather than `method = "Portnoy"` yields very similar results, a finding that is perhaps not very surprising in view of the similarity of the underlying Kaplan-Meier and Nelson-Aalen foundations of the two methods.

To see in a little more detail how the two methods compare we consider a small simulation experiment. Survival times are generated by the AFT model,

$$\log T_i = x_1\beta_1 + x_2\beta_2 + u$$

with the $u = \log(e)$ iid and e standard exponential; $x_1 \sim U[0,1]$ and x_2 is independent, Bernoulli with probability one-half. Censoring times are generated as $U[0, 3.8]$ if $x_2 = 0$ and $U[0.1, 3.8]$ otherwise. This configuration yields roughly 25% censoring. We consider 3 sample sizes n = 100, 400, 1600, and 8 distinct grid spacings, parameterized by $\gamma = .2, .3, \cdots, .9$ with grid spacing $h = 1/(n^\gamma + 6)$. Figure 3 presents scatterplots of the Portnoy and Peng-Huang estimates $\hat\beta_2(0.6) - \beta_2(0.6)$ for this experiment. The estimators behave very similarly, but for finer grids (larger values of γ) the correlation is clearly stronger.

5. Some One-sample Asymptotics

It is instructive to compare the performance of various quantile estimators in the simplest censored one-sample problem as a prelude to some simulation comparisons of estimator performance for the general regression setting.

Suppose that we have a random sample of pairs, $\{(T_i, C_i) : i = 1, \cdots, n\}$ with $T_i \sim F$, $C_i \sim G$, and T_i and C_i independent. Let $Y_i = \min\{T_i, C_i\}$, as usual, and $\delta_i = I(T_i < C_i)$. In this setting the Powell estimator of $\theta = F^{-1}(\tau)$,

$$\hat\theta_P = \text{argmin}_\theta \sum_{i=1}^n \rho_\tau(Y_i - \min\{\theta, C_i\}).$$

is asymptotically normal,

$$\sqrt{n}(\hat\theta_P - \theta) \rightsquigarrow \mathcal{N}(0, \tau(1-\tau)/(f^2(\theta)(1-G(\theta)))).$$

In contrast, the asymptotic theory of the quantiles of the Kaplan-Meier estimator is slightly more complicated. Using the δ-method one can show,

$$\sqrt{n}(\hat\theta_{KM} - \theta) \rightsquigarrow \mathcal{N}(0, \text{Avar}(\hat S(\theta))/f^2(\theta))$$

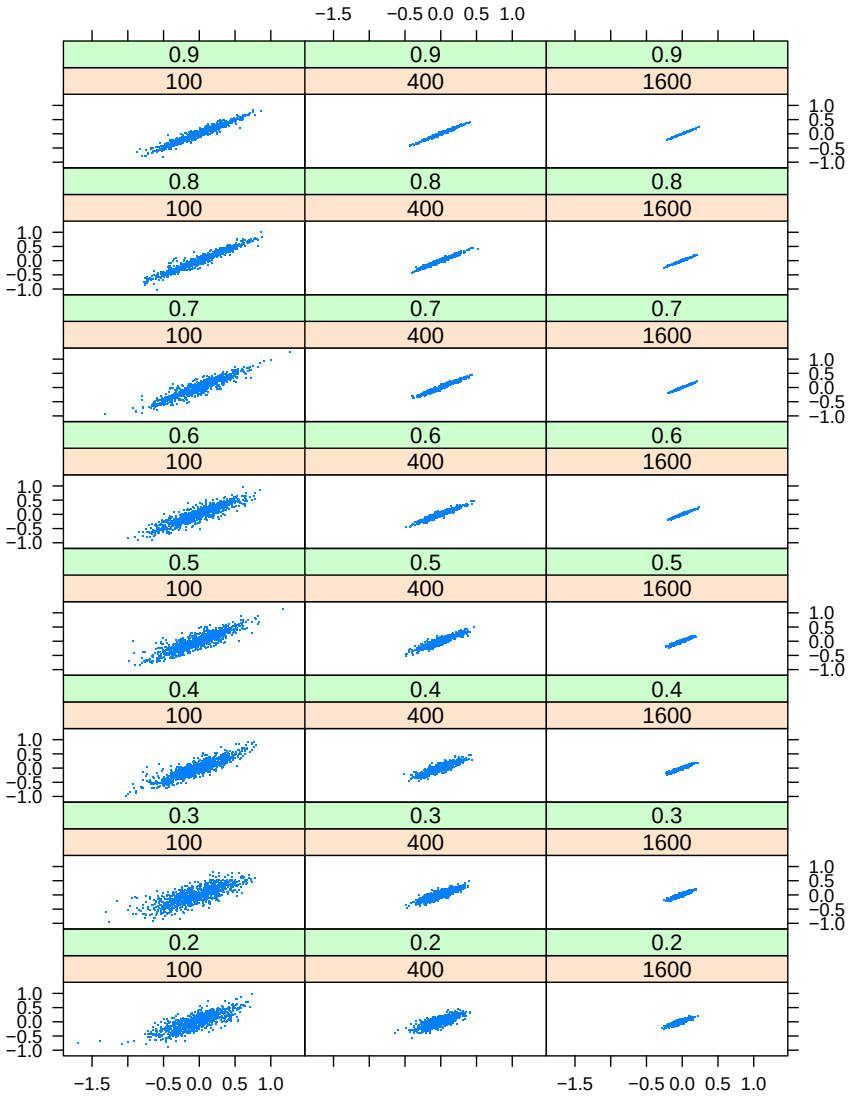

Figure 3: Scatterplots of the Portnoy vs. Peng-Huang estimators in a simple AFT censored survival model: For given sample size, finer grid spacing tends to strengthen the linear correlation between the two estimators.

where, see e.g. Andersen *et al.* (1991),

$$\text{Avar}(\hat{S}(t)) = S^2(t) \int_0^t (1 - H(u))^{-2} d\tilde{F}(u)$$

and $1 - H(u) = (1 - F(u))(1 - G(u))$ and $\tilde{F}(u) = \int_0^t (1 - G(u)) dF(u)$.
Since the Powell estimator makes use of more sample information than does the Kaplan Meier estimator it might be thought that it would be more efficient. This isn't true.

Proposition 1. $\text{Avar}(\hat{\theta}_{KM}) \leq \text{Avar}(\hat{\theta}_P)$.

Proof: Consider

$$\begin{aligned}
f^2(\theta)\text{Avar}(\hat{\theta}_{KM}) &= S(\theta)^2 \int_0^\theta (1 - H(s))^{-2} d\tilde{F}(s) \\
&= S(\theta)^2 \int_0^\theta (1 - G(s))^{-1} (1 - F(s))^{-2} dF(s) \\
&\leq \frac{S(\theta)^2}{1 - G(\theta)} \int_0^\theta (1 - F(s))^{-2} dF(s) \\
&= \frac{S(\theta)^2}{1 - G(\theta)} \cdot \frac{1}{1 - F(s)} \Big|_0^\theta \\
&= \frac{S(\theta)^2}{1 - G(\theta)} \cdot \frac{F(\theta)}{1 - F(\theta)} \\
&= \frac{F(\theta)(1 - F(\theta))}{(1 - G(\theta))} \\
&= \frac{\tau(1 - \tau)}{(1 - G(\theta))}.
\end{aligned}$$

■

Thus, not only is the use of the uncensored C_i's unable to improve upon the Kaplan-Meier estimator, it actually results in a deterioration in performance. Further reflection suggests why our initial expectation of an improvement was misguided: in parametric likelihood based settings a sufficiency argument shows that the C_i for the uncensored observations are ancillary. From a Bayesian perspective, the likelihood principle implies that they cannot be informative, see e.g. Berger and Wolpert (1984).

Having come this far it is worthwhile to consider a few other suggestions that have appeared in the literature regarding the use the uncensored C_i's. Leurgans (1987) considered the weighted estimator of the censored survival function,

$$\hat{S}_L(t) = \frac{\sum I(Y_i > t) I(C_i > t)}{\sum I(C_i > t)},$$

that uses all the C_i's. Conditioning on the C_i's, it can be shown that $\mathsf{E}(\hat{S}_L(t)|C) = S(t)$, and that the conditional variance is

$$\text{Var}(\hat{S}_L(t)|C) = \frac{F(t)(1 - F(t))}{1 - \hat{G}(t)}.$$

Averaging this expression gives the unconditional variance which converges to

$$\text{Avar}(\hat{S}_L(t)|C) = \frac{F(t)(1-F(t))}{1-G(t)},$$

and consequently quantiles based on this estimator behave (asymptotically) just like those produced by the Powell estimator. A remarkable feature of this development is that it reveals that replacing the empirical weighting by $1-\hat{G}(t)$ by the true value $1-G(t)$, yields *even worse* asymptotic performance, since in that event the limiting variance is

$$\frac{H(t)(1-H(t))}{(1-G(t))^2} = \frac{H(t)(1-F(t))}{(1-G(t))} \geq \frac{F(t)(1-F(t))}{(1-G(t))}.$$

It gets even curioser: if instead of replacing $1-\hat{G}$ by the true $1-G$, we instead replace it by an even worse estimator, the Kaplan-Meier estimator of the survival distribution of the C_i's, Wang and Li (2005) show that the resulting weighted estimator is even *better*. Indeed, the resulting weighted estimator achieves the same asymptotic variance as the Kaplan-Meier estimator given above, so the performance of the three versions of the weighted estimator becomes successively better as the estimator of the weights becomes worse!

To evaluate the reliability of these rather perverse asymptotic conclusions we conclude this section by reporting the results of a small scale simulation experiment comparing the finite sample performance of several estimates of the median in a censored one-sample setting. For this exercise we take T as standard lognormal, and C as exponential with rate parameter 0.25. We consider 6 estimators of the median of the lognormal: the (infeasible) sample median, the Kaplan-Meier median, the Nelson-Aalen (Fleming-Harrington) median, the Powell median, the Leurgans median, and finally the Leurgans median modified to employ the true rather than the estimated weights.

	median	Kaplan-Meier	Nelson-Aalen	Powell	Leurgans $\hat{G}$	Leurgans G
$n = 50$	1.602	1.972	2.040	2.037	2.234	2.945
$n = 200$	1.581	1.924	1.930	2.110	2.136	2.507
$n = 500$	1.666	2.016	2.023	2.187	2.215	2.742
$n = 1000$	1.556	1.813	1.816	2.001	2.018	2.569
$n = \infty$	1.571	1.839	1.839	2.017	2.017	2.463

Table 1: Scaled MSE for Several Estimators of the Median: Mean squared error estimates are scaled by sample size to conform to asymptotic variance computations.

The simulation results conform quite closely to the predictions of the theory. The Kaplan-Meier and Nelson-Aalen estimators perform essentially the same, sacrificing about 15% efficiency relative to the (unattainable) sample median. This is about half the proportion (30%) of censored observations in the simulation model. The Powell and Leurgans estimators also perform very similarly as predicted by the theory, sacrificing about 10% efficiency compared to the Kaplan-Meier-Nelson-Aalen. The worst of the lot is the omniscient weighted estimator that sacrifices another 20% efficiency. Beware of oracles bearing nuisance parameters!

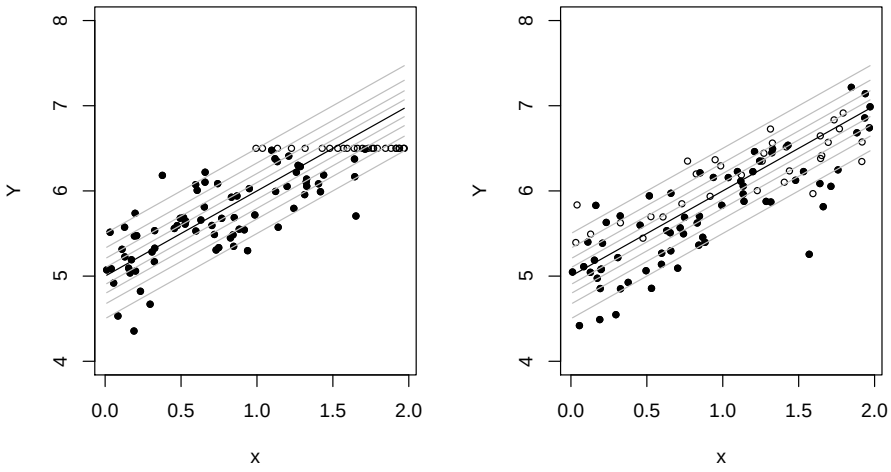

Figure 4: Two Censored Regression Models: The two panels illustrate configurations used in the simulation experiment. Both models have iid Gaussian error models conditional event times. On the left there is constant censoring of all responses above $Y = 6.5$, on the right there is random censoring according to the model given in the text. Censored points are shown as open circles, uncensored points as filled circles. The conditional median line is shown in black, the other conditional decile curves are shown in grey.

6. A Censored Quantile Regression Simulation Experiment

In this final section we report on a small simulation experiment intended to compare the performance of the Powell, Portnoy and Peng-Huang estimators of the censored quantile regression model. We consider four generating mechanisms for the data: two for generating event times and two for generating censoring times. Typical scatter plots of the four mechanisms with $n = 100$ observations are illustrated in Figures 4 and 5, censored points are plotted as open circles and uncensored points as filled circles.

Event times are generated either from the iid error linear model,

$$T_i = \beta_0 + \beta_1 x_i + \sigma_0 u_i,$$

or from the heteroscedastic model

$$T_i = \beta_0 + \beta_1 x_i + (\sigma_1 + \sigma_2 x_i^2) u_i.$$

Censoring times are either constant,

$$C_i = \kappa,$$

or generated from the linear model,

$$C_i = \gamma_0 + \gamma_1 x_i + \sigma_2 v_i.$$

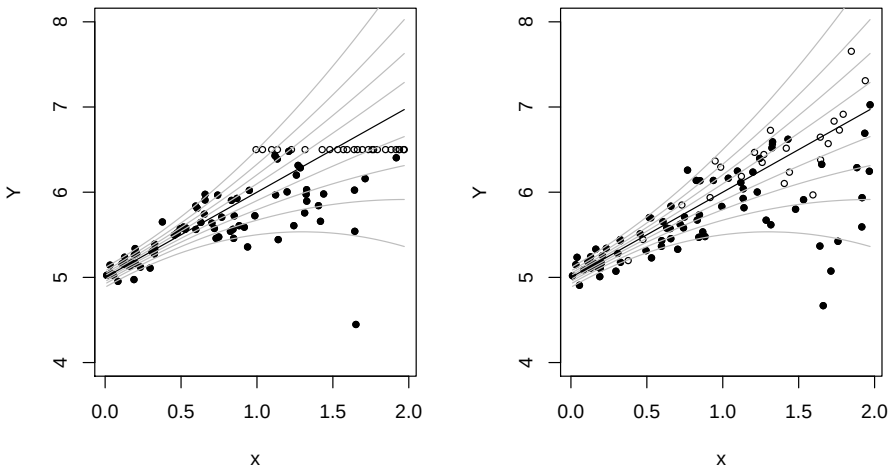

Figure 5: Two More Censored Regression Models: The two panels illustrate the other two configurations used in the simulation experiment. In both cases event times are generated according to the quadratically heteroscedastic model described in the text. On the left there is constant censoring of all responses above $Y = 6.5$, on the right there is random censoring according to the model given in the text. Censored points are shown as open circles, uncensored points as filled circles. The conditional median line is shown in black, the other conditional decile curves are shown in grey.

In each case the x_i's are iid $U[0,2]$, and u_i and v_i's are iid $\mathcal{N}(0,1)$. Parameters were selected so that the proportion of censored observations was roughly 30% in all cases: $\beta^\top = (5,1)$, $\sigma^\top = c(0.39, 0.09, 0.3)$, $\kappa = 6.5$, and $\gamma^\top = (5.5, .75)$.

We compare four estimators of the parameters of the conditional median function

$$Q_T(0.5|x) = \beta_0 + \beta_1 x,$$

for the two iid error models: the Portnoy and Peng-Huang estimators, the Powell estimator as implemented by the Fitzenberger algorithm, and finally the Gaussian maximum likelihood estimator for the conditional mean function, which in these cases happens to be identical to the conditional median function.

	Intercept			**Slope**		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy						
$n = 100$	-0.0032	0.0638	0.0988	0.0025	0.0702	0.1063
$n = 400$	-0.0066	0.0406	0.0578	0.0036	0.0391	0.0588
$n = 1000$	-0.0022	0.0219	0.0321	0.0006	0.0228	0.0344
Peng-Huang						
$n = 100$	0.0005	0.0631	0.0986	0.0092	0.0727	0.1073
$n = 400$	-0.0007	0.0393	0.0575	0.0074	0.0389	0.0598
$n = 1000$	0.0014	0.0215	0.0324	0.0019	0.0226	0.0347
Powell						
$n = 100$	-0.0014	0.0694	0.1039	0.0068	0.0827	0.1252
$n = 400$	-0.0066	0.0429	0.0622	0.0098	0.0475	0.0734
$n = 1000$	-0.0008	0.0224	0.0339	0.0013	0.0264	0.0396
GMLE						
$n = 100$	0.0013	0.0528	0.0784	-0.0001	0.0517	0.0780
$n = 400$	-0.0039	0.0307	0.0442	0.0031	0.0264	0.0417
$n = 1000$	0.0003	0.0172	0.0248	-0.0001	0.0165	0.0242

Table 2: Comparison of Performance for the iid Error, Constant Censoring Configuration

Tables 2 and 3 report mean bias, median absolute error and root mean squared error measures of performance for both the intercept and slope parameters for each of these estimators for three sample sizes. The Gaussian MLE is obviously most advantageous in these settings, but it is also noteworthy that the Portnoy and Peng-Huang estimators outperform the Powell estimator by a modest margin. Bias is generally negligible for all of the estimators in these iid Gaussian settings, so the MAE and RMSE entries can be interpreted essentially as measures of the dispersion of the respective estimators. The relative efficiencies of the estimators are quite consistent with the evidence from the one sample results reported in the previous section showing that the Portnoy and Peng-Huang estimators perform very similarly and exhibit a modest advantage over Powell. This advantage is somewhat smaller for the variable censoring model than for constant censoring, a finding that seems somewhat counter-intuitive. If one maintains the iid error assumption, but alters the form of the Gaussian error distribution then the superiority of the Gaussian MLE evaporates. For example, in simulations of a

	Intercept			**Slope**		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy						
$n = 100$	-0.0042	0.0646	0.0942	0.0024	0.0586	0.0874
$n = 400$	-0.0025	0.0373	0.0542	-0.0009	0.0322	0.0471
$n = 1000$	-0.0025	0.0208	0.0311	0.0006	0.0191	0.0283
Peng-Huang						
$n = 100$	0.0026	0.0639	0.0944	0.0045	0.0607	0.0888
$n = 400$	0.0056	0.0389	0.0547	-0.0002	0.0320	0.0476
$n = 1000$	0.0019	0.0212	0.0311	0.0009	0.0187	0.0283
Powell						
$n = 100$	-0.0025	0.0669	0.1017	0.0083	0.0656	0.1012
$n = 400$	0.0014	0.0398	0.0581	-0.0006	0.0364	0.0531
$n = 1000$	-0.0013	0.0210	0.0319	0.0016	0.0203	0.0304
GMLE						
$n = 100$	0.0007	0.0540	0.0781	0.0009	0.0470	0.0721
$n = 400$	0.0008	0.0285	0.0444	-0.0008	0.0253	0.0383
$n = 1000$	-0.0004	0.0169	0.0248	0.0002	0.0150	0.0224

Table 3: Comparison of Performance for the iid Error, Variable Censoring Configuration

variant of the foregoing models in which Student t_3 errors were used, the Gaussian MLE exhibits considerable larger variability than the other estimators as expected from regression robustness considerations, but also exhibits substantial bias as well. See Tables 6 and 7 for details.

Tables 4 and 5 report bias, MAE and RMSE for the quadratic specifications. Here, two versions of the Portnoy estimator are compared, one using a linear specification of all the conditional quantile functions, the other using a quadratic specification. Similarly, linear and quadratic specifications are compared for the Peng-Huang estimator. Note that while the conditional median function for our simulation model is linear, all the other conditional quantile functions are quadratic in the covariate x, so we might expect the misspecification of those functions by the linear model to cause difficulties for the Portnoy and Peng-Huang estimators. Consequently, for these models we must make some choice about how to evaluate and compare quadratic and linear specifications. For this purpose we have adopted the conventional strategy of evaluating the quadratic at the mean of the covariate, x.

The Gaussian MLE is severely biased in the quadratic settings since it assumes homoscedastic Gaussian error and the model is decidedly heteroscedastic. The Powell estimator performs quite well under both configurations. The differences between the Portnoy and Peng-Huang estimators are, as expected, almost negligible. However, the comparison of their linear and quadratic specfications is quite revealing. For both estimators bias is reduced by employing the (correct) quadratic specification, but this improvement is small and comes at a rather more substantial cost of variance inflation. Thus, from both MAE and RMSE perspectives the linear specification is preferable even though it suffers from a somewhat larger bias effect. Finally, comparing performance of the Powell estimator with those of Portnoy and Peng-

	Intercept			**Slope**		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy L						
$n = 100$	0.0084	0.0316	0.0396	-0.0251	0.0763	0.0964
$n = 400$	0.0076	0.0194	0.0243	-0.0247	0.0429	0.0533
$n = 1000$	0.0081	0.0121	0.0149	-0.0241	0.0309	0.0376
Portnoy Q						
$n = 100$	0.0018	0.0418	0.0527	0.0144	0.1576	0.2093
$n = 400$	-0.0010	0.0228	0.0290	0.0047	0.0708	0.0909
$n = 1000$	-0.0006	0.0122	0.0154	-0.0027	0.0463	0.0587
Peng-Huang L						
$n = 100$	0.0077	0.0313	0.0392	-0.0145	0.0749	0.0949
$n = 400$	0.0064	0.0193	0.0240	-0.0125	0.0392	0.0493
$n = 1000$	0.0077	0.0120	0.0147	-0.0181	0.0279	0.0342
Peng-Huang Q						
$n = 100$	0.0078	0.0425	0.0538	0.0483	0.1707	0.2328
$n = 400$	0.0035	0.0228	0.0291	0.0302	0.0775	0.1008
$n = 1000$	0.0015	0.0123	0.0155	0.0101	0.0483	0.0611
Powell						
$n = 100$	0.0021	0.0304	0.0385	-0.0034	0.0790	0.0993
$n = 400$	-0.0017	0.0191	0.0239	0.0028	0.0431	0.0544
$n = 1000$	-0.0001	0.0099	0.0125	0.0003	0.0257	0.0316
GMLE						
$n = 100$	0.1080	0.1082	0.1201	-0.2040	0.2042	0.2210
$n = 400$	0.1209	0.1209	0.1241	-0.2134	0.2134	0.2173
$n = 1000$	0.1118	0.1118	0.1130	-0.2075	0.2075	0.2091

Table 4: Comparison of Performance for the Constant Censoring, Heteroscedastic Configuration

	Intercept			**Slope**		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy L						
$n = 100$	0.0024	0.0278	0.0417	-0.0067	0.0690	0.1007
$n = 400$	0.0019	0.0145	0.0213	-0.0080	0.0333	0.0493
$n = 1000$	0.0016	0.0097	0.0139	-0.0062	0.0210	0.0312
Portnoy Q						
$n = 100$	0.0011	0.0352	0.0540	0.0094	0.1121	0.1902
$n = 400$	0.0002	0.0185	0.0270	-0.0012	0.0510	0.0774
$n = 1000$	-0.0005	0.0116	0.0169	-0.0011	0.0337	0.0511
Peng-Huang L						
$n = 100$	0.0018	0.0281	0.0417	0.0041	0.0694	0.1017
$n = 400$	0.0013	0.0142	0.0212	0.0035	0.0333	0.0490
$n = 1000$	0.0012	0.0096	0.0139	0.0002	0.0208	0.0310
Peng-Huang Q						
$n = 100$	0.0044	0.0364	0.0550	0.0322	0.1183	0.2105
$n = 400$	0.0026	0.0188	0.0275	0.0154	0.0504	0.0813
$n = 1000$	0.0007	0.0113	0.0169	0.0077	0.0333	0.0520
Powell						
$n = 100$	-0.0001	0.0288	0.0430	0.0055	0.0733	0.1105
$n = 400$	0.0000	0.0147	0.0226	0.0001	0.0379	0.0561
$n = 1000$	-0.0008	0.0095	0.0146	0.0013	0.0237	0.0350
GMLE						
$n = 100$	0.1078	0.1038	0.1272	-0.1576	0.1582	0.1862
$n = 400$	0.1123	0.1116	0.1168	-0.1581	0.1578	0.1647
$n = 1000$	0.1153	0.1138	0.1174	-0.1609	0.1601	0.1639

Table 5: Comparison of Performance for the Variable Censoring, Heteroscedastic Configuration

	Intercept			**Slope**		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy						
$n = 100$	-0.0020	0.0744	0.1122	-0.0002	0.0782	0.1167
$n = 400$	-0.0026	0.0371	0.0555	-0.0003	0.0377	0.0576
$n = 1000$	-0.0021	0.0226	0.0346	0.0006	0.0246	0.0356
Peng-Huang						
$n = 100$	0.0030	0.0750	0.1122	0.0074	0.0806	0.1193
$n = 400$	0.0042	0.0373	0.0563	0.0033	0.0377	0.0592
$n = 1000$	0.0015	0.0219	0.0345	0.0027	0.0244	0.0360
Powell						
$n = 100$	-0.0013	0.0806	0.1198	0.0083	0.0914	0.1427
$n = 400$	-0.0005	0.0390	0.0596	0.0035	0.0441	0.0700
$n = 1000$	-0.0006	0.0244	0.0375	0.0017	0.0292	0.0451
GMLE						
$n = 100$	-0.0420	0.0842	0.1437	0.0549	0.0848	0.1562
$n = 400$	-0.0401	0.0505	0.0816	0.0550	0.0538	0.1013
$n = 1000$	-0.0415	0.0407	0.0609	0.0560	0.0511	0.0765

Table 6: Comparison of Performance for the iid t_3 Error, Constant Censoring Configuration

	Intercept			**Slope**		
	Bias	MAE	RMSE	Bias	MAE	RMSE
Portnoy						
$n = 100$	-0.0026	0.0733	0.1071	-0.0020	0.0637	0.0986
$n = 400$	-0.0027	0.0364	0.0536	0.0003	0.0334	0.0496
$n = 1000$	-0.0013	0.0234	0.0353	-0.0008	0.0201	0.0312
Peng-Huang						
$n = 100$	0.0054	0.0729	0.1084	0.0001	0.0676	0.1002
$n = 400$	0.0061	0.0365	0.0545	0.0014	0.0335	0.0502
$n = 1000$	0.0033	0.0238	0.0356	-0.0001	0.0209	0.0314
Powell						
$n = 100$	0.0034	0.0763	0.1169	-0.0006	0.0740	0.1149
$n = 400$	0.0000	0.0364	0.0569	0.0025	0.0373	0.0557
$n = 1000$	0.0007	0.0247	0.0363	-0.0007	0.0221	0.0342
GMLE						
$n = 100$	-0.0107	0.0760	0.1204	0.0182	0.0726	0.1189
$n = 400$	-0.0119	0.0430	0.0668	0.0229	0.0410	0.0652
$n = 1000$	-0.0100	0.0265	0.0419	0.0217	0.0276	0.0443

Table 7: Comparison of Performance for the iid t_3 Error, Variable Censoring Configuration

Huang we see that for constant censoring the Powell estimator maintains a slight edge, while for the variable censoring model Powell performs slightly worse. In view of the one-sample results reported in Table 1 this is somewhat surprising, one might have expected to see more of an advantage for the Portnoy and Peng-Huang methods. This merits further theoretical investigation that lies beyond the scope of the present paper.

7. Conclusion

Censored data poses a diverse set of challenges in a wide range of applications. As was immediately apparent from the work of Powell (1984, 1986) quantile regression offers some distinct advantages over mean regression methods when there is censoring; departures from Gaussian conditions, or any deviation from identically distributed error, induce bias for least-squares based estimators. In contrast quantile regression estimation is easily adapted to fixed censoring of the type considered by Powell due to the "equivariance of quantiles to monotone transformations." Non-convexity of the Powell objective function can create some computational difficulties, however. Local optima abound and global optimization is far from being a panacea. In our experience, local optimization of the Powell objective via steepest descent, starting at the naive quantile regression estimator performs quite well.

Recently, Portnoy (2003) and Peng and Huang (2008) have introduced new approaches to quantile regression for randomly censored observations. These approaches may be interpreted as regression generalizations of the Kaplan-Meier and Nelson-Aalen survival function estimators, respectively. Although it is difficult to compute asymptotic relative efficiencies for the three estimators we have considered in general regression settings, asymptotics for the simplest one-sample instance suggests that there is a modest efficiency advantage of the new methods over the Powell estimator. This conclusion is supported (weakly) by simulation evidence. The martingale representation of the Peng-Huang estimating equation provides a more direct approach to the asymptotic theory for their estimator, but the simulation evidence suggests that performance of Portnoy's estimator is quite similar.

Software implementations of all three censored quantile regression estimators for the R language are available in the **quantreg** package of Koenker (2008b) using the function `crq`. Extensions to other forms of censoring and more general models remains an active topic of research and will be incorporated into subsequent releases of the package.

Acknowledgments

The author wishes to express his appreciation to Xuming He and Steve Portnoy for extensive discussions regarding this subject, to Achim Zeileis and an anonymous referee for helpful comments on the exposition, and to Stanislav Volgushev for pointing out an error in Section 5 This research was partially supported by NSF grant SES-05-44673. This is a slightly revised version of the paper that appears in the *J. of Statistical Software*.

References

Andersen PK, Borgan Ø, Gill RD, Keiding N (1991). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Barrodale I, Roberts F (1974). "Solution of an Overdetermined System of Equations in the ℓ_1 Norm." *Communications of the ACM*, **17**, 319–320.

Berger J, Wolpert R (1984). *The Likelihood Principle*. Institute of Mathematical Statistics.

Cerdeira JO, Silva PD, Cadima J, Minhoto M (2007). *subselect: Selecting variable subsets*. R package version 0.9-9992, URL http://CRAN.R-project.org/package=subselect.

Chernozhukov V, Fernández-Val I, Galichon A (2006). "Quantile and Probability Curves without Crossing." Preprint.

Efron B (1967). "The Two Sample Problem with Censored Data." In "Proc. 5th Berkeley Sympos. Math. Statist. Prob.", Prentice-Hall: New York.

Fitzenberger B (1996). "A Guide to Censored Quantile Regressions." In C Rao, G Maddala (eds.), "Handbook of Statistics," North-Holland: New York.

Fitzenberger B, Wilke R (2006). "Using Quantile Regression for Duration Analysis." *Allgemeines Statistisches Archiv*, **90**, 103–118.

Fitzenberger B, Winker P (2007). "Improving the Computation of Censored Quantile Regressions." *Computational Statistics and Data Anlaysis*, **52**, 88–108.

Fleming TR, Harrington DP (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, New York.

Fygenson M, Ritov Y (1994). "Monotone Estimating Equations for Censored Data." *The Annals of Statistics*, **22**, 732–746.

Hosmer D, Lemeshow S (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley, New York.

Koenker R (2005). *Quantile Regression*. Cambridge U. Press, London.

Koenker R (2008a). "Censored Quantile Regression Redux." *J. of Statistical Software*. Forthcoming.

Koenker R (2008b). ***quantreg**: Quantile Regression*. R package version 4.17, URL http://CRAN.R-project.org/package=quantreg.

Koenker R, Geling O (2001). "Reappraising Medfly Longevity: A Quantile Regression Survival Analysis." *J. of Am. Stat. Assoc.*, **96**, 458–468.

Koenker RW, D'Orey V (1987). "[Algorithm AS 229] Computing Regression Quantiles." *Applied Statistics*, **36**, 383–393.

Leurgans S (1987). "Linear Models, Random Censoring and Synthetic Data." *Biometrika*, **74**, 301–309.

Lindgren A (1997). "Quantile Regression With Censored Data Using Generalized L1 Minimization." *Computational Statistics and Data Analysis*, **23**, 509–524.

Peng L, Huang Y (2008). "Survival Analysis with Quantile Regression Models." *Journal of American Statistical Association*. Forthcoming.

Portnoy S (2003). "Censored Quantile Regression." *Journal of American Statistical Association*, **98**, 1001–1012.

Portnoy S, Koenker R (1997). "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-error Versus Absolute-error Estimators, with discusssion." *Statistical Science*, **12**, 279–300.

Powell JL (1984). "Least Absolute Deviations Estimation for the Censored Regression Model." *Journal of Econometrics*, **25**, 303–325.

Powell JL (1986). "Censored Regression Quantiles." *Journal of Econometrics*, **32**, 143–155.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Therneau TM, Lumley T (2008). ***survival**: Survival Analysis*. R package version 2.34-1, URL http://CRAN.R-project.org/package=survival.

Tobin J (1958). "Estimation for Relationships with Limited Dependent Variables." *Econometrica*, **26**, 24–36.

Wang J, Li Y (2005). "Estimators for Survival Function when Censoring Times Are Known." *Communications in Statistics (T&M)*, **34**, 449–459.

Affiliation:

Roger Koenker
Department of Economics
University of Illinois Champaign, IL 61820 USA
E-mail: rkoenker@uiuc.edu
URL: http://www.econ.uiuc.edu/~roger/